INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

The UNIVERSITY *of* WESTERN ONTARIO

Faculty of Graduate Studies

In the interests of facilitating research by others at this institution and elsewhere, I hereby grant a licence to:

THE UNIVERSITY OF WESTERN ONTARIO

to make copies of my thesis entitled

DARK MATTERS IN CONTEMPORARY ASTROPHYSICS:
A CASE STUDY IN ~~EVIDE~~ THEORY CHOICE AND EVIDENTIAL REASONING

or substantial parts thereof, the copyright which is invested in me, provided that the licence is subject to the following conditions:

1. Only single copies shall be made or authorized to be made at any one time, and only in response to a written request from the library of any University or similar institution on its own behalf or on behalf of one of its users.

2. This licence shall continue to the full term of the copyright, or for so longs as may be legally permitted.

3. The Universal Copyright Notice shall appear on the title page of all copies of my thesis made under the authority of the licence.

4. This licence does not permit the sale of authorized copies at a profit, but does permit the collection by the institution or institutions concerned of charges covering actual costs.

5. All copies made under the authority of this licence shall bear a statement to the effect that the copy in question "is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner".

6. The foregoing shall in no way preclude my granting to the National Library of Canada a licence to reproduce my thesis and to lend or sell copies of the same.

(signature of witness)	*(signature of student)*
24 January 2001	Ph. D.	Philosophy
(date)	*(degree)*	*(graduate program of student)*

DARK MATTERS IN CONTEMPORARY ASTROPHYSICS:

A CASE STUDY IN THEORY CHOICE AND EVIDENTIAL REASONING

by

William L. Vanderburgh

Graduate Program
in
Philosophy

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
December, 2000

THE UNIVERSITY OF WESTERN ONTARIO
FACULTY OF GRADUATE STUDIES

CERTIFICATE OF EXAMINATION

Chief Advisor

Kathleen Okruhlik

Advisory Committee

Examining Board

Bryan Brigic

Shantanu Basu

Wayne C. Myrvold

The thesis by

William L. Vanderburgh

entitled

Dark Matters in Contemporary Astrophysics:
A Case Study in Theory Choice and Evidential Reasoning

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date: January 19, 2001

Chair of Examining Board

ABSTRACT AND KEYWORDS

ABSTRACT: This dissertation examines the dynamical dark matter problem in twentieth century astrophysics from the point of view of History and Philosophy of Science. The dynamical dark matter problem (which should be distinguished from the cosmological dark matter problem) describes the situation astronomers find themselves in with regard to the dynamics of large scale astrophysical systems such as galaxies and galaxy clusters: The observed motions are incompatible with the visible distribution matter given the accepted law of gravitation. This discrepancy has two classes of possible solutions: either there exists copious amounts of some kind of matter that neither emits nor absorbs radiation (hence "dark"), or the law of gravitation must be revised.

Chapter 2 describes the physical and philosophical foundations of dynamical inferences—inferences *from* discrepancies between well-founded theoretical expectations and reliable observations *to* the characteristics of candidate solutions. Chapter 3 discusses the history of dark matter (beginning around 1930). Chapter 4 reviews the present evidence bearing on the dark matter problem. Chapter 5 evaluates the important candidate matter solutions in light of the available evidence. Chapter 6 evaluates two candidate gravitational solutions on evidential and methodological grounds, and addresses the problem of theory choice. I do not try to solve the dark matter problem, but to uncover and evaluate patterns of inference involved in evidential arguments for and against candidate solutions.

I show that Newton's "Reasoning from Phenomena" is a good framework from which to understand what is going on in this field. I argue that "higher order" and especially *non-dynamical* evidence is the best hope for solving this problem in dynamics. This is so in part because of "the dark matter double bind": the very existence of the dark matter problem means that we cannot be sure of the overall matter distribution in astrophysical systems, and this in turn means that the observed motions by themselves cannot provide relative confirmation of any theory of gravitational interactions taking place at these scales. I use Newton's Rules of Reasoning to argue that we should retain

General Relativity as our theory of gravity at galactic and greater scales, despite the lack of positive evidence to confirm it over its rivals at these scales.

KEYWORDS: Philosophy of Science, History of Astronomy, History of Physics, Twentieth Century Astrophysics, Newton, Duhem, Dark Matter, Gravitation, Dynamical Inferences, Underdetermination of Theory by Evidence, Theory Choice, Evidential Reasoning, Evidence, Methodology of Science

EPIGRAPH

And, hanging over it all, the brooding specter of Rudolf Carnap and Hans Reichenbach, the Vienna Circle of philosophy and the rise of symbolic logic. A muddy world, in which he did not quite care to involve himself.

—Philip K. Dick, *The Galactic Pot-Healer*, 126.

ACKNOWLEDGEMENTS

I am fortunate to have received assistance and support of many kinds from many people during my graduate studies. The Department of Philosophy at UWO has been a wonderful and enriching place to be. My fellow graduate students over the years, though I cannot name them all here, have provided me with a community of friends which I am going to miss terribly: thank you all for teaching me so much. I would like to give special mention to Gerry Calaghan, Todd Calder and Claire Nowlan, Tony Lariviere and Lyla Simon, and Aga Trojniak: their friendship has been very important to me, and I hope to be able to continue to lean on them. Professor Tracy Isaacs has also been a good friend, and a wonderful movie-going companion.

I owe a great debt to Bonnie MacLachlan, John Gilbert Thorp and especially Professor John Thorp, Chair of the Department. Their generosity and kindness in welcoming me into their family and providing me with a wonderful place to live (and eat!) for several important years of my graduate career is something I will never forget. Professor Thorp and I have had many conversations over the years which I think of as being foundational to my professional development, both as a philosopher and as someone who will take part in the life of a department. We collaborated on a CPA paper together, and John has supported me in innumerable other ways as well. I am more grateful than I can say.

Professor Kathleen Okruhlik, formerly Chair of the Department and now Dean of Arts, is someone I have admired greatly for as long as I have been at UWO. Professor Okruhlik combines philosophical acumen, administrative brilliance, an unparalleled capacity for work and a deep care for individuals in a way that leaves me awe-struck, and which I pledge to try to emulate. I am very fortunate that she agreed to be my supervisor—although I couldn't help but feel guilty for giving her yet more work to do! My thesis is much improved thanks to her. She is also largely responsible for the fact that I decided to do graduate work in Philosophy in the first place. Thanks, Kathleen!

I also owe a great deal to Professor Bill Harper, whose courses over the years (including "Kepler" in my MA year, out of which I eventually got my first publication,

and a reading course on Dark Matter in the same year) inspired my interest in methodology and evidential reasoning in the first place, and fed my interest in the history of astronomy. As is evident in the text below, Professor Harper's work on Newton has been an important influence on my philosophical development. A paper I wrote for his seminar in 1995 on the evidence for gravitation theories was the foundation for Chapter 6. Professor Harper's extensive comments on several drafts of my thesis have been extremely helpful, and it is much improved thanks to his insistence on getting things right; of course, any mistakes or misrepresentations that remain are entirely my own.

My third reader, Professor Aaron Sigut of the Department of Physics and Astronomy, provided extremely helpful, clear and detailed comments on some technical points in the thesis, and helped me to avoid some mistakes about the physics.

I would also like to thank Professor Robert DiSalle, who was instrumental in my being awarded a SSHRC doctoral fellowship, and who provided extensive comments on early drafts of several chapters of my thesis (and several things that did not make it into the thesis). I have benefited greatly from his input.

My good friend Dr. Francisco Flores (a UWO grad, now at San Luis Obispo) gave me very helpful feedback on Chapter 2, and provided me with a derivation presented in a footnote in the final section of that chapter.

I would like to thank Professors John Nicholas and Barry Hoffmaster, the former and present Graduate Directors, for their assistance and support.

And thanks to Dr. Elaine Landry (another former fellow UWO grad student) for helping me arrange a part time appointment at Concordia University this fall, which has made it possible for me to eat and pay rent (not to mention live in wonderful Montreal!) while finishing my thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS

AU	Astronomical unit, ~1.5 × 10^8 km, mean distance of Earth from Sun
c	Einstein's constant for the speed of light, c = 300 000 kms^{-1}
CBR	Cosmic Background Radiation
CDM	cold dark matter (dark matter consisting of particles moving at well below the speed of light)
COBE	Cosmic Background Explorer
CTG	Conformal Theory of Gravity (Mannheim)
DM	dark matter
DMP	dark matter particle
EBL	extra-galactic background light
eV	electron volt, a unit of mass for fundamental particles
FRW	Friedmann-Robertson-Walker: the standard cosmological spacetime model based on General Relativity
G	the gravitational constant
GeV	giga-electron volt (one billion electron volts, the rest mass of a proton)
GPE	gravitational potential energy
GR	General Relativity
GR+DM	combined theory of General Relativity plus a specific dark matter hypothesis
GUT	grand unified theory (a class of theories of the unification of all the fundamental forces except gravity)
Gyr	giga-year, 1 Gyr = 1 billion years
H_0	the Hubble constant for the cosmic expansion: H_0 = *50-100* $kms^{-1}Mpc^{-1}$; best current value: H_0 = ~*6* $kms^{-1}Mpc^{-1}$
H-D	hypothetico-deductivism
HDM	hot dark matter (dark matter consisting of particles moving at velocities a significant fraction of the speed of light)
HST	Hubble Space Telescope
H-R	Hertzsprung-Russell diagram/relation, an empirical relationship between spectral type and luminosity that can also be used to determine the mass of a star from its observed light and distance

HUD	Humean Principle of Underdetermination
KE	kinetic energy
kpc	kilo-parsec (1000 parsecs)
L	luminosity
$L_\odot$	solar luminosity (unit of measure)
ly	light-year
M, m	mass
$M_\odot$	solar mass (unit of measure)
MACHO	Massive Astrophysical Compact Halo Object
M/L	mass to light ratio, expressed in solar units: $M_\odot / L_\odot = 1$
MOND	Modification of Newtonian Dynamics (Milgrom)
Mpc	mega-parsec, 10^6 pc
N, n	number (of objects of some kind)
n	the exponent of r in the power law for the force of gravity, $F \propto r^n$
NM	Newtonian Mechanics
NTG	Newtonian Theory of Gravity
PBH	primordial black hole
pc	parsec, 3.26 light-years (the distance of an object whose heliocentric parallax is one second of arc)
PPN	Parametrized Post-Newtonian Formalism
q, q_0	the cosmological deceleration parameter, the rate of deceleration in the present universe
QUD	Quinean Principle of Underdetermination
R, r	radius, distance
RfP	Reasoning from Phenomena
SGR A*	Sagittarius "A-star", the powerful point radio source identified as a supermassive black hole at the dynamical centre of the Milky Way
SNO	Sudbury Neutrino Observatory
SR	Special Relativity
T	Period, as in Kepler's Third Law, T^2/R^3
UG	(Newtonian) Universal Gravitation
V, v	velocity

$\langle v \rangle$, $\langle v^2 \rangle$	average velocity, average of squares of velocities
VLBA	Very Long Baseline Array, a huge radio-interferometer telescope
WIMP	Weakly Interacting Massive Particle
γ	gamma, Lorentz correction factor, $(1 - v^2 c^{-2})^{-1/2}$
Φ	phi, gravitational potential
Λ	lambda, the cosmological constant: a cosmic force acting in effect as negative gravity, accelerating the Hubble expansion over time
ρ, ρ_0	density (usually mass density), density at the present epoch
Ω	omega, the mass density parameter: a ratio between the observed mass density (at some era) and the ("critical") mass density required to close the universe
Ω_{matter}, Ω_M, Ω_{baryon}	the contribution of matter or baryons to the overall mass density
Ω_Λ	the contribution of the energy of the cosmological constant to the overall mass density

CHAPTER 1

INTRODUCTION

There are more things in heaven and earth,
Horatio, than are dreamt of in your philosophy.
—William Shakespeare, Hamlet, I.v.

1.0 PRELIMINARIES

In this dissertation I use the dark matter problem in twentieth-century astrophysics as a case study of evidential reasoning in the physical sciences. There are two main goals. First, I aim to describe the dark matter problem—including its history, evidence and proposed solutions—in a way that is accessible to non-scientists. Despite the fact that many physicists and astronomers consider the dark matter problem to be the most important unsolved problem in the physical sciences, and despite the fact that it is a physical problem rife with philosophical implications, it has until now received no attention from philosophers, and little or no attention from historians of science. Second, I aim to examine in detail a subset of the philosophical issues raised by the dark matter problem, namely those related to evidential reasoning and the problem of theory choice.

There are in turn two main things that this introduction is designed to do. First, it describes the dark matter problem itself in outline, and later also says what the dark matter problem studied here is not. Second, it sets up the problem of evidential reasoning in general and discusses what philosophical issues arise in connection with it.

To begin with the first task then, the physical problem with which I am concerned is essentially this: the observed internal motions of galaxies and clusters of galaxies are inconsistent with the visible matter distribution, given the laws of physics familiar to us. The inconsistency arises from the fact that in every known case there is a systematic discrepancy between the two possible ways of measuring the masses of astronomical systems (such as galaxies and clusters of galaxies). The first kind of mass measure estimates the so-called "visible mass" of a system from the observed total flux of radiation emitted by it: using empirical relationships between total mass and total

luminosity, as established in a few well-studied nearby regions, one can infer the total mass of a previously unknown system from its observed luminosity and distance (this method is explained in Chapter 3). The second kind of mass measure uses the laws of motion and gravitation to calculate the total mass from the observed motions of bodies in the system in question (the physical and philosophical foundations of this sort of mass measure are discussed in Chapter 2, and cases of its use are discussed in Chapters 3 and 4). For systems larger than the local region of the Milky Way, the second or "dynamically determined" mass is always much higher than the visible mass—up to two orders of magnitude higher, depending on the type of system whose mass is being studied.

The systematic discrepancy between the visible and the dynamical masses of individual galaxies, clusters and other structures has two classes of possible solutions, one or the other of which *must* be true given the best available evidence and theories[1]:

> (1) There is much more matter present than is visible (up to 100 times more, so that the "dark matter" is by far the main constituent of the physical universe), and it is not distributed in the way the visible matter is. In this case the problem is to figure out what this unknown matter is, why it is invisible, and what causes the difference in the distributions of dark and visible matter.
> (2) There is no excess matter beyond what visible mass measures indicate, and the discrepancy arises because the theory of gravitation employed in the dynamical measures does not apply to objects the size of galaxies and larger. In this case the problem is to figure out what the new theory of gravitation ought to be.

The details, as we shall see, are more complicated, but in the most general terms the dark matter problem is (or arises from) a radical discrepancy between an apparently well-founded theoretical expectation compared against a set of apparently impeccable empirical findings. As it turns out, this is a discrepancy whose solution seems inevitably to require a radical change in our basic physical knowledge, either with regard to the

---

[1] Some combination of the two is also a possible solution, but not one that has received any attention in the scientific literature.

census of (types and amounts of) matter in the universe, or with regard to the law of gravitation.

If there is an over-arching philosophical thesis in this discussion, it is just that evidential reasoning is possible. Some recent commentators on science have denied the possibility of using evidence to arrive at objective or rational theoretical decisions, saying instead that theory choice is necessarily a matter of pragmatics or arbitrary decision. I believe that this is not the case. The proof of this, insofar as there is any given here, has several components. One component is in the success of the present account of the state of the evidence for the dark matter *problem*—which I show is undeniably a real scientific problem, that is, one the rationality of arriving at which is not in doubt and which cannot be ignored. Another component of the proof of the viability and rationality of evidential reasoning in science is in the success of the present analysis of evidential reasoning with regard to candidate *solutions*—where I show, contrary to some philosophers' claims, that it is *not* possible (or at least that it has not been *shown* to be possible) to reasonably retain any hypothesis you want in the face of any possible evidence. Still another component of the proof of the rationality of theory choice is in my description of what possible new evidence would allow us to solve the dark matter problem more or less definitively, were that evidence to become available. Of course, demonstrating an actual evidential solution to this problem would be a much stronger argument for the possibility of rational theory choice, but since the scientific problem on which I am focusing is as yet unsolved, that avenue is unavailable to us here. This is one reason why the argument for the rationality of science is not my main focus but rather an implicit theme.

Unfortunately some commentators have taken the lack of success of logical empiricist instance confirmation, hypothetico-deductive confirmation and deductive falsification programs as showing that no rational account of evidence is possible. The failure of any one or several attempts to explain evidential reasoning does not show the impossibility of the project. What these failures really show is just that evidential reasoning is harder than it looks. Evidential reasoning, like bicycle riding, is a skill humans perform regularly and successfully with relative ease, but which they find rather difficult to explain adequately.

I am an epistemic naturalist to this extent at least, that I think that if ordinary folks and scientists do it, and seem to do it successfully, then there is something for philosophers to try to explain. But I maintain the philosopher's prerogative to be prescriptive as well as descriptive about the foundations of scientific reasoning: just because a scientist makes a certain kind of inference does not mean that that inference is automatically correct. There is a role for philosophers in analysing just what is going on in cases of scientific inference, and in trying to determine which modes of inference are better than others.

The main theoretical resource I use in my evidential analysis of the dark matter problem is an interpretation of Newton's methodology which I owe to Bill Harper. Newton brought his Reasoning from Phenomena (**RfP**) to perfection while constructing his theory of Universal Gravitation (though its roots are in his earlier work on optics). It may at first seem to be somewhat perverse and anachronistic of me to try to apply this Newtonian approach to twentieth century astrophysics. But as Harper (1997a and other references) has argued, Newton's methodology and standards for empirical success still drive (implicitly) the testing of gravitational theories in the relativistic era. And in the present work, I am able to use RfP in the analysis of the use of evidence in the dark matter debates; it is interesting, but perhaps not surprising, that RfP applies even to this astrophysical problem. The link is that the dark matter problem is intricately tied up with gravitation. I am also able to use RfP as a guide for determining what kinds of possible evidence would permit us to give a fairly certain solution to the dark matter problem. RfP helps in this partly because it sets out an ideal for what a scientific theory ought to be like (that is, what an ideal theory in (gravitational) physics ought to be like: I make no claim that RfP is the only method of good science). RfP is also useful here because it sets out an ideal of empirical success that tells us when one hypothesis is evidentially superior to its rivals. RfP may not be the final or best way of understanding evidential reasoning in this sphere, but it is very useful.

By this return to Newton I do not mean to denigrate recent accounts of evidence in the philosophical literature, which indeed have their merits, but to point out that Newton's ideal of empirical success still bears examination, and that it can be fruitful in the analysis of contemporary physical problems.

One respect (I think an interesting and important one) in which the analysis given here goes beyond Newton's Reasoning from Phenomena, or rather elucidates and develops something implicit in it, is the account of what I call (for lack of a better name) "higher order evidence". (I am certainly not alone in discussing the idea of higher order evidence, but I do not remember where I learned of the idea.) As I describe, the dynamical evidence which makes us aware of the existence of the dynamical discrepancy, and which is our main resource for constructing solutions, is insufficient to distinguish between a huge class of very different rival solutions. Higher order evidence, I argue, and especially *non-dynamical* higher order evidence, is our best hope for an evidential solution to this problem in dynamics. As a consequence of the analysis of the dark matter problem I come to consider the empirical support of General Relativity (GR), and find that neither GR nor any rival gravitational theory is in fact supported, and perhaps none can be supported, by phenomena taking place at the distance scales involved in galaxies and larger dynamical structures.[2] This makes the task of deciding which dynamical law holds at these scales—an apparently necessary step in any solution of the dark matter problem—unexpectedly more difficult (and interesting) than we might have supposed.

As for the prospects of actually being able to acquire the kinds of evidence requisite for the rational solution to the dark matter problem whose possibility I advocate, it is hard to say. We cannot predict the future with certainty. But scientists are pursuing the right avenues of research. If the world turns out to have the right sorts of characteristics, we will eventually be able reach an evidential solution to the dark matter problem. Of course, any such solution we do actually reach is fallible, or rather corrigible in light of new evidence and new theories. It is also possible that the world will turn out to not have the right sorts of characteristics, in which case we will be unable to solve the dark matter problem. But we will thereby have learned something important about the world, and about the limits of knowledge.

---

[2] Note, this is a claim about gravitational interactions that take place over distances corresponding to the radii of galaxies or clusters, not about smaller scale interactions that happen to be very far away from us. GR has been tested for binary star systems, for example, but although they are distant from us these are not large scale interactions.

Because the dark matter case is essentially unknown to philosophers, I have included here, besides the main analysis of evidential reasoning in this case, chapters describing the physical and philosophical foundations of the kinds of reasoning involved in the discovery and solution of similar problems, the early history of dark matter (from about 1930 to the mid-1970s), a description of the contemporary astronomical evidence bearing on the case, and a "natural history" of the "zoo" of candidate dark matter hypotheses. The historical materials (see especially Chapter 3) are important because they situate the present debate, and show that there really has been little progress on the astrophysical dark matter problem since the 1930s (although there has been significant progress in determining what the *solution* to it is *not*). I also felt it necessary to give the historical treatment supplied here because almost nothing has been written about the history of dark matter (except a few reviews of the evidence by astronomers).

1.1 DISTINGUISHING TWO DARK MATTER PROBLEMS

Most discussions of dark matter blur an important distinction, namely the distinction between "dynamical" as opposed to "cosmological" dark matter. I use the phrase "dynamical dark matter" to refer to matter whose presence in galaxies and other structures we know about only in virtue of its gravitational influence on other, visible matter (that is, from its dynamical effects). I use "cosmological dark matter" to refer to matter that some cosmologists hypothesise to exist in order that the cosmic mass density be exactly enough to eventually exactly halt the expansion of the universe. Peebles (1993) is one of the few commentators who mentions the conceptual and evidential differences between the two:

> [T]he dark mass idea appears in two contexts. The first is the set of dynamical results that indicate [that] most of the mass in galaxies, and in systems of galaxies [that is, clusters and larger structures], is outside the bright central parts where the mass of the luminous stars dominates.... [T]he amount of dark mass estimated from these observations brings the mean density of the universe to about 10% of the critical Einstein-de Sitter value. The second context is the set of arguments... that leads one to give very careful consideration to the possibility that there is another factor of ten dark mass outside systems of galaxies, bringing the total to $\Omega = 1$. If Newtonian mechanics is a useful approximation on the scale of galaxies, the observations unambiguously establish the reality of the first effect, the presence of dark mass in galaxies. And if there is dark mass, it certainly is reasonable to

> consider the possibility that another factor of ten might be found. However, at the time this is written there is no compelling evidence that this last step follows. (Peebles 1993, 417)

I will shortly explain in more detail what makes these two dark matter problems significantly different, but I should first note that drawing the distinction serves a practical purpose here as well, namely that it allows me to divide the huge literature on dark matter (see the introduction to Chapter 4 for some rough statistics on the huge number of articles in this very active area of research) into a more manageable (but still huge) chunk. Thus I am mainly concerned here with dynamical dark matter, although I will have a little bit to say about the cosmological dark matter problem because some of its candidate solutions are also potential contributors to solutions of the dynamical dark matter problem. In what follows I attempt to justify this strategy by showing why the foundations of the cosmological dark matter problem are so weak as to make it uninteresting.[3]

I begin with a "first approximation" characterisation of the two dark matter problems, and then go into more detail about each; since the remainder of this dissertation is about dynamical dark matter, I focus in this section mainly on describing and critiquing the arguments for cosmological dark matter, and on some recent evidence that seems to make cosmological dark matter otiose. This is important to do because many people do not clearly distinguish the two dark matter problems, but their epistemic foundations and implications are quite different; in particular, many people automatically think of the cosmological dark matter when they hear "dark matter".

The dynamical dark matter *problem* arises from a discrepancy of the type described above between visible and dynamical measures of the masses of galaxies and clusters of galaxies. Dynamical dark matter is additional, otherwise unknown matter whose existence is hypothesised just in order to eliminate this discrepancy; the amount of dynamical dark matter supposed to exist in a given system is measured by the difference between the dynamical mass and the visible mass of that system, and the "problem" is to

---

[3] As I describe below and in the Appendix, recent observational evidence corroborates my view; it shows more or less definitively that cosmological dark matter is not what is responsible for bringing the overall mass density to the critical value, if the universe has the critical density at all.

determine just what this otherwise unknown matter could be. The cosmological dark matter *problem* is to find a model of some kind and distribution of matter that can bring the mass density to the preferred critical value while still remaining invisible. One complication that must be overcome if the existence of cosmological dark matter is to be established is the fact that reliable observations indicate that the actual mass content of the universe is only about 20-40% of the critical value. (Note, this *includes* all the dynamical dark matter: see below, and the Appendices).

With this "first approximation" to the two dark matter problems in hand, let me now describe each in more detail. Dynamical dark matter is known to exist because of a radical discrepancy between independent measures of the masses of individual astronomical systems such as galaxies and clusters of galaxies (from now on I will refer simply to "clusters" unless the context requires distinguishing them from globular star clusters, which are groups of stars within the Milky Way and other galaxies). Unless our dynamical theories are radically wrong, as much as 90% of the mass of these systems is detectable only in virtue of its gravitational effects on normal matter that emits or absorbs electromagnetic radiation (stars, gas, dust). Roughly, for *all* dynamical systems larger than the local region of the Milky Way—our own galaxy, other spiral, elliptical and dwarf galaxies, clusters, and cosmological large-scale mass distributions (superclusters, domain walls)—the dynamically inferred masses are much higher than the visible masses.

To the extent that we trust our dynamical theories, we must accept what the dynamical measures tell us, namely that about 90% of the mass of large astrophysical systems is both in some form that we do not know and distributed very differently from the visible matter. Another option is to say that the estimates of visible mass simply do not include all the ordinary matter that is actually present, but the idea that so much more ordinary matter is present is, as we shall see in Chapter 5, rather hard to reconcile with the observed lack of corresponding electromagnetic absorption or emission. The excess matter is called "dark" precisely because of this lack of an electromagnetic signature: it is much more likely, given the available evidence, that the dark matter is something quite unlike ordinary matter. The difficult part is to try to determine just what the dark matter is, from the rather limited and necessarily indirect evidence that is available to us.

Cosmological dark matter, in contrast, is assumed (not measured) to exist, in order to satisfy criteria that amount (in my judgement) to either philosophical prejudices or highly speculative and empirically under-supported theoretical arguments, arguments to the effect that the mass density of the universe is exactly sufficient to halt the Hubble expansion at some future time. The mass density parameter is represented by the symbol Ω (omega), which stands for the ratio of the actual mass density of the universe at the present epoch to the "critical" mass density, that is, the mass density that would be required to exactly halt the expansion in the infinite future. Universe models in which this critical mass density obtains have $\Omega = 1$, and are called "flat" because in General Relativity the overall mass density of the universe is related to the global space curvature. A value of $\Omega < 1$ corresponds to an "open" (positive curvature) universe that expands forever, and $\Omega > 1$ to a "closed" (negative curvature) universe that will recollapse in a "Big Crunch".[4]

Most cosmologists have (or until recently had) a preference for flat universes. As far as I am able to determine, besides what amounts to professional socialisation into this belief, this preference has three different though inter-related sources. First, the preference for flat universes is in part a result of a perceived "aesthetic" value of flat universe theories. (This goes along with a corresponding widespread belief among physical scientists that, for some never precisely spelled out and probably actually inchoate reason, the *beauty* of a theory is an indicator of its truth). Second, the preference for flat universes is in part a result of not being able to see any non-arbitrary reason for Ω

---

[4] $\Omega_{matter} = \rho_0 / \rho_{crit} = 8\pi G\rho_0 / 3H_0^2$, where ρ_0 and H_0 are respectively the mass density and Hubble constant at the present epoch and G is the gravitational constant. (See Peebles 1993, 98.) Note that the universe is expanding out of a much denser primordial state in order to see that the matter density parameter evolves with cosmic time. Also, it is important to note that the overall mass density parameter Ω need not come from *matter* alone; because of the mass-energy equivalence, a sufficiently pervasive energy background could by itself make $\Omega = 1$, even if the contribution of Ω_{matter} is only a small fraction of the total mass density. The BOOMERanG balloon-borne experiment has recently published the results of its detailed study of the cosmic microwave background radiation, which indicate that the universe is indeed flat. But observations described below also show that the *matter* contribution to this total mass density is at most about 40%. See < http://oberon.roma1.infn.it/boomerang/ > and de Bernardis, P. *et al.* (2000).

to have a value different than one—this is, in effect, a "Leibnizian" argument from sufficient reason. Third, the preference is in part due to the influence of Dicke's "fine-tuning" argument.

Dicke's argument (1970) that $\Omega = 1$ goes like this. In order for Ω to be as near to 1 as it is observed to be in the present epoch (within an order of magnitude), at the moment of the Big Bang Ω had to be *exactly* one: Dicke calculates that it must be the case that $\Omega = 1 \pm 1 \times 10^{-59}$ at the moment of the Big Bang, that is, the density parameter has to be fine-tuned to 59 decimal places in order for the universe to look as it does now![5] This degree of exactness is required because any early deviation of Ω from 1, however small, explodes as the universe ages: in the standard Big Bang model, $\Omega = 1$ occupies an unstable equilibrium point, which is to say that by the present epoch any initial deviation from 1 larger than one part in 10^{59} would produce an obviously open or closed universe in *much* less time than has in fact elapsed since the Big Bang, as opposed to a "flat-ish" universe like the one we inhabit. It would therefore be unreasonable, according to this argument, to suppose that Ω was initially anything but *exactly* 1—that is, not to 59 but to an *infinite number* of decimal places. The best explanation of this is that some (unknown) physical process *forces* Ω to 1. Guth's inflationary cosmology is one prominent attempt to explain what this physical process might be. And if Ω was initially exactly 1, it will be exactly 1 now, by the same argument.[6]

After Alan Guth's invention of inflationary cosmology, the flatness of the universe was taken by many cosmologists to be almost a necessary fact (see Guth 1981). Inflation

---

[5] Krauss (2000, 138-43) discusses fine-tuning problems in relation to cosmological parameters, and concludes that $\Omega = 1$ to one part in 10^{27}. This is considerably fewer decimal places than Dicke, but still a huge degree of fine-tuning. The reason for the difference between Krauss and Dicke on this point is unclear since Krauss does not discuss his calculation in detail (in fact he does not even mention Dicke, the originator of this argument—perhaps an indication of just how much the argument has now become a standard part of cosmology), but it is probably due to differences in the cosmological models used by each.

[6] I will not here offer any further comment on Dicke's rather questionable argument. I mention it only to show a commonly accepted motivation for believing that the mass-density of the universe *must* be much higher than it is observed to be.

was able to account for various observed features of the universe (including the observed abundances of the elements, the isotropy of the cosmic background radiation, and the paucity of magnetic monopoles), and did so in a way that *forced* $\Omega = 1$, *whatever* the initial density in the moment after the Big Bang.[7] The idea of an open universe had few adherents in any case. There is something aesthetically or intellectually pleasing about the notion that the universe is closed (and psychologically pleasing in that it implies the possibility of an unending cycle of Bangs and Crunches), whereas a flat universe has the advantage of being a "special" case (there is only one possible flat universe, but an infinite number of possible closed and open universes, corresponding to all the decimal values of Ω greater than and less than 1).

Both the closed and flat scenarios suffered from a long-standing discrepancy between their assumed values for the mass-density parameter, and the values inferred from observations: at most, the observations indicate the matter fraction of the overall mass density is less than about 40% of the critical density. The cosmological dark matter problem, then, is to figure out *where* the extra 0.6 or more of the present mass-density is, and *what* it is—assuming that it exists at all. A "zoo" of candidates has been proposed, including primordial black holes, neutrinos and other exotic fundamental particles that would have been created in the hot dense state immediately following the Big Bang, according to accepted theories of particle physics (some of these are also candidates for

---

[7] Peebles (1993, 365) notes that Dicke's argument was known in his circle at Princeton for at least a decade before the idea was published in 1970, and that before Guth's inflationary cosmology gave a reason to think the universe must have started exactly flat, Dicke's argument "was not generally considered compelling." Many cosmologists were swayed by the fact that inflation's apparent resolution of various problems such as the horizon problem (inflation's solutions to which were taken as evidence in its favour—essentially an argument from explanatory and unificatory power) required that the present space curvature be negligibly small (which is equivalent to $\Omega = 1$). The attractiveness of the inflation scenario in effect made people accept the force of the Dicke argument, and believe that $\Omega = 1$. As Peebles notes, however, "there is little objective evidence on which to decide whether the inflation scenario really is valid (1993, 366). Earman and Mosterin (1999) make the point quite a bit more strongly, and argue that not only is there little actual evidence in favour of inflation, but that its supposed explanations of various cosmological problems either no longer hold up now that the models of inflation are more realistic, or that those explanations do not provide any significant reason to believe that inflation is true.

the dynamical dark matter). If the cosmological dark matter exists, it is about ten times more abundant than even the total dynamical matter (which is already ten times the visible matter), so that ordinary matter would contribute only on the order of about 1% of the total mass of the universe.[8]

The cosmological dark matter problem is not an empirical discrepancy in the standard sense and is founded on theories for which there seems to be little independent warrant. And it turns out to be very hard to provide empirical constraints on or to perform any tests of the candidate hypotheses (most of which are fundamental particles that have never been observed, and which may interact so weakly with ordinary matter as to make them practically impossible for us to detect). Because the evidential basis of the cosmological dark matter problem is not merely so much weaker than that of the dynamical dark matter problem but in my view simply so weak that it should not be taken seriously as a cosmological hypothesis, the cosmological dark matter problem is not of much interest to me in this dissertation. The issues I want to investigate have to do with the uses of evidence in supporting scientific theories, and the cosmological dark matter problem in my view makes a poor example of this sort of reasoning.

But note that the recent evidence from surveys of supernovas (for example: Perlmutter, *et al.*, 1998; Alcaniz and Lima, 1999) fixes the observed value of the matter fraction of the overall mass-density, Ω_{matter}, with much more certainty than ever before. Furthermore, the technique used does not depend on estimates of visible mass, as do previous estimates of Ω_{matter}. This means that these observations measure the *true* value of Ω_{matter}, *including* any invisible matter the universe may contain. In recent studies, the best-fit value of Ω_{matter} is around 0.24 (and with very high confidence, $\Omega_{matter} \leq 0.45$) (Alcaniz and Lima, 1999, L89). Unless some rather complicated theoretical move is invented, this signals the death of the flat and closed models of the universe where *matter* is what is responsible for bringing the universe to the closure density. All the more so now that still more recent observations, discussed in the

---

[8] Some commentators have pointed out that this makes ordinary matter, which we used to think of as everything that exists, a mere "cosmic afterthought", and they have likened this shift in perspective to the Copernican revolution; see, for example, Krauss 2000.

Appendices, indicate that the Hubble expansion is in fact *accelerating*, not gravitationally decelerating as would be expected in a matter-dominated universe (Perlmutter, *et al.*, 1998; Alcaniz and Lima, 1999). Another recent suggestion is that the energy contribution of what Einstein called the "cosmological constant" (Λ (lambda), essentially an "anti-gravity" force whose effects are noticeable only over large distances) could raise the total energy density of the universe to the critical value, as the popular inflationary scenario seems to require, even though the matter contribution to the total is less than fifty percent.[9]

Anything seems possible in the discipline of cosmology, and startling results like these have been overturned on many occasions. But for now, the cosmological dark matter problem seems to have been dissolved. I shall have very little more to say about it here, except in Chapter 5 where some of its candidate solutions are considered as solutions to the dynamical dark matter problem.

1.2 PLAN OF THE WORK: DARK MATTER AND EVIDENTIAL REASONING

In contrast to the cosmological dark matter problem, the *dynamical* dark matter problem has a very solid empirical foundation. Highly reliable astronomical observations of the dynamics of various kinds of structures—our own galaxy, other galaxies, and clusters of galaxies—indicate that one of two surprising and important things must be the case. Either: (1) 90-99% of the total mass of these structures has never been detected by any other means, and the extra mass is not even a *type* of matter that we have ever before encountered; or (2) our theories of gravity are in need of radical revision.

There are various levels on which the dynamical dark matter problem is interesting from the point of view of history and philosophy of science, in addition to its interest as a physical problem. Its solution seems to require a scientific revolution, either in matter

---

[9] Einstein proposed the cosmological constant, Λ, as a component of his field equations so as to make it possible to have a static (non-contracting) matter-filled universe despite the attractive force of gravity. When Hubble discovered that the universe was actually expanding, Einstein called the cosmological constant his "greatest blunder"; as it turns out, calling Λ his greatest blunder may have been Einstein's greatest blunder.

physics or gravitation theory. Also, the dark matter, whatever it is (and if it exists), neither absorbs nor emits (noticeable) electromagnetic radiation, and so (as some people claim) it could turn out to be a kind of stuff that is in principle *unobservable* by direct means. These and other very interesting philosophical issues could be fruitfully discussed in light of the dark matter case. The particular issues on which this dissertation focuses have to do with scientific methodology and evidential reasoning. The goal is to explicate and analyse the ways in which scientists use evidence to formulate candidate solutions, and to argue for and against them.

In Chapter 2, I begin the first of several lines of investigation toward this goal. In Chapter 2, I examine the physical and philosophical foundations of inferences from dynamical effects, inferences which are used both to indicate the existence of the dark matter problem, and to provide constraints on its solution. In Chapter 3, I give a selective history of crucial episodes of reasoning from dynamical effects in which missing mass problems were discovered (and in some cases solved). These historical materials set the stage for the contemporary discussion, and show that the basic techniques involved in dynamical inferences have been used successfully for a long time in essentially unmodified form. In Chapter 4, I give a summary of the current evidence indicating the existence of the dynamical dark matter problem for astronomical systems of various length scales. In Chapter 5, I examine in detail many (though by no means all!) of the candidate solutions that have been suggested, with an eye to elucidating the ways in which evidence and arguments are brought to bear for and against these candidates. One thing this discussion shows is that none of the obvious, and even some of the not so obvious dark matter candidates are viable: finding a solution to the dynamical discrepancy will be harder than we would have hoped, and the solution (if it is a matter solution) will be stranger than we would have thought. In Chapter 6 I discuss issues arising in connection with the second class of possible solutions to the dynamical discrepancy, namely revising the laws of gravity. In particular I here evaluate the import of underdetermination arguments for theory choice in this evidential context.

Throughout, but especially in Chapters 2 and 6, I develop an account of the bearing of evidence on theory choice that is (I argue) both plausible and allows us to avoid the extreme relativist conclusions spawned by some interpretations of the underdetermination

thesis. I argue that it is possible, despite the very difficult epistemic position we are in with regard to the relevant astrophysical facts, to make significant progress towards a solution to the dark matter problem. As a component of the discussion in Chapter 6, I assess the evidence for applying General Relativity to dynamical systems larger than our solar system, and find it very much weaker than is commonly supposed. This means that there exists a set of rival theories of gravitation (theories which meet a certain set of very stringent criteria of empirical and philosophical adequacy) which are viable candidate solutions to the dynamical discrepancy. Nevertheless, I argue that the only two rival gravitation theories so far offered as solutions to the dynamical discrepancy are very unlikely to beat out General Relativity plus a model of particle dark matter as the account of the dynamics of galactic and larger systems. Further, I marshal methodological arguments (going back to Newton) for provisionally assuming that General Relativity is the correct theory to apply to these large scale systems despite the lack of evidence in support of this hypothesis; this in turn allows us gather evidence and narrow down the list of candidate dark matter hypotheses, and will hopefully eventually result in a decision in favour of one of them over its rivals. Importantly, I argue that the only way we will be able to decide upon one dynamical theory over its rivals is to make use of *non-dynamical* evidence. This amounts to a theses that in difficult evidential situations assumptions about the unity of physical science and attention to far-flung bits empirical facts can lead to an increase in our fundamental knowledge of the universe. In a related vein I give an account of "higher-order evidence" and its role and epistemic power in this evidential situation.

In short, in this dissertation I describe the astrophysical dynamical discrepancy and its possible solutions, give an account of evidential reasoning, and argue on the basis of that account that matter solutions seem to stand a better chance of success given the presently available and likely future evidence.

CHAPTER 2

FOUNDATIONS OF DYNAMICAL MEASURES OF MASS

If I burn 100kg of wood, how much does the smoke weigh? Weigh the ashes, the difference is all smoke.
—Demonax, 2nd Century A.D. (As quoted in Jammer 1997 [1961], 27.)

2.0 INTRODUCTION

Dynamical evidence is at present the main source of information available to us about the nature and distribution of astrophysical dark matter (or, alternatively, about the form of the law of gravitation at galactic and greater scales). Dynamical evidence is likely to remain an important source of information even if some "direct detection" scheme bears fruit: since most particle detection schemes, for example, try to count dark matter particles passing through instruments on Earth, we thereby gather information only about dark matter in the neighbourhood of the solar system. Making that sort of evidence bear on questions about distant astronomical systems will require inferences comparing the dynamics of those systems with the dynamics of the solar neighbourhood. Dynamical arguments about the quantity of unseen mass, and about what sorts of bodies carry that mass, therefore stand in need of clarification and justification, if we are to have any evidential understanding of the dark matter case. In hopes of moving toward that understanding, this chapter discusses the physical and philosophical foundations of measurements of the masses of distant astronomical objects on the basis of the motions of other objects with which they interact gravitationally. The point here is to elucidate the patterns of reasoning involved in what I will call "dynamical inferences" (inferences from dynamical effects to the existence and description of their causes).

The plan of the chapter is as follows. Section 2.0 lays the groundwork for discussing dynamical inferences in general, including their crucial role in the discovery of the dark matter problem and in the construction and testing of candidate solutions. Section 2.1 examines the Newtonian foundations of dynamical inferences, and describes

how to use Newtonian principles to measure the mass of a central body, given the radial distances and velocities of orbiting bodies. Section 2.2 describes the measurement of the masses of planets in orbit about a central body on the basis of their mutual perturbations of each other's orbits; this section also examines the "problem of inverse perturbations," or how to infer the mass and location of an unknown body from the unexplained perturbations of the orbit(s) of some known planet(s). Section 2.3 continues this discussion by briefly considering the "composition of causes" problem as it bears on the dynamical inferences important to the dark matter case. Section 2.4 gives an introductory discussion of the concept of mass, and section 2.5 summarises the chapter. One of my main concerns is with the role of "higher-order" phenomena, which are used as evidence to constrain the characteristics of candidate solutions to dynamically determined "mass discrepancies" (situations in which methods of determining the mass of a system disagree with other observed phenomena, or with each other).

Since no theory of gravitation has succeeded better than General Relativity (GR) in accounting for (among other things) the motions of massive bodies under gravitational forces[1], strictly speaking one ought to use GR to describe and explain astronomical motions. But for the sake of clarity and ease of explication, the discussion here is pitched in terms of Newtonian Mechanics and Universal Gravitation (UG). The points made here could be recast in relativistic terms fairly easily, but doing so is unnecessary. Besides this, however, pitching the discussion in non-relativistic terms is acceptable because

---

[1] GR is confirmed by a variety of solar system tests. These tests include: the gravitational redshift of light, precessions of the perihelia of solar system orbits, the deflection of light rays by the gravitational field of the Sun, the gravitational ("Shapiro") time delay of signals passing near the Sun, and time dilation for relatively moving clocks (an effect of SR, not GR). Other potential tests include measurement of the time-variation of the gravitational constant, the detection of gravitational radiation, and the measurement of frame-dragging (Lense-Thirring) effects in gyroscopes orbiting the Earth, among others. Most of these tests confirm the predictions of GR to within *very* narrow margins of error, and none show any inconsistency with GR. See Will (1993) and Jammer (1997 [1961]) for more on the tests of GR. Note that all of these are tests take place over relatively short scales, not greater than the radius of a stellar system. This turns out to be an important fact in Chapter 6, where I consider in detail the evidential basis for thinking that GR applies to system the size of galaxies and larger dynamical systems.

astrophysicists actually do use Newtonian physics to draw conclusions about the topics discussed here. They are justified in doing this because in the weak-field, low-velocity limit, the predictions of GR agree with those of UG. This means for example that, with the exception of Mercury (whose proximity to the Sun requires a very small relativistic correction to its equations of motion), Universal Gravitation is empirically adequate for the description of nearly all solar system motions.[2] Furthermore, the very small differences which do exist between the predictions of GR and UG for solar system motions are not significant as far as the dynamical inferences described here are concerned. More importantly, astronomers use Newtonian physics to describe the dynamics of galaxies and clusters because there is nothing in GR to suggest that relativistic effects will be significant in those systems as wholes.[3] Thus UG is an

---

[2] Laser ranging of the Moon and radar ranging of other planets (Venus and Mars) now fix the distances and positions of those bodies with such high precision that the very small relativistic corrections *are* in fact required to account for their motions with as much exactness as the measurements enable (more exactly: one gets better fit to the data when using the relativistic corrections). But the relativistic corrections for all bodies except Mercury are extremely small, and Newtonian Mechanics is adequate for telescopic studies of the solar system. (Harper, private communication; see Standish, 1992.)

[3] Although supermassive black holes are apparently present in the cores of most galaxies (see Chapter 5), the relativistic effects they cause are relatively short-scale compared to the overall size of a galaxy ($r \approx 10^4$ lightyears), so that their presence within galaxies has no extra relativistic effect (beyond the simple mass contribution) on the overall orbital motions. (That the relativistic effects of central black holes in galaxies can be ignored is clear from the fact that astronomers use Newtonian physics to study the motions of galaxies.) It is these overall orbital motions that are of interest in determining the total mass of galaxies and clusters, as we shall see. Furthermore, as I explain elsewhere (especially Chapter 4), the orbital motions within galaxies and clusters are *very* much less than the speed of light, so no relativistic effects enter the picture on that score either. (The rotation of the Milky Way at our radius is only about 220 kms^{-1}, and a typical cluster (Coma) has a velocity dispersion of only about 1000 kms^{-1} (Reid, et al, 1999; Colless and Dunn 1996).) So, perhaps surprisingly, even clusters of galaxies satisfy the low-velocity, weak-field Newtonian limit of GR. One *could* use GR to determine the masses of galaxies and clusters from their motions, but since the discrepancy between the dynamical and visible masses is so great, the very small increase in precision that would result would not yield any significant new information, and the increase in precision probably would not make up for the increase of difficulty.

approximation to GR that should (according to what we think we know about the theories and the systems in question) yield empirically equivalent results for all the dynamical systems relevant to this dissertation (although the relevant evidence from gravitational lensing *does* require a fully relativistic treatment).

Dynamical inferences of two kinds are important to the dark matter problem. The first kind of dynamical inference is involved in the discovery of the dark matter problem. Many astrophysical systems, for example galaxies and clusters of galaxies, can have their masses determined by two independent techniques, namely (1) adding up all the "visible mass" present, and (2) making dynamical measures of the total mass (which necessarily includes both visible and invisible matter).[4] The dynamical measures of mass in "(2)" are really inferences from observed dynamical effects to a conclusion about a particular quantity of mass being present. In almost every case where both techniques have been applied to astrophysical systems of these types, there is an extreme discrepancy between the results of the two methods. Roughly speaking for now, all systems larger than the local region of the Milky Way are found to have dynamical masses as much as 100 times greater than their visible masses. (This discrepancy is often called "the dark matter problem", but to avoid begging questions in favour of *matter* rather than *gravity* solutions, I will usually refer to it as "the dynamical discrepancy".) Note that this kind of dynamical inference is what was involved, for example, in the discovery of the discrepancy between the expected and actual motions of Uranus: here, one compares the known masses acting on Uranus with the total force acting on the planet, as inferred from observations, and one finds the known masses insufficient to account for the motions.

The second sort of dynamical inference involved in the dark matter issue is an inference *from* some dynamical discrepancy *to* the existence of some causal factor responsible for the discrepancy. There exist two broad classes of possible causes of

---

[4] Visible mass is an estimate of the total mass of a system made on the basis of the observed luminosity; see section 3.8 for an explanation of how visible mass is computed. Visible mass calculations include a proportionality factor for the expected presence of a certain (empirically determined) amount of unseen ordinary matter. In this chapter I will take it for granted that the visible mass estimates are reliable. (Gravitational lensing is a third technique for estimating mass, and it is discussed in later chapters.)

dynamical discrepancies, namely matter solutions and gravity solutions, corresponding respectively to whether the discrepancy arises because of errors in the visible mass estimate or in the dynamical mass measurement. The degree and kind of discrepancy between the expected (visible) mass and the dynamical mass of a system gives a measure of the total additional force that is acting on the system. This additional force can have its source either in the gravitational action of an unknown massive body (or set of bodies), or it can be due to the fact that the gravitational action is in fact governed by a different law than was expected. Adams's and Le Verrier's solutions of "the problem of inverse perturbations" with regard to Uranus concluded in favour of a matter solution, and proved to be correct, but gravitation solutions were also initially possible.[5]

Dynamical inferences of the first kind (detections of dynamical discrepancies) are, leaving aside practical difficulties particular to individual observational contexts, fairly uncomplicated so long as the theories invoked in them can be considered to be well confirmed. Dynamical inferences of the second kind (from discrepancies to causes) are more difficult to construct and to justify, in part because any given dynamical discrepancy is initially compatible with a whole host of possible causes. The initial evidence indicating the existence of a dynamical discrepancy underdetermines any possible causal hypothesis that we might fix on. Narrowing down the class of possible causes is a crucial part of dynamical inferences of the second kind: doing this necessarily

---

[5] George Biddell Airy, Astronomer Royal of England at the time Adams was investigating the Uranus problem, had earlier suggested that the Uranus discrepancy could be explained if the force of gravity were weaker than inverse square at such great distances as those of Uranus from the Sun, although (so far as I am aware) no one ever worked out a complete theory along these lines (see Grosser 1979 [1962], 48). Adams, in an address delivered to the Royal Astronomical Society in November 1846, suggested that Universal Gravitation was not to be abandoned except as a last resort.

> Now that the discovery of another planet has been confirmed. . .it is unnecessary for me to enter at length upon the reasons which led me to reject the various other hypotheses that had been formed to account for [the observed irregularities in Uranus's motion]. It is sufficient to say, that they all appeared very improbable in themselves, and incapable of being tested by any exact calculation. *Some had even supposed that, at the distance of Uranus from the Sun, the law of attraction becomes different from that of the inverse square law of distance. But the law of gravitation was too firmly established for this to be admitted until every other hypothesis had failed*, and I felt convinced that in this, as in every previous instance of the kind, the discrepancies which had for a time thrown doubts on the truth of the law, would eventually afford the most striking confirmation of it. (Adams 1896 [1847], 7; italics added)

involves employing a roster of inferential and evidential techniques, especially because in most cases considered here, we find ourselves in evidence-poor situations in which a major discrepancy screams for a solution. (The evidence *for* a discrepancy is very strong, but evidence available *for solving* the discrepancy—that is, for distinguishing among all of the possible solutions—is sparse and difficult to obtain.) What one needs in order to do this is some sort of "logic of theory choice under evidential poverty," something that can provide (if not definitive reasons in favour of one candidate theory then at least) guidelines for how to proceed in the search for a solution. Given that the dynamical discrepancy situations are so evidence-poor, the theory choices made at any stage of inquiry may quite possibly be wrong. This fallibility is not, however, a decisive argument against the possibility of finding a solution: rather, constructing and testing specific and detailed potential candidate hypotheses is an important way of marshalling new and otherwise unavailable evidence relevant to the ultimate solution of the discrepancy. When a theory is found to be empirically inadequate, the discrepancy between its predictions and empirically determined quantities is powerful new evidence for the construction and testing of successor theories that will approach more closely to the truth. In many cases, opting for a solution, however tentatively, is the best (or even the only) way to make progress.

Inferences from dynamical effects to their causes are *ampliative* inferences.[6] We need to find principles of *evidence*, *theory construction*, and *theory choice* with which to *perform* and *justify* these ampliative inferences from observed dynamical effects to their

---

[6] Salmon (1967) discusses the distinction between ampliative and non-ampliative inferences, which are parallel to non-demonstrative and demonstrative inferences respectively. Valid deductive inferences are demonstrative but non-ampliative, which is to say that it is impossible for their conclusions to be false when their premises are true, but this comes at the price of the conclusion saying no more than is already contained in the premises. Non-deductive inferences are non-demonstrative but ampliative, which is to say that even when their premises are true their conclusions are not guaranteed to be true (at best they are *probable* to some degree). The trade-off for the lack of certainty is that the conclusions of ampliative inferences can go beyond what is contained in the premises. (Ampliative inferences include, for example, inductive generalisations, statistical reasoning and inference to the best explanation—non-deductive inferences are often referred to generically as "inductive" arguments (in contrast to deductive arguments.)

causes. Dynamical inferences are ampliative in several senses. (1) Because the laws invoked dynamical inferences as premises are not known with certainty, but are only confirmed by instances (broadly speaking, they are inductively confirmed laws), one reason dynamical inferences are ampliative is that the application of these laws to new cases (of the same kind) is only certain up to some degree of probability.[7] (2) Dynamical inferences often apply laws to unfamiliar circumstances for which no tests of the law have been (or even can be) made, for example the applicability of *any* theory of gravitation (including GR) to galaxies and larger systems is not merely an *untested* hypothesis, but may in fact be an *untestable* hypothesis; this leaves room for solutions to dynamical discrepancies along gravitational lines (see Chapter 6). This use of the law in a new circumstance is an inductive extrapolation of it—one involving appeal to instances in which the law was successful, the assertion of an analogy between the new instance and old ones, and the invocation of a principle asserting that the same laws are expected to hold for merely analogous (not identical) situations. Thus it is an ampliative inference to the conclusion that the law applies to the new case. (3) Even given the law, the inference from observed effect to cause is not a deductive one, because more than one cause (in fact, many different possible causes) could be responsible for producing the same effect.[8]

---

[7] Alternatively (as Smith 1999a, has argued for Newton's law of gravitation), the dynamical laws are known (to some degree of probability, by inductive evidence) to be at least approximately correct given the available evidence (that is, within some margin of error). In either case, the laws are not certain, and therefore their application even to cases of familiar kinds requires an ampliative inference. This is normally a step we take without thinking about it, but the standard examples of the possibility of the failure of inductive generalisations apply here as well, even though the reasoning involved in establishing a set of dynamical laws is in fact much stronger than is possible with simple enumerative inductions.

[8] The fact that Newton requires his Rules of Reasoning in order to make many of his arguments work shows that the deductive system supplied by the axiomatic structure of the theory (definitions, laws of motion, mathematics) is *insufficient by itself* for the purposes Newton has in mind, which include dynamical inferences of the two types under discussion here. The Rules provide (purported) principles for making and justifying *ampliative* inferences, specifically inferences about causes, inferences which are not deductively justifiable by the formal part of Newtonian Mechanics alone.

There *is* a deductive component to these inferences, namely the inference from laws (treated *as if true*) and observed dynamical effects to the magnitude and direction of the additional force acting on the system. But to go from this general characterisation of the additional force to a specific model of its cause involves assumptions (hypotheses, principles of theory selection) which make the overall inference from effect to cause non-demonstrative. So these are not merely hypothetico-deductive (H-D) inferences—it is not just a case of inventing some hypothesis and then deducing that if true it would produce the observed effects. Rather, principles and assumptions are used to adduce evidence or reasons to choose one from among the multitude of possible hypotheses: as I will discuss in this chapter, by making weak and unobjectionable assumptions, it is possible in some situations to have solutions be *constructed* from empirical data, rather than being assumed or hypothesised and then later tested against experience.

There is, in addition, a strong element in these inferences of *approximative* reasoning. As Smith (1999a) has noted, Newton's arguments in the *Principia* are often approximative: they are constructed so that even if the data turn out to be incorrect, the conclusion will nevertheless still be close to correct. For example, although the phenomena of orbital precession in the solar system are inexact (the margins of error in the observations are relatively high), Newton's inference from the lack of observable precession to an inverse square power law for the force of gravity is guaranteed to be close to correct (Smith 1999a); that is, in the expression $F \propto r^n$, given the error bounds on the positional measurements, Newton's inference ensures that $n \approx -2$. Since the margins of error in most empirical data are well known, not only is it possible to construct arguments such as Newton's so that the conclusion is close to correct even if

the data are inexact, it is also possible to give (approximate) bounds on how far from the truth the conclusion could be.[9]

In Smith's words, what Newton gives us is a way to do "exact science via successive approximations": systematic discrepancies between the predictions of the initial theory and the observed phenomena[10] *themselves* become "second order phenomena," new evidence relative to the initial, approximative theory (Smith 1999a). In fact, these discrepancies lead to evidence which is more powerful than what is possible to obtain before we have the first approximative theory. With this new evidence in hand, we are able to construct a more exact successor theory. Ideally, by successive comparison of the predictions of better and better theories with new or more precise observations, we discover ever more fine-grained discrepancies, through which we construct ever better theories. This process is limited by our ingenuity at coming up with new theories, and by the margins of error in the data (which restrict the evidential significance of empirical discrepancies since they prevent us from making finer distinctions among theories in terms of empirical success).

Let me explain in more detail what I mean by "higher order" data, phenomena or evidence. By "higher order" data I mean evidence that goes beyond, or requires further analysis of, or involves juxtaposing additional background assumptions with, lower-level

---

[9] Thus, when Duhem and Popper notice that the observations of the orbits of the planets show that they do not actually obey Kepler's three laws, which are deductive consequences of Newton's theory of gravitation, they are wrong to conclude that this means the observations show that Newton's theory is *false*. As Smith (1999a) puts it, what Duhem and Popper fail to appreciate is that Newton's premises are *approximate* rather than *false*: I would add that they also fail to notice that the arguments involved are ampliative rather than demonstrative.

[10] Harper defines "phenomena" by noting that "Newton's phenomena are not just data. They are generalisations, or regularities, that can be expected to remain reasonably stable as new data are added" (1997b, 1). Phenomena, then, are generalisations fitting open-ended bodies of data; we may add, they support counterfactuals claims. An example of a phenomenon in this sense is the theoretical description of a planetary orbit. See also Woodward (1989) on the data/phenomena distinction, and my critique of Woodward below.

observations.[11] Higher-order evidence is gathered from the *subsequent analysis* of (a series of) first-order observations; we discover *patterns in the data* which are then themselves not just new evidence but a *new kind of evidence*. This is information that can be extracted from the data (or from relations between different data, or from relations between data and lower-order predictions), but is not identical with the bare data itself—in particular, higher order evidence does not manifest itself within a single piece of data or a data point. Rather, this information arises out of consideration of several pieces of data taken together, usually in relation to some theoretical background. (Data require arguments to turn them into evidence, and these arguments are mediated by theory.)

The difference between higher and lower order evidence is like the difference between a set of geocentric positions of some planet, and a characterisation of its orbit. For example, a set of geocentric positions of Mars might be the first-order "data", in which case the orbital path of Mars over time would be higher-order data, and might for example bear on the evidential status of some still higher-level hypothesis, say Kepler's Third Law or the mass of the Sun. Higher order evidence, then, is separated from lower order evidence by one or more layers of inference (usually ampliative).

My distinction between first and second order evidence is parallel to Woodward's (1989) characterisation of the relationship between data and phenomena. One difference is that whereas Woodward sees data as evidence for a phenomenon and stops there, I want to emphasise that a phenomenon in Woodward's sense can be used as evidence for yet another, more general (or more abstract, usually more speculative) phenomenon. So a "phenomenon" in Woodward's sense can function as (or *is*) "data" relative to some higher order phenomenon even though it is a hypothesis relative to lower order data. Furthermore, what Woodward calls "data" are often "phenomena" relative to the evidence and inferences that go into their construction. It is not the case that data are just

---

[11] My talk here of "higher order evidence" should not be confused with the notion of calculating some quantity "to higher order", which this has to do with calculating an approximation to a greater number of terms of a series, where subsequent terms are usually "higher order" in the sense that their exponents are higher numbers. Taking an approximation to higher order terms in general yields a more accurate result. See the note on the convergence of series in the section below on perturbation analysis.

"given": there is a relativity in the usage of "data" and "phenomenon": we call something data when it functions as evidence for something else, but many things that we *call* data are really phenomena relative to the (lower level) evidence that supports them. For example, the velocity-radius data points involved in the construction of galactic rotation curves are in reality the outcomes of a very long chain of evidential reasoning involving background theories of the camera and telescope, optics, electronics, and a whole host of other assumptions and hypotheses: it is only at the end of a long chain of ampliative reasoning that one can say that the photons received in the CCD camera provide velocity-radius data. In other words, these "data" are just as much *hypotheses founded on lower level evidence* as are phenomena.

Since the degree of uncertainty generally increases as one climbs this ladder of evidential inferences, I want to make a point of marking the differences between the steps. Also, and perhaps more crucially to what I think is distinctively important about higher order evidence, higher order evidence does not depend on gathering *new* observations but rather on analysing, comparing and looking for patterns within the observations already to hand. To take an example, data about the velocity of rotation of a galaxy at various radii can be turned into a rotation curve: the rotation curve in turn can be used as evidence about the overall distribution of matter in the galaxy. Note individual velocity-radius data points can only be used to determine the total mass interior to that radius. Only once we have the whole rotation curve, which is constructed as a curve-fitting problem given the velocity-radius data, can we acquire information about the overall mass distribution.

Note, too, that being a higher order phenomenon is not necessarily correlated with being more accurate or more reliable. The rotation curve for a galaxy permits only a fairly rough determination of the dynamical parameters of the galaxy (for example, it predicts the mass only to within about a factor of two). But higher order evidence *does* allow us to acquire knowledge that we could not otherwise have, given the lower-order evidence by itself. This is not a case of acquiring more, or more accurate, or more reliable evidence about the mass distribution than we had from the lower level evidence, since that lower level evidence provided us with *no* information about the mass distribution. The higher order evidence provides support for some still higher level

hypothesis depending in part on the strength of the inference from the data to the first order phenomenon, and of the inference from the first order phenomenon to the second order phenomenon, and so on. Evaluating the epistemic status of the information claim which is the outcome of such a reasoning process will involve paying detailed, case-by-case attention to the margins of error in and the epistemic warrant of each evidential inference in the chain.

A special case of the use of higher order evidence is discussed by Smith (1999a), who argues that Newton's method of reasoning from phenomena makes it so that discrepancies between first-order theoretical predictions and the corresponding observations count as second-order phenomena measuring the parameters of a successor theory to a higher degree of precision. That is, discrepancies between observations and a theory can themselves become (one type of) higher order evidence relevant to the construction and testing of a successor theory. Think of the Mercury example: the left-over precession in the Newtonian analysis became a key piece of evidence in confirming the successor to Newton's theory. We certainly do not observe the precession, let alone the excess precession—for that matter, we do not observe the orbit either: rather, we infer these things from lower level evidence and theories. (In the Mercury example, geocentric positions are reduced to heliocentric positions with the help of a theory of the Earth's orbit, and then a complex perturbation analysis can be undertaken, and so on.)

Smith's specific examples of discrepancies between theory and observation functioning as higher order evidence fall under the general pattern of higher order evidence I have described here. One interesting point that Smith makes is that since (by definition) successor theories become successively better and better, the discrepancies between observations and predictions will become smaller and smaller with each generation of theory. And thus the next theory must be even more precise in order to be able to save the higher order evidence that is the discrepancy. (In this respect, Smith's examples are of an especially epistemically powerful case of higher order evidence.) However, I want to emphasise a different aspect of higher order evidence, namely the fact that constructing higher order phenomena makes available (kinds of) evidence that could not otherwise be obtained. Both aspects of higher order evidence test theories more stringently than lower order evidence, in the sense that it is harder for a theory to save

more (kinds of) data as well as *more precise* data: both the kinds of evidential improvement one gets from higher order evidence restrict the class of viable rival hypotheses to a smaller group, and thus help in the task of theory choice and testing. But sometimes an increase in the precision with which a theory must save the phenomena will be a less serious test of the theory than would introducing a new kind of evidence. As I will discuss in Chapter 6, it seems as if the (first order) dynamical evidence will be insufficient for us to make progress on the choice problems involved in the dark matter debates, but higher order evidence of various kinds does promise to improve our epistemic situation with regard to the dark matter problem.

Under this conception, the astrophysical dynamical discrepancies become a source of higher order evidence with which to construct a better theoretical model of galaxies and clusters. Assumptions built into the first-level hypotheses used in the inference determine *which type* of solution (matter or gravity) the discrepancy becomes evidence *for*. This might seem to be a major flaw in the procedure, a kind of begging the question in favour of one class of solutions, but it is not, for two reasons. First, since what Smith calls "the method of exact science via successive approximations" yields self-correcting or self-improving theories (Smith 1999a), we will in principle be able to find out when we have opted for a mistaken hypothesis—if the hypothesis is mistaken, subsequent investigation will reveal a new discrepancy, or the hypothesis will not be able to account for some significant part of the totality of the available evidence. (All empirical hypotheses are of course fallible. But, more importantly, on this conception empirical hypotheses are *corrigible*.)

Second, since the initial evidence is consistent with *both* kinds of solutions, we can perform the inference with both kinds of assumptions, and use the discrepancy as evidence for constructing and testing solutions of both kinds, whose relative empirical success we can then compare against each other. In the dark matter case, the available evidence is so sparse and so radically underdetermines a solution, that both of these procedures are likely to be used at the same time (though probably by different workers). All of the things we learn this way become higher-order evidence for the construction and testing of subsequent solutions (which are successively better approximations). A key point here is that in evidential situations such as we are in with regard to the dark matter

problem, we are unlikely to make *any* progress toward a solution *unless* we construct and use approximative solutions as an interim step in the manner just described.[12] The information we acquire via these approximative solutions remains even once the approximative theory has been superseded or falsified.

This pattern of exact science by successive approximation, furthermore, enables us to say with confidence what must be the case (within certain bounds) if the solution we choose turns out to be wrong. (For example, "unless there exists dark matter with such-and-such characteristics, gravity must be so-and-so".) The further we pursue the chain of approximation-discrepancy-new evidence-new approximation, the more (and more exact) information about the world we get, and therefore the more detailed and exact the "unless..." statement becomes, which is to say that the class of viable candidate solutions gets continually narrowed down by this process. Note, though, that however far we carry this process, we are not thereby able to escape the underdetermination problem. We merely end up with a choice between rival hypotheses all of which account for the available evidence to higher precision than before.

As Harper has said, the real "Newtonian revolution" was in scientific *methodology*. In contrast to what Kuhn has claimed, Newton's ideals of empirical success *do* persist across the radical theory change of the Einsteinian revolution: these same ideals govern experimental practices in gravitational physics even today (see Harper 1997a, Will 1993). An important point in Smith's account of Newtonian methodology is that being able to do exact science via successive approximation means that almost all the epistemic risk is in the inductive generalisation (of the laws or parameters measured by phenomena) to other (analogous) systems. Interestingly, this is exactly where the Newtonian theory of gravity fails with regard to Mercury, and it is also where General Relativity potentially fails with regard to galactic and cosmic phenomena.[13] Knowing this fact about the epistemic

---

[12] We may do this in the first instance by employing an H-D inference, but we achieve better empirical success by looking for and exploiting for evidential purposes discrepancies between the first order theory and observational data.

[13] What I mean is that while the parameters of Newton's gravitation theory are measured to high precision by various phenomena in the solar system, the inductive extension of the theory (with those parameters) to

"weak link" in the chain of scientific inferences tells us where to look for new severe tests of current hypotheses—that is, it tells us where new discrepancies (which will lead to new knowledge and better theories) are likeliest to be found.

2.1 THE FOUNDATIONS OF DYNAMICAL MEASUREMENTS OF MASS

Newton's First Law of Motion, "Every body perseveres in its state of rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by forces impressed" (Newton 1999 [1726], 416), *defines* what happens to bodies not subject to forces (thus defining inertial motion at the same time): a body not subject to forces moves in a straight line with a uniform velocity; any body in non-rectilinear or non-uniform motion must (therefore) be subject to forces. The Second and Third Laws tell us how bodies subject to forces accelerate, thus simultaneously defining what a force is, and enabling us to make inferences from observed accelerations to the forces that cause them. Newton proves in Corollaries One and Two to the Laws that the overall acceleration experienced by a body on which multiple forces are acting simultaneously is found simply by adding up the accelerations that would be produced by each force as if it were acting independently. These corollaries validate the application of the technique of vector addition of forces to actual physical systems. Using vector addition with the theory of Universal Gravitation we are able to distinguish observed effects due to known bodies from effects which must be accounted for by hypothesising the existence of additional, unknown sources of force (the problem of the composition of forces notwithstanding: I will ignore this problem in what follows and treat it separately in section 2.3). Since, on Newton's theory of Universal Gravitation, all gravitational forces have their sources in bodies, and these gravitational forces are proportional to the masses and to the inverse of the square of the separation (by $F = GMm\,r^{-2}$), we can use Newton's laws to determine for a given dynamical system the total force acting whenever we observe a body experiencing gravitational acceleration. Furthermore, when we know the

---

the new case of Mercury, which as it turns out is significantly different from the other planets in that it is close enough to the Sun for relativistic effects to be important to its motions, ends up yielding incorrect predictions. As this dissertation develops the similarity of this case to the case of galactic motions will become clear.

location of the gravitational source as well, we can determine the mass of the source. The more complicated case of reasoning to the mass of a source when its location is *not* known is discussed below.

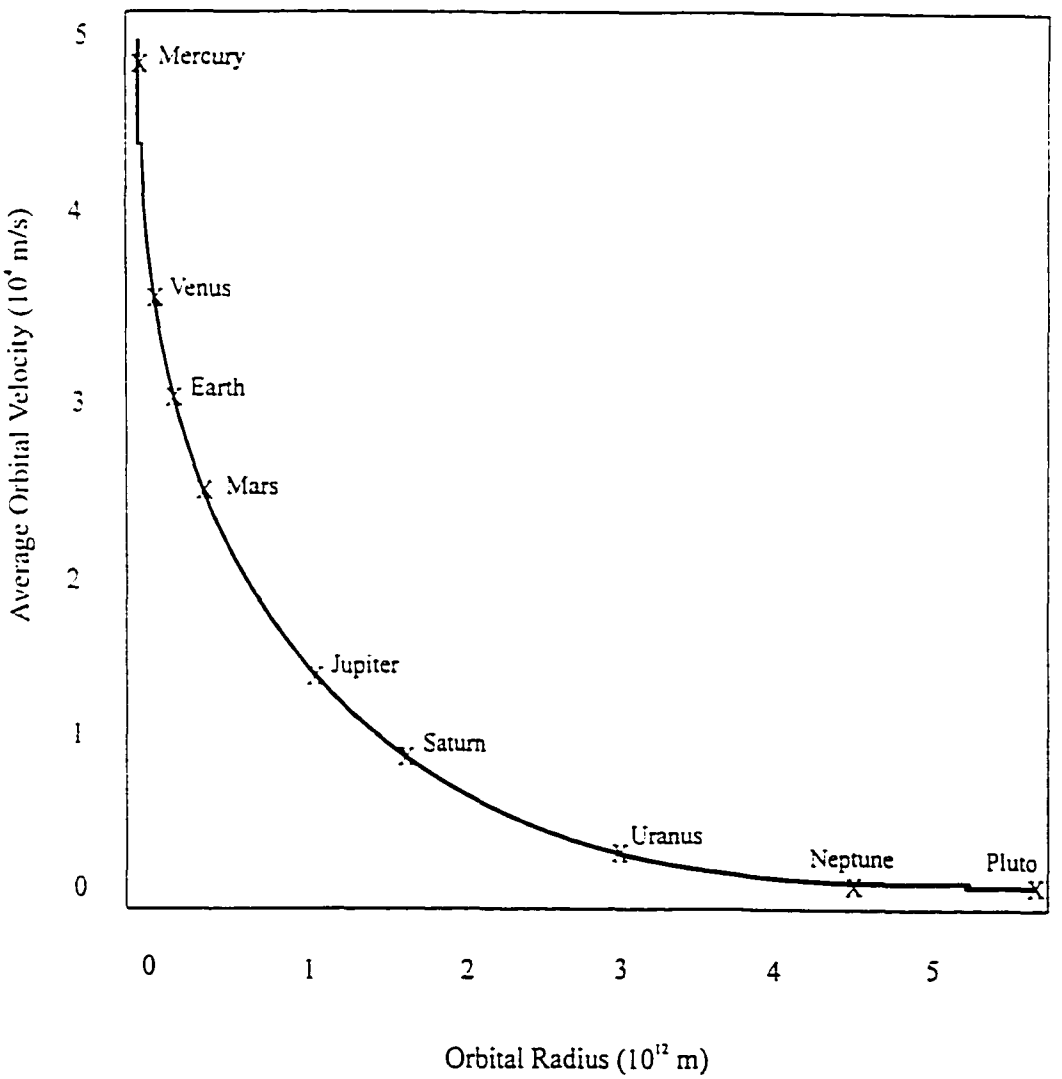
"Sourceless forces" are ruled out in Newtonian Mechanics by the very definition of "force" (where that definition is taken to be given by the three Laws of Motion together). Actions on bodies (which are mediated by forces) are taken to always produce equal and opposite reactions. The reaction involves the motion of another body, namely the source of the force (its motion is in proportion to the force impressed and inversely proportional to the inertial mass of the reacting body). Hence, any time we find a force, there must exist some *body* responsible for that force. (A force without a source is not a force at all: Newton's treatment has it that centrifugal forces are pseudo-forces precisely because there exists no body which suffers an "equal and opposite" reaction.)

Because there are circumstances in which Newtonian Mechanics directs us to assume the existence of sources, whether or not we can observe them, Newtonian Mechanics *requires* us to admit into our scientific account of the world the existence of unobserved and even so-called *unobservable* objects (objects for which we have no other evidence of their existence and nature, and of which we have not made or cannot make independent detections) when empirical conditions of a certain kind arise. Thus, methods originally developed by Newton for use with respect to interactions among bodies already known to exist, are adaptable to consideration of as-yet-unobserved and even unobservable entities, at least in the special case of invisible bodies which have a discernible gravitational influence on visible matter. I will argue that our epistemological grounds for accepting assertions about theoretical entities of this kind are not different from our grounds for accepting assertions about ordinary objects, at least to the extent that the inferences involved in acquiring knowledge of both kinds have exactly the same structure (though they may differ with respect to the margin of error or degree of confidence we attach to their conclusions).

Newton showed that Kepler's Third (or "Harmonic") Law ($T^2/R^3 = k$, that is, the ratio of the square of the period of an orbit to the cube of its mean radius is a constant for all planets in the solar system), which Kepler discovered empirically, is a deductive consequence of the inverse square centripetal gravitational attraction (Cohen 1985, 164-

65; Newton 1999 [1726], especially Propositions I through VI). Thus Universal Gravitation predicts that a graph or "rotation curve" of the velocities of the bodies orbiting the Sun, as plotted against their respective distances from the Sun, will decrease to zero as the radius goes to infinity: $v_r \propto r^{-1/2}$. This form of the rotation curve is called "Keplerian" because of the Harmonic Law relation just mentioned. (See Figure 1.) If we were to use this rotation curve prediction as a test of Universal Gravitation for orbital motions around the Sun, we would find that all nine planets fit the Keplerian form for the solar system rotation curve rather exactly, and hence we would find that the predictions of Universal Gravitation are confirmed to a high degree by the orbital motions in our solar system.

Figure 1



Rotation curves are interesting for our purposes because they can be used to calculate the mass of the central body. The rotation curve for the planets in our solar system, for example, at one and the same time depicts nine independent and precisely agreeing measurements of the mass of the Sun. As Lawrence Krauss notes, the independent measures agree to a very high degree of precision:

> In my entire career in physics, I have only once otherwise seen such good agreement between data and theory. . . . Newton's law of gravity works! From data such as these, we find that the mass of the sun is about 2×10^{30} kilograms. The accuracy of this value is limited by our uncertainty in Newton's constant, G. Were it not for this uncertainty, the planetary data would in fact allow one to determine the mass of the sun to better than one part in a billion. (Krauss 2000, 64)

The fact that rotation curves can be used to measure the mass of the central body follows from the second part of the Newtonian principle stated above, namely, that the gravitational attraction between any two bodies is directly proportional to the product of their masses. Newton's pendulum experiments provided the first good evidence for a hypothesis famously proposed by Galileo, namely that all bodies in a homogeneous gravitational field fall at the same rate, regardless of their masses. Given the truth of this hypothesis, and the fact that the strength of the gravitational field in question (which defines the rate at which all bodies in the field will fall) is determined by the mass of the body generating the field, it follows that the mass of a body falling in the field disappears from the calculation of the mass of the body generating the field. So the radial distance and velocity of an orbiting body together are sufficient to allow one to calculate the centripetal acceleration deflecting the orbiting body from its inertial motion, and thus to find the strength of the gravitational field, and therefore the mass of the central body. The "normalisation" of the rotation curve (its magnitude, or height above the origin of the graph, which is given by the actual velocity of the rotation at each distance), then, gives us a direct measurement of the mass of the Sun. The most general equation one can use here is $F = GMm\,r^{-2}$; or, since $F = ma$, we could use $ma = GMm\,r^{-2}$, which reduces to $M = ar^2\,G$ (but see the final section of this chapter for an important note about this).[14]

---

[14] As Harper, Bennett and Valluri (1994) describe, Newton showed how to use Kepler's Harmonic Law ratio, $a^3/t^2 = k$, where a is the semimajor axis of a solar system orbit and t is the period, to measure the

In the case of trying to determine the mass of a galaxy or cluster from its rotational motions, it is practical to switch from the "force-based" account described above to an "energy-based" approach (derived in its original formulation from statistical mechanics). (The explanation that follows is based on Tayler 1991, especially Appendix 3; see also the more technical discussion in Binney and Tremaine 1987, 211-14.) The energy-based approach I am speaking of here involves the (Scalar) Virial Theorem, which for a system whose overall properties are time-independent is $2KE - GPE = 0$: twice the kinetic energy is equal to the negative of the gravitational potential energy (and gravitational potential energy is defined to be negative, so the two negatives cancel).[15] For a spherical system of total mass M and radius r, the gravitational potential energy is given by $GPE = -\alpha GM^2 r$. Its kinetic energy is given by $KE = 0.5M\langle v^2 \rangle$, where $\langle v^2 \rangle$ is the mean of the squares of the velocities of the particles in the system (where we consider the system from a frame of reference in which its centre of mass is at rest relative to us). By the Virial Theorem, then, $M\langle v^2 \rangle = \alpha GM^2 r$, and we find $M = r\langle v^2 \rangle \alpha G$, where M is the mass interior to the radius r, G is the gravitational constant and "the value of α depends on the mass distribution in the system but is usually of order unity" (Tayler 1991, 194). Note that in astronomical studies the equation $M = r\langle v^2 \rangle \alpha G$ is much easier to use than $M = ar^2 G$ because it is very much easier to measure the velocities of a distant group of

---

mass of the Sun. The relevant equation is $a^3 t^2 = (k^2 4\pi^2)(1+m)$, where 1 is the mass of the Sun, m is the mass of the planet (together with all its moons) in solar masses, and k is Gauss' constant. This can be made more accurate by taking account of the fact that periods and planetary masses vary with time (because of accretion of solar system material, or destructive collisions). Harper, Bennett and Valluri (1994, 133-36) show how the five other planets known to Newton give measures of the mass of the Sun that agree with the result one obtains from the Earth's orbit by this method. This is certainly an interesting technique, but it is useful only for the solar system.

[15] The Scalar Virial Theorem was first proved by R. Clausius in 1870 (Binney and Tremaine 1987, 213). Note that the Virial Theorem applies only to systems whose global properties are time-independent: this is to say that the forces involved must be conservative, and that the system is closed and in equilibrium. Thus when we apply the Virial Theorem to galaxies and clusters we are treating them as frictionless, collisionless gases in which gravity is the only important force. This is a good approximation given what we know about the matter distribution in these systems and the short range of non-gravitational forces.

particles (using the Doppler effect) than it is to measure their centripetal accelerations. This equation makes it easy to calculate the mass of an astronomical system once the average velocity of its particles (stars in a galaxy, or galaxies in a cluster) and the radius of the system are known.[16]

On both the force-based and energy-based approaches, finding the mass of a central body given the distance and velocity of an orbiting body requires being able to assume that the system in question is in "gravitational equilibrium". A system is in gravitational equilibrium when the velocities of bodies in the system are exactly balanced by the gravitational forces: in effect this means that the configuration of the system is *stable*. Unstable systems either collapse onto the central body or evaporate into space: if the gravitational attraction over-balances the velocities, the orbiting bodies will eventually spiral into the central body; if the attraction under-balances the velocities, the orbiting bodies will eventually escape into space.

For most scenarios of the formation and evolution of gravitationally bound astronomical systems, there will be an initial "equilibrating period", during which the orbits of bodies in the system may change radically, but eventually the system will "relax" to a stable form (this may involve a certain degree of evaporation or collapse). Imagine a body moving too quickly to be held in orbit at its present distance from the

---

[16] We can choose units so as to set $G = 1$, and since (according to Tayler) the factor α is usually close to unity, its contribution to the mass is swamped by the dynamical discrepancy and therefore can be ignored in the cases of interest here. It is easy to see that $r\langle v^2 \rangle$ is therefore always close to an exact measure of the mass interior to r. And both r and v are empirically determinable to high precision: v is nearly exact, thanks to the high precision of Doppler shifts derived from astronomical spectra, and the greatest error in r is the uncertainty in the Hubble constant (Krauss 2000, 49; the Hubble constant comes in since we need to know the distance to the astronomical system under study in order to be able to convent its angular size to a diameter in light-years). The precise value of G only becomes important when we want to convert relative masses determined dynamically to kilograms; astronomers avoid this step by using "solar units" for mass and luminosity.

central body but below the escape velocity for the system.[17] Such a body will move to a higher orbit. A corresponding process occurs for a body whose velocity is too low for its initial distance but above the value on which it would eventually impact on the central body. Such a body will move to a lower orbit. In order to accurately apply the dynamical techniques for determining mass that are described here, we must be able to assume that this equilibrating epoch is over. Otherwise, we cannot be sure that the velocities observed at given radii correspond in the right way to the centripetal accelerations at those radii, that is, we cannot be sure that the velocities measure the central mass. The margin of error this assumption introduces, if false, is small. Systems very far from equilibrium will evaporate or collapse on very short time scales, cosmically speaking; given the age of most astronomical systems, the fact that those systems are still *apparent* systems indicates that they must be quite close to gravitational equilibrium, if not actually in equilibrium.

For systems in gravitational equilibrium, then, there is a unique circular orbital velocity corresponding to each radial distance at which a body could be in orbit. What this velocity is, is determined by the strength of the gravitational field at that radius, and therefore depends only on the mass of the central body (the gravitational power law is here assumed to be Newtonian). So, in fact, what is going on in the solar mass measurement described above is that the orbital distance gives us an orbital velocity, and that orbital velocity *measures* the force of gravity that must be acting on the body to keep it in orbit, on the assumption that the system is in equilibrium. Since the force of gravitational attraction varies directly with the mass of the gravitating body, measuring the force is measuring the mass. In sum, only three things are required, in conjunction with Newton's Laws of Motion, in order to measure the mass of the Sun. We need to know the radial *distance* and *velocity* of an orbiting body, and we must be able to make an assumption of *gravitational equilibrium*.

Note that we have reasonably good evidence for the hypothesis that the solar

---

[17] It is worth mentioning that there need not be a "central body": an individual star, say, moves under the collective potential of all the mass in its parent galaxy. But for the sake of the example I will continue to speak about systems that do have central bodies in which most of the mass of the system is found.

system is in gravitational equilibrium. Several thousand years of observations demonstrate the relative stability of the planetary orbits. Furthermore, considerations from planetary formation, geological evolution, and the existence of life on Earth make it clear that our orbit has been at least relatively stable for *billions* of years. (If the orbit were unstable, Earth would have drifted out of the narrow "habitable band" of the solar system, where the amount of sunlight received per unit area is in the range sufficient for life.) Stability over such a long a long period can only reasonably be explained by the solar system really being in gravitational equilibrium.[18] We make an inference from the *apparent* long-term stability of the orbits of the solar system to the fact that the solar system *is* in gravitational equilibrium. The bounds on stability (how unstable the orbit *could* be, given the length of time for which we have observations, and considering the precision of those observations) give bounds on how far from equilibrium the orbits could be, which in turn contributes to the margin of error in the dynamical mass measures. Given the available evidence, and the degree of confidence we have in our background theories, this error is taken to be very low. A parallel argument applies to galaxies and clusters.

The possibility that a given galaxy (or perhaps even cluster) under study has recently been knocked out of equilibrium by a transient massive body (say, a near collision with another galaxy, a large black hole, or another cluster) cannot be ruled out. From the number of galaxies and their line of sight velocities one can calculate rough odds of (for example) galaxy-galaxy collisions, and one finds that the ratio of the number of collisions at a given time to the number of galaxies will be low. Taking an average over many systems of similar type makes the possibility of disequilibrium have a negligible effect on estimates of dynamical mass: *all* spirals, for example, show a huge dynamical discrepancy, but our best reasoning indicates that only very few of them are likely to be out of gravitational equilibrium at a given time. Trimble (1987, 441) notes

---

[18] The stability of planetary orbits also puts limits on the number and size of "transient" bodies that have passed through the neighbourhood of the solar system in the time since the formation of the planets; this can be used as evidence relevant to some dark matter candidates, see Chapters 4 and 5.

that if we just wait a billion years, we will be able to determine empirically whether a given cluster is in equilibrium.

Under what conditions would the method described here for determining the mass of a body from the motions of a body (or set of bodies) orbiting in it, *fail*? One condition has already been mentioned, namely if the system is not in gravitational equilibrium. In that case the velocity is not a measure of the gravitational potential (and therefore of the mass), and so the mass we infer from the motions will be mistaken. A second condition that would defeat the dynamical mass measure is if a significant proportion of the mass of the system is not interior to the orbit in question; for, as Newton showed, a body anywhere within a spherical shell of matter feels no net gravitational force from that shell of matter. (See Binney and Tremaine 1987, 34-6. Note, as will be discussed in subsequent chapters, that our best evidence indicates that the mass distribution in galaxies and clusters *is* spherical, so the following comments apply to those structures although they do not apply to structures that are more like the solar system.) This means that given a smooth spherical distribution of matter, the velocity of a body orbiting the centre only measures the mass interior to that orbit. If in fact the system has a significant proportion of its mass outside that radius, the dynamical measure will underestimate the total mass of the system—the circular velocity does not measure the total mass if there exists significant mass exterior to the radius studied. (Let me emphasise that this means there could be much *more* dark matter in galaxies and clusters than our dynamical measures indicate—the dynamical mass is a minimum mass: see Chapter 4.) A third way in which dynamical measures could fail is if the force law of gravity is not what the Newtonian limit of General Relativity says it is—obviously we will get the dynamical mass wrong if we use an incorrect dynamical law.

Suppose we discover a system whose rotation is non-Keplerian. At least one of three background assumptions must be incorrect. Either the system is not in gravitational equilibrium, or the mass is not mainly concentrated interior to the orbits being considered, or the force governing its dynamics is not an inverse square attraction. (Several kinds of astrophysical systems, notably galaxies and clusters, have radically non-Keplerian rotation patterns; see Chapter 4.) The discovery of such a non-Keplerian system does not, however, merely imply the simple falsification of all or some of these

three background assumptions. Rather, the kind and degree of the discrepancy between the Keplerian expectation and the observed dynamics yields information about the system, information which can be used to construct alternative causal hypotheses or new sets of dynamical assumptions. That is, when we know that the velocity at r is not the value e (expected given the distribution of visible matter and a dynamical law), we usually also know that it has some *specific* value: we know, for example that $v_r = e - \Delta$, not just that $v_r = e$. The background theories and the phenomena are chosen or constructed so that "systematic dependencies" are present. This means that we are able to construct arguments to support detailed counterfactuals of the form "if the phenomena have alternative values, then they measure certain theoretical parameters to have different values". (For more on the role of systematic dependencies in evidential reasoning see below, Harper 1997a, and Chapter 6.) In this way, the difference Δ either measures the degree to which the force law differs from $n = -2$ (as it would have done in the Mercury case), or it provides information about the mass distribution in the system (as in the Neptune case).

Which horn of the dark matter dilemma we should pursue is not obvious; this is a variation on the problem in H-D falsification of where the "arrow of disconfirmation" is to point (see Chapter 6 for a discussion of this). In actual observing situations, the three requirements for making a dynamical mass measurement of a central body (radius, velocity and equilibrium) are not known with certainty. We know the radius and velocity only to within some margin of error; similarly we merely have plausibility arguments in favour of equilibrium, arguments that establish likely maximum differences from equilibrium. Given these margins of error, one can calculate maxima and minima for the enclosed mass. This means that (as is always the case for any kind of measurement) the empirical data on the dynamics of a system actually measure a range of possible values for the enclosed mass. If the margins of error in such a mass measurement were quite large, it could turn out that an apparent discrepancy between a theoretical expectation and a dynamical measure was illusory. There is no risk of the dark matter problem being an illusion of this kind, even though we do not know the dynamical mass with perfect precision: the dynamical mass is roughly 100 times the visible mass, and the likely error in the dynamical mass is only about a factor of two.

2.2 DYNAMICAL MASS MEASURES VIA PERTURBATIONS OF ORBITS

I have so far discussed dynamical measures of the mass of a central body orbited by one or more other bodies. Let us now consider how the masses of the orbiting bodies themselves can be dynamically determined. There are three possibilities, only the third of which is really of interest here. First, in the case of a planet with its own satellites, the techniques already described apply: the motions of the satellite(s) measure the mass of the planet. Second, in a case where the ratio of masses of the orbiting body to the central body is high enough, one can measure the mass of the orbiting body by observing the amount by which it displaces the central body according to the law of action and reaction. But unless the ratio is quite high the effect on the central body will be too small to measure reliably. Third, where there are two or more bodies in orbit about a central mass, we can determine the masses of the orbiting bodies by their gravitational interaction, as manifested through their effects on each others' orbits. (For bodies without satellites, this is the only method available.) Such interactions are called *perturbations*. A perturbation on an orbit, since it is caused by a gravitational attraction that is inversely proportional to the separation and directly proportional to the mass of the body causing it, is a measure of that mass provided that the relative positions of the two orbiting bodies are known. The precision of the mass measurement from a given perturbation depends on the precision with which the basic Keplerian orbit is known, on how well other perturbations can be separated out, on the degree of exactness with which the excess deviation from the expectation is known once the other perturbations have been subtracted, and on how well the distances between the bodies over time are known. The margins of error in these quantities determine the margin of error in the perturbational mass measurement.

Since all bodies attract one another gravitationally, whenever an additional massive body is present, it will have an effect on the shape of the orbit. In some cases, however, the perturbation of one body's orbit caused by another body will be too small to be measurable—just how small a deformation is measurable is defined by the margin of error in the positional observations (which determines the degree of precision with which the Keplerian orbital parameters are known, and which thus defines the smallest empirically significant deviation from the Keplerian orbit). Conversely, given the

Newtonian theory of gravitational interaction, any deviation from a Keplerian orbit must be caused by the gravitational action of some other body or bodies. The magnitude of the perturbation is determined by the mass of the perturbing body (or bodies) and is proportional to the inverse square of the separation; the way in which the perturbation manifests itself determines the direction of the centre of mass of whatever is causing the perturbation. If the direction and distance of the body causing the perturbation are known, therefore, the discrepancy between the (two-body) expectation and the actual orbit of one body can be used to measure the mass of the perturbing body. In this way the individual effects of the various planets in the solar system on each other's orbits can be decomposed. We know we have a complete catalogue of the (dynamically significant) objects in the solar system when all of the observed orbital motions are accounted for by assigning a consistent set of masses and distances to all the known bodies in the solar system. Since the distances between bodies at specific times can be determined trigonometrically from geocentric observations and the masses of most bodies can be dynamically determined in multiple ways (by systems of multiple perturbations, and by the orbits of moons), the evidence rather tightly constrains our account of the dynamics of the solar system.[19]

---

[19] The mathematical details of perturbation analysis are beyond the scope of this dissertation, but I should say a few words about what is involved in general. Using the Newtonian equations of motion one can in principle predict all future and past states of a system provided that one knows the momenta and positions of all the particles of that system at a given time. However, n-body integrations (for $n > 2$) cannot be solved analytically. Mathematicians therefore approximate a given n-body solution by doing the sum of a series with an infinite number of terms. Obviously one cannot actually complete an addition of an infinite number of terms. However, the approximation can be taken to arbitrarily many terms, which means that the various dynamically important quantities can be specified to an arbitrary degree of precision. Taking just a few terms is often enough to get results that are more precise than the margins of error in the observations. (Here is a case of "exact enough" science by approximation.) The catch is that in order to be sure the approximation yields results that will be accurate over very long periods, one must know that the sum of the series converges to a specific value rather than summing to infinity; it is not always known that the series being used will in fact converge. This means that long term predictions may be incorrect (in the short term we can test the adequacy of an n-body approximation by comparing it to observations). As Peterson writes,

We now come to a case of special interest to the present project, namely situations in which we observe perturbations that cannot attributed to the effects of known bodies. What can be done in such situations? The fact that a perturbation exists tells us that there is a force acting on the perturbed body; furthermore, the characteristics of the perturbation can be used to precisely determine the strength and direction of the total force producing it. (Recall that I will treat the problem of the composition of forces later.) The question then is to find the cause of the force. (Whereas discovering the unexplained perturbation involves dynamical inferences of the first kind, determining the cause of the perturbation involves a dynamical inference of the second kind.) The problem of determining the masses and positions of such unknown bodies on the basis of their gravitational effects on other bodies orbiting a common centre is called the problem of "inverse perturbations".

Two especially notable instances of this have arisen in planetary astronomy. One involved Adams's and Le Verrier's independent successful predictions (in 1846) of the existence and geocentric position of the previously-unknown planet Neptune on the basis of unexplained perturbations of Uranus's orbit. The other was Mercury's famous 43" century excess perihelion precession that could not be explained by Newtonian Mechanics on the basis of interactions with known bodies.[20] Both cases prompted attempts to perform an inverse perturbation analysis to account for the respective discrepancies by inferring from those discrepancies the existence of hitherto unknown bodies. In one case, the prediction of the unobserved body was a stunning success; in the other, because no such body could be detected observationally or even described in a way

---

> In celestial mechanics, approximate solutions of the. . .equations of motion yield infinite series expressed in terms of such variables as the orbit's eccentricity or some other orbital parameters. Mathematical astronomers evaluate such expressions to as many terms as they believe necessary to make predictions of a certain accuracy. In some instances, however, they have no proof that the series they used actually converges. (Peterson 1993, 146)

For more on perturbation analysis, including chaotic perturbations in the solar system, see Peterson 1993.

[20] Le Verrier discovered this discrepancy in 1843, and put it at about 39" century; Simon Newcomb recalculated the problem in 1882 with better values for the distances and masses of the Sun and planets, and came to the now accepted value of 43" century; see Roseveare (1982).

consistent with all the available evidence, it was eventually concluded that the discrepancy counted as a falsification of the theory, which was replaced by General Relativity.[21]

It seems, from the Neptune case, that dynamical measures of mass can (at least sometimes) be useful for the detection or prediction of the existence of previously unknown bodies. Obviously such predictions are to be counted as successful when an independent observation finds the predicted object(s). But what about in cases where we have not observed—or cannot observe—any object in the location demanded by the deduction from the phenomena (as was the case with regard Mercury's perihelion precession discrepancy)? What then is the epistemic status of the claim about the existence of the otherwise unknown body? This question is important because we are in such a situation with regard to dark matter.

In the case of Mercury, persistent attempts to characterise a mass distribution capable of producing the excess perihelion precession failed because no matter candidate could be described which was consistent with all the available evidential constraints.[22] There was, therefore, an excess acceleration on Mercury which (it appeared) no possible matter distribution was capable of producing while remaining consistent with the known

---

[21] The story of attempts to solve Mercury's perihelion precession discrepancy, and its evidential bearing on the shift from Newtonian to Einsteinian gravitation, is more complicated than can be discussed in detail here. See Roseveare (1982), or Earman and Janssen (1993).

[22] For example, Simon Newcomb argued against Le Verrier's proposal of an intra-Mercurial ring of planetoids on the grounds that, given plausible assumptions about the density and albedo of the particles, the zodiacal light was insufficiently bright to be consistent with a total mass capable of producing the correct effect on Mercury. Newcomb did not take this to be a definitive refutation of the hypothesis because he recognised that the background assumptions, although plausible, were not in fact well established (Roseveare 1982, 41-2). A stronger argument, he realised, was that in order for such a ring to cause the required precession of Mercury, it would have to be inclined to the ecliptic in a way that would also cause an unobserved motion in the nodes of Mercury's orbit (Roseveare 1982, 34-5). The so-called Vulcan hypothesis (of planet interior to Mercury) was likewise eventually ruled out because of a failure to observe any such body, and because it would have produced a perturbation in Venus' orbit that was not present (Roseveare 1982, 24ff.; Earman and Janssen 1993, 133).

facts about the solar system. It is an interesting question whether, had this not been the case, General Relativity would have been taken to have received an increment of confirmation in virtue of its ability to account for Mercury's motions without the need for extra matter. Of course it is impossible to answer questions about counterfactual histories, but my intuition is that were GR introduced before the matter hypotheses had been exhausted, saving Mercury's perihelion precession would not have been such powerful evidence for GR.[23] A principle that seems to be operative in science, and which would support both the historical events and my intuitions about the counterfactual history, is the principle of theoretical conservatism: retain well- and broadly-confirmed theories until all plausible options for resolving discrepancies within that framework are exhausted.[24]

By this principle, not every instance of an acceleration whose source is unknown should lead us to abandon or revise our dynamical theory. In some cases, then, we should hypothesise the existence of an unobserved (or even unobservable) body instead of revising the background theory. But exactly what conditions must be met in order for assertions of the existence of unobservable bodies to be empirically justified?

In the case of detecting an unknown planet on the basis of the perturbations of a known planet's orbit, we must first of all account for *all other* accelerations due to known bodies (for example, the Sun and other planets), before we can use the left-over or unexplained acceleration as an indication of the existence of—and measure of the mass of—a previously unknown body. It follows from the definitions and laws of Newtonian Mechanics and the existence of an acceleration, that *there must exist a mass distribution*

---

[23] The question of the evidential impact of solving the Mercury discrepancy is muddied by the fact that physicists at that time already expected Newtonian gravitation to be replaced because it was incompatible with Special Relativity (**SR**) (there was no *prima facie* reason to suppose that mere compatibility with SR would make the theory capable of explaining Mercury). The fact that GR is consistent with SR in the weak field limit worked together with its novel correct predictions to give it eventual widespread acceptance.

[24] See below and Chapter 6 for discussion of Newton's Fourth Rule of Reasoning in Philosophy, on which a theory is to be maintained just until a rival does better with regard to Newton's distinctive ideal of empirical success.

of some particular magnitude the centre of mass of which is at some particular distance that is responsible for the unexplained accelerations. By this I mean that Newtonian physics directs us that *there exists* some *specific* configuration that is in fact responsible for the motions in question. What we do not know immediately is *which* of the (indefinite number of) possible specific matter configurations is the actual one.

We are forced to this conclusion by acceptance of Newtonian Mechanics: in turn, we ought to accept Newtonian Mechanics because of its independent empirical success in many similar circumstances, in particular, in cases where we *do* directly observe the perturbing body. The empirical success of Newtonian Mechanics in accounting for all the motions in the solar system, Mercury's small discrepancy aside, provides grounds to believe that the claims about Neptune made on the basis of Uranus's motions would be justified even if the planet were not independently observed. This is important because we will later need to evaluate dynamical inferences about the existence and characteristics of dark matter in light of the fact that no independent evidence about dark matter has so far been obtained (such independent evidence may even be impossible to obtain: see Chapters 4 and 5). Furthermore, independent measurements of the mass of the perturbing body from its effects on several different planets will convince us that the mass and distance we have determined are correct. Of all the possible solutions consistent with the observed acceleration in a single case, very few will be able to account simultaneously for *all* the effects in all orbits, since the directions and relative strengths of the perturbations at a given time caused by the unknown source on all the planets, taken together, constrain quite severely the possible positions and possible masses of the unknown body.

In sum then, *that there exists* an unseen body follows without need of independent verification from Newton's laws and the existence of an unexplained acceleration, but giving a precise description of its location and mass is susceptible to the problem of the composition of forces (discussed below). As I shall argue, appeals to other orbits and perturbational patterns emergent over long periods (higher order evidence), rather than just to single data points, permit an answer to this problem and enable us to make a specific determination of the mass and location of the body we have strong reason to believe exists. For example, the fact that Uranus was unexpectedly perturbed, but the

other planets were not, is an evidential constraint on the mass distribution causing the Uranian perturbation: the body responsible has to have the right mass and the right distance from the Sun so as to produce the required effect on Uranus's orbit but no discernible effect on the other planets. (As Le Verrier showed, there were no unaccounted for perturbations of Saturn, for example; this proved that the unknown planet had to be outside rather than inside the orbit of Uranus: see Grosser 1979, 100-01.)

This sort of "consilience of inductions" (to use Whewellian language) provides us with strong (indirect) empirical evidence for the mass and location of the body in question, even when we cannot obtain telescopic or other corroboration of its existence. When we achieve this level of unification and confirmation, we are justified in introducing a mass whose existence is not verified by visual detection, without that being a merely *ad hoc* defence of the theory of gravity. Thus the inference to the existence of the planet or other "missing" body, is justified independently of direct observation of that body. As John Herschel said in an address to the British Association in 1846 (in fact, just 13 days before Galle's optical detection of Neptune), the predictions made by Adams and Le Verrier give us evidence of the existence of a trans-Uranian planet "hardly inferior to ocular demonstration" (as quoted in Jones 1956, 832). Direct observation does indeed count as confirmation of the hypothesis of the existence of the body, but that hypothesis already had a high (enough) degree of probability (for belief) because of the other successes, in similar circumstances, of the dynamical theory from which the hypothesis follows. Because of this, we are warranted in drawing the inference to the missing body whether or not its existence is later confirmed by direct check.[25]

Under what conditions, then, *will* we have empirical grounds for taking a dynamical theory to be falsified? I suggest that we have such grounds whenever no model even of an unseen mass distribution can both save the phenomena at hand *and* be consistent with other empirical considerations (as was the case with trying to account for

---

[25] Rather than merely confirming the *prediction* of the unknown planet (though of course it does that too), I claim that the visual identification should be taken also, and more importantly, as confirming the theoretical structure (laws, assumptions, and so on) used to make the prediction. Thus UG itself increases its epistemic warrant by its successful application in the trans-Uranian realm.

Mercury's excess perihelion precession by using missing mass hypotheses, for example). The condition for consistency with additional (including higher order) empirical considerations is important, and the empirical considerations invoked need not be straightforwardly connected with the phenomena we are trying to save. The case of Mercury's perihelion precession illustrates this point: the unacceptability of the various "missing mass" hypotheses did not hinge on the fact that they were unable to account for the excess precession, but rather on the fact that no matter distribution that could save the motion of Mercury could also be made consistent with, for example, the lack of corresponding perturbations of Venus.[26]

The history of the importance of the Mercury case to the acceptance of a new theory of gravitation is of interest here if any general lessons can be learned from it about how to proceed with regard to the modern dynamical discrepancies. Under what conditions should we opt to revise or replace a dynamical theory instead of hypothesising distributions of unseen matter? Physicists' attitudes in the early part of this century toward the Mercury problem—sticking with the old theory until a fully-fledged rival with *better* empirical and explanatory credentials was at hand, and remaining agnostic about which sort of solution was going to turn out to be correct—seem to be supported, for example, by Newton's Rules of Reasoning in Philosophy. Below I consider how accepting Newton's Rules would lead one to act as physicists in fact did act with regard to Mercury, and as they are now acting with regard to dark matter.

Let me quote Newton's Rules of Reasoning in Philosophy (not including the explanations he gives of them):

> Rule 1: No more causes of natural things should be admitted than are both true and sufficient to explain their phenomena.
> Rule 2: Therefore, the causes assigned to natural effects of the same kind must be, so far as possible, the same.
> Rule 3: The qualities of bodies that cannot be intended and remitted [i.e., qualities that cannot be increased and diminished] and that belong to

---

[26] See Earman and Janssen (1993). In fact, some early attempts at revision of the theory of gravity failed for similar reasons: the modification of the power law that would have allowed it account for Mercury's motion would not have been consistent with the lunar motions.

> all bodies on which experiments can be made should be taken as qualities of all bodies universally.
>
> Rule 4: In experimental philosophy, propositions gathered from phenomena by induction should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions. [Newton comments:] This rule should be followed so that arguments based on induction may not be nullified by hypotheses. (Newton 1999 [1726], 794-6)

Rule 1 is obviously a version of the principle of parsimony, an injunction to keep our theories as simple as possible; in this version it is an injunction to admit as few different kinds of causes into our physical theories as possible. In his comment on Rule 1, Newton refers to a justification for the principle offered by some commentators (though it is not clear that he is endorsing that justification, just using it for rhetorical purposes), to the effect that since Nature itself is simple, our theories ought to be simple as well. This is clearly an unjustified (if not unjustifiable) claim, since we conduct empirical enquiry precisely because we do *not* know what Nature is like; this will not do as a reason to adopt Rule 1. Nevertheless, simplicity is a popular and important principle of theory choice, one for which other (more plausible) justifications have been offered (see the discussion of simplicity in Chapter 6), and so we may consider it here. Rule 2 enjoins us to preserve explanatory unity in our theories, so that the same effect is always attributed to the same cause; this requirement Newton seems to think follows from Rule 1, presumably because Nature will not multiply causes unnecessarily in order to produce similar effects.

Rule 3 is crucial to the argument to Universal Gravitation, in that it tells us to attribute to very distant bodies upon which we can perform no experiments the properties we discover by experiment in bodies nearer to hand. Without such a rule there is no reason to think that distant galaxies obey the same dynamical laws obeyed by bodies in the solar system; likewise, we need such a justification for thinking that the spectroscopic qualities of distant bodies indicate the presence of temperatures and chemical compositions that bodies near the Earth displaying those spectra would have (see Chapter 3). Rule 3 provides warrant for thinking that celestial and terrestrial matter obey the same laws not so much by asserting that we have evidence for that proposition, or that our experiments on nearby bodies provide inductive support for it, but by pointing out

that while experiments on bodies within our reach all lead to the conclusion that matter obeys a certain set of laws, we have absolutely no evidence at all (at the beginning of Book III of the *Principia* at least) about *what* laws are followed by matter beyond the reach of our experiments. Given this, we have absolutely no warrant for asserting, as Aristotle and his followers asserted for so long, that there is a fundamental difference in kind between terrestrial and celestial matter. Newton seems to be suggesting that since all the information we have about matter is that it follows a certain set of laws and not others, we ought to attribute those laws to *all* matter, even matter presently beyond the reach of experiment: we have no (experimental) *reason* to believe that celestial bodies are *different* in kind from terrestrial ones, and so we should make the conservative assumption that they are the same. Thus we see also that Rules 1 and 2, advocating simplicity and unity in our theories, are also effectively operative here: unless we have (adequate) reason to believe otherwise, we should hold that apparently similar things are in fact similar, and thus that celestial matter beyond our experiments also gravitates, has inertia, and so on, in the same way that local matter does. Rule 3, then, provides the rationale for excluding the possibility that some different law of gravitation operates at galactic scales—this comes up again in Chapter 6, where gravitational alternatives to dark matter are considered.

Finally, Rule 4 tells us not to doubt theories derived by Newton's method unless we encounter some strong empirical reason to revise them. Harper's interpretation of Rule 4 suggests that "direct empirical support for a theory from measurements of its parameters can be undercut when a rival theory clearly does better" (Harper 1997a, 77, n.16).

By "doing better" what Harper has in mind is doing better with regard to meeting what Harper argues is Newton's ideal of empirical success. Meeting this ideal of empirical success involves having a very strong sort of empirical support. The sort of empirical support involved is best illustrated by an example from Newton himself: take the example of measuring the power law of the force of gravity from the amount of precession of the planetary orbits, as mentioned in an earlier discussion. Newton begins by making some rather weak and plausible assumptions (including, for example, his Laws of Motion), and goes on to use these assumptions together with certain phenomena that he selects to deduce the values of important fundamental parameters of his

gravitational theory. This process is known as "Reasoning from Phenomena" (RfP).

The deduction of the parameters of the theory from phenomena ends up giving a high degree of empirical support because Newton is able to prove that "systematic dependencies" hold between possible states of the phenomena and possible values of the parameters. That is, Newton is able to prove that, given that the background assumptions hold, alternatives to the observed phenomena *would* measure the parameter in question to have different values, values systematically correlated with specific alternative states of the phenomena. For example, Newton shows that while lack of precession measures $n = -2$, positive and negative precession would respectively measure n to have a value greater than or less than -2, where the exact value depends on the amount as well as the direction of the precession. The consequence of this from the epistemic point of view is that we have high confidence in the value of the theoretical parameter.

When systematic dependencies are present, alternatives to the phenomena would measure the theoretical parameter in question to have different values. Furthermore, Newton is able to make several apparently disparate phenomena give agreeing measures of the same parameter. This, Harper argues, is a much stronger form of empirical success than is possible merely from hypothesising some law and its parameters, and finding that it is able to correctly predict the phenomena. Newton is able to show for bodies falling near the Earth, the Moon in orbit about the Earth, the planets around the Sun, and the moons of Jupiter around Jupiter that all of these bodies gravitate according to a centripetal force of attraction which is proportional to the inverse square of the distance between the orbiting body and its primary. So in order to "do better" in the sense required of a potential successor theory on Harper's interpretation of Rule 4, the potential successor cannot merely correctly predict the observations. It must also have its fundamental parameters measured to at least the same degree of exactness from at least the same breadth of different phenomena as the entrenched theory has.

This ideal of empirical success and the role of Rule 4 in theory choice will receive more attention in Chapter 6. For now let me note that in the argument to Universal Gravitation, Rule 3 and Rule 4 together set the standard that must be met by any competitor to Newtonian gravitation with regard to celestial bodies, since Rule 3 tells us to consider the properties of gravitation as discovered by Newton to be characteristics of

all bodies whatsoever, and Rule 4 tells us what a contrary hypothesis must be able to achieve in terms of empirical success in order to be taken seriously as a rival theory.

I read Newton's Rules of Reasoning as together providing implicit support for the claim that in the case of a dynamical discrepancy it is *prima facie* preferable to introduce a hypothesis about the existence of an otherwise unknown massive object (or set of objects) rather than falsifying the well- and broadly-confirmed dynamical theory through which we discover the discrepancy. Something like this understanding of Newtonian theory and the evidential support of that theory seems to have been held by both Adams and Le Verrier, although neither explicitly endorsed this reading of the Rules. For although both Adams and Le Verrier were aware of proposals to solve the Uranus discrepancy by revisions to the Newtonian law of gravitation[27], neither took such proposals seriously, and instead used the available evidence to construct matter solutions to the dynamical discrepancy.[28] Le Verrier, also the discoverer of the Mercury

---

[27] For example, George Biddell Airy developed such a gravitational suggestion; Airy later became Astronomer Royal and was closely involved in the sad story of Adams' loss of priority in the discovery of Neptune. See below for more.

[28] Given that Adams and Le Verrier each worked from the same data and used the same Newtonian theory to construct their respective matter solutions to Uranus's dynamical discrepancy, a hypothetico-deductivist might expect that each should have arrived at the same result simply by deduction from the theory and data. The impression that they actually did so is exaggerated by the close agreement between the geocentric positions they predicted and the actual position of the planet when it was discovered: in fact, their descriptions of most of the important characteristics of the trans-Uranian planet (average distance, average speed, orbital parameters, mass) differ significantly from each other, *and from the truth* (see the discussion in Grosser 1979 [1962], 140-41). There are two things to explain here: Given that both men were working from the same data, why are their respective matter hypotheses so different from each other's? And why, given this difference, are their predicted geocentric positions so close to each other's, and to the truth?

In answer to the first, note that the inference from the dynamical discrepancy is necessarily ampliative because the data underdetermine the solution. Various assumptions and hypotheses have to be introduced (for example, about the inclination of the orbit, the mean distance, and so on) in order to able to concoct any answer; furthermore, even the inference from these assumptions and the data is ampliative. It is no surprise, then, that Adams and Le Verrier describe Neptune very differently, and that neither of them comes very close to describing what the planet is actually like, since both of them are forced to essentially

discrepancy, was confident that it too would be solved by a matter hypothesis, because his confidence in Newtonian gravitation was very high. In support of this attitude, Newton's Rules of Reasoning enjoin us to first attempt to solve apparent empirical discrepancies (such as the Uranus, Mercury and dark matter problems) *not* by revising well- and broadly-confirmed theories (such as Universal Gravitation), *nor* by introducing new kinds of causes, but by treating the objects in which the discrepancies are found as if they were analogous to other objects whose properties are well known (dynamical systems obeying UG), and to search for an explanation of the discrepancy in terms of theories and causes which are familiar to us from experimental evidence. It is easy to see, why a set of principles such as Newton's Rules would lead one to search for matter solutions to dynamical discrepancies, to the initial exclusion of gravity solutions.

The moral of the Mercury case is that sometimes this conservative strategy will fail.

"make up" some important input conditions. In answer to the second question, Herschel (1872) notes that it has been suggested that the near coincidence of the three geocentric positions (Adams's prediction, Le Verrier's prediction, and the observed position) was, given the facts just mentioned, a mere accident. Herschel remarks, however, that taking this line involves misconceiving the problem to be solved. This, he says, was just to find some way to tell *where to look* for the new planet, and this the calculations were perfectly capable of doing. That is, the deductions from the phenomena accurately succeeded in determining the *direction* of the body perturbing Uranus's orbit (at the time it was to be searched for). This deduction was robust in the sense that even though many of the other assumed quantities of the unknown planet were quite wrong (for example, the mass Adams assigned to the new planet is more than six times smaller than it should be), the inference to its direction was nevertheless close to correct. (Note that Grosser [1962] 1979 attributes this line of argument to Smart 1947.) There was a certain amount of luck in the fact that astronomical science (including celestial mechanics, observing methods and technologies) had reached a stage of development sufficient to detect the discrepancy in Uranus's orbit and to make the inference to the existence and geocentric position of the perturbing body at exactly the time that the planets were in a configuration that produced a noticeable disturbance of Uranus. Herschel notes that before 1800 the effect of Neptune on Uranus was too small to be measurable, that the maximum was reached around 1822 (when the planets happened to be in conjunction), and that by 1846 the effect was again near its minimum; the preceding and following conjunctions were in 1781 and 1953 respectively. This makes the discovery of Neptune seem to me to be even less like an accident, in the sense that the prediction of its geocentric position was successfully made at the very first opportunity for making it. (See Herschel 1872, pp. 533-550: §§760-776.)

in the sense that the best explanation of a discrepancy is to be found by replacing an otherwise well-confirmed theory. General Relativity meets the Newtonian ideal of empirical success even better than does Universal Gravitation, so GR is to be preferred (Harper 1997a). The general philosophical or methodological issue then is to characterise so far as possible the conditions under which it becomes acceptable to opt for "revolutionary" solutions to discrepancies. The historical or factual question that follows is whether or not we find ourselves in a situation in which the conditions that would make pursuing non-matter solutions permissible are yet met with regard to the dark matter problem. Note, though, that it is plausible to characterise the transition from UG to GR as also meeting the criteria for theory choice of Newton's Rules, since GR is arguably simpler and more unified, and is certainly empirically more successful, than UG.

This methodological attitude suggests that *ad hoc* hypotheses—*mere* hypotheses, for example, adopting a new power law *just* in order to save Mercury's motions—are not to be accepted. What is required in order to overturn a well-established theory is another theory, one at least as complete and comprehensive, that introduces "systematic dependencies" (Harper 1997a) which allow us to reason from phenomena to the values of new theoretical parameters, and thereby save the phenomena in the same strong way, and to at least the same degree, as the formerly-accepted, well-confirmed theory. The evidence available for the Newtonian power law was diverse and robust, and therefore the induction from the fact that all known phenomena measured the power law to be inverse square to the universal generalisation that all gravitational interactions involve an inverse square attractive force, is not to be discarded lightly, even in the face of discrepancies such as those found in the Neptune and Mercury cases. Since these cases are parallel to the dynamical discrepancies indicating a dark matter problem in contemporary astrophysics, the methodological prescription ought also to apply to the dark matter case.

Note, however, that on this reading, missing matter explanations of dynamical discrepancies are *not* necessarily *ad hoc*, contrary to the claims of some scientists (for example, Mannheim 1994, 493; see Chapter 6 for a fuller discussion). In fact, the situation is quite the opposite, since the well-confirmed dynamical theory *demands* that such bodies be present whenever unexplained accelerations are observed, in order that the

theory itself be preserved: in addition, as is discussed in Chapters 4 and 5, some dark matter candidates are "well-motivated" in the sense that they were originally hypothesised for other reasons. To reject missing matter hypotheses in these circumstances is to ignore strong evidence based on well-confirmed theories.[29]

But we cannot avoid revising our theory if after exhaustive investigation no matter model (including models involving unobservables) can be made consistent with the available empirical data.[30] The upshot is that indefinite protection of the gravitational theory in the face of contrary evidence is ruled out. We should introduce a dark matter

---

[29] The phrase, "strong evidence based on well-confirmed theories," may seem strange (or worse) to some readers, but the theory of evidence implicit here has it that empirical information can only be acquired through an ampliative inference which necessarily involves (more or less well-confirmed) background theories. What distinguishes good (that is, reliable/highly probable) empirical information from bad is that the background theories involved in producing the latter will have a high degree of probability, whereas in the latter they may not (there are other reasons as well why empirical information may be unreliable). In Smith's account as well, the very existence of higher order evidence—for example, discrepancies between a theory and the world—depend on background theories, in that we would not have the higher order evidence if we did not first have a theory to compare against the world. Thus, I would say, we acquire evidence through theories, and we acquire better evidence through better theories; hence the phrase above.

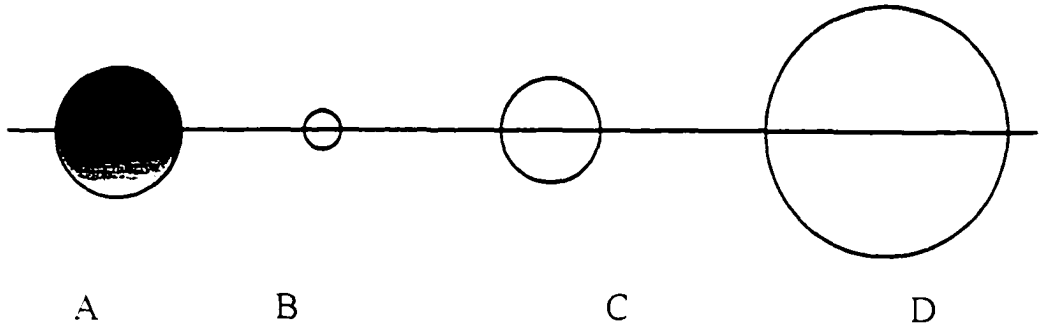
[30] I suggest that scientists were nearing this point of exhausting all plausible matter explanations for the Mercury discrepancy around the time GR came on the scene. In these circumstances, any new gravitational theory capable of explaining Mercury's motions, even in the absence of other reasons for adopting it, would have been the best theory of Mercury available, and should/would have been accepted. Of course, GR also had additional empirical and explanatory successes in its favour. With regard to the Mercury problem, once the confusion regarding the (apparently spurious) Dicke and Goldenberg solar oblateness measurement was cleared up (Will 1993, 181-83; Earman and Janssen 1993, 163), GR stood alone as the broadest, best confirmed theory of gravitation available. We are, at present, a long way from exhausting the plausible matter candidates for solving the astrophysical dynamical discrepancy. If the principles of reasoning under evidential poverty invoked here are correct, this suggests that we ought not yet seriously pursue gravitational solutions to the astrophysical dynamical discrepancies. Note that although we have good reasons to think that GR must ultimately be replaced by a theory of quantum gravity, there is no reason to think that a theory of quantum gravity, which deals with the very small, will have any effect on gravitational action at large distances—in fact, theories of quantum gravity are designed to exactly duplicate the large scale predictions of GR. (Robert Geroch, private communication.)

hypothesis rather than opting to revise the law of gravity, unless we have strong evidence that a revision to the law of gravity will be much more empirically successful than (not just the matter hypothesis but) the whole theoretical structure which is replaced by introducing a new law of gravity.

2.3 THE COMPOSITION OF CAUSES

I have left until now discussion of an important and difficult aspect of inferences from dynamical effects, namely the problem of the composition of forces. The measurement of the total unexplained force acting on a body (say, in a perturbation situation) is a straightforward deduction from phenomena that can fail only if the input data are wrong or the dynamical laws invoked are incorrect. The inference from the specification of that force to some hypothesis about its cause or source is quite a bit more difficult, both to make and to justify, because the dynamical evidence indicating the *existence* of the total unknown force does not distinguish among the indefinite number of distributions of matter capable of *producing* the same force. The observed acceleration by itself only tells us the direction of the *centre of mass* of the perturbing object; the acceleration by itself tells us nothing about the *number or distribution of the components* contributing to the total force. We cannot be sure that what is responsible for the acceleration in question is a *single* body to be found on the line of the acceleration vector, since the same total acceleration could be produced by any number of bodies in any number of positions. The only empirical constraint on the mass distribution is that the total acceleration must equal the observed value. Even assuming that it is a single body whose mass is being measured (and which therefore must lie on the same line as the acceleration vector), we cannot (from the first order phenomenon of the perturbation) decide whether the acting body has a certain mass at a certain distance, or a slightly larger mass slightly farther away, or a lower mass at a nearer distance. (See Figure 2.)

Figure 2.



What the perturbations measure exactly is a quantity we may call "mass-position": the value of this quantity is completely and precisely determined by the total force acting on the perturbed body: it refers to the set of possible masses and their positions that could produce the total force measured by the deviation of the planet from the path it was expected to follow. This quantity can be known to very high precision (as precisely as the perturbing force is known) with a high degree of confidence. What is at issue is how to disentangle the components of this quantity, and to make a prediction about a *specific* mass distribution in a specific location. It is in this step that the degree of epistemic warrant decreases, because it is necessary to employ assumptions whose warrant is less than that of our knowledge of the mass-position quantity.

One way to think about this is to note that the observed acceleration, since $a = (GM)m/r^2$ and G and M are known constants, measures the ratio m/r^2 but tells us nothing about the values of m or r independently. An unexplained perturbation, therefore, merely tells us that there is *some* unknown mass distribution acting on the perturbed body. The magnitude and direction of the acceleration (as these quantities change over time) define a range of possible masses and distances each of which could

account for the observed motions.[31] In this sense the composition of forces problem is a special case of the problem of theory choice given evidential underdetermination: the observations pick out a range of possible hypotheses, but do not provide grounds for choosing one particular hypothesis from within the range. How can we get from information about the total force (or the mass-distance quantity) to a causal hypothesis about a body or set of bodies with a specific mass at a specific distance?[32]

This is easiest to do if we can identify some previously unknown body on the line of acceleration. If we actually observe a body (or set of bodies) whose direction (or the direction of whose centre of mass) corresponds to the direction of the acceleration, it is a good bet that it is responsible for the perturbation we are trying to explain. It is unlikely (purely on probabilistic grounds, given our knowledge of what the solar system is like) that there are other bodies (which bodies are "really" responsible for the perturbation) whose centre of mass also happens to fall on that line. And because bodies at different orbital distances orbit in different periods, the coincidence of the direction of the visible body and of the centre of mass of the matter really responsible for the perturbation will be a transitory accident unless the visible body happens to occupy the centre of mass of this other matter distribution. Further, given the earlier remarks about the solar system being old enough that it must have reached gravitational equilibrium, it is also old enough that such a cloud of matter would likely have amalgamated into a single body (centred on its centre of mass) in the time available, unless we happen to be unlucky and observe the perturbation near the time of some collision or explosion involving the perturbing mass.

More importantly, there are various higher order phenomena we can examine in order to check the hypothesis that a body visible along the line of acceleration contains

---

[31] The range in fact contains an indefinite number of possible mass distributions, but as is argued below, most multi-body configurations can be ruled out by taking into account considerations regarding the stability of the system, and other higher order evidence.

[32] What follows is an over-simplified account of perturbation analysis, but it does not differ from the actual details of perturbation analysis in any way that is philosophically or methodologically significant. Determining the existence and cause of a perturbation is sometimes more difficult than this account may make it seem, but the process still only involves Newtonian Mechanics and mathematical analysis.

all the mass causing the perturbation. Since every orbital distance has its unique velocity, there will be various periodicities in the pattern of perturbations over the long term, corresponding to maxima and minima of interaction at conjunction and opposition respectively. Using these periodicities in the observed perturbations, one can infer the orbital distance of the perturbing body. If this distance (and direction) correspond to the visible body we have detected, we can be sure that that body occupies the centre of mass causing the perturbation. The frequency of the periodicity of orbital perturbations is, then, a higher order phenomenon that constrains the possible mass distributions to a quite narrow range: only bodies occupying a small range of orbital distances will be capable of producing the periodicity observed. Thus, it is unlikely that the perturbation in question could be due to the body visible along the line of acceleration, *plus* some other(s). In turn, this range of distances constrains the range of possible masses of the perturbing body quite narrowly, since the amount of the acceleration on the perturbed body is known (up to some margin of error). When the location of the visible candidate is consistent with this higher order information, we have strong reason to accept that body as responsible for the perturbations in question, and that it therefore has the mass calculated from the strength of the perturbation and the distance of the perturbing body.

One can see, moreover, that the higher order evidence is still useful for determining the position (and therefore mass) of the perturbing matter even when we have no direct visual detection of any body or set of bodes that might be causing it. The hypothesis about this otherwise unknown body which we arrive at through the higher order evidence will probably be less certain than it would be if we had a direct detection to back it up, but its probability may nevertheless be high enough for us to accept the hypothesis. In short, appeals to the totality of the evidence available to us, in particular to higher order evidence, will enable us to constrain the solutions to "decomposition problems" rather narrowly in many scientifically important cases, even though we may not be able to escape the underdetermination problem altogether.

As we shall see in subsequent chapters, there is no exact parallel in the dark matter search to the case of inferring from *perturbations* the existence, location and mass of otherwise unknown bodies. But while the exact *kinds* of higher order evidence (for example, about periodicities in the perturbations, and so on) are not applicable in the dark

matter case, the notion of using higher order evidence to constrain potential solutions via dynamical arguments *is* retained.

2.4 THE CONCEPT OF MASS

Since so much of the discussion in this dissertation involves the use of the notion of "mass", a few words ought to be said about it. In astrophysics and cosmology it was formerly common to talk about "the missing mass" as a synonymous for "the dark matter", but of course it is not the *mass* that is missing—dynamical measures tell us exactly how much mass is present. Rather, what is missing is the *light* that we would expect to go along with so much mass.[33] What then is "mass", this quantity that the dynamical phenomena measure?

The first systematic "historico-critical" treatment of the concept of mass was Max Jammer (1997 [1961]). The following is the penultimate paragraph of Jammer's book:

> Throughout its long history in human thought, from its early adumbrations in Neo-Platonic philosophy, its mystic and still inarticulate presentation in theology, to its scientific manifestation in the physics of Kepler and Newton, to its carefully thought-out redefinitions in positivistic and axiomatic formulations, up to its far-reaching manifestations in modern theories of physics—nowhere does science seem to get full command and control over all the conceptual intricacies involved. One has to admit that in spite of the concerted effort of physicists and philosophers, mathematicians and logicians, no final clarification of the concept of mass has been reached. (Jammer 1997 [1961], 224)

This pessimistic conclusion (even if correct) need not bother us in this section, since we are here concerned only with mass insofar as that concept is relevant to dynamical measurements of mass, and dynamical mass is something about which we are able to say something more or less definitive. We need not enter into, as Jammer does, consideration of the more difficult problem of the concept of mass from the point of view of

---

[33] The phrase "missing mass" is perhaps apt with regard to the *cosmological* dark matter problem, which claims that there really is *mass* missing: about 0.6-0.8 of the mass density required for "closure" of the universe is completely undetected by any measure (optical, dynamical, or other). Of course, this mass is only "missing" if the reasons for thinking that the universe ought to be closed are good ones, which is doubtful according to the latest evidence, as I discuss in Chapter 1.

electromagnetics or quantum field theories.[34] This brief section is not meant to give a complete analysis of the concept of mass, but a preliminary discussion sufficient to make sense of the use of the concept in the dynamical dark matter episode.

As Jammer's book shows, the history of the concept of mass is a messy business, and I shall not try to give a historical treatment here. Note, though, that Newton was the first to demonstrate the need for a distinction between mass and weight (in his *Principia*, 1999 [1726]). Newton refers to some observations made by Richer in 1671 of the fact that a pendulum clock the mass of whose bob is constant runs at different rates in different locations on the Earth (because of the variation of the strength of the Earth's gravitational field). (See Jammer 1997 [1961], 73-4.) Richer's experiment was shown by Christian Huygens to demonstrate that although the mass and weight of a given body are proportional to each other (at a given location on the Earth), the two concepts are nevertheless distinct: the mass of a body is a constant, whereas its weight depends on its mass and on the strength of the gravitational field. Weight is the measure of the gravitational force acting on a body, whereas mass is a measure of its "quantity of

---

[34] The fact that energy has gravitational mass may be important with respect to the cosmological dark matter problem. As our best evidence now seems to indicate (see Chapter 1 and the Appendix), the greater part of the energy density of the universe as a whole may be contributed by "dark energy", probably the energy of the vacuum, which is causing the expansion of the universe to accelerate over time. To be more clear, dark energy is *possibly* a solution to the cosmological dark matter problem. An effective cosmological constant, which would explain the observed acceleration of the Hubble expansion (see the Appendix), would contribute to the energy density of the universe, and because of the mass-energy equivalence, this cosmological constant (which recently acquired the name "dark energy") would also contribute to the overall value of Ω (the mass density parameter). This is interesting if one wants to hang onto the idea that $\Omega = 1$, since the observed contribution of matter (including dynamical dark matter) to Ω is now known with high confidence to be less than 0.4. If there is enough dark energy, the overall value of Ω could still be 1 (although our best evidence also seems to go against *that* claim as well; see the Appendix). But he local contribution of dark energy to the effective mass density is too small to notice (it is only because the intergalactic spaces are so huge that the small local contribution of dark energy to the energy density can add up to an appreciable fraction of the closure density). This means dark energy has no noticeable effect on the scales of galaxies, where the dynamical discrepancy is already huge. Moreover, the dark energy would be perfectly homogeneously distributed within and around (to effectively infinite radius) all the dynamical systems considered here, and it would therefore be dynamically undetectable.

matter". Newton made fundamental use of this distinction and its consequences in constructing his physical theory.[35]

Drawing a distinction between weight and mass was part of the continuing and fruitful drive in physics to understand the world in terms of quantities that are *conserved* through reactions of various kinds and across different frames of reference. But Einstein (in his Special Theory of Relativity) showed that in fact mass is *not* strictly conserved, although *mass-energy* is. This is to say that there are some reactions in closed systems in which mass is converted into energy, and *vice versa*, but that the total quantity of mass-energy in a closed system is always constant. One corollary of the interconvertibility of mass and energy is that the velocity at which a body moves in a given frame of reference, which is related to its *kinetic energy*, affects the measured value of its mass in that frame, and in different frames of reference moving relatively to one another the mass of a given body will be measured to have different values. This means that mass is a *frame-dependent* quantity, although *within* a given frame of reference mass-energy is conserved. The closest thing we have in relativity to the "true" mass of a body is its *proper* or *rest mass*, the mass of the body as measured from a frame of reference at rest relative to (co-moving with) the body. But since, according to the principle of relativity, no inertially moving frame of reference is privileged, we cannot really say that the rest mass is the true mass: the rest frame is not the true frame, since there is no true frame. Rest mass is just a convenient quantity, calculable from within any frame from the body's velocity relative to that frame and its relativistic mass as measured in that frame.

Despite Einstein's results about the velocity-dependence of mass, the Newtonian (velocity-independent) account of mass is perfectly adequate to the dynamical cases of importance to the dark matter problem. This is because unless the relative velocity of a

---

[35] I thank my external examiner, Brian Baigrie, for asking me to emphasize something that really should be emphasized in this context: Newton defined matter as that which moves and is capable of resisting any change of motion. Mass (quantity of matter in the inertial sense) is the measure of this resistance. Newton offered a new means for measuring mass by showing that matter is not only that which offers resistance to change of motion but also that which causes change of motion in other portions of matter (both by its gravitational action and transfer of momentum in collisions).

body is very high (a significant fraction of the speed of light), its relativistic mass will be empirically indistinguishable from its Newtonian mass. This is easy to see by some heuristic manipulations of the equation $E = mc^2$ (energy is equal to mass times the square of the speed of light). The total (relativistic or proper) mass of a body, m, is equal to its rest energy E_r plus its kinetic energy KE, all of this divided by the speed of light squared: $m = (E_r + KE)/c^2$. (Note that the rest energy of atoms and molecules includes binding energy and the rest energy of the simple constituent particles.) Rest energy is by definition a constant for any given body, and the kinetic energy of a given body is determined by its velocity (a frame dependent quantity). It is easy to see, then, that since the speed of light squared in the denominator is a very large number, the kinetic energy (velocity) of a body has to be very high indeed in order for it to make any significant contribution to the total mass. Another way of saying this is that if we add energy to a body (say, by increasing its temperature or its velocity), we thereby increase its mass, but unless the change of energy is enormously large, the change of mass will be insignificant, since $\Delta m = \Delta E/c^2$. The exact equation relating the relativistic mass of body to its velocity, v, is the following, where m_r is its rest mass: $m = m_r / (1 - v^2/c^2)^{1/2}$ (see, for example, Cohen 1985, 181).[36] As I will discuss in later chapters, the velocities involved in the dynamical measurements of the masses of galaxies and clusters are *very* far below

---

[36] I owe the following rigorous derivation to Francisco Flores (private correspondence). The relativistic mass or proper mass of a body, m, can be expressed as [1] $m = (E_r + KE)/c^2$, where E_r is the rest energy of the body and KE is its kinetic energy. This expression can be derived from the definition of the relativistic or proper mass, m, in the following way. Begin with the definition of m: [2] $m = m_r \times \gamma$, where m_r is the rest mass and γ (gamma) is the usual Lorentz factor, $(1 - v^2/c^2)^{-1/2}$. Using the binomial expansion, one can approximate γ by the following series: [3] $\gamma = 1 + (1/2)v^2/c^2 + \ldots$ *(higher order terms)*. Substituting this expression for γ in [2], and neglecting higher order terms, we get: [4] $m = m_r [1 + (1/2)v^2/c^2]$. Multiplying through the m_r, we get [5] $m = m_r + (1/2)m_r v^2/c^2$. Finally, we use the fact that $m_r = E_r/c^2$ (a rearrangement of $E = mc^2$ for a body at rest) in [5] to obtain the desired result: $m = (E_r + KE)/c^2$, using the usual definition for the kinetic energy KE. Note that this expression for the mass proper mass m is correct only for velocities much smaller than the speed of light: approximating gamma by the expression in [3] while neglecting higher order terms works only if $v << c$. Where v is a significant fraction of c, we can take the approximation of γ in [2] to a greater number of terms.

the level required for the relativistic contribution to the mass of these systems to become significant (typically, the average velocity in a galaxy is only a few hundred kilometres per second).

In Newtonian and Einsteinian physics, three *kinds* of mass are distinguished:

> (1) *inertial mass*, which by Newton's second law of motion is determinable through its reaction to a mass-independent force; (2) *active gravitational mass*, defined as the material source of the gravitational field or the mass that "induces" gravitation. . .; and finally (3) *passive gravitational mass*, defined as the mass susceptible to and receptive of gravitation. (Jammer 1997 [1961], 125)

There turns out to be a "universal proportionality" between the three kinds of mass. Newton's experiments with pendula with bobs made of different materials established the proportionality between inertial and passive gravitational mass (Newton claims, to one part in 1000; Newton 1999 [1726], 807); modern "Eötvös experiments" have proved the equivalence to an extremely high degree of precision ($\approx 10^{-12}$; Will 1993, 27). In contrast,

> The universal proportionality between the active and passive gravitational masses of the same body. . .is a consequence of Newton's third law (action = reaction), as can be seen from the following considerations. If m_{a1} and m_{p1} denote the active and gravitational masses of body 1 and m_{a2} and m_{p2} those of body 2 respectively, the gravitational force exerted on body 2 is given by the expression $F_2 = Gm_{a1}m_{p2}/r^2$ and the gravitational force exerted on body 1 by the expression $F_1 = Gm_{a2}m_{p1}/r^2$. . . . [A]ccording to Newton's third law $F_1 = F_2$, which implies $m_{a1}/m_{p1} = m_{a2}/m_{p2}$. (Jammer 1997 [1961], 125-26) [See note[37].]

---

[37] Stein 1991 (see 215-9, especially 218-9) points out that the sort of use of the Third Law involved here is actually an invocation of a special case of it, namely where we take the forces involved to be *forces of interaction*, that is, where we assume as Newton does that the force of A on B and the force of B on A is in fact one organic interaction governed by "a law in which the interacting bodies enter altogether symmetrically" (Stein 1991, 218). The assumption that gravitational attractions are due to interactions between the bodies suffering them does quite a bit of the work in Newton's inference to Universal Gravitation, as Stein shows. Without it, the reaction force to the force on a planet accelerating it toward the Sun, for example, could be found in vortical particles exterior to the planet pushing it toward the Sun. In fact, some of Newton's contemporaries objected to this very part of his argument, since they believed that all transmission of force happens by contact, and in that case the reaction force would *not* be expected to be

This points to an interesting difference between the two proportionalities, which are really equivalences, so far discussed: "while the proportionality between inertial and passive gravitational masses is a purely empirical and accidental feature of classical physics, the proportionality between active and passive gravitational mass is deeply rooted in the very principles of Newtonian mechanics" (Jammer 1997 [1961], 126). Since the relationship of proportionality is transitive, once we know that inertial mass is proportional to passive gravitational mass, and that passive gravitational mass is proportional to active gravitational mass, we also know that inertial mass is proportional to active gravitational mass. The only laboratory test of the equivalence of active and passive gravitational mass was performed by Kreuzer in 1968: he compared the active and passive gravitational masses of fluorine and bromine in a version of the torsion balance test, and found equivalence up to tolerances of 5 parts in 10^5 (Will 1993, 214).

In the shift from Newtonian gravity to General Relativity these fundamental concepts and their relations are retained. To them we must add, however, the Special Relativistic equivalence of mass and energy discussed above, which implies that the energy contained in a body will have inertial and gravitational effects. That this is the case is verified by experiment, but the contribution of the energy of the astrophysical systems of interest here to the quantity of their inertial and gravitational masses is insignificant, as I have said.

We are now able to state with more precision something mentioned in section 2.1 above, namely the fact that the mass of a planet orbiting a central body drops out of the calculation of the mass of the central body. Take m_i to denote the inertial mass of a body orbiting at distance r a central body whose mass is M. Where the gravitational mass of the orbiting body is m_g (note that we do not need to specify active or passive because of the fundamental equivalence noted above) and the gravitational constant is G, the gravitational equation expressing the force F between the central body and its satellite is $F = Gm_gM/r^2$. The inertial equation is $F = m_ia$ (where a is the acceleration on the satellite). Combining the equations we find that $m_ia = G\,m_gM/r^2$, and rearranged this

---

found acting on the Sun. But insofar as Jammer is taking Universal Gravitation as given here, there is no objection to his way of putting the argument above.

gives $M = (m_i\, m_g)\ (a\, Gr^2)$. Because of the equivalence of the inertial and gravitational masses, $m_i\, m_g = 1$, so that the acceleration on the orbiting body is a direct measure of the mass of the central body given G and r, and we do not need to know the value of m in order carry out this measurement.

In a certain sense the explanation of mass in modern physics has not been able to answer the most fundamental question. We do not know *why* bodies have these three kinds of mass, nor can we explain why the three kinds of mass should be equivalent, in that we do not have an account of a causal mechanism responsible for producing the three mass effects in bodies, an account which would naturally explain why the three kinds of mass are equivalent. Mach proposed that inertial effects on individual bodies arise as a result of bodies' interactions with all of the matter in the universe (see Dicke 1970). This implies that anisotropies in the universal mass distribution should lead to differences in the inertia of a body depending on the direction in which it is measured; so far, any directional effect that might exist in the inertia of bodies is well below our ability to detect it.[38] Particle physics has proposed that the gravitational interaction is mediated by the exchange of fundamental particles known as "gravitons", but these particles have never yet been detected in particle accelerators, and in any case even if correct this account seems merely to push the explanatory question down to the next level of

---

[38] The discovery of the dark matter problem may make the possibility of measuring directional differences in the inertia of bodies even more remote. The best hope for detecting such directional differences was to try to measure a difference of inertia in the direction of the galactic centre as opposed to away from the galactic centre or perpendicular to the galactic plane. The information we have about dark matter, however, points to the fact that it is not only 10 to 100 times more prevalent (by mass) than ordinary matter, it is also distributed in a spherical halo around the galaxy that extends to several times the radius of the visible matter. If correct, this means that the mass distribution in the galaxy, as viewed from our position within it, is much more homogeneous and isotropic than we would have thought given the distribution of visible matter, and this means that the expected differential inertial effects (if they exist) will be very much smaller, and therefore even farther below our ability to detect them.

structure.[39] In this sense Jammer is right that we do not yet have a perfectly adequate theory of mass.

Finally we must ask, in what sense do dynamical measures measure mass? Take the case of measuring the mass of a spiral galaxy from the average velocity of some group of stars or gas clouds orbiting that galactic centre at a given radius. To do this we adopt a frame of reference in which the centre is at rest, and the remainder of the galaxy rotates around it. This centre in fact has a recessional velocity relative to us due to the cosmic expansion, but since this speed is small compared to the speed of light in all cases that we study in this way (extremely high redshift galaxies, that is those with recessional velocities approaching c, are generally too dim for the detailed spectroscopic analysis required to obtain detailed rotation curves), the relativistic mass as it would be measured from our frame differs only by a very small proportion from the mass as measured in a frame at rest relative to the galactic centre. And in any case, dynamical techniques measure the forces acting on the bodies *within* that galaxy: any relativistic effects due to *our* motion relative to that galaxy will have no effect on those bodies (their frame is not our frame). Similar considerations show that whatever proper motion the galaxy might have across the line of sight will also be of no consequence in the dynamical measures. Since the velocity of rotation of galaxies and clusters, and even of the sub-systems within them, is relatively slow compared to c, although the kinetic energy of the system will contribute to the system's relativistic mass, that contribution will be a very small proportion of the total mass. In short, techniques for measuring dynamical mass determine the value of a quantity that is, for the kinds of systems considered here, nearly indistinguishable from the rest mass of the system.

---

[39] Particle physicists seem confident that they are now close to being able to detect the "Higgs boson", a fundamental particle whose existence would confirm theories that a Higgs field permeating space is the causal mechanism that generates mass. If this hoped-for discovery comes to pass, we would for the first time have a confirmed theory of *how* mass is generated.

2.5 CONCLUSION

In this chapter I have described the philosophical and physical foundations of dynamical inferences. I have categorised dynamical inferences into two kinds, and have argued that both are essentially ampliative and therefore cannot be analysed in purely deductive terms. I have shown that dynamical inferences of the first kind—inferences from dynamical effects to the existence of a dynamical discrepancy—while possible only relative to a set of background theories and assumptions, are nevertheless relatively straightforward. Their reliability depends only on the epistemic quality of the theories invoked in them and on the epistemic quality of the data to which they appeal. In contrast, dynamical inferences of the second kind—attempts to infer the cause of a dynamical discrepancy—are more difficult to make and to justify because philosophical principles of theory choice (which themselves stand in need of justification) must be invoked in order to be able to argue from dynamical effects to hypotheses about their causes. This means that discovering a dynamical discrepancy is much easier than solving it. I have given the beginnings of an account of the role of higher order evidence in constraining dynamical inferences of the second kind, and have shown how higher order evidence can help to decrease the epistemic significance of the composition of forces problem. In subsequent chapters the basic principles involved in dynamical inferences will be invoked and further developed in relation to possible solutions to the astrophysical dark matter problem.

CHAPTER 3

A SELECTIVE HISTORY OF DYNAMICAL MEASURES OF MASSES: EXTRA-SOLAR AND EXTRA-GALACTIC STUDIES, TO 1970

Unseen bodies may, for ought we can tell, predominate in mass over the sum-total of those that shine: they supply possibly the chief part of the motive power [i.e., gravity] of the universe.

—Agnes Clerke, 1903

3.0 INTRODUCTION

This chapter examines some selected cases in the history of dynamical measures of mass in astronomy. The principle of selection is relevance to the contemporary dark matter problem, either in the sense that the episode was important for the realisation of the existence of the dark matter problem as presently conceived, or in the sense that the episode illustrates some important philosophical or foundational thesis in the area of dynamical inferences to the existence of distant, unseen masses.

The episodes discussed include the first detection of invisible binary companions of stars, and the first studies leading to the claim that astronomical systems of various levels of structure are composed mostly of dark matter: notably, Oort (Milky Way), Babcock (M31, the Andromeda Galaxy), Smith (Virgo Cluster), and Zwicky (Coma Cluster). This historical study ends around 1970, the threshold of the "modern" period of dark matter astronomy, which is discussed in detail in Chapter 4.

3.1 BEFORE DARK MATTER

In principle, the techniques (described in Chapter 2) for the determination of the masses of gravitationally bound systems were available from the time of Newton's *Mathematical Principles of Natural Philosophy* (first published in 1687). However, the observational basis of astronomy needed to be considerably augmented, and various technical problems in the theory of the solar system solved, before these techniques could profitably be applied to the discovery of unknown masses, especially for systems beyond the neighbourhood of the Sun. In part, the breadth but especially the accuracy of long-

term records of telescopic observations was lacking. Even William Herschel (1738-1822), the builder of telescopes not bettered for their size and quality until the mid-nineteenth century and arguably the greatest observational astronomer after Tycho Brahe (1546-1601), could not resolve most galaxies into stars (because of this, his early studies of "nebulae" lump together objects we now know to be galaxies external to the Milky Way with gas clouds and star clusters within the Milky Way). Certainly, Herschel's observations of galaxies lacked the detail to be able to determine their structural features with enough precision or distinctness to be able to say anything about whether or not they rotated, let alone about their rates of rotation. (See Crowe 1994.)

Several things were needed, in addition to bigger and better telescopes, and a longer record of accurate observations, before it became possible to apply dynamical techniques to the measurement of the masses of distant systems. Among the most important developments were the invention of spectroscopy and its application to astronomy[1], and the use of photography in astronomical investigations[2]. It is only with

---

[1] In the early 19th century Josef Fraunhofer did significant studies on solar spectra, among other work; Bunsen and Kirchhoff established spectroscopy on a firm basis in 1859, in particular showing that bright and dark line spectra enable the determination of the chemical composition of the source; William Huggins undertook spectroscopic analysis of astronomical objects, and in 1868 proposed the Doppler shifting of spectra of bright objects as a method of determining their velocity along the line of sight. See Crowe (1994, 178-83) and sources therein, and Meadows (1984).

[2] The daguerreotype was first used in astronomical photography in 1840; by the 1880s, new photographic techniques had been developed that greatly increased the sensitivity of the plates and the ease and fruitfulness of applying photography to astronomical investigations. Among the most important advances brought to astronomy by photography were the ability to obtain highly accurate indications of the positions of large numbers of bodies in a short time, which could be analysed later, and with much greater precision, under better conditions than a cold, dark observatory dome. Thus both the volume and exactness of the work were increased, at the same time making the work actually easier to carry out. Furthermore, these highly accurate records could be referred to a long time after they were taken, and could be compared to more recent photographs of the same region to look for any changes that might have occurred: it thereby became possible to detect *very* small changes that would have been impossible to see otherwise. See Lankford (1984) for details. The comparison of plates taken at different times was vitally important to

photography that the minute changes of position of stars perturbed by invisible companions can be reliably detected, or that the rotation of galaxies becomes possible to notice and quantify. (A few true binary stars were in fact discovered by inference from naked eye observations or telescopic observations without photography (see below), but most could not have been discovered in this way because either their brightness or angular separation were below human thresholds of perceptibility.) And spectroscopy made possible the study of the rotation of galaxies, through the Doppler effect. The study of dynamics through spectroscopy depends on the fact that the characteristic light signature of each chemical substance (most useful in astronomy are the signatures of hydrogen and some metals) is red- or blue-shifted depending on the motion of the source along the line of sight, either away from or toward the observer—the amount by which the spectral signature of the object is shifted depends on the speed of the motion of the object along the line of sight. Thus it becomes possible to give a precise measurement of the component of the rotation along the line of sight for astronomical systems such as galaxies and clusters, from which it is possible to make inferences about their true rotational velocities and therefore (using the techniques outlined in Chapter 2) about the mass contained in each of those systems. These inferences are such that any uncertainties in them (due either to inaccuracies in the observations or doubtful assumptions) correspond to margins of error in the final mass estimate, and these uncertainties can be quantified. Furthermore, various important facts can be learned by constructing higher-order phenomena out of the raw data, without needing to know the "true" value of the rotation. (For example, the "rotation curve" for a galaxy can be known perfectly well even though we only know the component of velocity along the line of sight and not the true velocity, and yet from the rotation curve alone we can discover that the distribution of matter in the galaxy does not have the same form as the distribution of light.)

In what follows I trace the early history of the discovery of the dynamical dark matter problem. The presentation is in certain places slightly out of historical order, but this is in order to describe first the results for smaller and nearer dynamical systems and

---

various stages of the early dark matter investigations, notably in determining the internal motions of galaxies and clusters, and also in Oort's studies of star motions in the Milky Way.

then those for larger and farther ones. What we see, especially in the modern versions of these results but also here, is that the quantity of dark matter required increases with the scale of the system under consideration.

3.2 THE FIRST INVISIBLE BINARY STELLAR COMPANIONS

A key step in applying the techniques of dynamical mass measurements to stars was to see stars as involved in local gravitational interactions. Newton and his followers had suggested that the fixed stars must attract each other gravitationally, but this suggestion was made under the assumption that the fixed stars were essentially stationary relative to one another because the *total* force acting (due to an infinite number of bodies distributed homogeneously through space) was zero. If that were true, it would be impossible to apply dynamical arguments to the measurement of stellar masses. There was, however, another route available. John Michell (1724-1793), in 1767 and in more detail in 1784, argued on probabilistic grounds that there are far too many apparent double stars for all of them to be accidental configurations. In other words, the odds are "astronomically" high against finding so many pairs of stars so close together on the sky, on the assumption that stars are scattered randomly.[3] Michell's argument amounts to the claim that there is likely to be some *physical connection* between (most, or at least many) apparent binary companions, which is to say that they are *physical* and not just *optical* pairs. The obvious mechanism for this physical connection was Newtonian gravitation. William Herschel catalogued apparent binaries between 1779 and 1784, in an effort

---

[3] Michell argues, for example, that the odds of the six brightest stars of the Pleiades star cluster being an accidental configuration are 500 000 to 1 against (see Crowe 1994, 112-13, 116). Michell's other contribution to the dark matter story is his invention (in his 1784 paper to the Royal Society) of the idea that some bodies might be so massive that no light can escape their gravitational fields—he thus anticipated the General Relativistic notion of black holes. He actually proposed the idea in the course of discussing a possible method for determining the distances, sizes and masses of stars through the gravitational retardation of light corpuscles emitted from their surfaces. (See Israel 1987, 201-02.) With the rise of the wave theory of light and especially Young's discovery of interference in 1801, there seemed no longer to be any reason to expect light to be affected by gravity since it was not a particle, and the idea of invisible or dark stars fell out of fashion until the relativistic era (Israel 1987, 204).

initially undertaken in order to determine whether annual parallax could be observed in them (this in turn was part of his attempt to determine the size of the Milky Way). It turned out that no parallax was visible, and furthermore that the apparent motions actually observed within binaries had to be attributed to the *proper motions* of their members. Over time, the stars did not drift apart from one another (as one would expect for optical binaries, in which the partners are really very distant from each other and moving with independent proper motions), but instead appeared to *orbit* one another. This basic result was not clear until 1803; and it was not until 1827 that Felix Savary (1797-1841) made the first calculations to show that the motions of one pair of stars were indeed explicable as elliptical motions about a common centre of mass.[4] The analysis of the motions of binaries (culminating with their subsumption under the Newtonian theory of Universal Gravitation) proved Michell correct: most binaries are physical (dynamical) systems, not merely accidental optical configurations.[5] Identifying binaries as dynamical systems makes it possible in principle to apply the techniques described in Chapter 2 to the stars in binary systems to discover the component stars' masses, sizes, orbital velocities, and so on. As Berry notes, the precise analysis of the motions of binaries "may be regarded as the first direct evidence of the extension of the [Newtonian] law of gravitation to regions outside the solar system" (Berry 1961 [1898], §262: p. 343).[6]

Once this step had been taken, it became possible not only to measure the masses of known binary companions, but also to infer the existence, mass and orbits of *invisible*

---

[4] Herschel 1872 (note, this is William's son John Herschel (1724-1793)), §§833-842: pp. 606-615. See also Berry 1961 [1898], §§263-4: pp. 341-44 and §309: pp. 398-400; and Crowe 1994, pp. 112-3ff.

[5] Although Michell's initial assumptions were wrong (he assumed many too few stars in each category of magnitude), the argument is still essentially correct when those assumptions are modernised. In any case, probabilistic arguments of this type were superseded by observations demonstrating an orbital motion in many binaries, which by itself proves the existence of a physical connection between the components.

[6] In only very few cases, however, has it been possible to perform the detailed calculation to find the masses of binary systems; Seeds (1987, 165) says that accurate masses have been determined for fewer than 100 binaries. No doubt this number has increased significantly in the past 13 years, but the practical difficulties of performing the required observations with sufficient precision remain.

companions (stars or other massive bodies too dim to see) on the basis of the observed motions of visible stars. Among the practical difficulties of employing dynamical measures in such cases, aside from the problems of trying to obtain accurate positional measurements, and the impossibility of knowing in advance which apparent binaries are real binaries, is the fact that pure telescopic measurements can only determine proper motions across the line of sight (in arc seconds per year) while the orbital plane of the binary could be oriented in any direction. Furthermore, without an accurate estimate of the distance of the binary from us, it is difficult to translate the observed motions (in arc seconds per year) into the speeds and distances needed in order to employ the techniques of dynamical mass measurement. Fortunately, these difficulties do not render the mass measurement impossible, they just mean that the mass estimate must be given with a fairly large margin of error.[7]

From observations of the motions of binaries it is a small step to considering the oscillatory motions of single stars to be due to the gravitational influence of invisible companions. According to Trimble,

> The first detection of nonluminous matter from its gravitational effects occurred in 1844, when Friedrich Wilhelm Bessel announced that several decades of positional measurements of Sirius and Procyon implied that each was in orbit with an invisible companion of mass comparable to its own. The companions ceased to be invisible in 1862, when Alvan G. Clark turned his newly-ground 18½-inch objective toward Sirius and resolved the 10^{-4} of the photons from the system emitted by the white dwarf star Sirius B. (Trimble 1987, 425)

In 1896, J. M. Schaeberle reports observing the companion of Procyon in roughly the position it was predicted that it ought to be found (Schaeberle, 1896). The techniques used by Bessel and others for the prediction of invisible companions to bright stars are

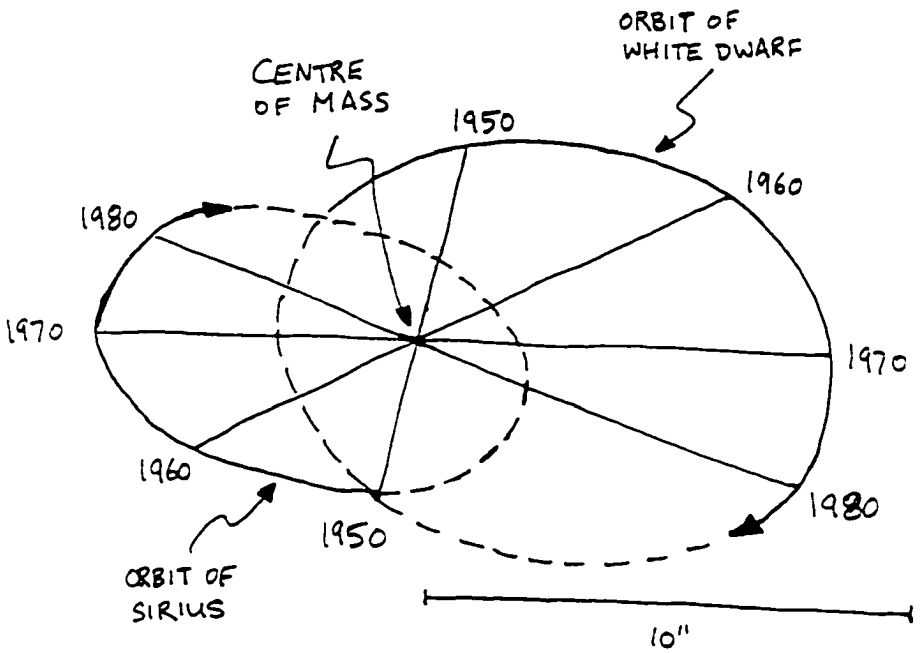
---

[7] Leaving aside the uncertainties inherent in estimates of the distance of a binary system from us, which corresponds to uncertainties in the diameter of the system, the fact that we often cannot be sure of the inclination and orientation of the orbit of a binary system to the line of sight when one of the companions is invisible to us means that any mass determination will give a *minimum* value for the dynamical mass of the system (Seeds 1989, 162-3). When applied to larger systems such as galaxies and clusters, even this minimum mass is often much greater than the visible mass: thus we can be entirely sure of the existence of large amounts of dark matter despite the uncertainties of the dynamical mass determinations.

essentially those described in Chapter 2. Photographic and other records of the positions of the stars in question are compared, and the star is found to move in some non-inertial way, which implies that its proper motion is influenced by some exterior acceleration. Given that the only known force able to affect a star in this way is gravity, one can use Newtonian Mechanics to deduce the position and mass of the body causing the acceleration. The results are uncertain because of the limited records of the position of the star, errors in the positional estimates, and because the star's distance from Earth is not known exactly. Even though a precise estimate of the position and mass of the perturbing body will therefore sometimes be impossible, it will nevertheless be possible to ascertain that an invisible body *is present*, at the very least: there is no (oscillatory or rotational) motion without an acceleration, and no gravitational acceleration without the presence of some mass.

Telescopic searches (such as Schaeberle's) will be able to check for the existence of the previously unknown companion; observational verifications count as confirmations of Newtonian Mechanics if the size and position of the new body can be put into calculations which recover the motions of the primary body. More exactly: If the distance to the binary can be determined, then the apparent separation of the two stars can be turned into a true distance. From the observed period of revolution and this distance, the mass of the system as a whole can be estimated (the complication is in determining the inclination of the orbit to the line of sight, and the degree of eccentricity of the orbit, both of which are needed to allow the distance to the binary to be used to turn the angular separation of the members into a radial distance), as follows. "According to Newton's laws of motion and gravitation, the total mass of two stars orbiting each other is related to the average distance between them, a, and their orbital period, P. If the masses are M_1 and M_2, then $M_1 + M_2 = a^3 / P^2$" (Seeds 1989, 161). (A line joining the two stars always passes through the centre of mass, so that both stars have the same period: see Figure 3.)

Figure 3.



The so-called "H-R diagram", named after the two main contributors to its development, Hertzsprung and Russell, appears to supply us with an independent way of determining the so-called visible mass of a star given only its spectral characteristics. (See the final section of this chapter on the H-R relation and calculations of so-called "visible mass".) If the dynamical and visible mass techniques were truly independent ways of learning the masses of stars, the convergence of these two measures would provide confirmation of the assumptions implicit in the dynamical measures. However, the appearance of independence in the case of measuring the masses of binaries is illusory since the H-R relation is calibrated by appeal to dynamical measures of the masses of binary stars.

Modern dynamical measurements of the masses of stars involved in binary systems, as well as the detection of invisible companions on the basis of otherwise unexplained motions of stars, are helped by spectrographic techniques. Observations of cyclical patterns of red-shifting and blue-shifting of the light of a star give precise orbital periods, as well as the component of velocity along the line of sight, which make calculations of the mass of the system easier, and less affected by the sources of error mentioned above. In this way the masses of many binary systems have been measured, and the discovery of the existence of invisible stellar companions around many stars has been made possible. The first claimed detection of an extra-solar planet was made in 1995, and as of this writing about 32 extra-solar planets have been discovered in this way (almost all are at least several times more massive than Jupiter, because the observations are not yet sensitive enough to detect motions caused by less massive planets, although a very recent report says that a few Saturn-sized planets, about one third of Jupiter's mass, have now been detected (Sky & Telescope News Bulletin, March 31, 2000).[8]

---

[8] See http: www.physics.sfsu.edu ~gmarcy planetsearch planetsearch.html and links therein for more information. The 11 August 2000 Sky & Telescope News Bulletin reports that the total of known extra-solar planets has now risen to 50. Note that there was a brief controversy over whether the Doppler shift of a star's light really indicates the existence of extra-solar planets when David F. Gray of the Department of Physics and Astronomy at the University of Western Ontario offered a competing hypothesis. He suggested that the Doppler shifting of the light could be due to intrinsic motions in the star itself, such as

The detection of invisible companions, and the measurement of the dynamical mass of individual stars is, however, of relatively little importance for the modern dark matter problem, except perhaps as a way of testing the dynamical techniques themselves although no one really doubts the validity of the techniques at this scale.[9] One benefit of this way of detecting invisible matter, though, is that it points to the existence of low-mass, low-luminosity objects such as brown dwarfs and jupiters, which are now thought to exist in fairly large numbers on their own, that is, not in binary systems. These low-mass objects will be evaluated as candidates for the dark matter in Chapter 5: if some mechanism for producing them in large enough numbers can be found, or if observation reveals a sufficient number of them, these ordinary objects could contribute a significant fraction of the galactic dark matter.

3.3 JAN OORT (1932, 1960 AND 1965) AND THE MASS OF THE MILKY WAY

Another sort of dynamical measure of mass for many-body systems was developed by Jan Oort in 1932, drawing on earlier studies by Jeans and by Kapteyn (both

---

cyclical risings and fallings of the surface (Gray, 1997). Recent observations (Gray and Hatzes, 1997) seem to rule out Gray's hypothesis: there are no harmonics in the frequency shift, as we would expect from the vibration of body like a star, and there are no brightness variations that would go along with changes in the or shape of the stellar surface. Still more recent observations indicate massive planets (several times Jupiter's mass) orbiting other stars: it seems impossible to explain the periodic variations except as due to the gravitational influence of orbiting bodies. (The April 1999 issue of *Scientific American* describes a new, non-dynamical technique for detecting extra-solar planets. Observations of "banded" accretion disks around young stars are taken as indicating the presence of massive planets, on the hypothesis that the gravitational field of a planet is responsible for sweeping clean a path through the dust of the disk. Also, at least one eclipse of a distant star by one of its own planets has now been observed.)

[9] Observations of binary pulsars and some other exotic systems provide a way to test GR. Pulsars emit extremely regular radio pulses; the Doppler shift in the frequency of these pulses gives the orbital speed along the line of sight very precisely, and the cyclical pattern of Doppler shifting gives a very accurate measure of the period of the orbit. Over time these orbits are found to decay in a way that is explainable by a loss of energy from the system that exactly corresponds to the amount of gravitational radiation GR predicts ought to be emitted by such systems.

in 1922; the phrase "dark matter" is already in use in these works, albeit as a description rather than as a name picking out a new kind of matter). In these studies the mass of the Milky Way is determined from statistics of observations of stellar velocity dispersions; in Oort's version, the motions and distances of stars perpendicular to the galactic plane are used to calculate the local mass density of the disk, which can then be compared against observations of the disk luminosity to yield a disk mass-to-light ratio.

The main motion of stars in spiral galaxies such as our own is a circular orbit around their galactic centre, but a small proportion of their total velocity may be an oscillation above and below the galactic plane. One can imagine a carousel on which the horses bob up and down while the whole thing rotates: this method of measuring the mass of the galactic plane has by analogy been called the "carousel technique". If we could track the height of a star above the galactic plane together with the component of its velocity perpendicular to the plane through an entire oscillation, it would give a reliable measure of the mass of the plane in the local region. However, "The time taken for a star to perform one vertical oscillation is measured in millions of years. It is not therefore possible for any direct observation to be made of the [path] of a single star" (Tayler 1991, 54). We are forced, then, to rely on statistical measures. Once again assuming that the stars observed are indeed gravitationally bound to the galactic disk and are not in the process of escaping from or collapsing onto it, we can use "A count of the density of stars at different heights above the galactic plane, together with a study of the distribution of [their vertical components of velocity]. . .[to measure] the gravitational force that is slowing down their motion" (Tayler 1991, 54), and thereby obtain a value for the mass of disk.[10] The basic idea is that the higher and faster a star is moving, the

---

[10] Modern observations show that there are relatively few "free stars" in intergalactic space. Those that are present are thought to have been ripped from galaxies by tidal forces in close approaches or actual collisions between galaxies. Ferguson, *et al.*, 1998, find that about 10% of the stars in the Virgo cluster are intergalactic; Theun and Warren, 1997, find that up to 40% of the stars in the Fornax cluster are free (although their error bounds are larger); since stars make up a relatively small proportion of the mass of galaxies, the contribution of intergalactic stars (a still smaller fraction of the stars belonging to those galaxies involved in collisions) to the overall cluster mass will be negligible. This is strong evidence to suggest that the stars in velocity studies such as Oort's are indeed gravitationally bound to the system.

greater the mass of the plane must be in order to keep the star from escaping the system. Technical limitations mean that we can actually perform this kind of observation only for our own galaxy (it would be extremely difficult to measure the motions of individual stars inside distant galaxies); this is nevertheless an important sort of test because our position within our galaxy makes it very difficult to study its overall rotation curve, particularly external to the solar circle.

Oort's research led him in 1936 to the conclusion that the disk of our galaxy contains about twice as much mass as is indicated by the observed flux of light. Oort there finds a mass-to-light ratio (**M/L**) for the local disk of 1.8 (Oort 1936, 286). By the time of his follow-up study in 1960, the observations were reliable enough that Oort was willing to assert the existence of a large amount of dark matter: he claims that low-mass, faint stars below the luminosity threshold then observable "must make up the unidentified 40 per cent of the total mass density near the sun" (Oort 1965, 500). "The result of Oort's original discussion was that the local mass density was 0.092 $M_{\odot}\ pc^{-3}$ [solar masses per cubic parsec] and that the mass of known stars was only 0.038 $M_{\odot}\ pc^{-3}$. When he gave a further discussion almost thirty years later, the total density was estimated to be 0.15 $M_{\odot}\ pc^{-3}$ with 0.08 $M_{\odot}\ pc^{-3}$ known in the form of gas and stars" (Tayler 1991, 55). That is to say, in the first study the total mass was about 2.24 times greater than the mass of known stars, and in the second study the total mass was about 1.88 times greater than the visible mass. This corresponds to M/L ratios for the central plane of the disk of about 2.2, and for a cylinder of about 3.8 (Oort 1965, 473). However, as we will see in Chapter 4, more recent studies of this type now find a much smaller dynamical discrepancy in the disk of the Milky Way, in part because of increased

---

Otherwise, one would expect to find more stars in the inter-galactic spaces, either escaping from a system or moving in to maintain the overall apparent structure of the galaxies, without actually forming gravitationally bound systems. The "free star" counter-example will not apply to consideration of the rotation of external galaxies in any case. Since there are few free stars, and *all* the stars in all the observed galaxies are moving too quickly to be gravitationally bound to their systems by the visible mass in them alone, it would be too much of a coincidence for all the stars to be "accidentally" arranged as galaxies.

accuracy and in part because of increased technical capability to detect dim electromagnetic emissions across the whole spectrum.

Although Oort's "carousel technique" can only be used to measure the mass of the plane of the Milky Way, if we are willing to grant that other spiral galaxies are fundamentally similar to our own (there are good reasons to grant this, although I will not discuss them here), it nevertheless provides an interesting possible constraint on the *distribution* of mass in other spiral galaxies, in particular giving us some indication of what proportion of the total mass is to be found in the plane. And, of course, being able to determine the overall distribution of mass is in turn an important constraint on the *nature* of the dark matter—that is, we learn thereby that the dark matter has to be a kind of stuff capable of having a given distribution.

3.4 BABCOCK (1939) AND THE ROTATION OF THE ANDROMEDA NEBULA

The rotation of galaxies outside our own is amenable to investigation using the same techniques as are used in the study of stellar companions, although there is little hope of reliable optical detection of the motion of individual stars or gas clouds and as a result spectrographic observations of large parts of each galaxy are used. Observers look for a characteristic pattern of Doppler shifting of star light, from which the rotation rate of the galaxy as a whole can be determined. In this case, the light will be red-shifted on the side of the galaxy receding from us, and blue-shifted on the approaching side by the same amount. This difference allows the red-shift due to the Hubble expansion and any peculiar motion of the galaxy as a whole along the line of sight to be factored out, so that we get a true value for the rate of rotation.[11] That is, we get the true value of the

---

(The existence of various kinds of structure common to many different galaxies with a great range of ages also points to their being stable, causal structures rather than merely accidental conglomerations.)

[11] The observed redshift on one side of the galaxy is due to both the Hubble expansion and the intrinsic rotation of the object; the rotation on the other side is indicated by the blue-shift due to rotation towards the observer, plus the Hubble recession. (Note that the contribution of the component of a galaxy's peculiar motion along the line of sight is swamped by the Hubble motion, so we may ignore it.) For the Doppler shift of the receding side and the approaching side we have respectively $s_a = v_r - H_0$ and $s_b = -v_r - H_0$.

component of the rotation along the line of sight: this means we are unable to use this technique to study galaxies which rotate perfectly in a plane perpendicular to the line of sight, but this is a rare configuration and there are plenty of other galaxies to study. In all other cases, we are able to measure that fraction of the rotation that is along the line of sight: for galaxies that are exactly edge-on, the observed rate of rotation is the true rate. (To find the masses for systems not viewed edge-on we can combine the following equations: $\Phi = -\alpha G m^2 r$, $2KE - \Phi = 0$ and $KE = 0.5m\langle v^2 \rangle$, where Φ is the gravitational potential, and KE is the kinetic energy. This yields $m = r\langle v^2 \rangle \alpha G$, where $\langle v^2 \rangle$ is the mean of the squares of the velocities and α is a factor depending on the mass distribution of the system, but which is usually of order unity. See Tayler 1991, 194.) For almost all inclinations of the plane of rotation to the line of sight, then, the value of the rotation determined from the Doppler shift will be *less* than the true value (a portion of the rotation will *not* be along the line of sight, and is therefore unavailable to us). This, in turn, means that the dynamical mass calculated from the rate of rotation will be the *least possible* mass given the observed Doppler shift due to rotation. The true mass of a galaxy is a function of the measured Doppler shift and the inclination of its plane of rotation to the line of sight. (It is possible to estimate the inclination and, by assuming that the orbits are circular and lie in the plane of the disk, to use the inclination to reduce the error in the dynamical mass estimate.)

Babcock, in his 1938 doctoral dissertation, was the first to attempt this sort of detailed study of the rotation of another galaxy. Some earlier work had been done on the radial and internal motions of galaxies, but the results were quite uncertain besides showing *that* there was both radial recession and possibly internal rotation. Vesto Slipher in 1914 "reported the detection of a curvature in the spectral lines of some [galaxies] seen edge-on", and in 1915 Slipher published a spectroscopic investigation of the radial motions of galaxies in which in general he found that they were receding (Crowe 1994,

---

Therefore, $s_a - s_b = v_r - H_0 - v_r - H_0 = 2v_r$. Thus the Hubble factor drops out even without our needing to be sure what its value is, and we get the true velocity of rotation (the component of rotation along the line of sight, that is) by taking half of the difference between the Doppler shifts measured on opposite sides of the galaxy (at the same radii).

239). The refinement of this technique for spectrographically determining the radial recessional velocities of distant bodies was all that was required before it could be used to investigate the motions of different parts of a single galaxy, that is to say, to study the rotational motion of galaxies. Adriaan van Maanen in 1916 published a study of the proper motions of stars in the galaxy M101, a face-on spiral in which he claimed to have detected rotational motions across the line of sight, although it was soon shown that van Maanen's claimed motions were simply not present.[12] Arthur Stanley Eddington (1994 [1917]) summarises the very early studies of radial and internal motions of spiral galaxies.[13]

Babcock picked as his object of study the Andromeda Nebula, M31, the closest independent galaxy, a spiral (like the Milky Way) which we see essentially edge-on. His results were surprising—so surprising that the astronomical community completely disbelieved them, and he decided to switch from extra-galactic to solar astronomy for the rest of his career. (Trimble 1993, 149) In fact, it was not recognised until the 1970s that

---

[12] It turns out that van Maanen's results were entirely spurious. Given our present knowledge of the distance and actual velocities within M101, we know that no positional shifts of stars could possibly have been observable by him, given the relatively short time between plates and the maximum precision of the plate measuring techniques available to him. Further, were the positional shifts as large as van Maanen had claimed, they would have corresponded to internal velocities within M101 in excess of the speed of light (van Maanen reported much lower velocities because he took M101 to be a feature of our galaxy and therefore much closer to us than it really is; in fact, he used his measured velocities to argue that nebulae must be internal to the Milky Way). The van Maanen controversy is discussed in Hetherington 1988 (83-110). Hetherington's conclusion that the incident is an example of the failure of scientific objectivity and an illustration of the fact that observers see what they want or expect to see, seems to me to be quite overstated. See van Maanen 1916, and Hetherington 1988, 83-110.

[13] James Jeans and Johannes C. Kapteyn, two giants of early twentieth century astronomy, both published works in 1922 in which they calculated the local mass density of the Milky Way from stellar velocity dispersions and distributions above the galactic plane (respectively they found values of 0.143 and 0.099 solar masses per cubic parsec; see Trimble 1990, 356). It is an interesting historical question—one to which I do not have an answer—who was the first person to suggest that motions of stars in a galaxy could be used to determine the masses of the host galaxy.

the rotation pattern discovered by Babcock is typical of almost all galaxies. (Trimble 1987, 432) Babcock's detailed optical study of M31's pattern of rotation, which involved taking the "rotation curve" for the galaxy (the velocity of rotation as determined by the Doppler shift of the 21cm hydrogen line, at various radial distances from the centre of the galaxy) revealed two surprising results. First, the absolute rate of rotation of the galaxy is much higher than would be expected given estimates of its mass from the amount of visible light. Second, while the "Keplerian" expectation (see Chapter 2) is that the speed of rotation should drop off asymptotically to zero once the radius considered exceeds the radius which encloses most of the mass, in contrast the rotation curve of M31 was found by Babcock to be essentially flat, and perhaps even still *rising* at the visible extremity of the galaxy.[14] In other words, the rotation velocity was found not to decrease with distance, to the limit of the visible light. These results are robust, and still hold in modern observations of almost every spiral galaxy (see Chapter 4).

As in the case of Mercury's excess perihelion precession, the discrepancy between the observed and expected patterns of motions in M31 (and in other spirals) demands solution in one of two general ways, either by admitting the existence of more matter than we thought was present, or by revising the law of gravitation. Granting that the law of gravity used to calculate the Keplerian expectation is correct, what Babcock's results show is that (1) there is much more mass present than is indicated by what is visible in the galaxy, even including reasonable estimates of the quantity of unseen gas and dust that ought to be present, and (2) the mass is not distributed where the light is. This second result follows from the fact that the rising rotation curve cannot be sustained in a gravitationally stable system in which the mass is distributed as the light is in M31 (that is, almost entirely in the plane of rotation, with the highest mass concentration near the galactic centre—the stability result was proved in numerical simulations by Ostriker

---

[14] To be more clear: If the mass in M31 (or other spirals) were distributed just where the light is, that would mean that the mass is concentrated towards the central bulge. If this were so, we would expect the rotation curve to be highest near the centre of the galaxy, and to drop off toward zero at the limit of the light. Since this rotation pattern is not what is observed, we know that the mass in M31 (and other spirals) is not distributed as the light (or the Newtonian limit of General Relativity does not apply to galaxies).

and Peebles, 1973.) This information about the *distribution* of the dark matter (more than information about its *amount*) in galaxies and clusters is important for determining just what the dark matter is. (In particular, it has to be a kind of stuff whose exclusion from the galactic plane can be naturally explained.) Trimble (1987, 432) recalibrates Babcock's data to take account of modern estimates of the distance of M31, and finds a value for the mass of M31 of 3×10^{11} solar masses, and a mass-to-light ratio of 17, out to 18 kpc (kiloparsecs—a parsec is the distance of an object whose annual parallax is one arcsecond, or about 3.26 light years). The negative reaction to Babcock's conclusion is perhaps less surprising when we note that at around the time of his study the accepted mass-to-light ratio was about 2 or 3 (Oepik (1922) gives M/L = 3.2 for the Milky Way, and as we saw above Oort's 1936 study found a value of 1.8).

3.5 SINCLAIR SMITH (1936) AND THE MASS OF THE VIRGO CLUSTER

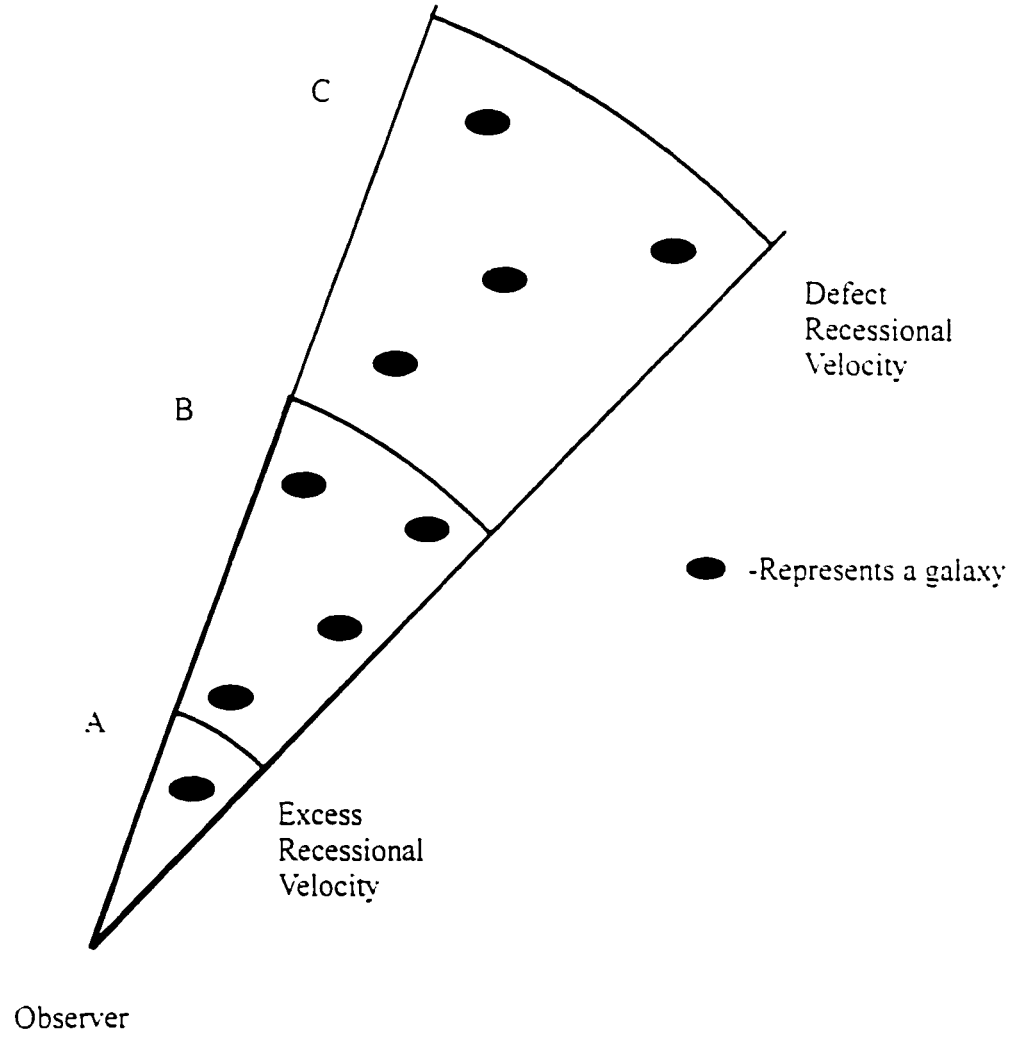
As Smith points out, the mass of a cluster of galaxies can in principle be taken without introducing whatever uncertainties there may be in calculations of visible mass: just compare the *dynamical mass of the cluster as a whole* with the *sum of the dynamical masses of the component galaxies*. Whatever excess there is must be due to the presence of invisible intra-cluster material, not gravitationally bound to any one galaxy in the cluster. (This intra-cluster material will include any additional member galaxies too faint for us to detect.) It follows that the dynamical dark matter problem for clusters is even more firmly established than is the dynamical dark matter problem for individual galaxies, since the discrepancy does not depend on assumptions about visible mass or mass-to-light ratios.

Smith is somewhat mistaken here, although he is also right in an important sense. While it is true that establishing the relative abundances of dark to visible matter requires a comparison of the type mentioned, it is however also true that *by itself* the fact that the rotation curve of a galaxy (or the velocity dispersion in the case of elliptical galaxies or clusters of galaxies) does not drop off asymptotically as the radius goes to the limit of the visible light shows that there is a great deal of unseen mass in the galaxy. (This is because if the visible mass were all the mass present, the rotation curve or velocity dispersion *would not* remain flat out to arbitrarily high radius.) However, from this

information alone we cannot quantify the amount of unseen mass. That is to say, the shape of the rotation curve, considered in concert with Newtonian assumptions, tells us that most of the mass of the system cannot be distributed as the light is, in the plane of the galaxy with density varying roughly as the inverse square of the radius. So, merely from the *shape* of the rotation curve we can know *that* there exists a large proportion of dark matter in the galaxy; it is important to emphasise the fact that it is *not* the case that we can know about the existence of dark matter *only* by comparing the dynamical mass with the visible mass. But we do need additional information and background theory in order to determine just how much extra mass there is. Smith is exactly right, however, that his method provides an absolutely certain measurement of the total dynamical dark matter that is special to the cluster, as opposed to the dynamical dark matter contained in the individual galaxies that make up the cluster. In any case, it turns out that the errors in the estimates of visible mass (introduced by faulty background assumptions or in other ways) will be swamped by the sheer size of the discrepancy between the visible and the dynamical mass. So, we do not need perfect knowledge of the quantity of either visible mass or dark mass in order to know that the dark matter problem exists. It is only later, when we are trying to solve the discrepancy by finding its exact cause, that we need to make the calculation of the ratio of visible to dark mass as accurate as possible, and to know just what are the error bounds on our estimates of visible and dynamical mass.

There are, nevertheless, various indeterminacies in Smith's cluster procedure. Most notably, we must assume that the cluster is gravitationally bound, that it is not merely an accidental conglomeration of galaxies in one place on the sky. This assumption cannot be conclusively proven, for the simple reason that we do not have time enough to determine whether the cluster is collapsing or evaporating (collapse or evaporation will take place over extremely long time scales, and any given snapshot of this process will be effectively indistinguishable from a similar gravitationally bound system). This sort of objection can still be posed even with regard to the modern dark matter evidence. We have to use the very same observations of redshift in order to determine both cluster membership and peculiar velocity. (See Figure 4.)

Figure 4.



When we observe galaxies, our main clue to their distances is the amount by which their light has been red-shifted (a few of the closest galaxies can have their distances estimated by other means). Almost all clusters appear to have a largest and brightest member near their centres. Taking these to define the dynamical centres of clusters, as seems reasonable given models of cluster evolution (Dubinski, 2000), allows us to use their redshift as a baseline for determining cluster membership. It would seem, then, that in order to know the true distance of a galaxy, and thus whether or not a group of galaxies are in close enough proximity to be a true dynamical system, we just need to know the value of the Hubble constant precisely. Even if we do not know the Hubble constant precisely, we could know that all galaxies with the same redshift are at the same distance. This convenient way of proceeding is unfortunately complicated by the fact that galaxies (especially those in clusters) have large peculiar velocities as well as cosmic-expansion-induced motions. If a galaxy happens to have an exceptionally large peculiar velocity toward or away from us, its redshift will make it seem (according to the Hubble relation) closer or farther than it really is. Thus galaxies that are actually in the foreground or background of a cluster, and not actually members, could be mistaken for cluster members in some situations. Considering the velocities of these interlopers, as computed by their redshift, to contribute to the kinetic energy of the cluster will artificially increase the Virial mass of the cluster. The question, then, is whether this problem is enough to significantly affect the dynamical discrepancy. It seems as likely that exceptionally fast galaxies that are members of the cluster but which are moving toward or receding from us will be considered to *not* be part of the cluster, leading to an underestimate of the Virial mass. So long as interlopers are only a small proportion of the total number of galaxies taken to comprise a galaxy, the Virial mass will only be off by a small factor. And since the dynamical discrepancy for clusters is of the order of a factor of 100, the discrepancy swamps all reasonable possible error in the Virial mass measurement: the discrepancy is real, and large.

According to Smith we can be reasonably confident that the Virgo Cluster is gravitationally bound. He notes that if the cluster *were* collapsing or evaporating, then we should expect to observe a concentration of the highest velocity members at the centre of the cluster: his results—which rely on observations of the radial velocities of 25

galaxies taken by Humason, five by Slipher, as well as nine of his own—show that the distribution of velocities is fairly even across the cluster. There is, furthermore, no dependence of velocity on brightness: the absence of a "magnitude effect" indicates that the cluster is likely to be a true cluster. If there were a dependence of velocity on brightness, the best explanation of it would be that foreground galaxies (brighter because they are closer) have been counted as cluster members when they really are not.

Smith's data lead him to conclude that, probably, "the outermost particles [of the cluster] move in circular orbits with a speed of 1500 km/sec" (Smith 1936, 29). If this is correct, we can calculate the total mass using either $m = v^2 r / 2G$ or $m = v^2 r / G$. (Either equation will do: the factor of one-half is negligible because the discrepancy between the dynamical cluster mass and the sum of the dynamical masses of the component galaxies is so much greater than double). Assuming a circular orbit with a radius of 2×10^5 parsecs (one tenth of the cluster's distance from us—a reasonable assumption which can be checked by refining the Hubble distance estimate), Smith calculates the mass of the cluster as 2×10^{47} grams, or 10^{14} solar masses. Then, "Assuming 500 [galaxies] in the cluster and no inter[galactic] material, we find for the mean mass of a single [galaxy]... 2×10^{11} [solar masses]. This value is some two hundred times Hubble's estimate of 10^9 [solar masses] for the mass of an average [galaxy]" (Smith 1936, 29). According to Smith, the only large source of error is the assumption that the cluster is not merely a statistical fluctuation in the background distribution of galaxies, but the probability of this assumption being incorrect, he says, is "extremely small" (Smith 1936, 29). The argument here is of the same sort as Michell's argument that double stars are very likely to be physically bound systems. The discrepancy between the mass of the galaxy as determined dynamically and as determined by the sum of the masses of visible component galaxies is extremely large. It is also unexplained: Smith's calculation should be approximately correct—certainly, the margin of error is much less than a factor of 100—and therefore either Hubble's estimate of the average galactic mass is wrong, or there is a large amount of unseen material either throughout the cluster or perhaps distributed "in the form of great clouds of low luminosity surrounding the [galaxies]" (Smith 1936, 30).

3.6 FRITZ ZWICKY AND THE MASS OF THE COMA CLUSTER

Fritz Zwicky is one of the most important figures in the early history of the dark matter problem. Not only did he improve and extend known dynamical techniques for measuring the mass of astronomical systems, he developed new ones, and he was one of the first advocates of the presence of large amounts of dark matter in astronomical systems. In his earliest studies (for example, Zwicky 1933), like everyone else in the period, he did not expect that the missing mass in galaxies and clusters was anything other than ordinary matter in some dim form, such as low-luminosity stars, dust or gas clouds. This was a sensible position in that at the time instruments and observations were in a primitive enough state that it was impossible to rule out the chance that the extra mass was of this ordinary, merely dim but not dark, character.

Zwicky (1937) is a *tour de force*: in it, he reviews and critiques everything that was known up to that point about how to determine the mass of distant astronomical systems, gives a sophisticated account of how to extend to large, complex systems and apply to real situations the dynamical techniques for measuring mass, and invents the idea of using gravitational lensing to measure galactic mass. Furthermore, he reports results showing that the Coma Cluster must contain much more mass than would be expected from standard mass-to-light ratios.

One of the most significant facts about Zwicky (1937) is that the different techniques of dynamical mass measurement that it discusses are *independent*, and therefore can in principle be used to double-check one another:

> Each of the three new methods for the determination of masses of [galaxies] which have been described makes use of a different fundamental principle of physics. Thus method iii is based on the Virial theorem of classical mechanics; method iv takes advantage of the bending of light in gravitational fields; and method v is developed from considerations analogous to those which result in Boltzmann's principle in ordinary statistical mechanics. Applied simultaneously, these three methods promise to supplement one another and to make possible the execution of exacting tests to the results obtained. (Zwicky 1937, 245)

The three methods apply under different circumstances or to dynamical systems of different kinds, however, and this limits the practicality of comparison between the methods. Method iii, says Zwicky, is useful only in the investigation of clusters (modern thought on the applicability of the Virial Theorem to elliptical and even spiral galaxies

disagrees with Zwicky's judgement here—various additional assumptions and correction factors must be considered, but even including these factors the ratio of missing mass to margin of uncertainty is still so large as to make Virial mass estimates of spiral galaxies quite usable). For method iii to give very accurate measurements, we must also know with high precision the radial velocities of the particles, and the real size of the cluster. Not knowing these factors with high precision does not preclude the possibility of using the technique, but only makes its results somewhat less reliable. The margins of error in the mass estimate depend on the precision of the inputs—what we learn, then, is that the Virial mass falls within some well-defined range (provided that the weak and plausible input assumptions are not far from correct). In most dark matter situations, with respect to detecting the discrepancy at least, the degree of the discrepancy swamps the probable error (the discrepancy is of the order of a factor of 100, while the errors are of a factor between 2 and 10 or more: Tayler 1991, 60). Greater precision is needed when the goal is to describe the dark matter as precisely as possible in order to constrain possible solutions.

Method v, "enables us to find the masses of all types of [galaxies], provided the absolute mass of a single type of [galaxy] is known" (Zwicky 1937, 245). It does so by giving us a way to rank the relative masses of different types of galaxies; once the relative masses are known, a single absolute mass calibrates the whole scale. This depends on there being an appropriate physical significance to the morphological characteristics of galaxies, such that all galaxies with similar morphologies have similar mass profiles (both magnitude and distribution). There seems to be no *a priori* reason to expect this to be true—why should it be the case that galaxies that have similar appearances cannot come in a very large range of masses?—but it is nevertheless still an assumption of modern astronomy that morphology is a reliable guide to the real physical characteristics of astronomical systems.[15] One plausible reason to accept the idea that

---

[15] Older work in particular takes seriously the idea that objects with similar appearances have similar physical characteristics, for example, stellar classification systems based on spectrographic observations, Hubble's quasi-evolutionary classification scheme for galaxies based on appearance, etc. Zwicky (1957) developed a (somewhat influential) philosophy of science which he called "the morphological approach."

morphology reliably indicates kinds of physical structures is that each galaxy of a given type is likely to have been formed by the same physical (causal) process over roughly the same period of time, and it seems plausible that the differences between types can be explained by different processes operating, where each such process operates on a certain range of initial masses, or perhaps where each process automatically forces the mass to a characteristic range.[16]

Method iv, using gravitational lensing, depends on finding cases of gravitational lensing and on being able to determine accurately the deflection angle and the absolute distance of the lensing body. At the time of Zwicky's writing, instances of gravitational lensing of background by foreground galaxies or clusters were purely theoretical (the first case was observed in 1979, and as of recently, about 50 cases of gravitational lensing were known; see Chapter 4):

> Since method iii gives only the average masses of [galaxies in the cluster] and method v furnishes only the ratios between the masses of different types of [galaxies], much depends on whether or not a single image of a [galaxy], modified through the gravitational field of another [galaxy], can be found. A single case of this kind would, so to speak, provide us with the fixed point of Archimedes in our attempt to explore the physical characteristics of [galaxies]. (Zwicky 1937, 245)

As will be discussed in later chapters, it is still true today that instances of gravitationally lensed background objects are perhaps the most crucial piece of missing evidence in the quest to solve the dark matter problem. This is because gravitational lensing measures of mass are based on different fundamental assumptions than, and are independent of, dynamical measures. Gravitational lensing therefore provides an independent check on dynamically determined masses. The gravitational lensing cases studied so far confirm the order of magnitude of the total mass of typical spiral galaxies, though it is important to note that no single galaxy has had its mass measured in both

---

[16] According to Dubinski (2000), it seems likely that all or most galaxies start out as spirals, and some of them evolve by gravitational interactions or mergings with other galaxies into ellipticals. His computer simulations (though still quite rough—individual mass elements in the simulation are five orders of magnitude greater than a solar mass!) start from cosmological parameters and end up with clusters of about the right size and with about the right proportion of galaxy types and sizes.

ways (galaxies that are gravitational lenses are usually too dim for the detailed spectroscopic work necessary to obtain rotation curves, though this may change with improved technology). On the assumption that General Relativity (**GR**) is correct, the two kinds of measurement are mutually supporting of each other, and this in turn provides an interesting constraint on alternative theories of gravitation offered as solutions to the dark matter problem: they must explain why, if GR is wrong, these two measures agree. And they must account for the gravitational lensing effects as well as the rotation without recourse to dark matter. Zwicky's work is remarkable both for its prescience, and because it points out to a modern reader how little fundamental progress has been made in the last sixty years in developing new techniques of astrophysical mass determination.[17]

3.7 THE STATE OF THE EVIDENCE FOR DARK MATTER TO 1970

Virginia Trimble (1990) notes that the early work (up to the end of the 1930s) indicating the existence of a dark matter problem was essentially ignored by the mainstream astronomical community at the time, partly because there was no clear way to proceed towards a solution, partly because it seemed possible that the results were due to unknown systematic errors, and partly because the results diverged so greatly from astronomers' expectations. It was not until the mid-1970s that a "critical mass" of astronomical opinion was achieved, and the astronomical community as a whole began to recognise that the dark matter problem was real and important.[18] Work did continue,

---

[17] No new tests for the detection of missing mass have been introduced since Zwicky—the dynamical resources of our current physical and gravitational theories appear to have been exhausted. Progress in this arena has consisted merely of refinements of precision of measurement (thanks to improvements in telescopes, spectroscopes and image recording and measuring devices), and of applications of these techniques to new problem situations. (With regard to *candidate particle solutions*, however, many new kinds of tests have been developed over the last 50 or so years; see Chapters 4 and 5.)

[18] It is an interesting historical question why, although the evidence itself was essentially no different in 1970 as opposed to 1940, opinions about the status of the dark matter problem were so different at those times. The following strikes me as a likely explanation. The fact that astronomers were in this period

however: between 1939 and the mid-1970s several new studies were done that essentially confirmed the early results of Babcock, Oort, Zwicky and Smith. The catalyst for the shift of opinion was a pair of papers (Ostriker, *et al.*, 1974, and Einasto, *et al.*, 1974) reviewing the (by then fairly extensive) evidence for dynamical discrepancies in astrophysical systems. (Trimble 1990, 359) By 1961,

> opinion had crystallized around two strongly opposing views. The rich clusters must either be bound by dark matter, associated more with the clusters as a whole than with the individual galaxies, or the clusters must be short-lived and currently expanding out of some fairly violent explosion. (Trimble 1990, 358)

There were, however, some dissenters:

---

being overwhelmed by new and startling discoveries would have set up a kind of expectation that discrepancies of the sort discussed above would eventually be naturally resolved by future observations. The picture of the universe and its constituents was also in this period relatively unsettled, so that an attitude of skeptically waiting for new information would have been natural. Furthermore, to give up very basic theories such as Newtonian gravitation and matter theory as then understood would have made progress in other areas impossible.

Some other relevant factors were more sociological or accidental, rather than evidential. Babcock, as a neophyte astronomer, did not have the reputation to be able to back up the radical change in astronomical theory that his results implied, and the fact that he left galactic astronomy altogether (and thereafter worked in solar astronomy) because of the negative reaction to his dissertation made it even easier to dismiss or ignore his results. Oort was a very well established astronomer, but he downplayed the implications of the results of his first study. Zwicky had a reputation as something of a maverick and dabbler (as it turns out a brilliant one), and even he expressed his initial results conservatively and suggested that the discrepancy would likely be solved by the discovery of dim but ordinary matter. No doubt the Second World War interrupted astronomy generally, and by the time things got going again after the war the hot topics were elsewhere: cosmology and large scale structure, General Relativity, stellar evolution and composition, quasars, and radio astronomy, to name just a few things that captured attention in the post-war period. Finally, it was not until the 1970s that particle physics suggested (however roughly) some ideas about what the unseen matter could possibly be if it were not gas, dust or stars. Given the state of astronomical knowledge generally, the available evidence for dark matter, and the sociological factors just mentioned, it is perhaps no surprise that the significance of the dynamical discrepancies went unappreciated for so long.

> Holmberg (1961). . .believed that observational errors, substructure in clusters, and foreground/background interlopers could account for the large apparent velocity distributions with no need either for dark matter or explosions, and Lemaître (1961). . .proposed that rich clusters might be constantly exchanging galaxies with the field, so that the configurations were permanent but the individual members not gravitationally bound. (Trimble 1990, 359)

These possibilities remain, but are widely held to be either implausible or inadequate as explanations for the vast extent of the mass discrepancy.

The first suggestions that the Newtonian limit of General Relativity might not be the correct way to model the forces holding the clusters and galaxies together were made by van den Bergh (1961) and Finzi (1963). This history, and especially the views of modern adherents of this kind of solution—notably Milgrom and Sanders, who have both written extensively on "MOND" (Modification of Newtonian Dynamics), and Mannheim, who has developed an alternative to GR which I call the Conformal Theory of Gravity (CTG)—will be discussed in Chapter 6. For now, it suffices to note that since there is at present no independent check on which law of gravity operates at large scales, alternative theories of gravity, ones that differ with regard to their predictions on scales larger than the solar system (or even alternative accounts of what forces hold galaxies and clusters together), are quite possible. Therefore, non-Newtonian force laws or other gravitational theories will be viable alternatives to the existence of large amounts of dark matter unless and until some independent reason to exclude them can be devised.

In summary, significant evidence for the dynamical discrepancies in the Milky Way, in other galaxies, and in clusters, began to accumulate in the late 1920s. This evidence, though actually reliable, was for various reasons not widely accepted as significant in the astronomical community until the mid-1970s. Objections to the existence of the dynamical discrepancies ultimately failed because the degree of the discrepancies is so much greater than the probable error in the measurements. As subsequent chapters will show, the various kinds of tests described here are still important in the modern era; the basic outlines of the problem are also the same now as they were in the period discussed in this chapter. This early period lacked plausible candidates for what the dark matter might be: as Chapter 5 discusses, stars, gas and dust,

the only matter candidates proposed in this period, are now known to make up no significant proportion of the dark matter.

3.8 "VISIBLE MASS", M/L, AND THE HERTZSPRUNG-RUSSELL DIAGRAM

Let me conclude this chapter with a discussion of "visible mass", an important concept in the overall discussion here because it is the visible mass of astronomical systems against which we compare the dynamical mass, and thereby discover the dark matter problem. Visible mass is of course a misnomer—we see light, not mass—but it is a useful term nevertheless. It is meant to refer to the total mass we can reasonably infer to be present given a certain observed flux of radiation. It is the purpose of this section to describe how the mass-to-light ratios through which this inference is made are established.

It is an interesting fact that the telescopically observable characteristics of a star (namely, its absolute luminosity and the spectrum of its radiation) can be predicted completely given just three parameters, namely the mass, initial chemical composition, and age of the star. That is to say, the theory of stellar evolution is so far advanced that the luminosity and spectrum of a star with given mass, initial composition and age can be known *a priori*; *vice versa*, given empirical knowledge of any two of these three fundamental parameters for a given star, and the characteristics of its observed light, the value of the third fundamental parameter can be calculated. Thus, if we could determine the age and initial chemical composition of a star with given observed characteristics, we could know its mass. And knowledge of the masses of individual stars would go a long way toward enabling us to acquire knowledge of the proportion of the mass of a galaxy that is due to visible bodies.

It seems at first, however, that we have no way to gain access to the information we require in order to perform this calculation. Clearly, since stars burn for millions or billions of years (depending on how massive they are: more massive ones burn faster), we have no direct access to information about the formation of any star now visible, nor can we hope to track the evolution of any star now forming. For reasons of temporal screening-off, then, we might think that we have no way of knowing any star's age or initial chemical composition. But in fact we *can* know a star's initial composition: since

nuclear burning takes place only at the core of a star and there is no evidence of significant convective mixing of layers, present-day observation of the chemical composition of the outer layer of a star, which is available by spectroscopy, is a reliable indicator of the initial composition of that star.

We have, then, knowledge of a star's initial composition through analysis of its spectrum. We can also easily measure its apparent luminosity: a star's luminosity is also called its "magnitude". Rather than working with apparent visual magnitudes, astronomers prefer to use "absolute bolometric magnitudes". The visual magnitude of a star is the intensity of its light in the limited visual wavelength band, and its bolometric magnitude is the total intensity integrated over all wavelengths. The bolometric magnitude can be measured directly, or calculated from the visual magnitude and a so-called "bolometric correction factor" (Swarzschild 1965, 7-9). One can also easily calculate a star's absolute luminosity once its apparent luminosity and distance have been determined.[19] The remaining information we need in order to compute the masses of stars from the characteristics of their observable radiation is obtained through the Hertzsprung-Russell (**H-R**) diagram.

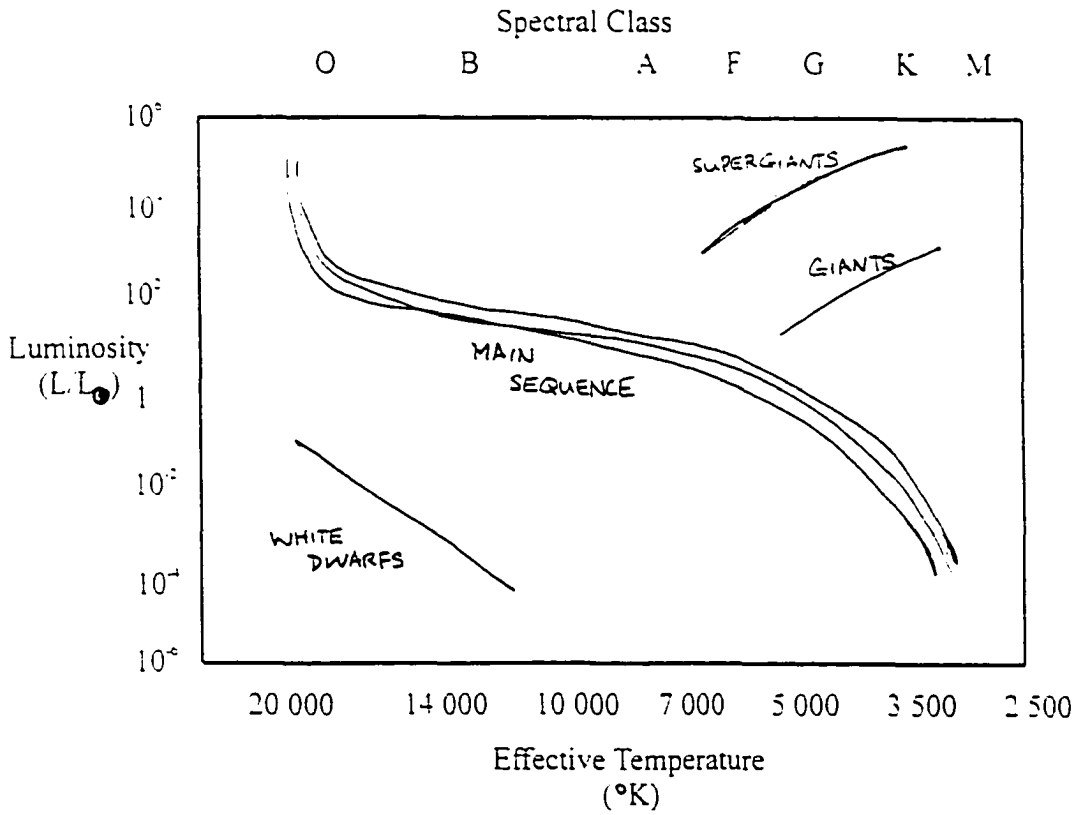
The H-R diagram graphically represents an empirical relation between *spectral type* (a system of categorizing stars according to the characteristics of their spectra, most notably dominant colour) and *absolute luminosity*. This relationship was discovered by Henry Norris Russell in 1913—Russell was anticipated in some respects by Ejnar Hertzsprung, in acknowledgement of which fact the relation and its corresponding diagram are referred to by both astronomers' names (Crowe 1994, 238 and 245; for more on the history of the H-R diagram, a good place to start is Philip and DeVorkin 1977). The distance to a star (as determined, for example, by its annual parallax) calibrates the apparent luminosity and yields an absolute magnitude for that star. Once the empirical relation is established between absolute magnitude and spectral type for a large enough sample of stars, it becomes possible to estimate the distance of a star not in the sample just from observations of its spectral type and *apparent* luminosity. (See Crowe 1994,

---

19 "The apparent magnitude of a star at 10 parsecs distance is its absolute magnitude" (Peebles 1993, 57). The magnitude scale is dimensionless, whereas luminosity is usually expressed in solar units.

especially 243-254.) The diagram that follows is a sketch of an H-R diagram, showing its features relevant to the present discussion. Note that since stars are black-body radiators, the colour of a star (determined by the wavelength at which most of the star's light is emitted) is a measure of its effective surface temperature. (There are also other important features of the spectrum of a star, namely the absorption and emission lines which indicate the chemical composition of the surface layer.) The luminosity of a star with given effective surface temperature varies as its surface area, and thus is a function of radius. (See Figure 5.)

Figure 5
The Herzsprung-Russell Diagram



As it turns out, there exists an additional empirical relationship between a star's position in the H-R diagram and its *mass*. This empirical relation was discovered and calibrated by looking at binary star systems in which both members are stars of roughly the same luminosity and temperature (spectral class), and calculating from their orbital motions the dynamical mass of the pair by the equation mentioned above, $m_1 + m_2 = a^3 P^{-2}$ (where a is the average separation and P is the period). Since the systems are chosen so that the stars are similar in light profile, we may assume they are intrinsically alike, and we can then find the mass of a single star by dividing the dynamical total by two. (More than half of all stars are in binary systems, so systems useful for this purpose are not hard to find (Seeds 1989, 161).) By inductive generalization, this relation between spectral type, luminosity and mass is extended to other stars whose dynamical masses have not been determined but whose spectral types and absolute luminosities are known. Stars on the main sequence of the H-R diagram are ordered by mass, with the most luminous being the most massive, from about 40 to about 0.5 solar masses (Seeds 1983, 166).[20] This relation is known as the stellar mass-luminosity relation, and it is used as the basis from which to calculate mass-to-luminosity (**M/L**) ratios for other kinds of systems composed of stars.[21] M/L ratios for groups of stars, for example, are estimated by counting the number of stars of each spectral type, and making estimates of the total additional mass of unseen matter (gas, dust, dim stars) by extrapolating typical

---

[20] The mass-luminosity relationship for an individual star is roughly described by $L = M^{3.5}$, in solar units. A star four times as massive as the Sun radiates 128 times as much energy. (Seeds 1989, 167.)

[21] The H-R diagram is interpreted as representing the evolutionary history of stars, that is, as a star ages and burns more and more of its hydrogen, its surface temperature (which determines its spectrum of emission by the law of blackbody radiation) and radius (which, with the temperature, determines the total light emitted) evolve, so that at different points in its lifetime the same star will occupy different positions in the H-R diagram. Once a star begins to burn its helium, it leaves the main sequence to become a red giant: depending on its mass it either ends in a supernova or stops burning nuclear fuel and contracts and cools to form a white dwarf. The specific path of a star through the H-R diagram is determined by its initial composition and mass. Since stars spend such a small fraction of their total lifetimes off the main sequence, the contributions to M/L ratios of stars not on the main sequence to galactic M/L ratios are commonly ignored (that is, at any time a very small percentage of the stars will be off the main sequence).

values found in similar systems that have been studied in detail. Thus M/L ratios are always (theory-bound, error-prone) estimates, but nevertheless their reliability is fairly high (within a factor of 2; the largest single source of error for very distant systems is the uncertainty in the Hubble constant), and certainly their reliability is high enough that the error in the estimate of visible mass cannot account for the large discrepancy (many times the probable error in visible mass estimates) discovered by comparison with dynamical measures of mass.[22]

Note that M/L ratios built on the H-R relation are calibrated using dynamical measures of mass. This means that the discrepancy between the visible and dynamical masses is unlikely to be due to errors in calculating the visible mass of the systems in question (provided that all of their light is observed)—the main source of possible error is the inductive generalization to stars whose masses have not been individually checked by dynamical means. But given the quantum mechanical explanation of stellar structure and evolution, and therefore of the mass-luminosity function, the possibility that the generalization is a source of error becomes less significant.

---

[22] Israel (1983, 209) credits Eddington (1924) as the discoverer of the stellar mass-luminosity relation.

CHAPTER 4

MODERN DARK MATTER: EVIDENCE AND CONSTRAINTS

Cosmological observations are always right at the hairy edge of the possible.
—James Gunn, as quoted in Lightman and Brawer (1990), 262.

4.0 INTRODUCTION

This chapter and the next serve three main functions. They review the modern evidence for the dynamical mass discrepancy in astronomical systems of various scales (from the local part of the Milky Way to superclusters), they review the most important of the matter solutions that have been proposed, and they discuss evidential constraints that any such solution must meet in order to be minimally acceptable. I here ignore the second horn of the dark matter dilemma, the possibility that the correct solution to the dynamical discrepancy is a revision to gravitation theory: that issue will be treated in Chapter 6.

The vastness of the literature on dark matter precludes this from being a comprehensive survey—the article on dark matter in the *Encyclopedia of Cosmology* (Trimble 1993, 156), reports that over 200 papers were published in the area in 1990 alone; Trimble (1987) lists 777 important historical sources; an electronic search of just four prominent astronomical journals gives over a thousand "hits" for the last five years; and this is by no means an exhaustive list. The aim of the present chapter, therefore, is to give a sketch of the whole territory while focusing on particular facets that show especially well the patterns of reasoning about evidence that emerge in this field.[1] My

---

[1] An important source of information for this survey was a series of review articles by Virginia Trimble over the last decade or so (in Hetherington's *Encyclopedia of Cosmology* (1993), in the proceedings of a conference on history of twentieth century cosmology (Bertotti, *et al.*, 1990), and most importantly in an issue of the *Annual Review of Astronomy and Astrophysics* (1987). I have also made important use of

approach pays more careful attention to the distinction between dynamical and cosmological dark matter than is usually done. I also provide details of some relevant observational and theoretical studies that have appeared since Trimble's reviews.

The astronomy and cosmology community did not widely accept the existence of the dark matter problem until the mid-1970s. There was some earlier reaction to the increasing number of studies that seemed to show the existence of a radical mass discrepancy, but most workers thought that the studies in question were too doubtful (the likely errors too high or the conclusions too surprising) to be taken seriously, or that improved observations would show that the "hidden" mass was really there all along (that is, that the missing mass was not "dark" but merely *dim*). Trimble notes that this situation changed beginning in 1970 with Freeman's discussion of the fact that galactic rotation curves are commonly flat, and with the publication in 1974 of two review articles that brought clearly into view all of the various kinds of evidence that support the idea of there being a large amount of unseen mass in most astronomical systems. Since then, non-believers have been decidedly in the minority, although as we shall see their challenges to the new orthodoxy raise interesting issues in the philosophy of evidence. (Trimble 1990, 359; Trimble 1987, 426; Trimble 1993, 150)

It seems as though every article published recently about the dark matter problem proclaims its solution. There are rhetorical and practical reasons for this, I suppose, but reports of the death of the dark matter problem have been greatly exaggerated. The very fact that so many supposed "solutions" appear in any given year is by itself an indication that the problem is alive and well. There is no consensus in the field of astrophysics even about in what direction the correct solution lies—whether it should be sought in known or unknown fundamental particles, sub-stellar objects, gas and dust, or new laws of gravity—although there are recent studies which give extra weight to the hypothesis that it is fundamental particles, which seem to rule out sub-stellar objects, and which make the gas and dust hypothesis look implausible. And, so far, every proposed solution faces significant observational and/or theoretical difficulties.

---

books and articles by Bartusiak (1993), Krauss (2000), Rubin (1989), Riordan and Schramm (1991), Tayler (1991), and others.

Although the solution to the dark matter problem is not in sight, more and more relevant evidence is accumulating, and that evidence is useful in constraining the possible solutions. The outlines of the problem are essentially what they were after the early studies discussed in Chapter 3, although the continued (and marked) increase in our technical ability to acquire very-hard-to-obtain data, and to tease useful information from that data, means that a greater quantity and breadth of evidence has been brought to bear on the problem, including some new kinds of evidence. What we learn from this body of evidence is that nearly every astronomical system at every scale that can be studied (beyond the scale of the solar system and perhaps the local galactic plane) must contain very large amounts of dark matter. We have good data about the total mass of these systems, we are getting a better idea about what fraction of the total mass is merely dim, and therefore we are able to say with greater precision what fraction of the total mass is truly dark.[2] Estimates of the ratio of mass to luminosity (the "M/L ratio"), and thus of the proportion of dark matter, increase as the scale of the system considered increases, and the discrepancy between the visible and dynamical masses of almost all astronomical systems studied is huge: depending on the type of system in question and taking account of the probable error in the calculations, 90 to 99% of the mass is unaccounted for.

The present evidential situation with regard to dark matter differs from the historical situation in one significant respect, namely that in addition to the unexpected mass discrepancy, present evidence (derived from theories of Big Bang nucleosynthesis and observed elemental abundances in the universe) also indicates that the dark matter, whatever it is, cannot be ordinary matter. Not only is the dark matter not in "luminous" forms such as stars, gas and dust, there is a strong likelihood that it is not even *baryonic* matter (matter made up of protons and neutrons). Arguments for this will be considered

---

[2] Inferences to the existence of dark matter involve the idea that the dynamical mass is *so* high that it cannot all be ordinary matter, or it would be easily visible (and then there would be no discrepancy). This form of argument does not specify the margin of error in the estimate of visible mass, but simply claims that the degree of the discrepancy far exceeds any reasonable estimate of the error in the value for the visible mass. Determining exactly the mass fraction of ordinary matter is nevertheless important in order to know exactly the proportion of dark matter in various kinds of systems, which is a very important factor in constraining the *character* of the dark matter.

below. But this result, if correct, taken together with the basic fact that the dark matter neither emits nor absorbs noticeable amounts of electromagnetic radiation, indicates (or so some have claimed) that the dark matter is *unobservable*. Dark matter theories have, on this account, been compared to ether theories and Aristotle's quintessence (see for example Bartusiak 1993 and Krauss 2000).

4.1 THE EVIDENCE FOR DARK MATTER

The obvious way to organise this discussion of the evidence for the existence of a dynamical discrepancy is to look first at the local galaxy, and to move outwards to progressively larger astrophysical systems. As we shall see, the fact that the fraction of the total mass contributed by dark matter increases as the size of the system increases, has interesting implications both about the *distribution* of dark matter in various kinds of systems, and about the *nature* of the dark matter.

There are several different ways to measure or estimate the dynamical mass of astronomical systems: velocity dispersions of stars perpendicular to the plane of the Milky Way, rotation curves for spiral galaxies, the motions of globular star clusters and satellite galaxies, velocity dispersions for elliptical and irregular galaxies, velocity dispersions for clusters, gravitational lensing of distant background objects, and microlensing of background sources by objects in the halo of our or other galaxies. Almost all of these methods rely on the dynamical inferences discussed in Chapter 2, including the assumption of gravitational equilibrium; some of them also include a component of statistical reasoning (partly dependent on some general results from thermodynamics, such as the Virial Theorem). But interestingly, "For many of the cases considered, $GM/R \sim V^2$ is all the physics needed" (Trimble 1987, 426; the main exception is gravitational lensing). "Where several methods can be applied to the same part of the same galaxy, results are frequently, but not always, in reasonable agreement" (Trimble 1987, 427). In general, systems of the same kind are usually found to have similar fractions of dark mass (that is, spiral galaxies all have similar dark matter profiles, and so on for other types of astrophysical systems). Furthermore, the gravitational lensing check, where it has been applied, yields results that agree at least to within an order of magnitude with the dynamical measures (greater agreement cannot be established since

no galaxy has been measured both ways: but in general galaxies of the same type are found to have similar dark mass fractions and total masses, and the lensing results can therefore be taken to confirm the typical total masses for systems of those kinds). Even leaving the lensing check aside, we do have "quasi-independent" evidence in the fact that nested structures (solar systems to galaxies to clusters to superclusters) have dark matter fractions that steadily increase with the scale of the system considered. The dynamical measurements of mass at larger scales therefore provide independent evidence for dark matter at smaller scales (if there is dark matter on large scales, there must be at least as much as there is on smaller scales).

4.1.1 Dark Matter in the Milky Way

In order to properly address the question of whether there is any dark matter in our own galaxy (a typical spiral galaxy), we should distinguish between the local disk, the disk as a whole, and the galaxy as a whole (including a diffuse spherical halo extending to several times the radius of the optical disk). Present results suggest that there is little or no dark matter in the plane of the galaxy beyond what is contributed by a massive halo whose mass scales as $1/r$, and that this dark halo has a very high mass in total: the halo accounts for almost all of the mass of the galaxy as a whole and none (or nearly none) of the light, which is concentrated in the disk and bulge. Thus the available evidence indicates that dark matter is more or less homogeneously distributed within the Milky Way (indeed, its extent defines the dynamical boundaries of the galaxy). This distribution is much different than the distribution of ordinary matter as traced by visible radiation, which follows a thin circular disk centred on a spherical bulge. The fundamental difference in the amount and distribution of dark as opposed to visible matter is one of the main things to be accounted for by a theory of dark matter, in addition to accounting for the observed dynamics.

Chapter 3 discussed Jan Oort's studies of the motions of stars perpendicular to the plane of the Milky Way. This type of study is especially important because it is the only source of information about the detailed distribution of dark matter *within* galaxies (it is impossible to conduct similar studies on the disks of other spiral galaxies, because the motions of individual stars are impossible to discern, so the results are applied to other

spirals by analogical argument). Oort's later study indicated that the disk is about forty percent dark matter (Trimble 1990, 356). As it turns out, however, more recent studies tend to support the more conservative conclusion of Oort's original paper, to the effect that the local disk mass can be fully explained by dust, gas, and stars that are merely dim. The more recent studies repeat the velocity and height measurements—the longer time over which to compare astrometric photographs and the more precise measuring techniques now available mean that the errors in estimated velocities are lower (the displacements are a greater multiple of the minimum measuring error, so the ratio of measured length to probable error is greater)—and the present estimates of the distances to the observed stars are also somewhat better. Such studies conclude that the local disk mass is actually less than Oort thought. Furthermore, the maturation of non-optical (infrared, microwave, X-ray, radio, etc.) astronomy has enabled us to be sure that if any significant excess quantity of "dim" gas, dust or stars were present, we *would* be able to detect it. (This is, of course, based on the assumption of the completeness of our knowledge of the possible forms such matter may take, and of the corresponding spectra of emission.)[3] These results increase the expected mass-to-light ratio of our galactic disk somewhat. Overall, the discrepancy between the visible mass and dynamical mass of the local galactic disk is less than Oort thought, and is quite close to zero.

In fact, as Trimble reports, Bahcall (1984) finds a ratio of dark to luminous matter in the solar neighbourhood of 0.5 to 1.5: "The total disk M/L is then about 3 and the local density of dark matter 0.1 [solar masses per cubic parsec]" (Trimble 1987, 428). Trimble points out, however, that rather similar tracer populations have been used in all these

---

[3] This assumption may be problematic in the sense that it seems to involve an argument of the following form. We have never seen a form of baryonic matter whose spectra we could not detect, therefore because we see no emission there must be no baryonic matter. (The actual argument is of course more sophisticated, in that it relies on highly confirmed theories about blackbody radiation, and so on.) It is a fairly regular occurrence for some new form of interstellar matter to be proposed as galactic dark matter (the latest of which I am aware is cold molecular hydrogen, which apparently would have escaped the earlier detection schemes)—but note that, as I will discuss in more detail later, nucleosynthesis predictions restrict the total number of baryons available for such candidates, probably to a level insufficient or just barely sufficient for galactic dark matter, but certainly insufficient for dark matter at larger scales.

studies, and "stellar brightnesses may have been overestimated and distances underestimated, resulting in an *overestimate* of the local [dynamical] mass density" (Trimble 1987, 428; italics added). An even more recent study finds that "the local [total] mass density may be less than 0.10 solar masses per cubic parsec, and essentially all accounted for by stars and gas" (Trimble 1993, 151). Thus, "The existence of a separate dark component belonging specifically to the Milky Way disk should probably. . .not be accepted without reservation" (Trimble 1987, 428). "A moderate dark matter component [in the disk of the Milky Way] is not excluded, but it is likely to be explicable as simply part of the total galactic dark matter supply, gravitationally concentrated into the galactic plane as the disk formed" (Trimble 1993, 151-52). Note that the contribution of the dark halo itself can be present in the disk but not have a dynamical effect because the halo is close to spherically symmetric about our position, and so the gravitational pull from the halo matter on either side of the disk is nearly exactly balanced out. (This is similar to the notion that cosmological dark matter could be present in extremely vast quantities and yet be dynamically undetectable provided that it has a completely homogeneous and isotropic universal distribution.)

Peebles (1993, 432-3) describes the recent results through an idealised model of the Milky Way, where the disk is imagined to be a two-dimensional plane onto which all the mass and light is projected. From an Oort-style study of stellar velocity dispersions above and below this plane, Peebles gives an "order of magnitude" calculation of Σ, the mass per unit area, in the following way. The general equation for the gravitational potential as a function of distance from the plane z is : $d^2\Phi/dz^2 = 4\pi G\rho(r)$, where $\rho(r)$ is the density as a function of radius. From this we get the gravitational acceleration g, where $\Sigma(z)$ is the mass per unit area between sheets at $\pm z$ from the plane: $g = d\Phi/dz = 2\pi G\Sigma(z)$. The mean square velocity normal to the disk, $\langle v^2 \rangle$, is a measure of the gravitational potential. Thus from the measured value $\langle v^2 \rangle^{1/2} = 18.8 \pm 1.3 kms^{-1}$ we calculate $\Sigma = 55 \pm 10\ M_\odot\ pc^{-2}$. The observed contributions to the local mass density by hydrogen-burning stars, neutral and ionised gas, and white dwarf stars are respectively: $\Sigma_{stars} = 30 \pm 5\ M_\odot\ pc^{-2}$, $\Sigma_{gas} = 12 \pm 3\ M_\odot\ pc^{-2}$, *and* $\Sigma_{wd} = 4\ M_\odot\ pc^{-2}$. "The sum is $46 \pm 6\ M_\odot\ pc^{-2}$, less than one standard deviation from the dynamical mass. This is in line with much more careful analyses by Kuijken and Gilmore (1991) and

Bahcall et al. (1992) in indicating that if dark matter is present in the local disk of the Milky Way it is sub-dominant: the main ingredient of our neighborhood is baryons" (Peebles 1993, 433). (Obviously, I have not given all the details of the calculations: see Peebles 1993, 432-3, and references therein.)

If, as these results suggest, there is no local dark matter component, we need somehow to account for the exclusion of dark matter from the disk region—if almost all the visible matter condensed into the disk and bulge, why did the dark matter not do so as well? This can only be explained in terms of the nature of the dark matter itself and different processes of gravitational evolution in galaxy formation for dark and visible matter structures. For example, one popular possibility is that the dark matter is some kind of fundamental particle that interacts only by the weak nuclear force, and gravity. Weak interactions are rare as well as being of low strength, which means that dark matter of this sort would be effectively "dissipationless". This means that the dark matter would have no way to dissipate its initial kinetic energy, and therefore would be unable to drop down to lower orbits along with the visible matter. Visible matter is able to collect in the centres of gravitational wells because it can dissipate energy by releasing electromagnetic radiation. So on the hypothesis that dark matter is only weakly interacting, one can explain why the distributions of dark and visible matter should be so different. (Other explanations might be possible as well, of course.)

Thus one thing the observed lack of dark matter in the local disk suggests is that the thermodynamical properties of dark matter are likely to be quite different than those of ordinary matter. The point I want to make here is just that information about the non-luminous component of the disk mass can be used to acquire information about the nature of the dark matter.

Beyond Oort-style studies of the local Milky Way, one might hope to be able to determine a rotation curve for the disk as a whole, and from this calculate its dynamical mass. However, it is difficult because of our position inside the disk of the Milky Way to accurately measure the entire velocity curve of our own galaxy, which means that it is hard to obtain a dynamical mass for the Milky Way as a whole from its rotation. (It is not too hard to get data on the rotation curve interior to the orbit of the Sun around the galactic centre, but measurements of rotation outside the solar circle are difficult.) This is

partly due to the fact that we cannot view the galaxy from outside, and therefore have to rely on assumptions about the rotational motions we observe (in particular, about how to convert heliocentric motions to galactocentric motions); it is also partly due to the fact that our view to other parts of the galactic plane is obscured by bands of dust. The limits of the reliability of the rotation curve we are able to construct put limits on the kinds and reliability of inferences we can make about the quantity and nature of dark matter in our own galaxy. In part we are forced to rely on analogical reasoning: other galaxies similar to ours in relevant features are shown to have large dark matter components, so ours must too. It is now possible, however, to make more or less direct measurements of the velocity of our own solar system around the centre of the galaxy. According to Trimble (1987, 428-9) the main difficulties here are determining our radial distance from the galactic centre and our rotational speed. But within the uncertainties for these values, we can calculate the dynamical mass interior to our orbit.

The most interesting and reliable results of this type were announced in June 1999. A group using part of the Very Long Baseline Array (**VLBA**: a system of ten 25-metre radio telescopes stretching from Hawaii to the Virgin Islands linked together as a single interferometer) has made a very accurate measurement of the orbital velocity of our solar system around the galactic centre. The huge effective dish diameter of the VLBA means that its angular resolution is extremely good (the group reports being able to resolve objects with up to 0.1 *milliarcsecond* accuracy!). This extremely high angular resolution allowed the group to look for the very small secular motion of a powerful radio source in the core of the Milky Way known as Sagittarius A* ("A-star"), relative to distant extragalactic radio sources, due to the orbit of the solar system around the galactic centre.[4] The initial results, although model dependent, fairly reliably fix the solar system's orbital velocity at 219 ± 20 kms^{-1}, give its radial distance as about 26000 light-

---

[4] Even though it takes about a 100 000 years for the solar system to complete a single orbit, the VLBA was able to detect this parallactic motion in observations taken only about a month apart (Reid, *et al.*, 1999). Note that the dust that obscures visible and infrared light from the core is transparent to radio waves. The Sgr A* radio source has long been thought to be a supermassive black hole (see below) at or very near the dynamical centre of the galaxy.

years, and fix the minimum mass of the radio source Sagittarius A* as at least 1000 times the mass of the sun, confined to a region smaller than the solar system.[5] (Reid, *et al.*, 1999; see also space.sci.news (01 June 1999) and <http://www.nrao.edu>.)

This result is important in that it fixes accurately the galactic orbital speed of bodies at our radial distance: this tells us how much total mass there is interior to our galactic orbit. It does not, however, provide the same amount of information as a rotation curve would, since it does not tell us the rotational velocity at *other* radii, and so we cannot (on this basis at least) tell whether or not the rotation curve of our own galaxy is Keplerian, and in turn this means we have less information about the overall distribution of matter in our own galaxy than we do for similar spirals whose rotation curves we know in detail.

Other methods have to be tried in order to determine the dynamical mass inside orbits exterior to our own: this is something that is required in order to be able to determine just how far the dynamical structures related to our galaxy extend. Estimates of the distances and orbital velocities of stars exterior to our orbit increase in uncertainty as distance increases. Looking at the radio signature of hot gas in the disk allows us to extend our knowledge of the gravitational potential of the galaxy even farther than the limit of visible stars, out to about twice our distance from the galactic centre.[6] Looking at stars and gas in the halo gives similar results, although in this case an additional uncertainty comes in since there is no strong reason to think that the assumption of nearly circular orbits will hold for bodies far out in the halo. Nevertheless, these studies give mass estimates out to 5-10 times our distance from the galactic centre. And in general these studies confirm the results found for other spiral galaxies, to the effect that

---

[5] The mass result for Sgr A* is important because it proves that the object is a black hole and not some other unusual radio phenomenon. The result discussed here is consistent with the idea that Sgr A* is a supermassive black hole containing all of the dark mass deduced from stellar motions in the core (~2.6 million times the mass of the sun): "it is likely, but unproven, that most of this mass is contained in. . .Sgr A*" (Reid, et al. 1999, 7).

[6] Hot gas can only stay hot, and continue to emit energetic X-rays or radio waves, if it is continually heated. The energy source for this heating is taken to be the gravitational potential of the galaxy (or cluster) in which the gas finds itself. The temperature of a gas cloud is thus a measure of the strength of the gravitational field acting on the cloud.

rotational velocity is constant out to high radius, and that the dynamical mass of the galaxy as a whole is (one to two orders of magnitude) higher than the visible mass. (Trimble 1987, 428-30) Finally, the velocity dispersions of globular clusters and companion galaxies, and observed upper limits on the diameters of globular clusters (interpreted as due to tidal forces imposed by the parent galaxy), all probe the total mass of our galaxy (interior to the greatest distances for which it is possible to obtain measurements: the dark matter in the galaxy probably extends much further than the visible objects do). Results from these kinds of studies, as summarised in Trimble (1987, 430-31), suggest that the mass of the our galaxy interior to 100 kpc is close to 10^{12} solar masses, which corresponds to a mass-to-light ratio greater than or about equal to 30 (or about ten times that for the disk).

> Despite the uncertainties and reservations. . .it seems safe to conclude that (a) within R_0 [our galactic radius], there is about as much mass in a spheroidal dark halo as within the luminous disk; (b) outside R_0, there is at least 2 and probably 3-10 times as much matter as inside. (Trimble 1987, 431-32)

4.1.2 Dark Matter in Other Galaxies: Spirals

It is possible to construct arguments from the fact that only some distributions of matter are stable over sufficiently long periods, to the conclusion that the amount of mass and its overall distribution in galaxies must be very different than that of the visible matter. The visible spiral structure found in many galaxies, for example, cannot by itself be stable—that is, the spiral shape cannot persist if gravity is the only force acting and the matter is distributed as the light. Ostriker and Peebles (1973) were the first to discover the "morphological instability" of spiral mass distributions, in a computer simulation of a galactic disk. Given this, if we grant the assumption that spirals are indeed long-lived structures held together by self-gravitation,[7] we are forced to conclude that the mass

---

[7] Since the look-back times for some spirals are a significant fraction of the total age of the universe, and since spirals are found in every direction and at every redshift (and very few observed in the process of *acquiring or losing* their spiral structure), there is reason to think that spirals are indeed long-lived structures: galaxies having a range of ages, sizes, luminosities and (dynamically determined) masses exhibit this morphological characteristic. But galaxy formation and evolution are by no means settled topics in

distribution in these galaxies is very different from the distribution of the light. (Astronomers sometimes express this by saying that "mass does not follow light" or "light does not trace mass".) The only plausible way to make a matter distribution in which a spiral structure is stable is to embed the visible spiral in a (nearly-) spherical halo of (unseen) matter whose total mass is several times that of the spiral and whose extent is several times that of the visible disk.[8] (Bartusiak 1993, 213; Ostriker and Peebles 1973.)

More definitive than the argument from morphological stability is the argument from observed rotation curves. (It is also more general since a version of this argument also applies to elliptical galaxies and even to clusters of galaxies, in the form of velocity dispersions plotted against radius.) In the case of spiral galaxies, the velocities of stars (or more commonly clouds of hot interstellar gas) are plotted against their respective radii from their galactic centre. In order to use this technique, we need to make an assumption of gravitational stability. It is, however, a somewhat weaker assumption than that of *morphological* stability, in that we do not require that the exact *configuration* of stars be preserved over the long term, but only that the system as a whole neither entirely collapses nor evaporates. The assumption actually need not require permanent stability, but only an approximation to stability (no considerable collapse or evaporation) over some significant fraction of the typical lifetime of a galaxy. Galactic mass estimates that depend on this assumption can be given a margin of error whose range takes in a range of approximations to gravitational stability that meet this weaker criterion. Clear evidence

---

astrophysics, so it is still just a hypothesis that these systems are in gravitational equilibrium, albeit a plausible one with some empirical arguments to support it.

[8] Note that one can calculate a galaxy's gravitational force per unit mass at any radius, on the assumption that the mass has a spherical distribution and its density is a function of radius, by the equation $g_r = -GM(r)\, r^{-2}$. Since we do not *see* mass distributions in spirals as spherical, this equation may be in error; however, even when "a sphere of uniform density is replaced with a very flat spheroid, also of uniform density and of the same total mass, the gravitational field inside the body in its midplane is only increased by a factor of $3\pi/4$" (Tayler 1991, 59-60). So we may use the equation for g_r even though we do not know that galactic mass distributions are really spherical, without fear of being too far wrong. Note that the stability arguments in Ostriker and Peebles (1973) give good reason to think that spirals are in fact embedded in roughly spherical invisible halos.

of the fact that galaxies are long-lived structures is that despite their higher-than-expected rotational velocities, galaxies do not shed stars in large numbers: the inter-galactic spaces are nearly empty of free stars or other significant mass (Knapp 1995). Gravity is the only known force that can act powerfully enough at the distances involved to keep the fast-moving gas and stars within the galaxy.

On the assumption that galaxies are approximately gravitationally stable, the velocity of rotation at each radius will balance the attraction of gravity from the mass interior to that orbit. A standard theorem of mechanics says that if all the mass interior to a given sphere is distributed spherically symmetrically, we can treat the mass interior to the sphere as being concentrated at the centre point for the purposes of calculation. If it is fair to assume that the matter in the disk of a spiral galaxy is roughly symmetrically distributed (with mass decreasing in density with distance from the centre, in step with the light)[9], it follows from the inverse square action of gravity that the rotation curve for bodies orbiting in a spiral galaxy ought to have its highest value near the centre and decrease as the radius increases. In general, since the light in galaxies is observed to fall off roughly as the inverse of the radius, on the assumption that mass traces light (the most reasonable initial assumption) we should expect the speed of rotation to be lower near the outskirts of the galaxy (that is, beyond the limit of the visible light). However, this is not at all what is observed: in fact, galactic rotation curves tend to remain flat or even to *rise* as radius increases beyond the limit of the visible light (Rubin 1989). The observations, then, yield two surprises: the velocity of rotation is at all points higher than would be expected from the visible mass alone, and the *pattern* of rotation revealed in the rotation curve does not at all fit what would be expected from the distribution of the visible mass.

It is impossible to account for this observational result on the assumption that mass traces light in spirals. This has two consequences. It forces us to conclude, first, that there is much more mass present in spirals than can be accounted for by the visible matter

---

[9] This is obviously only an approximation in the case of spirals, but it is a workable one since their centres of mass are very close to their geometric centres. Since galaxies do have significant internal (gravitational) structure, we can expect local velocity inhomogeneities due to interactions among neighbouring galactic components, but these can be safely ignored for the purposes of the kind of analysis of interest here.

alone (because the rotation is faster than it would be on that assumption)[10], and second, that the majority of the mass must be distributed in a way quite unlike the way in which the visible matter is distributed (because the shape of the rotation curve is inconsistent with that distribution and an assumption of gravitational quasi-stability). There are only two possible mass distributions that can produce the observed rotation curves. A matter distribution in which mass density is proportional to $1/r$ where mass traces the light, with a significant dark matter component, could produce the observed rotation curve (see Tayler 1991). The trouble with this solution is that such a system is unstable to bar-like collapse on a time scale much shorter than the lifetime of a typical galaxy (see Ostriker and Peebles 1973). The only other option is that the visible part of the galaxy is embedded in a spherical halo of dark matter. This configuration is stable and can fully account for the observed motions of the visible part of the galaxy. Thus arguments from morphological stability and gravitational stability both lead to the same conclusion: spirals are surrounded by dark matter haloes.

---

[10] Using $GM/R \sim V^2$ it is possible to calculate the mass interior to a given radius of the galaxy from observations of the redshifting of the light on either side of the axis of rotation. This gives the component of rotation along the line of sight. Estimating the angle of the plane of rotation to our line of sight allows one to calculate the true velocity of rotation, and thus the true dynamical mass. "A [spiral] galaxy observed face on would look circular. If it is observed at some angle to the face-on position, it appears elliptical and the ratio of the major to the minor axes of the ellipse is directly related to the angle. Once the angle has been determined we know what fraction of the rotation velocity we are observing and we can deduce [the true circular velocity] from the observations" (Tayler 1991, 57). Furthermore, estimates can be made of the range of possible values for such things as the oblateness of the halo. Taking all these factors into account, it is possible to calculate the total possible error in the mass calculation from the velocity observation. It turns out that the difference between the visible mass and the dynamical or total mass is so great that it swamps all the possible errors, even taken together. Thus we know with certainty that a mass discrepancy exists for spiral galaxies. This result is quite independent of the morphological stability arguments mentioned earlier, although the calculation does rely on the weaker assumption of near-equilibrium. And we also know that the discrepancy between the visible mass and the dynamical mass of these structures is very large. In other words, there is no way that the discrepancy can reasonably be attributed to errors in the dynamical mass calculation, at least not if the assumptions regarding gravitational equilibrium and the form of the law of gravity are close to correct.

As discussed in Chapter 3, Babcock (1939) was the first to study the rotation curve of a spiral galaxy. Freeman (1970) was the first to notice that flat rotation curves were common in spirals. But few rotation curves were studied until the 1970s, when Vera Rubin and her colleagues began systematically obtaining rotation curves as the first step in understanding the structure and evolution of galaxies. By 1989 rotation curves for over 300 spirals had been obtained (Rubin 1989, 90), and all showed the same general, non-Keplerian form: the velocity of gas at high radius does not decrease, and in many cases *rises* well beyond the visible-light limit of the galaxy. "The conclusion is inescapable: matter, unlike luminosity, is not concentrated near the centre of spiral galaxies. In short, the distribution of light in a galaxy is not at all a guide to the distribution of matter" (Rubin 1989, 90).

Many of the general arguments (as opposed to specific observations) that apply to spiral galaxies can be extended by analogy to the Milky Way. For example, since other spirals are shown (by rotation curves, etc.) to contain significant dark matter, and because we have no reason to think that the Milky Way is an atypical spiral, we can infer that the Milky Way must also have significant dark matter. Such arguments provide some independent confirmation for the direct measurements of the mass of the Milky Way mentioned above. Similarly, the results of measurements of the density and distribution of dark matter that are possible within the Milky Way but not for other galaxies can be extended by analogy to other spirals.

There are several other "ancillary" lines of evidence that provide corroboration of the existence of dark matter haloes around spiral galaxies. These include arguments based on observations of *flaring* at the edges of some visible disks and *warps* in others: the best way to account for such observations is to hypothesize the existence of a massive halo exerting a gravitational influence on the visible disks of these spirals, producing the deviations from a flat disk. Massive haloes also seem to be required in order to explain the existence of *polar rings*, thin bands of stars and gas that encircle a few known spiral galaxies at the outer radius of and perpendicular to their disks.

Additional support for a dark matter halo comes from micro-lensing observations in the halo of the Milky Way (these observations are discussed in some detail below). Although further observational work is required before the argument can be firmed up,

the available data suggests that there is a population of massive, low-luminosity objects surrounding the Milky Way (and by extension, other spirals). (On present evidence it seems very unlikely that this population is sufficient to account for all of the missing mass, but see below for more detail.) We also know that there is a visible spheroid, a low-density cloud of ordinary stars (and gas and dust) surrounding the Milky Way (cf. Trimble 1993, 152).

None of these ancillary arguments is, on its own, very strong evidence for dark matter haloes—the evidence is sparse, the available theoretical explanations are not detailed enough to be robust, and alternative explanations for each of these phenomena are certainly possible. But taken together, and especially in combination with morphological stability and dynamical arguments, these considerations are strongly suggestive. A case could be made that the existence of a dark matter halo (of the right sort) would give a unified explanation for all these diverse facts, and would thereby acquire a higher degree of confirmation.

4.1.3 Dark Matter in Other Galaxies: Ellipticals, Irregulars and Dwarf-Spheroidals

Galaxies of other types are also shown to have large dark mass fractions, although in some cases the evidence is less certain than it is for spirals. Stars in elliptical galaxies, unlike those in spirals, do not orbit in a coherent pattern, so rotation curves are of no use. Instead, studies of velocity dispersions (employing the Virial Theorem to calculate the mass of these galaxies) are used to show that these galaxies, too, contain a large fraction of dark matter, similar in proportion to the dark matter in spirals. The motions of globular clusters (tight spherical clusters of stars found within galaxies) and companion galaxies, where these objects exist and their motions are observable in detail, provide another way to estimate the mass of elliptical and irregular galaxies. Finally, and perhaps most importantly, the X-ray emission of high temperature gas can be used to infer that a greater mass than is visible must be present in these galaxies in order to keep this gas from evaporating out of the system. While good data are hard to come by, the measures of the ratios of dark to total mass for these galaxies essentially confirm what we learn from spirals. Most significantly, the different measures for different classes of systems agree among themselves. This tells us that the dynamical discrepancy is not a quirk of

spirals and their evolution. The agreement of measures here is similar to the consilience of inductions we find in Newton's argument to Universal Gravity. This kind of agreement of diverse measures is an especially strong form of unification, which in turn provides a significant degree of confirmation to the (nevertheless fallible) hypothesis of the existence of dark matter. This conclusion is fallible because, given different background assumptions than Newton's, these same phenomena can (or could) be turned into agreeing measurements of the parameters of a different theory, for example an alternative theory of gravity which explains astrophysical dynamics without the need for dark matter. The epistemic difference between these two competing measurements is in the warrant for their respective background information: in Chapter 6 I show that the present evidence is too weak to decide with much weight one way or the other, but that methodological considerations seem to indicate that our best hope for future epistemic progress on this issue is to provisionally accept and use General Relativity at galactic and greater scales—which is to say, to pursue matter solutions—in which case the phenomena mentioned above provide agreeing RfP measures of the parameters of the dark matter distribution.

The smallest of the galaxies mentioned here are useful in another sense because "they have the potential for telling us the smallest configuration that can have a dark halo and thereby [have the potential for] constraining the minimum particle mass possible in the halos" (Trimble 1987, 439). That is, observations of the minimum halo size, and of the relation between the density of dark matter and the radius considered, provide constraints on the nature of the dark matter. Dark matter begins to dominate somewhere between the scales of a solar system and a dwarf spheroidal galaxy: if we can determine just where this happens, the result will help to constrain dark matter hypotheses.

The dwarf galaxies in orbit around the Milky Way were shown in early studies by Mark Aaronson (1983) to have internal motions that imply the existence of much more mass than is visible in them. An additional reason to suspect that the dwarf companions of large galaxies ought to have plenty of dark matter is that without it globular clusters within such companion galaxies ought to have been torn apart by tidal forces induced by the large galaxy. However, data is sparse, and these results are somewhat controversial. (Bartusiak 1993, 214; Trimble 1987, 440-41; Trimble 1993, 153)

X-ray emissions of high-temperature gas in elliptical galaxies are useful because,

> The X-ray spectrum reveals gas temperature as a function of position and thus [yields] velocities for use in the equation $M = V^2 R\, G$, because a gas in [thermodynamic] equilibrium has $kT = m_e V^2$, where m_e is the rest mass of the electron and k is the Stephan-Boltzmann constant. With the exception of galaxies whose rotation speeds drop with a radius outside 10 kpc or so (apparently because interactions with other galaxies have stripped them of their halos), M/L rises with radius, and dark matter must be less centrally condensed that the luminous stuff. (Trimble 1993, 152)

The two important points here are the explanation of why temperature indicates velocity of rotation, and the fact that almost all galaxies display rotation curves that indicate an extended halo in which the dark matter is distributed differently than the light.

Finally, we come to gravitational lensing as a method of determining the mass of the lensing body. There are some known cases where background objects (quasars and galaxies) are gravitationally lensed by foreground galaxies and clusters. In principle, in such situations it is possible to calculate the mass of the lensing body from the disposition of the images. So far, the uncertainties are so great that this method only tells us that the galactic lenses are of about the same mass (within an order of magnitude) of other typical galaxies of their type. That is, when a spiral galaxy is a gravitational lens, its mass is found to be of the same order of magnitude as spirals are generally found to have on dynamical measures. On its own, this piece of information does not provide much evidence about dark matter, but as will be discussed in Chapter 6, these results, even as imprecise as they are, may play a crucial role in debates over which law of gravity has the best empirical support on the evidence of the dynamics of astrophysical systems.

4.1.4 Dark Matter in Clusters of Galaxies

Studies of the motions of binary galaxies and of the motions of the Local Group (a small cluster with a diameter of about six million lightyears, of which our galaxy, M31 and about two dozen other galaxies are members) indicate the existence of large dark matter fractions in these systems. One difficulty is that it is impossible to be sure that the systems in question are gravitationally bound as a unit, but the observed motions in the Local Group are "very difficult to understand" unless the mass-to-light ratio is 20-60

(Trimble 1990, 152). That is to say, explanation of the motions of the Local Group, *whether gravitationally bound or not*, seems to require vast amounts of dark matter.

At the next level of structure, the study of the dynamics of clusters of galaxies reveals once again that there is a large discrepancy between the visible mass and the mass required in order account for the observations on the assumption that these systems are gravitationally bound. In fact, the observations show that there must be as much as an order of magnitude more dark matter in clusters than there is in individual galaxies: the dark mass of these systems is 10-100 times the visible mass. Roughly speaking, the mass discrepancy for clusters is ten times worse than it is for individual galaxies. This suggests that there is "cluster-specific" dark matter, which is a possible restriction on the nature of the dark matter: either the dark matter has some characteristic that makes it clump less strongly at galactic scales than at cluster scales (small particles moving at relativistic speeds are one possibility here), or there is more than one kind of dark matter, one kind (at least) for each of the two levels of structure.[11]

There are essentially three methods for assessing the masses of clusters dynamically, all originating with Fritz Zwicky in the 1930s, as described in Chapter 3. The first two are based on the method of examining velocity dispersions in clusters and applying the Virial Theorem, which Zwicky used to investigate the mass of the Coma cluster. They are, (1) using the Doppler shift of visible light and or of the 21 cm emission of hydrogen gas to measure the velocities of member galaxies along the line of sight and applying the Virial Theorem to calculate the cluster mass; and (2) using the spectrum of X-ray emissions of hot intra-cluster gas to probe the dynamical mass of the cluster, as was described for elliptical galaxies above. This second method is a useful check because gas at the extremely high temperatures observed would surely have evaporated away into inter-cluster space unless the cluster mass were very high; furthermore, this gas is generally not associated with any particular member galaxy, but rather is distributed throughout and extends beyond the visible light boundaries of the

---

[11] The cosmological dark matter problem requires that there be yet another order of magnitude more mass than even dynamical studies of clusters reveal: this cosmic dark matter, if it exists at all, might have to be yet some other kind of matter. (See Chapter 1 and the Appendix.)

cluster. The gas itself, given its luminosity, can only make up a small fraction of the total mass: in early studies the X-ray satellite ROSAT revealed a cluster that had to have 30 times more dar' mass than visible mass (Bartusiak 1993, 279). Both these methods find that the velocities of cluster components are constant out to high radius: in this respect, the velocity dispersions for clusters look like the rotation curves for galaxies, and this fact has similar implications for the shape of the dark matter distribution in clusters. The modern results are more robust than Zwicky's in several senses: the velocities are known more accurately, more members of each cluster are included in the calculation (Zwicky's original paper relied on just seven members of Coma!), and a greater number of clusters have been studied, with similar results for all of them.

(3) The third method of assessing cluster mass is to use gravitational lensing observations. This is possible when a cluster lenses the light of a background quasar into multiple images, or when a cluster lenses background galaxies into arcs. The details of gravitational lensing measures are discussed in the following paragraphs. Importantly, all three kinds of measures find roughly the same proportion of nonluminous mass in clusters (at least within an order of magnitude).

Multiple images of quasars are produced when a background quasar is lensed by a foreground cluster, and arcs are produced when background galaxies are lensed by foreground clusters or galaxies (this latter phenomenon was first observed in 1987; the former were first identified in 1978).

> Gravitational lensing was predicted by Eddington, Lodge, Zwicky, Einstein and others, long before the first convincing example of this phenomenon... was discovered [the report of this discovery by Walsh, Carswell and Weymann appeared in *Nature*, 1979: Vol. 279, 381]. Since this date, the tally of lenses has increased to (in my [Blanford's] subjective opinion) nine secure plus six probable instances of multiple quasar imaging, five secure and two possible cases of radio rings and 25 secure plus at least 10 probable cases of rich clusters exhibiting arcs and arclets. (Blanford 1997, 94)

From the number and relative positions of the quasar images, or the distortion of the light of a background galaxy into an arc, it is possible to calculate the mass of the lensing body. To do the calculation accurately requires knowledge of the distance from here to the lensing body, and from the lensing body to the lensed body; in the absence of other information and for ease of calculation, the mass of the lensing body is assumed to

be distributed spherically (correction factors can be calculated to take account of the possibility of a non-spherical distribution).[12] The distances are only known to within twenty percent. Even rough calculations, where the error in the relative distances is high, give masses for the lensing bodies that are consistent with other mass estimates for typical systems of those types. As far as I know, no system whose mass has been estimated in virtue of its being a gravitational lens has also had its mass measured dynamically—getting accurate spectra for rotation curve work requires long exposures for even the brightest galaxies, and it simply is not possible to obtain them for most galaxies.[13] But in principle gravitational lensing provides a consistency check on the dynamical mass estimates; furthermore, the gravitational lensing results show that the dark matter in clusters is not just the dark matter in its component galaxies (more mass is required to produce the lensed images than is available in the masses of the component galaxies). (cf. Trimble 1993, 152) Typically, gravitational lensing calculations for clusters yield an M/L of up to 100 (Trimble 1993, 152). The gravitational lensing results, because they are derived from a method that is independent of the Newtonian power law which is assumed in the other dynamical mass measures, play a crucial role in Chapter 6, since theories of gravity offered in order to save the rotation curves without invoking unseen matter must also be able to account for the gravitational lensing observations.

Further potential evidence about dark matter from gravitational lensing involves the microlensing of the images of quasars already lensed by clusters. When a mass element of the cluster halo dark matter passes in front of an image of a quasar, that image (and not the others) brightens in a way that depends on the position and mass of the interposing

---

[12] J. Anthony Tyson and colleagues (1990) were the first to do a computer simulation that mapped the dark matter in a cluster on the basis of observations of gravitational lens arcs; they found that the dark matter, which is most of the mass and therefore mostly responsible for the lensing of a cluster, is spherically distributed and concentrated in the centre of the cluster. (Bartusiak 1993, 215-16.) A more recent study (Meillier, *et al.*, 2000) has mapped the large-scale distribution of dark matter from a study of the gravitational lensing of 200 000 distant galaxies over a two-square-degree area of the sky.

[13] All-sky redshift surveys currently underway, such as the Sloan Digital Sky Survey (see www.sdss.org), will catalogue the redshifts of millions of galaxies in an effort to map large scale structure, but these spectra are for each galaxy as a whole, and do not give the variation of the redshift with radius within each galaxy.

body. Thus, the microlensing of individual quasar images may tell us the unit size of the dark matter (Trimble 1990, 152); the calculation is the same as described below for the microlensing of MACHO candidates (although because of the greater distances to the lensing body the even may occur over months instead of days (Bartusiak 1993, 237)).

The study of the X-ray temperature of intra-cluster gas in rich clusters provides another reason to think that there is cluster-specific dark matter. "As with single galaxies, the results are consistent, imply large total masses with $M/L = 100h$, and are not explained by the mass of the X-ray gas itself, which is, at most, comparable with that of the luminous parts of the galaxy" (h is related to the uncertainty in the Hubble constant and takes values between 0.5 and 1; Trimble 1990, 152). (According to more recent results, the Hubble constant is 76 $kms^{-1}Mpc^{-1}$ to within 10%, instead of the factor of two here.) But the important thing is that because the galaxies in a rich cluster are closer together than the radius of the dark matter halo of a typical spiral, and because otherwise dynamical evolution would have led to a hierarchical distribution by mass (with higher mass galaxies *always* taking more central positions) which is not observed in these clusters, it cannot be that the dark matter is primarily bound to individual galaxies: the cluster itself has its own dark matter. (Trimble 1990, 152) The observations also indicate that the dark matter does not bind with ordinary matter, or else it would have pooled in the centres of galaxies and clusters in the way that ordinary matter does: in fact, dark matter haloes extend to several diameters of the visible matter in most dynamical systems, as the arguments from rotation curves and stability show. (Bartusiak 1993, 277) This leads to the idea mentioned above that dark matter is "dissipationless"—if dark matter could efficiently dissipate its energy by radiation or collision, it would drop to a centre-weighted distribution like that of the visible matter in less time than the age of a galaxy. This fact puts constraints on both the emission properties of dark matter and on the strength of its non-gravitational interactions with ordinary matter. These facts are crucial explananda to be accounted for in the evaluation of dark matter candidates.

4.1.5 Dark Matter in Superclusters and Arguments from Large-Scale Structure

At the level of superclusters, which are conglomerations of clusters that stretch huge distances (50-100 Mpc or more) across the universe, we have little dynamical

evidence that is of any use. The most important factor is that present surveys only cover a very small volume of the whole observable universe, so we cannot be sure that we see the largest structures there are, and we cannot be sure that the sample we have is typical of the rest of the universe. Observations do indicate that there are "large scale streaming motions" in the universe, but it is not possible to turn these motions into dynamical measures of mass. The most secure such result is the so-called Virgo-centric in-fall of the Local Group, which is taken to imply the existence of an extremely massive "Great Attractor"—but since we do not observe it but rather infer its existence from the observed streaming motions of the Local Group on the assumption that those motions have a gravitational origin, we do not know what portion of its mass is dark.[14]

The only other thing we can say about dark matter at large scales is that given the minimal fluctuations in the cosmic background radiation as observed by COBE[15], there has not been enough time for gravity to create clusters unless there is a very large proportion of dark matter in them (either that, or some other—completely unknown—force or process was responsible for structure formation). Thus considerations about the dynamical evolution to the observed level of structure and mass clumping from the nearly

---

[14] In one sense it is *all* dark, of course, since we cannot observe it, but since our view in the direction of the Great Attractor is blocked by nearer objects, we do not know how far away it is; nor do we know whether our motion relative to it is a gravitationally stable orbit or a non-equilibrium collapse toward it, so we cannot apply the Virial Theorem. Lynden-Bell, *et al* (1988) estimated the mass of the Great Attractor at $5 \cdot 4 \times 10^{16} M_{\odot}$, on the basis of observed large-scale streaming motions of galaxies in the Local Group. Matthewson, *et al.* (1992) have argued, however, that since there is no observed "back-side in-fall" (that is, no group of galaxies streaming toward us, on the other side of the supposed Great Attractor) it may be that the observed $600\ kms^{-1}$ streaming motion is just a bulk flow on scales $> 60\ h^{-1} Mpc$. The absence of back-side in-fall suggests that there might not be a Great Attractor after all, that the observed motion of the Local Group is not caused by a common acceleration toward a very large gravitational source.

[15] The Cosmic Background Explorer satellite was launched in 1989, and its results announced in 1992. The cosmic background radiation (**CBR**) is thought to have be released at the moment of matter-radiation decoupling, and its wavelength has been redshifted to an effective temperature of about 2.7 degrees Kelvin. The inhomogeneities in the matter distribution at that time produced the fluctuations in the CBR by differentially absorbing the radiation in different locations. But the initial inhomogeneities must have been very small, because the CBR is found to be smooth now to 1 part in ~100000.

smooth CBR presents us with the need for dark matter: without it, there is not enough time for mass fluctuations to grow into galaxies, let alone superclusters, by the gravitation of the visible matter alone. "[T]he very existence of galaxies and clusters is a strong indicator of the presence of dark matter. Galaxy formation is not at all well understood, and although it is difficult to model the process even with copious amounts of dark material, it is nearly impossible without" (Trimble 1990, 153).

The study of the evolution of large scale structure is problematic in part because it involves doing computer simulations in which the parameters of the models used are unavoidably much different than the expected parameters of the early universe. Besides the fact that we have to guess or infer significant features of the initial conditions and the physics of the early universe, the main problem is that even the most powerful computers available are limited in the number of mass elements they can deal with in a reasonable amount of time, and therefore each mass element in the simulations is very massive if any reasonable portion of the universe is to be modelled: in many cases, the mass elements are more massive than individual galaxies, and at best they are many times the mass of a typical star. Thus the simulations start with a smooth distribution of roughly galaxy-sized clumps, which is not really how the universe is thought to have started. (The COBE measurement of the fluctuations of the CBR, mentioned in note 14, constrains the mass density fluctuations at the decoupling epoch to be very much smaller than this.)

Models of large scale structure formation have several constraints:

(1) The power spectrum of the initial mass density fluctuations (inferred from the very small fluctuations from smoothness in the COBE observations of the 3° K cosmic microwave background radiation).

(2) The power spectrum of the observed structure in the universe (known only very partially from several surveys, including the Huchra-Geller survey (Geller and Huchra, 1989) that first demonstrated the existence of voids bounded by interconnected strings of clusters: these surveys cover only a very small fraction of the total universe, so extrapolating general large scale structure from them is somewhat risky, but it is the best information we have).

(3) The total amount of matter in the universe: recent simulations include both the observed quantity of visible matter and lots of dark matter (more than the

measured amount of the dynamical dark matter). (3') The assumption is generally made that gravity is the main force that sculpts large scale structure, so only the total mass and its initial distribution matters. An increase in computational power in recent years means that gas dynamics (hydrodynamics) is now also included in the simulations.

(4) The age of the universe (inferred from measured values of $1/H_0$ and q_0, the deceleration parameter: there is still some uncertainty about the correct values for these parameters, but a range of ages (between 10 and 15 Gyr) can be accepted with confidence; see Appendix A.1 for a discussion of some recent results constraining the age of the universe).

One way to proceed with modelling large scale structure is to take the measured values for (1) (which are highly certain), (2) (which is not well known), and (4) (which until recently was not known to within better than a factor of two), and adjust (3) until the end state of the simulation "looks like" the present universe. The characteristics of the assumed dark matter in these simulations is an important factor in how they turn out. "Cold dark matter" and "hot dark matter" models, on which the dark matter moves at much less than the speed of light or some significant fraction of the speed of light, respectively, yield very different results, holding all other factors constant. The best fit to the observed structure is achieved with "mixed" hot and cold models. Most simulations of large scale structure either assume a mass density close to the critical value, or find the best fits when the mass density is higher than the observed mass density of the universe. (See Appendix A.2 for discussion of recent measurements of the overall mass density of the universe, and the contribution of matter to it). This suggests that either the laws governing the simulations are not those governing the universe, or that the total *effective* mass density of the universe is only partly determined by the mass contribution. This second option is one additional reason to consider the possibility that there exists cosmological dark matter, but it is not a very strong reason because uncertainties in the simulations are large. The fact that we need "mixed" models in order to be able to mimic the observed large scale structure is also a possible, though weak, reason to think there might be dark matter in addition to the amount measured to exist dynamically.

As mentioned, a problem with trying to make conclusions about the evolution of structure from computer simulations done this way is that mass elements in the models are too few and individually much too large (advances in computing will go some way toward alleviating this problem, but it is impossible to completely overcome it). It is also hard to come up with an objective standard for judging whether the end state of the model "looks like" the present universe (a problem compounded by the fact that our data on large scale structure is limited). Further, the uncertainty of our knowledge of the initial conditions means a "philosophical" objection can be posed to the conclusions derived from these simulations. A successful run of the simulation is supposed to be taken as evidence for the amount (and kind) of dark matter used in the simulation. But we should be wary of the inference from the "success" of the simulation to the correctness of its input conditions for several reasons, not least of which is the fact that there will be *many* possible ways to produce roughly the same end state. This is especially true when physics (perhaps unknown physics) in addition to the what is allowed for in the simulations is possibly important to the actual evolution of structure. So the simulations provide little more than checks on the consistency of their assumptions with the state of the present universe. (For a recent commentary on how numerical simulations work and what they (can) show, see Ostriker 1997.)

Despite the problems with numerical simulations, this much seems clear: the evolution of large scale structure cannot be accounted for without hypothesising the existence of lots of nonluminous matter if we allow that gravitation as described by General Relativity is the main force responsible for producing structure out of the initial smooth state. Thus at every scale there is evidence (some weaker, some stronger, but together highly convincing) that there is much more mass in the universe than is indicated by the visible matter, and that this dark matter is not distributed in the way that the visible matter is. Just exactly how much dark matter there is, and how exactly it is distributed, is still a matter for further research, but we already have some fairly good information.

4.2 SUMMARY OF THE EVIDENCE FOR DARK MATTER

The discussion above sketches the general problem of determining the dark mass fraction at various astronomical scales. What the evidence shows is that the larger the

system, the greater the proportion of dark matter it contains. The evidence also provides constraints on the distribution and nature of the dark matter. For example, dark matter cannot be distributed as the light in galaxies and clusters: it most likely takes the form of a spherical halo extending several times farther than the radius of the visible matter in galaxies and clusters. This difference in the distributions of the dark and visible matter is an important explanandum for any theory of what dark matter is. In the next section I discuss the challenges to the existence of the dark matter discrepancy, and in Chapter 5 I assess the main candidate solutions. Let me conclude this section by summarising the dynamical measures M/L for systems of various scales. Note first that,

> The Sun, a very average star, [by definition] has a mass-to-light ratio of 1.0.... Because massive stars are not only much brighter (roughly $L \propto M^3$) than little ones but also much rarer (N proportional to about M^{-2} over most of the range 0.3-60 [solar masses]), a typical stellar population will also have M/L near 1. Values of 0.5-3.0 are, in fact, observed for star clusters of varying ages. The mass in gas is less than or, at most, equal to the mass in stars in all common varieties of galaxies and clusters of galaxies. *An object with M/L much greater than 3 must, then, be regarded as containing significant dark matter.* (Trimble 1993, 151; italics added)

The following table, with a few small changes, is taken from (Trimble 1993, 150). The "Ω" column lists the fraction of the total mass density of the universe contributed by each class of objects; each Ω value includes the contributions above it (the total dynamical mass contribution is the value on the final line, not the sum of the column).

Table of Contribution to Mass-Density by Scale

Object: ***size scale***	**Methods**	***M/L***	**Ω-contribution**
Binary stars, star clusters: *AU to a few pc*	Orbit velocities; velocity dispersions; stellar structure models	0.5-3	~ 0.003
Galactic disks and nuclei: *1-10 kpc*	Stellar velocity dispersions; rotation curves	3-10	0.003-0.01
Binary galaxies; small groups: *0.1-2 Mpc*	Velocity differences; X-ray temperatures; orbit modelling	20-50	0.02-0.05
Rich clusters: *1-10 Mpc*	Virial theorem; X-ray temperatures; kinematic models	50-200	0.05-0.20
Superclusters: *~100 Mpc*	Virial theorem; models of Virgocentric in-fall, kinematic evolution, and mergers	100-400	0.10-0.40

4.3 CHALLENGES TO THE EVIDENCE

Nearly every astronomer accepts that there has to be *some* dark matter: given our restricted observational position it makes sense to accept that there is mass that we have not yet identified, whose radiation we have not seen, and which therefore is not included in estimates of the mass-to-light ratios for the various kinds of structures. What is not agreed upon by everyone is *how much* dark matter must there be, and whether it needs to be some *exotic* kind of matter. In this section I enumerate some of the possible objections to the evidence for the existence of a very large dynamical discrepancy, and show why most of these objections do not stand up—in the next section I address the evidence relevant to specific dark matter candidates. Countering the objections in question is usually a matter of assessing a balance of probabilities about which we have little direct information (and little hope of soon improving our epistemic situation), so the judgements reached are indeed fallible. But given the current evidential situation, we ought to accept the existence and degree of the dynamical discrepancies

Speaking abstractly for now, there are several ways in which it is possible to call into question the conclusions of dynamical measurements of mass and inferences from them to the existence of unseen bodies. The first way is to challenge the dynamical assumptions involved in the dynamical measurements. This challenge takes two forms.

(A.i.) It can be a denial that the *laws* used are valid for the phenomena studied, which amounts to a proposal to adopt new laws—this possible solution to the dynamical discrepancies is studied in detail (and provisionally rejected) in Chapter 6. Note that this category includes both purely gravitational laws, and proposals to the effect that other, non-gravitational forces are acting.

(A.ii.) The second form of the challenge to the dynamical assumptions is a denial of the satisfaction of the conditions that must be met in order for the *application* of the laws

to be valid. Several of the objections considered below are of this type, and it is this type of challenge that poses the biggest threat to the existence of a dynamical discrepancy.[16]

(B) The second way of challenging dynamical inferences to the existence of unseen masses is to deny that the discrepancy between the visible mass and the dynamical mass is as large as dark matter advocates suppose it to be. Type-(A.i.) objections often also involve a claim to this effect. (It is often argued that the proposed new law is evidentially supported by the fact that according to the new law the dynamical mass is much closer to the visible mass. And it is supposed that avoiding the need to introduce large amounts of unobserved matter counts in favour of the new law.) It is also possible to argue that estimates of visible mass and the correlated mass-to-light ratios are artificially low. This objection turns out not to be powerful enough to solve the mass discrepancy: it is highly unlikely, given our present observational capabilities and background knowledge, that the mass-to-light ratio could plausibly be as far wrong as it would need to be in order to fully account for the dynamical discrepancy.

(C) A third type of challenge is to say that the margins of error in calculations of dynamical mass are very large because of the many uncertain assumptions that must be made in order to complete the calculation. It is true that many assumptions have to be made, and that the margin of error in the mass calculations is fairly high as measurements in physics go, but it is unlikely that the error can possibly be great enough to make the discrepancy disappear. One of the striking things about the degree of the dynamical discrepancy is that it entirely swamps any reasonable estimate of the error in the calculation (by as much as one to two orders of magnitude).

(D) Finally, specific objections can be levelled against any attempt to argue from the dynamical evidence to a specific matter solution. Such objections will be considered in the next chapter as part of the account of the candidate solutions so far proposed.

---

[16] But one worry raised by these challenges is that it may be *impossible* to dynamically determine the masses of the various kinds of structures in the universe: unless we can assume that the systems in question are in gravitational equilibrium, we cannot apply the dynamical techniques to them. This is equivalent to saying that we must give up large parts of key projects in astrophysics (e.g., on the evolution of structure).

One can see these four abstract kinds of objections exemplified in Trimble's list of the four main classes of alternatives to dark matter that had currency in the 1960s (before the so-called dark matter revolution), and some of which still have some adherents now:

> (1) short-lived, expanding clusters of galaxies, (2) observational errors, substructure, and foreground/background objects, (3) deviations from standard gravitational theory, (4) light between the galaxies that had somehow not properly been counted and would imply M/L ratios like those expected for stellar populations. (Trimble 1993, 155)

Determining the velocity dispersion for a cluster depends on observing the Doppler shifts of the light of its component galaxies—this Doppler shift is interpreted as being due to the motions of those galaxies along the line of sight. The Doppler shifts are then converted to velocities (and then to velocity dispersions) in order to calculate the cluster mass using the Virial Theorem. This has been challenged on four grounds. First, some observers (notably Halton Arp) have claimed to have discovered that the redshifts of galaxies are "quantizied", that is, they do not present a continuum of values. Arp has argued that this implies that the redshift is not due to motion but has some physical cause intrinsic to the objects studied, so that the redshift is also not a measure of distance via the Hubble relation; without this, we have no way of even determining cluster membership. If this is so, obviously the velocities determined by redshift are spurious, as are the masses calculated from those velocities. (The "tired light" explanation of redshift, similar in some respects to Arp's claim, has been ruled out.)

Second, there is the foreground/background problem. (See Figure 4 in Chapter 3.) Redshifts are used to determine the velocity dispersions of clusters, but in order to do this the redshift due to the general Hubble expansion has to be subtracted off. The Hubble recession, however, is determined by taking an average of the observed redshifts; obviously the velocities in the cluster might conspire to make this average an inaccurate indication of the Hubble recession at that distance. (And if we do not know the distance, we cannot determine the radius of the cluster, or map velocity dispersions as a function of cluster radius.) A more serious problem is that it is impossible to tell whether all the galaxies in the observed region are truly part of the cluster. Since we have to rely on the redshift due to Hubble expansion to determine cluster membership, it is easy to see that foreground galaxies with unusually high recession velocities, or background galaxies

with unusually low recession velocities, could be mistakenly counted as members of a cluster when really they just lie along the same line of sight and are physically unassociated with it, and this would lead to erroneous mass estimates.

Third, there is a selection bias in favour of bright (therefore large, or nearby) galaxies; if this is accompanied by an over-estimate of the distance, this will lead to supposing that the dynamical mass of the cluster is higher than it ought to be. More importantly, leaving the dimmer galaxies out of the visible mass estimate obviously increases the degree of the dynamical discrepancy (since the gravitational contribution of the dim galaxies will be noticed even though their light is not).

Fourth, the mass calculated by the Virial Theorem will be wrong if the cluster is not actually in gravitational equilibrium. We have no way of checking the assumption that clusters are gravitationally bound, because we cannot possibly watch a cluster for long enough to tell whether it is collapsing or evaporating. It has even been suggested (see Trimble 1990, 358) that while clusters themselves are relatively permanent structures, their members are not gravitationally bound: the cluster could constantly be exchanging galaxies with the field. (One initial proposal, now debunked, was that the observed velocities of galaxies in clusters could be accounted for if the clusters were expanding out of some explosion; the main problem was that no mechanism for producing such an explosion could be devised.) It has also been suggested that galaxies and clusters are "dissipative structures", energy sinks which maintain their morphology by "spending" energy acquired from outside the system. So far as I know this option is not taken seriously. Attempts to construct models which rely on non-gravitational forces (which would therefore not require extra gravitating mass) have also met with a cool reception. (See Parker 1993 for a definitive rebuttal of Alfvén's "plasma" model.)

As it turns out, none of these four objections is fatal to the idea of large quantities of dark matter being present in galaxies and clusters. At worst, these objections force us to be conservative in making mass estimates, which means incorporating the possibilities mentioned here into the margins of errors for the masses calculated. The velocity interpretation of redshift is very widely accepted, so the first objection mentioned above is not taken seriously—but it should be mentioned that there is no way to *prove* the hypothesis that all redshift is due to recession (Hubble or other), though it fits a broad

collection of facts and theories. Note that if the recessional velocity interpretation of redshift is dropped, we could not be sure about galactic rotation curves or cluster velocity dispersions; moreover, we would not have reliable distance estimates to these systems and therefore could not convert angular size into true size, a crucial step in deciding how much mass has to be present to hold a system together given its velocity of rotation at specific radii. The second objection is more serious, but can be dealt with using statistical techniques; in any case *most* of the galaxies that appear to belong to a cluster, certainly do belong (conditions have to conspire in a very particular way for the foreground/background problem to have a serious effect on dynamical measures of mass). The possibility that some galaxies that are not truly part of a cluster have been included in its Virial mass can be taken care of by increasing the margin of error attributed to the mass value. The third objection can be taken care of in a similar way, and better observational studies, using more sensitive instruments, are able to put upper bounds on the amount of "visible" mass that could be missed. The likelihood of the fourth possibility is greatly lessened by observations of the very hot gas halo in clusters; if clusters were accidental (and not gravitational) conglomerations, there would be no explanation for why clusters generally have gas envelopes at all, and no explanation of why the gas does not dissipate despite its high temperature. Of course, these four objections apply only to cluster dark matter, and not to the galaxy-specific dark matter.

It has also been suggested that we should not accept the assumption that stars are gravitationally bound to their galaxies. This is even less plausible than the corresponding claim for clusters, in that it seems impossible to account for star formation except in the presence of a very large quantity of mass (in fact, some accounts of star formation depend on shock waves from supernovae in previous generations of stars; this obviously requires that lots of dense gas and stars be near together, as they are in galaxies—and as they are *not* in intergalactic space). Still we can admit that some of the fastest moving stars observed in the Milky Way have exceeded the escape velocity, without thereby being forced to throw out the hypothesis of dark matter altogether; other kinds of observations (including the VLBA study and others that measure the mass interior to our galactic orbit) indicate the need for dark matter even if Oort-style studies do not.

It has been claimed that the velocity measurements (for both clusters and galaxies) suffer very large margins of error, which of course correspond to large errors in the dynamical estimates of mass. This is surely correct. But in order for there to be *no* mass discrepancy—that is, for the discrepancy to be due entirely to overestimates of the velocities—the velocities would have to be *very* wrong indeed. The implausibility of this is compounded by the fact that the existence of the dark matter problem does not depend simply on seeing that the rotation curves are *faster* than expected, but rather primarily derives from the rotation curves having a form decidedly incompatible with the distribution of visible light in these objects. Note that the *shape* of the rotation curve is accurate even if the velocities are not, since velocities at every radii will be in error by the same amount and in the same direction. Nevertheless, improving the accuracy of velocity determinations (limiting the sorts of measurement error discussed in this section) is very important to the eventual solution of the dark matter mystery, since a decision between two candidate solutions may depend on minute details of the dynamical evidence.

Similarly, it has been claimed that M/L ratios have been inaccurately estimated, and that were we to take proper account of the dim mass present in the systems studied, the mass discrepancy would decrease. This is so, but it seems impossible that the M/L ratios could be adjusted enough to make the dark matter problem go away entirely. The table above shows that the dynamical mass (depending on the scale of system considered) is roughly 20-100 times the mass expected given the number and type of stars observed; it is extremely unlikely that this much extra mass is really there but not properly accounted for by our calculations from the radiation flux.

While it may be true that lots of light from spirals, and other galaxies, will be absorbed by dust, that light will necessarily later be re-emitted (in the infrared). So we necessarily have a complete census of the luminosity L if we study all possible wavelengths. Trimble (1993, 156) mentions a study that fixes the maximum effect of dust absorption to a factor of two in underestimating the real value of L, which makes estimated "visible mass" values half what they should be. But the dynamical discrepancy is much larger than this, and is even much larger than this combined with the uncertainty in the Hubble constant (which is important in judging the distance to other galaxies,

which in turn is required in calculating their true diameters from their apparent sizes). Elliptical galaxies have much less dust, so they are less affected by these considerations and yet their dynamical masses are still very much higher than their visible masses.

There are various problems with drawing inferences about dark matter from observations of large scale structure (including problems similar to the ones discussed above for clusters). However, on the whole the errors involved make it more likely that the total mass will be *underestimated* rather than overestimated, because the largest structures might be larger than we now think. What remains a matter of debate is the scale of the largest gravitationally bound structures. An acknowledged limitation of our knowledge is that no survey we could hope to complete could map more than a tiny fraction of the universe as a whole, which means that we just do not (and cannot) know what the universe is like on the largest scales. (Trimble 1993, 153) Of course, not much depends on estimating the quantity of dark matter at these extremely large scales; we have enough other evidence to show that we need to take the dark matter problem seriously.

4.4 OBSERVATIONAL CONSTRAINTS ON DARK MATTER CANDIDATES

This is the evidential situation in which we find ourselves. One difference between astrophysical investigations and other kinds of physical inquiry is perhaps the fact that the evidence which indicates the existence of the problem is (almost—see below) the only evidence available with which to try to construct and test candidate solutions. Candidate matter solutions to the dark matter problem tend to be constructed with this evidence in mind, that is, the models are fitted to (some subset of) the available information. Some writers, for example Philip Mannheim, object to matter solutions on the grounds that every matter solution is *ad hoc* for this reason (see Chapter 6). But it seems that finding a matter solution to the dark matter problem is just a question of describing some physical entity (or a combination of several different ones) capable of producing the observed dynamical masses while remaining consistent with the (lack of) observed energy flux.

This is actually quite a bit less easy than it may sound. As Trimble is fond of pointing out, the masses of candidate dark matter particles range from 10^{-38} grams

(fundamental particles) to 10^{-39} grams (supermassive black holes) (Trimble 1993, 153, and elsewhere). This range, she says, is a measure of our ignorance about the nature of the dark matter, and of how little the available evidence tells us about its properties. The situation is not hopeless, however, in that observational and explanatory constraints (at least in principle, and likely also in fact) are capable of eliminating some candidates and of standing as evidence for others. It seems likely that the continued extension of the process of developing and rejecting candidate solutions will eventually lead to success. This process, even when it results in the rejection of a candidate, provides detailed information about what the ultimate solution must be like, or rather about both what the candidate must be like in order to save the observed motions, and what it must not be like in order that it be consistent with all the other empirical and theoretical constraints. The process of reasoning about evidence in order to provide information that will allow us to constrain the class of possible dark matter solutions is exactly analogous to the process of reasoning about matter solutions in response to the anomalies in the motions of Uranus and Mercury, although in the dark matter case the evidence and constraints are quite a bit more complex in detail. Thus Trimble's mass range for dark matter candidates is somewhat misleading, since nearly every candidate in the range of proffered dark matter models has been ruled out in one way or another.

Various gross-matter hypotheses and a virtual "zoo" of fundamental particles have been considered as dark matter candidates. The typical pattern is for these candidates to be introduced in order to solve a dynamical discrepancy at one level of astronomical structure, and then to be eliminated because they are inconsistent with the evidence at some other level of structure, or because they turn out to have properties that would render the candidate particles observable when in fact they are not observed. I list here an incomplete but informative catalogue of some of the candidates that have been considered as dark matter solutions (namely, those I discuss in Chapter 5). Note that in most cases fundamental particles (WIMPs) are originally proposed as solutions to the *cosmological* dark matter problem or to problems of the formation of large scale structure; these candidates are included because whether or not there is cosmological dark matter, such particles (if they exist) could contribute to the dynamical masses of galaxies and clusters.

Table of Matter Candidates

BARYONS
- dust
- gas
- MACHOs:
  - jupiters
  - brown dwarfs, red dwarfs
  - old white dwarfs
  - stellar mass black holes
  - primordial black holes
- supermassive black holes

NON-BARYONS
- WIMPs
  - neutrinos (electron, muon, tau)
  - axions
  - "-inos"
  - other fundamental particles
- TOPOLOGICAL DEFECTS
  - point defects, monopoles, some kinds of primordial black holes
  - line defects, cosmic strings
- OTHER?

The candidates divide into two main classes, baryonic and non-baryonic solutions. The choice between baryonic and non-baryonic dark matter,

> depends heavily on whether you are satisfied with a universe having $\Omega = 0.2 \pm 0.1$, as suggested by most of the observations [more recent observations set $\Omega = 0.4$, but the point is unaffected by this change]. . . . The alternative, $\Omega = 1$, is favored by theoretical considerations of galaxy formation and inflation. (Trimble 1993, 153)

Some of the candidates proposed in order to solve the cosmological dark matter problem would also help with the dynamical dark matter problem, and so those candidates are considered here in that light. In the present context, I take it as legitimate to invoke cosmological considerations to rule such candidates out, but not to use such considerations as evidence *for* a candidate (thus I leave the cosmic dark matter question completely open). Many of the cosmic dark matter candidates are ruled out because they fail to account for, or are inconsistent with, the available dynamical evidence.

CHAPTER 5

MODERN DARK MATTER: ASSESSING THE CANDIDATES

We see it as Columbus saw America from the shores of Spain. Its movements have been felt, trembling along the far-reaching line of our analysis, with a certainty hardly inferior to ocular demonstration.

—John Herschel speaking to the British Association, 10 September 1846, about the imminent discovery of Neptune, as quoted in Jones (1956, 832).

5.0 BARYONS

The name "baryon" refers generically to protons and neutrons. To say that the dark matter is baryonic is to say that the dark matter problem is simply a "booking keeping" error: dark matter is not some exotic unknown type of matter, we just did not count all the ordinary matter. There are two main constraints on baryonic solutions. The first is that if the dark matter is in baryons, we have to be able to account for why it is *dark*. The second constraint (philosophically more interesting) is that arguments based on the production of fundamental particles in the early universe can put limits on the baryon content of the present universe. The evidence in support of these arguments is very strong, and therefore leads to constraints that cannot be ignored. If we were to find that the total dynamical mass of the universe is greater than the amount of baryons in the universe, this would be definitive evidence that (at least some of) the dark matter is non-baryonic. Limits on the number of baryons therefore provide information about the nature of the dark matter. As it turns out, present evidence is ambiguous about whether the dynamical mass is above or below the baryon mass: at a maximum, though, there are just barely enough baryons to account for the dynamical masses of galaxies and clusters (according to dynamical measures, $\Omega_{total} = 0.2$ *to* 0.4, while according to calculations based on helium abundances $\Omega_{baryon} \leq 0.2$; according to recent probes of large scale mass density (for example, Alcaniz and Lima 1999, L89), $\Omega_{matter} \leq 0.4$). Some commentators take the close agreement between the dynamical mass of the universe and

the predicted baryon mass to be strong evidence in favour of the hypothesis that all the dark matter is baryonic. (Bartusiak 1993, 223) That is, they take the coincidence of arriving at (almost) the same number (albeit with fairly large margins of error) by two different methods to be a sign that the number is correct. But the numbers are not known precisely enough for this agreement to be decisive, and so other arguments against purely baryonic dark matter are still important.

Let me begin by outlining the arguments that lead to $\Omega_{baryon} \leq 0.2$ (the following account is adapted mainly from Weinberg 1993). According to the big bang theory of cosmogenesis, the early universe was an extremely hot, extremely dense state, initially just pure energy expanding outwards (but without any particular centre of expansion). As the universe expanded, it cooled, as required by the gas law: the total energy being held constant, an increase in volume leads to a decrease in temperature. As the universe cooled, it became possible first for quarks (and other very light particles such as neutrinos), and later for larger particles made up of quarks, to exist for definite periods, although at first these particles were quickly annihilated by interaction with the extremely dense sea of very high energy photons. Eventually the photon temperature cooled enough that any baryons that formed could survive without being dissociated by impacts with the photons. Soon thereafter, it was cool enough for atomic nuclei to form. (The era when it became possible for electrons to remain bound to nuclei came later, and corresponds to the moment when the universe became transparent to radiation: the cosmic background radiation (CBR) is a relic from that moment.)

> [T]he nuclear reactions that produced light elements in the first few minutes of the "big bang" are affected by the ratio of the number of these atomic particles to the number of photons. . .present at that time. A relatively high ratio of atomic particles to photons would allow the nuclear reactions that convert hydrogen to helium to proceed more nearly to completion, reducing the amount of matter left over in the form of less tightly bound light elements like deuterium or lithium. These light elements are not believed to be produced in the stars, so measurement of their present abundances informs us about the ratio of atomic particles to photons in the first few minutes. But this ratio has not changed appreciably since, so we can infer something about its present value, and hence (since we know the number of photons per cubic centimeter in the cosmic microwave radiation background) about the present abundance of atomic particles. (Weinberg 1993, 182-83)

Before a certain time the universe was too hot for it to be possible for helium nuclei to persist: the average energy and density of photons was so high that if any helium had existed it would have immediately been dissociated into its constituent atomic particles by that radiation. Eventually the universe cooled enough that it became possible for protons and neutrons to form small groups that were stable over long enough times that a series of reactions leading from hydrogen to helium nuclei became possible. Since helium is a more stable nucleus that the others involved in this chain, the limiting constraint on the conversion of hydrogen to helium is the average lifetime of the particles involved in the intermediate steps: the photon temperature below which helium is stable is greater than the minimum energy required to dissociate the nuclei of the intervening steps. It follows that whether or not it is possible to produce helium from hydrogen in the conditions of the early universe depends on whether the intervening nuclei can survive long enough for the reaction to come to completion. This, in turn, depends only on two things, the *density* (which, like the temperature, is merely a function of time, and so can be ignored since at some time or other after the big bang the right temperature and density conditions will exist for nucleosynthesis) and the *photon-baryon ratio*. A higher number of baryons as compared to photons means that the baryons are closer together, and that a smaller fraction of the energy of the universe is in photons capable of dissociating the nuclei in question. A higher number of baryon-baryon collisions means that a greater quantity of hydrogen is ultimately converted to helium. Thus the ratio of helium to hydrogen in the present universe is a measure of the ratio of baryons to photons in the early universe. But since mass-energy is neither created nor destroyed after the moment of origin, and since the baryon number is almost constant after the decoupling of matter and radiation (inaptly named the "*re*combination" time, when it first became possible for electrons to remain bound in atoms), it follows that the ratio of baryons to photons now is the same as it was at recombination. Thus the observed relative abundances of the light elements (particularly hydrogen, helium, deuterium and lithium, which except for helium are not produced as final products by stellar fusion and hence are truly primordial) is a measure of the present ratio of baryons to photons. The observed energy flux thus fixes the number of baryons, and it turns out that the total mass of baryons is probably not enough (or perhaps just barely enough) to account for the dynamical masses of galaxies

and clusters. Once the calculations of the baryon mass and the total mass of the universe are firmed up, and if it is found that the former is less that the latter, this will be the strongest evidence we have, though not the only evidence, that the dynamical dark matter is not "ordinary" matter. Even the present calculations are robust enough that it is definitive that the *cosmological* dark matter (if it really exists) cannot be ordinary baryonic matter. If there *are* enough baryons to account for all of the dynamical mass, we need to explain how it is that 90% of these baryons are neither absorbing nor emitting sufficient electromagnetic radiation that we can detect them.

Weinberg remarks of the nucleosynthesis inference just described,

> It is truly impressive that with the plausible choice of a single free parameter, the ratio of atomic particles to photons, it is possible to account for the observed present abundances not only of ordinary hydrogen and helium (H^1 and He^4), but also the isotopes of H^2 (deuterium), He^3, and Li^7. (Weinberg 1993, 183)

In fact we ought to view the logical order as running in the opposite direction: given plausible and minimal assumptions about the early universe (thermal equilibrium, temperature before decoupling, and so on), the observed abundances of the light elements (in conjunction with theoretical and empirical reasons for thinking that light elements above helium are never synthesised in stars, only destroyed), *measures* the photon-baryon ratio: since the CBR gives the photon number, it is relatively easy to calculate the baryon number from this ratio and hence the total baryon mass of the present universe. This is the direction of inference in which the observed abundances and the theory of nucleosynthesis bear on the dark matter issue: these things together provide us with a maximum value for how much of the dynamical mass can be baryonic.

The calculations of the relative abundances of the light elements "agree best with the observations when the average baryon density is 1-20% of the closure density (that is, $\Omega \leq 0.2$). This is just barely enough to account for the dynamical masses of galaxies and clusters" (Trimble 1993, 154). Note that this fact does not necessarily tell us much about the relative baryon-to-dark matter abundances in galaxies and clusters, since quite a lot of *ordinary* matter might be hidden.

The mechanism of nucleosynthesis also predicts an equal number of anti-baryons, but these have not been observed. In fact, observations of background radiation limit the

antimatter abundance to one particle in a million in deep space (Bartusiak 1993, 255). If there were no way to account for a bias in favour of matter creation (as opposed to antimatter creation), the observations just mentioned would give a reason to mistrust the nucleosynthesis determination of determining the baryon fraction. However, as Weinberg (1993, 183) notes, a 1964 experiment showed that the laws of particle physics are not perfectly symmetric between matter and antimatter. This explains the slight excess of matter in the early universe, which led to the universe we presently observe: Grand Unified Theories (GUTs) also predict a very slight excess of quarks over antiquarks (one in ten billion) in the moment of matter-radiation decoupling, which means that after the annihilation of all the antimatter there was a slight bit of matter left over, which became the baryons. (This annihilation scenario also explains the photon-to-baryon ratio of 10 billion to one.) If this is the correct explanation of the fact that we do not observe an amount of antimatter in the universe equal to the amount of matter, then the observed light element abundances are indeed strong evidence for the big bang, and good reason to trust the derived value of the baryon mass-fraction.

To summarise, if the fact that the dynamical mass is measured to be about equal to the predicted mass of baryons (from observed light element abundances, given a theory of nucleosynthesis), is confirmed in more accurate studies, this will be the strongest evidence we have in favour of the dark matter being entirely baryonic. But if this turns out to be so, two problems related to the distribution of mass in the universe will have to be solved. First, we need to find some mechanism for excluding significant numbers of baryons from the voids,[1] so that they all fall just where they are needed in order to account for the measured dynamical masses of galaxies and clusters (if the density of baryons in the voids is at all far above zero, the baryons will be too spread out to account for the missing mass in galaxies and clusters). Second, if the dynamical mass in those systems is all the mass there is in the universe, then there is not enough time or gravity to evolve galaxies and clusters from the observed primordial density fluctuations. (We

---

[1] Observations of large scale structure indicate that above the scale of superclusters (~100Mpc) visible matter is distributed in a foam-like network of sheets and filaments that intersect, bounding regions nearly empty of visible matter, called "voids". See Huchra and Geller 1989, or Geller 1989.

reach a similar conclusion if we want a *gravitational* explanation of the exclusion of the baryons from the voids). In other words, additional (non-baryonic) matter is still required *even if* the dynamical mass of clusters and galaxies is accounted for by all the baryons (and if this additional matter exists, it could contribute to the dynamical mass of galaxies and clusters, so that there would be little reason to insist that all the dynamical mass is baryonic).[2] One might be tempted to take this simply as a problem with the theory of structure evolution that has nothing to do with the mass content of the universe, but it is more plausible to conclude that some of the dynamical mass is non-baryonic.

In any case, another kind of objection that leads to the same result can be posed: it seems impossible to find a way to prevent such a large quantity of baryons from taking forms which would be visible because of their absorption and especially emission of radiation. Showing why this is so is part of the task of the next sub-sections, which review some of the baryonic candidates.

5.0.1 Baryonic Candidates: Dust and Gas

The nucleosynthesis predictions—based on the observed photon flux (from the CBR) and the observed abundances of light elements not synthesised in stars—indicate that it is just possible that all the dark matter is baryonic. The question then is what form these baryons take that results in their having no observable electromagnetic signature, even though in these forms the baryons would have to make up at least 90% of the total mass of galaxies and clusters. In what follows I consider several of the possible arrangements of baryons that might be able to do the job.

One initially plausible configuration of baryons as dark matter is interstellar dust and gas. Both Oort and Zwicky, for example, at first took it to be likely that the

---

[2] Thus the idea that all the dynamical mass is baryonic can only just be saved if a theory of the formation of large scale structure can account for how the baryons were *all* sequestered in the galaxies and clusters, but no theory can account for the observed degree of clumping at large scales evolving from the initial density fluctuations in the time available *without* incorporating *large* amounts of additional mass (far too much for all of it to be baryonic). Since we need large amounts of non-baryonic matter for the formation of large scale structure anyway, there seems to be little reason to insist that the dynamical dark matter cannot or should not be non-baryonic as well.

dynamical discrepancies they discovered could be solved in this way. But several things make it implausible that dust and gas can be the whole solution. First, in order for dust to make up the missing mass, there would have to be so much of it that our view out of our galaxy would be obscured, and other galaxies would have a very different appearance. Light passing through a dust cloud is reddened (by selective absorption of short wavelength, blue light) in a way that is not observed in other galaxies. Furthermore, and more definitively, any body whose temperature is above absolute zero radiates at a wavelength determined by the temperature of that body: we now have sufficient technological capabilities and extensive enough surveys in the infrared, ultraviolet, radio, visible and X-ray portions of the electromagnetic spectrum that we can be sure that if a large amount of gas and dust existed in our galaxy, we would be able to detect it no matter what its temperature.[3] Since the observations do not indicate an energy flux in any of these wavelengths sufficient to account for the huge amounts of gas and dust that would be required in order to explain the dynamics of galaxies, dust and gas cannot be the dark matter. A cold gas halo would have fallen into the galactic centre by now, and a hot gas halo would radiate in a very easily detectable way.[4] Finally, since every chemical substance has a distinctive spectrum, when looking at another other bright object we would also find bright emission lines not due to the bright object if there were some cloud of hot material intervening between us and it, or we would find dark absorption lines in the spectrum of the bright object if the intervening material were cold. Both these effects are present in some observations, but neither in sufficient quantity for dust or gas to be all the dark matter. Of course, there is a large amount of gas and dust in our galaxy and others, and it makes up a significant portion of the total mass. The observations just indicated, then, put an upper bound on how much of the galactic mass can be attributed to

---

[3] A 10^8 K gas between the galaxies was ruled out in 1990 because it would have distorted the spectrum of the CBR (by scattering) more than the COBE observations allowed. (Trimble 1993, 153)

[4] In January 2000, the FUSE satellite group announced observations detecting for the first time a halo of hot gas around the Milky Way (thought to have been produced by supernovae explosions). The density of this hot gas is, however, too low for it to contribute significantly to the missing mass. (space.sci.news, 13 Jan 2000)

gas and dust. Knapp (1995, 20) notes that interstellar gas and the stellar population of the Milky Way each make up about half of the *visible* mass of our galaxy. Bartusiak (1993, 161) notes that interstellar gas makes up perhaps 5 to 10 percent of the total mass of the Milky Way. This puts constraints on the fraction of the mass that is truly dark and not merely dim, and this in turn constrains the nature of the dark matter. Note that if gas and dust were the dark matter in galaxies, it would have to be present in quantities ten to twenty times greater than other observations indicate it is.

A recent announcement (space.sci.news, 17 August 1999) describes new infra-red observations of distant galaxies using the European Space Agency's ISO satellite. The spectral signature of molecular hydrogen (H_2) was detected in all eight places studied along the radius of a distant spiral (NGC 891, some 30 million light-years distant). The energy flux seems to indicate that molecular hydrogen is in the range of 5-15 times the atomic hydrogen in these regions. This is a much higher fraction of molecular hydrogen than is typically expected. This is significant, according to investigator Edwin Valentijn, because "it is well established that if there is about 10 times as much molecular hydrogen as atomic hydrogen in the disks of spiral galaxies, then the missing mass problem is resolved" (as quoted on space.sci.news, 17 August 1999; see also http://sci.esa.int). Of course, the results depend on a single set of observations of a single galaxy, albeit one that otherwise seems perfectly typical.[5]

5.0.2 Baryonic Candidates: Machos

As discussed above, arguments from gravitational stability indicate that the visible disk of the Milky Way, and the disks of other galaxies whose rotation curves remain flat out to high radius, must be surrounded by a massive "halo". This halo will most likely be spherical, or close to spherical (see Tayler 1993, 59-60, on the oblateness of the halo), and will contain most of the total mass of the galaxy in question. This halo is necessarily

---

[5] I have been unable to track down the "well established" argument that this quantity of molecular hydrogen solves the galactic dynamical discrepancy. Even if this claim is right the results mentioned in the text have nothing to say about the dynamical discrepancy for clusters, or on the issue of the formation of large scale structure. What is also so far not explained is why, if this quantity of molecular hydrogen is typical, its signature has not been observed in our own galaxy, or in our nearest galactic neighbour, M31.

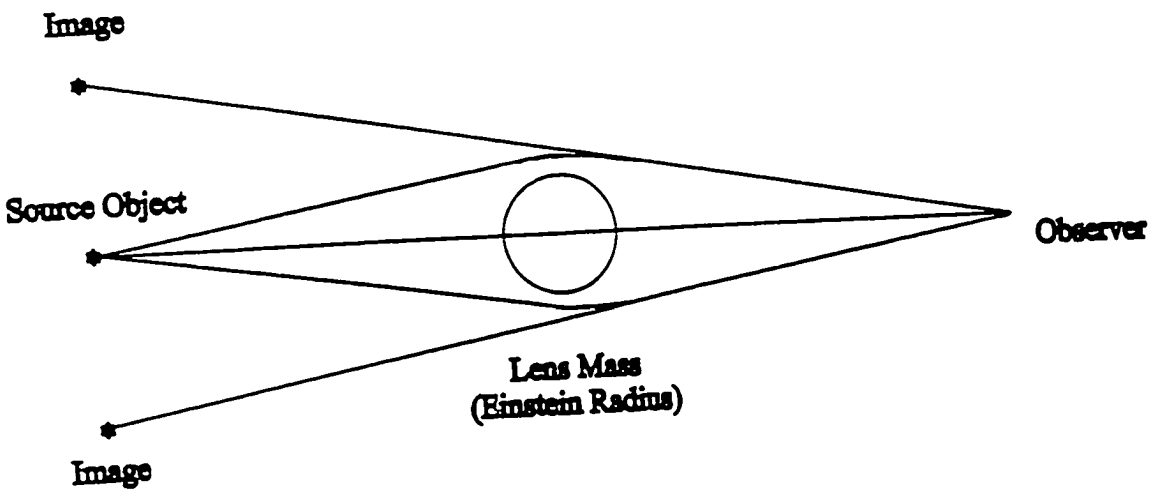
mostly dark matter, since if it had a significant visible component, we would be able to detect its light rather than having to infer its existence and mass from dynamical effects. One possibility is that the halo objects are stellar- or sub-stellar-bodies—"jupiters", old white dwarfs, red dwarfs, brown dwarfs, neutron stars or stellar-mass black holes. These objects, known as MACHOs (Massive Astrophysical Compact Halo Objects), though quite different from one another in most respects, are alike in that they emit little radiation although they are massive. A cloud of them could quite possibly be missed by telescopic searches (in the visual, infrared or ultraviolet bands), and so they have exactly the right characteristics to be considered as dark matter candidates.

Several versions of an observing program to search for jupiters or dim stars in the halo of the Milky Way have been tried. These programs look for "micro-lensing events"[6], that is, characteristic changes in the light profiles of background stars: when the brightening has a specific pattern that preserves the spectrum of the light, is symmetrical in time, and never repeats) it is argued that the cause cannot be an intrinsic brightness variation but must be due to an otherwise unseen, relatively nearby, massive body passing across the line of sight. The duration of the event and the maximum increase in the brightness of the background object are used to calculate the likely mass, distance and velocity across the line of sight of the lensing body. From these facts inferred about the body, further inferences about its nature can be made. Given that the lensing bodies are too dim to be detected from Earth, and given the range of masses to which the observations are sensitive, any detected MACHOs are likely to be failed stars (brown and red dwarfs, jupiters), dying stars (old white dwarfs, neutron stars) or stellar-mass black holes. Less massive objects at typical halo distances will not produce detectable microlensing effects: more massive or hotter objects would produce a detectable radiation flux of their own.

---

[6] To date, approximately 300 microlensing events have been detected by the MACHO group (space.sci.news 14 Jan 2000). (See Figure 6.) Perfect alignment in gravitational lensing situations for a point-source and a point-lens yields a ring with Einstein radius, r_E; imperfect alignment yields images separated by $2r_E$. Where the amount of lensing is small (as when the mass of the lens is low), the images cannot be resolved and what we see is an increase in brightness of the background star as the lens crosses its line of sight.

Figure 6.



The number of microlensing events detected is lower than expected, but statistical arguments from the number of stars in the field (in the millions in each study) and the number of lensing events observed over a given time provide constraints on the number of MACHOs in a given volume of space. And from this, the minimum radius of the galactic halo, the range of masses of the lensing bodies and the total fraction of the mass of the halo due to MACHOs can be calculated. The microlensing proponents at first had hoped to show that the halo was mostly made of MACHOs: had this been true, there would have been no need for non-baryonic dark matter. However, there have been fewer observed lensing events than expected on the MACHO-as-dark-matter models. The results show that only a very small fraction of the total halo mass required for gravitational stability can be found in MACHOs. This is very important evidence with which to try to figure out what the dark matter is.

Besides the failure to detect MACHOs via microlensing, there are other reasons to reject these objects as the dark matter. No one has proposed a plausible mechanism for the formation of a sufficient number of white dwarfs, neutron stars or stellar-mass black holes (which are end-states of massive stars).[7] Clouds of red dwarfs in sufficient quantities to make up the dark matter would be easily detected. In fact, a dedicated search with the Hubble Space Telescope turned up many fewer red dwarfs than expected on models of star formation, and *many* fewer than would be required for them to be the dark matter. That leaves brown dwarfs, which are even cooler and dimmer, and therefore much harder to detect. If brown dwarfs are the dark matter, there should be one in every 30 cubic light-years of our galaxy (Bartusiak 1993, 232). But few brown dwarf candidates have been turned up in observational surveys, and far too few microlensing events have been observed for them to be common enough in the halo of the Milky Way or in the halo of Andromeda. Some theorists have suggested that brown dwarfs are so cool that smoke-like particles could form in their atmospheres, which would make their apparent luminosities even lower, and this might account for why they have not been seen as often expected (Bartusiak 1993, 231-32). Note, though, that this cannot explain the

---

[7] Note that the number of stars of a given mass in a typical population is proportional to the inverse square of the mass (Trimble 1993, 151).

microlensing results. Taken together, then, these two lines of evidence seem to rule out red dwarfs and other minimally luminous, roughly stellar-mass bodies as the dark matter.[8] In short, the observational evidence presently available is strongly against the galactic dark matter being MACHOs, but this candidate is not yet definitively ruled out. If there *were* a sufficient number of MACHO events, characteristics of the events (increase of brightness, duration) could be used to determine the mass range of the lensing bodies. In combination with formation scenarios, this could go a long way toward establishing the nature of these objects even if direct detection remains impossible.

5.1 BLACK HOLES

Black holes are *prima facie* excellent candidates for dark matter: they are massive and by definition no radiation escapes from them. Some models of black hole formation also provide an excellent mechanism for sequestering baryons so that their gravity but not their luminosity has an impact on the universe. But as I will describe here, there are strong reasons to think that no significant portion of the dark matter can be made up of black holes. To begin, we may distinguish several types of black holes. The only intrinsic difference between the different "types" is their mass (though they also differ with regard to how they are formed, which is important to how they are distributed in space), but this in turn affects the ways in which a given black hole can be detected, or inferred to exist in a given location. I will first consider so-called "supermassive" black holes, and later turn to "stellar-mass" and "primordial" black holes.

Observations of several sorts now make it a near certainty that supermassive black holes (with masses in the region of 10^{6-9} solar masses) are to be found at the centres of most if not all galaxies. They can be shown to exist there in two ways. The first way relies on the fact that matter falling into a black hole will heat up and emit radiation. The peak wavelength of radiation emission is determined by the temperature of the matter, and the temperature is determined by the orbital velocity. Thus radiation emitted at a

---

[8] Note that although we now know that jupiter-like planets exist around a few dozen other stars, and therefore are likely to exist around most stars, planets cannot make up a significant fraction of the galactic mass: even Jupiter is only one-thousandth the mass of the Sun.

given radius can be shown to originate in matter orbiting at a certain velocity, and this in turn is the basis for a dynamical determination of the central mass. The central regions of most galaxies emit high energy X-rays, which can only be produced by very hot, and therefore very fast, matter, which implies a very high central mass. The second way to determine the existence (and mass) of a central supermassive black hole is to study the velocity dispersion of nearby stars. Mass determinations based on these two methods show that the amount of mass that would have to be present is so "astronomically" high that the only plausible candidate for the central body is a supermassive black hole. Additional evidence comes from the fact that large disks of material rotating at high speeds have been observed at the centres of some galaxies. There is no way to account for these observations except as accretion disks of supermassive black holes.

So black holes are an attractive dark matter candidate, and there is reason to think that supermassive black holes exist at the centre of all or most galaxies. Unfortunately, although such objects no doubt do contribute significantly to the total mass of galaxies, it is not possible that they are the solution to the dark matter problem. The main obstacle is that although increasing the central mass will affect the "normalisation" or *height* of the rotation curve for the galaxy, it will have no effect on the *shape* of the rotation curve. The Keplerian expectation is still violated, so we still need a massive dark halo, even if central supermassive black holes do contribute significantly to the overall dynamical mass of galaxies. If the halo were made of supermassive black holes, some of them would eventually migrate to the galactic core, and then we should expect the galactic core to be much more massive than it is. Alternatively, matter falling into such objects in the halo (as it surely would if there were enough of them to make up the dark matter) would radiate in a very easily detectable way.

Supermassive black holes also cannot possibly be the answer to the mass discrepancy for *clusters*, although no doubt they do contribute a lot of mass, and we know they exist. But the fact that more X-ray signatures are not found puts limits on the contribution to the total mass density in galaxies and clusters that can be expected to exist in such forms, and this total is not enough to solve our problem.

Stellar-mass black holes (which take a range of masses within a few times the mass of the Sun) are thought to form when some stars that are more massive than the Sun

reach the end of their fusion-burning lifetimes and can no longer support themselves by thermal pressure against their own gravity: if the mass is sufficiently high (above the Chandrasekar limit, or 1.44 solar masses) the star will collapse to a singularity (slightly less massive stars are supported against total collapse by neutron degeneracy pressure and form neutron stars). Stars capable of ending as stellar-mass black holes are quite rare, according to the relation mentioned above ($N \propto M^{-2}$; see note 7).

Stellar-mass black holes do likely make up some fraction of the galactic mass, although few candidates have so far been identified. But several factors make them implausible as the entire solution to the dark matter problem. First, given the model of their formation, it is difficult to see how to produce enough of them in the lifetime of a galaxy. Second, such black holes will mostly be found where the stars are, that is, in the disks of galaxies, not in the haloes where most of the dark matter has to be. Third, even though dark matter is diffuse in the halo, there would have to be so many such black holes that some baryonic matter (gas or whatever) would inevitably fall into them, and they would collide with each other, fairly frequently: the distinctive radiation signature of such events is not observed. Fourth, the MACHO microlensing studies (discussed above) are sensitive to bodies in this mass range, so the paucity of microlensing events also means a low number of stellar-mass black holes in the halo.

The third type of black hole is the primordial black hole (PBH). These are supposed to be created in great numbers in the early universe: individually they are not very massive. Let me begin by noting that PBHs are really a proposed solution to the *cosmological* dark matter problem, which as I mentioned in Chapter 1 is perhaps no problem at all. (Hawkins, 1993 and 1997, is an enthusiastic if somewhat bitter advocate of the PBH solution to the cosmological dark matter problem.) Provided a convincing mechanism for their production can be established, PBHs could also possibly be (or contribute to) the dynamical dark matter. But it seems that primordial black holes suffer the same defects as other black hole dark matter candidates: if they exist in sufficient quantities to be the dynamical dark matter, they either would be detected in virtue of the radiation of in-falling matter, or by microlensing.

Primordial black holes are interesting in part because they remove baryons from the universe before nucleosynthesis takes place (Trimble 1993, 154-5): on some models

at least. PBHs form out of free protons and neutrons before the era when it is possible for atomic nuclei to persist. Thus, they can explain why the dark matter is *dark* without forcing us to postulate otherwise-unknown kinds of particles. The trick is to find a way to produce enough of them in the early universe, and to hide their presence now.[9]

It has been suggested (see Chapter 2) that the relative long-term stability of planetary orbits in our solar system puts limits on the mass and space density of black holes and other MACHO candidates, where the "particles" of dark matter are individually very massive. A better (and more general) way to put this is to say that the long-term stability of the solar-system puts constraints on a quantity expressing a relationship between the mass and distribution of all dark matter candidates: it has to turn out to be improbable that particles of mass sufficient to disrupt the solar system passed near enough to do so in the time since the solar system formed. What is fixed by this evidence is a statistical description of the space density and mass of individual particles of matter (the total mass is determined by the dynamical studies): the space density can be higher if the particles are less massive, and must be lower if the particles are more massive.

5.2 WIMPs

Weakly interacting massive particles (WIMPs) are the most popular of the non-baryonic candidates. This is partly because some such particles are known to exist, and others are predicted to exist by the same theories that successfully account for particles that are known to exist. And by their very nature, as the name suggests, these particles fit the bill as dark matter candidates: they have mass, but interact only "weakly" with ordinary matter, and are therefore very difficult to detect. To be more precise, such particles are only affected by the weak nuclear force and gravity, so they do not interact electromagnetically (and therefore have no direct visible effects): weak interactions

---

[9] Various models have been proposed; for some recent ones see *Physical Review D*, Vol. 59 (1999), 124013 and 124024. These give a mechanism for PBH formation from initial adiabatic density fluctuations during the radiation-dominated phase of the universe; the density fluctuations give limits on the PBH mass spectrum. Note that it is generally assumed that dark matter candidates must have a plausible "creation story" that stands independently of their worthiness as solutions to the dynamical discrepancy: this methodological constraint in effect rules out some "overly *ad hoc*" hypotheses.

between particles are rare, so it is also hard to detect them in this sense. As David Schramm remarks (Lightman and Brawer 1990, 443), it is ironic that two *incorrect* particle accelerator experiments in the early 1970s which claimed to have detected the neutrino mass were responsible for WIMPs becoming (sociologically) viable as dark matter candidates. Although these initial accelerator experiments were wrong about the neutrino mass, it has recently been established that the neutrino does indeed have a non-zero rest mass (see below). Many of the WIMPs are "well motivated" in the sense that particle physicists invented them for other purposes, and yet they seem (or seemed) to also solve some cosmological problems (Bartusiak 1993, 333).

Observations put limits on the amount of matter present in the universe that is capable of interacting in ways other than just weakly and gravitationally: "If most of the hypothetical dark matter particles were capable of electromagnetic or [strong force] interactions, we would see them easily. Thus their only interactions must be gravitational, weak, or a combination of the two" (Trimble 1993, 155). The observed radiation flux at all wavelengths is the maximum emission of electromagnetic radiation that can be predicted by the final model of the total matter distribution (including visible as well as dark matter). Given that the known visible mass accounts for the observed radiation, it seems that the dark matter must not interact electromagnetically. That leaves gravitational and weak interactions, which leave no electromagnetic traces. This is an important piece of information for the dark matter search, in the sense that matter candidates fitting this description are certainly many fewer than we would otherwise have to consider. This argument does not by itself rule out baryonic dark matter because there are several configurations of baryons (as discussed above) which could potentially also satisfy this description. But in some ways it is easier to account for the "darkness" of dark matter with WIMPs than it is with strange configurations of baryons.

Most of the candidates from the WIMP particle "zoo" are not known to exist, although usually they are predicted from not-implausible theories of fundamental particles. Ideally, it is often supposed, such candidates should be observed in accelerator experiments before they are considered truly viable dark matter candidates; particle physics is not so well established that we can automatically accept its untested

predictions.[10] One important thing to note is that most of the WIMPs are invoked in order to solve, and are better suited to solving, the *cosmological* rather than the dynamical dark matter problem. Neutrinos in particular were at one time a very attractive candidate for the cosmic dark matter because they were known to be produced in the early universe in very great numbers: if there were enough of them, and if their individual masses were high enough, the total could provide the extra mass necessary to "close" the universe. (The same can be said for axions: see below.) Electron neutrinos eventually fell out of favour for this purpose once it was shown that their rest mass is too low to do the job; although the muon and tau neutrinos are thought to be (relatively) much more massive, they were not visible to the previous generation of detectors. The newly operational Sudbury Neutrino Observatory is able to detect all three types; other detectors have also recently come on-line, including a detector at the South Pole. Recent studies indicate, however, that the universe does not have a closure density of matter after all; if this result is correct, the proposed neutrino solution obviously becomes otiose. Note that if the universe *is* at the critical density, then the majority of the dark matter must be exotic, because of the nucleosynthesis limits on the number of baryons (see above). But regardless of the impact of WIMPs on the question of the overall mass density of the universe, *if* such particles exist, they *will* contribute to the dynamical masses of galaxies and clusters. What I aim to do in the remainder of this section, then, is to evaluate the WIMP candidate solutions of (or potential contributions to the solution of) the dynamical dark matter problem.

The mass of the neutrino is not settled but because of "phase space constraints", not enough of them can fit into dwarf spheroidal galaxies to solve the dynamical discrepancy in those systems (Trimble 1987, 440). (Phase space constraints are quantum mechanical limits, related to the exclusion principle, on how tightly one can pack neutrinos together.) Nevertheless, since neutrinos are known to exist and are expected to have been created in large numbers in the Big Bang, neutrinos *are* part of the dark matter.

---

[10] However, some WIMPs are thought *not* to be producible in presently available or even projected particle accelerators, nor to be detectable by any passive detector we could hope to build.

we just do not yet know what part. Determining the neutrino fraction of the dynamical (and cosmological) dark matter depends on settling the neutrino mass and abundance.

A strike against neutrinos is that computer simulations of large scale structure formation in neutrino-dominated universes turn out to yield universes in which the voids and filaments form "too efficiently" (Bartusiak 1993, 307). Particles like neutrinos that move at relativistic speeds are known as "hot dark matter" (**HDM**). HDM cosmological models form large scale structure first (the HDM "washes out" small scale mass density fluctuations in the early universe). In contrast, in cold dark matter simulations (**CDM**: particles that move at slower speeds), structure forms "bottom-up" instead of "top-down". This enables structures as small as dwarf spheroidal galaxies to form with dark matter halos bound to them, as the dynamical evidence discussed in Chapter 4 indicates must be the case. CDM models in which $\Omega = 1$ give the best fit to the observed structure in the universe, though according to some commentators these models are just barely adequate even using this exaggerated value for the cosmic mass density (see Bartusiak 1993, 317).

The nucleosynthesis predictions discussed above also involve a theory of primordial neutrino production: the early universe will produce a certain number of neutrinos (depending on which laws of particle physics actually held in the early universe), and the presence of those neutrinos affects the rate of production of light elements. Thus, the success of the nucleosynthesis predictions, which involves establishing the baryon ratio, also establishes the ratio of neutrinos to photons. (Weinberg 1993, 183) Recent work has shown that the neutrino does indeed have some (small) mass: the Super Kamiokande experiment in Japan[11] has shown that neutrinos oscillate between three types (electron, muon, tau) as they travel through space, and this is only possible if they have some mass. Though the presently operating neutrino experiments will eventually settle the question of the mass of the three types of neutrino, accelerator experiments have already set an upper bound for the rest mass of the electron neutrino, and the "standard model" of particle theory predicts that the muon and especially tau neutrinos will be much more massive. The neutrino-synthesis prediction

---

[11] See <http://www.phys.washington.edu/~superk/sk_release.html>.

and the measured upper bound on the rest mass of the electron neutrino together set bounds on the mass fraction of the universe that could possibly be contained in neutrinos.

The evidence suggests that *if* the universe is closed, there is cosmological dark matter, because there are not enough baryons to close the universe (by an order of magnitude of mass at least; see Chapter 1 and the Appendix). One popular account of the source of this cosmological dark matter is primordial massive neutrinos. Successive accelerator experiments have driven the maximum possible mass of the neutrino farther and farther down without being able to say positively what the neutrino mass is, although the Super Kamiokande experiment shows that neutrinos definitely do have mass. The extremely low upper limit for the neutrino mass now accepted means that even though there are as many primordial neutrinos as photons, the mass fraction of neutrinos is insufficient to close the universe.

> Laboratory limits from tritium beta decay rule out. . .an electron neutrino more massive than 4.4 eV. Present cosmological bounds on the masses of other neutrino species are stricter than those from laboratory experiments; a 45 eV neutrino would lead to $\Omega = 1$, so for a universe at less than critical density the neutrinos must all be lighter than this. The exception to this is if a neutrino is so massive ($\geq$ 1 MeV) that it was non-relativistic during freeze-out, i.e. Cold Dark Matter. (Gawiser 2000, 1)

The paper just quoted derives an upper limit on the neutrino mass from cosmological simulations (using up-to-date parameters, for example the matter fraction is assumed to be just 40% of the critical density, baryons making up ten percent of that, the Hubble constant is set to 65 $kms^{-1}Mpc^{-1}$, and the basic model is CDM plus a cosmological constant. The result is an upper bound on the neutrino mass of less than 4 eV. This seems to rule out neutrinos as a significant part of the cosmological dark matter.

Despite the impossibility of getting $\Omega = 1$ with neutrinos, Weinberg notes, "the [cosmological] missing mass is also possibly contained in particles that are much heavier but also much less abundant" (1993, 185). According to some theories, particles of every possible mass would have been created in the early universe. If some of these did not annihilate with their antiparticles and are stable to decay over long enough periods of time, they would still be present today. If their numbers and individual masses are high enough, they could be, or contribute to, the mass needed in order to close the universe.

Another candidate for the cosmic dark matter is the axion, a fundamental particle that would have been produced in extremely large numbers (many times the number of photons) in the early universe—if it exists at all. Its mass is predicted to be very small (a few hundred micro-eV; Blout, *et al*, 2000, 1), but there are (or could be) so many of them that their total provides the excess mass required for critical density. They are also a popular Cold Dark Matter candidate in structure formation scenarios, where they "could dominate [the] potential wells of most astrophysical systems" (Blout, *et al*., 2000, 1). These and other attempts to use particle physics to solve the cosmological dark matter problem are beyond the scope of this dissertation, but the candidates are interesting here insofar as if they exist, they could potentially clump together in galaxies and clusters in a way that accounts for the dynamical discrepancy.

Axions were the next most popular WIMP dark matter candidate after neutrinos fell out of favour. The axion was originally introduced by Wilczek and Weinberg in 1978 to account for a discrepancy in certain strong force interactions (for a non-technical account of this background, which does not bear on the question of whether axions are plausible dark matter candidates, see Bartusiak 1993, 277-79). The axion as originally described (a particle about five times less massive than an electron) was soon ruled out by its failure to be detected in dedicated accelerator searches. But the later development of Grand Unified Theories (**GUTs**) allowed the axion to be retained as the solution to its original problems: putting the axion in this framework resulted in its mass being radically revised downwards, to billions of times less than an electron. In the GUT scheme these extremely low mass axions are produced in extremely large quantities when the GUT symmetry is broken as the early universe cools. Wilczek calculated that axions would be present in very large numbers in the universe today (a billion per cubic inch!). With its low mass and extremely weak interactions with other matter, the axion is a perfect candidate for the cosmological dark matter.

But do axions actually exist? If axions pass through an extreme magnetic field, it is predicted that they will decay, emitting microwaves (at a frequency dependent on the axion's rest mass). So far, detectors built to search for this effect have had no positive results, but this may simply be due to the fact that the detectors have to be tuned to exactly the right frequency (one out of a million possible channels), and the axion's mass

is not known well enough to tell us what range of frequencies to search first. (Bartusiak 1993, 337-38) Another constraint on the axion is that its mass *cannot* be much more than a billionth of the mass of an electron: if axions were heavier, their presence would cool the cores of stars, one effect of which would be a lower flux of neutrinos from supernovae than was observed in SN1987a (Bartusiak 1993, 338); thus observed supernova neutrino fluxes provide information about the possible mass of the axion, and this applies whether or not the axion is the dark matter.

One very interesting attempt to constrain the possible particles and masses of the dark matter comes from analysis of the "extra-galactic background light" (**EBL**). The idea here is to search for a "background" of electromagnetic emissions corresponding to the decay signature of candidate particles. The decay signature of a fundamental particle depends on many factors (including rest mass, the characteristic of primary importance for the viability of a given WIMP as a dark matter candidate), but the important thing is to calculate the emission spectrum and the decay rate for a specific model of the particle. Then, one can put limits on the total mass contributed by those particles by observing the flux of energy at those wavelengths. Trimble (1993, 155) mentions that WIMPs with lifetimes of $10^{8\text{-}9}$ or 10^{23} years would have distinctive decay products including observable photons and "inos" (photinos, neutralinos, gravitinos, . . .). Note that the fact that dark matter is important in structure formation and the dynamics of galaxies provides a constraint on the decay rate of the dark matter particles: it must be long-lived enough that it is still the most significant mass fraction ~15 billion years after the Big Bang.

Overduin and Wessen (1997), Sciama (1998) and Overduin, *et al.*, (1999) discuss the impact of observations of the ultraviolet background on the possibility of the existence of a cosmic dark matter neutrino and other particles, and derive some limits. Unfortunately a very recent (Edelstein, *et al.*, 2000) re-analysis of the ultraviolet background observations used to construct limits on dark matter shows that the original data reduction introduced systematic errors. When the data is properly analysed, it seems that the new upper bound on the EBL flux will no longer support some of the conclusions drawn from the original analysis. As Edelstein, *et al.*, remark in their conclusion, the Overduin and Wessen results "may now be compromised".

In principle it is possible to impose constraints of a similar type on the neutralino and other such particles from the flux of gamma radiation and electrons arising from the annihilation of such particles in the region of the Sun, but predictions and observations are presently such as to leave this indeterminate. Likewise, some WIMPs would mutually annihilate in the halo of the Milky Way, so calculations of this sort in conjunction with observations could be used to put limits on the quantity of a given particle *within* galaxies. But the practical difficulties of distinguishing the resulting gamma rays, antiprotons and positrons from the normal galactic background makes the task nearly impossible (Bartusiak 1993, 340). This is why the studies mentioned above focus on *extra-galactic* radiation backgrounds; even so, efforts have to be made to account for the effect of intergalactic dust. There are various other uncertainties in the method as well. For example, the calculated decay rates and emission spectra for particles that have never been observed are necessarily speculative; this problem can be alleviated in part by checking for the emissions that would be produced by a range of values for each kind of particle, but this will clearly still be model dependent. This method also relies on thinking up the right sort of particle in order to conduct the test. Given the recent results mentioned above, it seems clear that the scientific side of things is not at a stage where any conclusive statements can be made, but nevertheless the idea of using the EBL (at all wavelengths, not just ultraviolet) as evidence for or against various dark matter candidates is potentially powerful, especially if we want to evaluate candidates that, when they are not decaying, have no electromagnetic signature.

Let me conclude this section with some remarks on the viability of CDM as a contribution to the dynamical dark matter. Gribbin and Rees (1989, 147) note that CDM could possibly explain the solar neutrino discrepancy, as well as making the universe flat and accounting for galactic evolution and dynamics. If this were correct, one would expect that the ability of CDM to give a unified solution to such a diverse set of phenomena would give CDM a large confirmational advantage over competing dark matter candidates which cannot supply the same or similar levels of unification. However, recent evidence suggests that the CDM is not the correct explanation of the solar neutrino discrepancy. Interestingly, this fact may itself disconfirm CDM models of the missing mass. (I have not seen this argument given in the literature, but I suppose it

must be known.) Let me explain by beginning with the solar neutrino discrepancy. The radiation flux from the Sun together with the laws governing the fusion of hydrogen into helium yield an expected flux of solar (electron-) neutrinos. Given the interaction cross-section for neutrinos, an expected rate of detection in various types of Earth-bound detectors can be computed (since the interaction cross-section depends partly on the mass of the neutrinos, a range of expected detection rates must be computed). But only about one-third of the expected number of neutrino events were recorded in the previous generation of neutrino detectors. This is either evidence that our knowledge of fusion reactions is very wrong (which seems unlikely given the huge quantity of confirming data we have from reactors and accelerators), or that some unknown process is affecting the production or emission of neutrinos in the Sun, their transmission, or their reception in Earth-bound detectors.

Now, if the CDM model of galaxy formation is correct, the dark matter halo of galaxies that is detected by dynamical tests is composed of particles with mass comparable to the proton (that is, within several orders of magnitude):

> a star like the Sun should gather [CDM particles] up as it orbits around the Galaxy. Over the lifetime of the Sun, perhaps as much as one-trillionth of its mass ...could have built up in the form of WIMPs trapped in its core by gravity. ... The effect...would be to lower the temperature at its centre, because ...the WIMPs would spread the warmth at the heart of the Sun out over a broader region. (Gribbin and Rees 1989, 147)

This lower core temperature would mean that the Sun and other stars would have lower fusion rates, and therefore a lower production of neutrinos. This is an interesting speculation, but recent evidence from the Super-Kamiokande detector (see above, especially note 11) suggests this is not the correct explanation of the solar neutrino discrepancy. The accepted solution at present is that neutrinos oscillate between the three types (electron, muon, and tau) as they travel through space: this explains the discrepancy between the predicted and observed neutrino detection rates because only the electron neutrino was detectable by early neutrino observatories. As mentioned above, the Japanese Super-Kamiokande detector has recently obtained evidence that neutrinos (in this case, produced in a nuclear reactor) truly do oscillate. However, the oscillation solution to the solar neutrino discrepancy will not be definitive for a few years, that is,

until the Sudbury Neutrino Observatory (**SNO**) and other new projects that can detect all three types of neutrino have an adequate stock of data.

If the SNO results do confirm the oscillation solution (it is widely accepted already), then obviously the cooling of the solar core by capture of cosmic WIMPs is otiose as an explanation of the solar neutrino discrepancy. But since the Sun *must* have captured halo WIMPs if they exist, an observed *lack* of cooling (indicated by the corresponding lack of decrease in neutrino production) would be evidence to suggest that the halo is in fact *not* composed of CDM. At the very least, such a result would provide constraints on the particle mass and space density of CDM particles, and those constraints might turn out to be inconsistent with other observed features of the dark matter. Furthermore, the possibility remains open that neutrinos of all three types will be detected, but that the *total* observed solar neutrino flux will still be less than the prediction: then the margin of difference here would provide bounds on the total mass of WIMPs that could possibly have been captured by the Sun, and would thereby constrain the mass and space density of halo particles. We will have to wait until the data comes in before we can decide amongst these alternatives and divine their implications for the dark matter, but what this shows is that the present consensus about the solution to the solar neutrino problem is *prima facie* incompatible with some versions of the CDM model of galaxy formation. The CDM particles thereby ruled out would consequently no longer be candidates for the dynamical dark matter either.

As a final note, it should be remarked that *mixed* "C + HDM" models have lately been favoured in simulations of the evolution of cosmic structure. A recent paper mentioned above (Gawiser 2000) concludes that adding a neutrino hot dark matter component to a CDM plus cosmological constant model does *not* improve the fit of simulations to observed large scale structure. In any case, since the "mixed" scenarios merely conjoin candidates considered in either the CDM or HDM scenarios, they have no new consequences for the dynamical dark matter.

5.3 MONOPOLES, SUPER-STRINGS AND TOPOLOGICAL DEFECTS

Topological defects are predicted to be produced in the "symmetry breaking" epochs of the early universe (when each of the four forces "freezes out"), in the transition

from GUTs to the laws of physics in effect now. The details of the formation of topological defects in GUT symmetry breaking are beyond the scope of this work. It suffices to note that scientists use the metaphor of defects forming in the phase transition from liquid water to ice—point, line and plane defects appear in the ice as it forms. During the phase transition, the system goes from a state of higher symmetry to a state of lower symmetry. Topological defects in space are said to form in the symmetry-breaking epochs of the early universe "in just the same way." (For a slightly more informative description see Guth 1997, 136ff.) What is important for present purposes is that point, line and plane defects *can* form, and that *if* they do they will be extremely massive (because they are relics from earlier stages of the universe when the energy density, and therefore the effective mass of a given volume of space, was very much greater). Their contribution to the mass density of the universe could be important in providing the excess mass to close the universe and thus to solve the cosmological dark matter problem; some kinds of defects could also be useful in seeding large scale structure formation, and thus in helping to solve the problem of how the structure observed in the present universe could form, in the limited time available, from the almost perfectly homogeneous state revealed by the cosmic background radiation.

One kind of point defect is a magnetic monopole. According to some GUTs a huge number of monopoles (equal to the number of baryons) will be produced in the early universe, and each will be very massive. Monopoles do not normally emit electromagnetic radiation (unless they are travelling through electromagnetic fields), so they make a decent dark matter candidate. "Since there should be almost as many monopoles as baryons, and their unitary mass is 10^{16} GeV, i.e., 10^{16} times greater than the mass of the proton, the mass density of monopoles should be 10^{14} times as large as the baryon density!" (Earman and Mosterin, 1999, 15). However, besides the fact that such a quantity of monopoles would have caused the universe to recollapse long ago, observing programs have failed to detect monopoles. This casts doubt on the GUTs that predict them, and especially on the idea that the monopoles could contribute to the solution of either dark matter problem. This detection failure led to the so-called "monopole problem" that Alan Guth was thinking about when he came up with the original model of inflationary cosmology in 1979 (see Guth 1997, for an account of the

history, and Earman and Mosterin 1999, for a scathing critique of the evidential status of inflation, especially pp. 14-7 on monopoles). If monopoles are produced in the pre-inflationary universe, they could possibly account for the cosmological dark matter, or even for the dynamical dark matter if they swarm in haloes around galaxies and clusters. But if monopoles could solve either of these problems, their density in the local Milky Way would be high enough for us to be sure of making many detections of them. The fact that no detections occur is therefore definitive evidence that monopoles are not the dark matter (whether they actually exist or not). I should note that Earman and Mosterin (1999) point out (following Penrose) that inflation may be a solution to a self-made problem:

> [S]uppose that (a) independently of inflation, we had good reason to believe that GUTs are correct, and (b) there is no reasonable way to resolve the monopole problem within GUTs without invoking inflation. Then we would have good reason to believe in inflation. However, at the present time neither of these suppositions seems correct. As for (b), Langacher and Pi (1980) and others have offered resolutions of the monopole problem within GUTs without invoking inflation. As for (a), there are some theoretical and aesthetic reasons for believing GUTs, but. . .there is hardly sufficient evidence to take them as more than a serious possibility and there is some evidence to indicate that they are badly wrong. (Earman and Mosterin 1999, 17)

It seems, then, that with or without inflation there is at present little theoretical reason to suppose monopoles exist, and no empirical reason to do so. Monopoles have therefore fallen out of favour as dark matter candidates.

Line defects formed in the GUT symmetry breaking are usually called cosmic strings (these are to be distinguished from superstrings, which are invoked by supersymmetry, a theory created to explain fundamental particles in terms of something even more fundamental; *cosmic* strings are just topological defects of spacetime). Trimble (1993, 155) notes that monopoles are the only type of topological defect likely to be able to contribute to the dark matter, while the others may be useful for other purposes such as seeding galaxy formation. This is because line and plane defects could not form the halo structures we need in order be able to account for galaxy and cluster dynamics. Also, strings and other defects are likely to dissipate by gravitational radiation (they are very massive and vibrate very quickly), so that they would have radiated away to nothing by this stage in the evolution of the universe; they could still have seeded galaxy

formation before they dissipated, by providing a centre of mass around which other matter would gather. (If this is what happened, we do not need there *now* to be dark matter in excess of what is present in galaxies and clusters in order to explain large scale structure formation.) Cosmic strings would produce a distinctive kind of gravitational lensing (where two equally bright images are produced; in ordinary cases of gravitational lensing, an odd number of images is produced and all have different brightnesses), so they are in principle detectable. There is at present no observational evidence to suggest that cosmic strings exist, but this need not imply the disconfirmation of the theories predicting them provided that they do in fact dissipate by gravitational radiation.

5.4 CONCLUSIONS ABOUT DARK MATTER

The available evidence gives us limits on how much of the total dark matter can be contributed by each type of candidate. What we know as a result is that we do not know what most of the dark matter is. The main thing we do know is that, "The fact that [the dark matter] is less centrally condensed than the luminous matter in galaxies (and probably in clusters) suggests that it does not readily dissipate energy the way ordinary gas does when collisions excite atoms and molecules, which then radiate away the energy" (Trimble 1993, 153). Of course, the fact that we detect no electromagnetic signature from dark matter, which we know must exist in haloes, already tells us that something like this must be the case.

There are many fundamental particles predicted by our best theories, some of them known to exist and many of them massive, and there are more than 90 naturally occurring elements (and several more that can be artificially created, though some of these persist only for brief periods), and countless kinds of molecules. Each of these substances has very different electromagnetic, spectroscopic and other properties, properties that determine how we are able detect the presence of the molecule or element and distinguish it from others. Given this plethora of kinds of "ordinary" matter (most of it massive), there is perhaps little reason (besides our ignorance of what dark matter is, or some principle of parsimony designed to guide our thinking under conditions of evidential poverty) to assume that dark matter, which makes up more than ninety percent of the mass of dynamical systems, comes in only one type. We should therefore be

willing to accept that more than one kind of dark matter is present, that some "mixed model" is the right answer. The difficulty then is in determining the individual contributions of the various dark matter components to be included in the mix, and in finding some way to justify the scheme. Since we do not yet have even *one* dark matter model (mixed or not) that is consistent with or can explain the known evidence, we do not yet need to worry about what principles of theory choice to employ in deciding between equally empirically adequate mixed models of dark matter. We might, therefore, be a long way from finding a solution to the large and obvious dynamical discrepancies at various scales of structure.

The interesting thing, given the limited evidence available and the lack of any completely adequate theoretical description of dark matter, is how much we *can* know about the dark matter (its quantity, distribution, and other properties). Perhaps the most important constraint discussed above is the limit on the fraction of the total mass that can be baryonic. Eventually determining this proportion precisely and accurately is of the utmost importance, because this in turn determines the fraction of the total mass that must be of an exotic type.

According to some writers, baryonic dark matter candidates have a decided advantage over other candidates in that baryons are known to exist, whereas many of the other candidates have not been detected in the laboratory (for some of them, we have reason to suspect that we could *never* detect them in the lab). Two philosophical questions arise with regard to such claims. (1) Does a candidate's being known to exist count as evidence for that candidate being the correct solution? (To put the question in probabilistic terms, does being known to exist increase the likelihood of the candidate's being the solution?) (2) Does being (directly) "observable" confer an evidential advantage? (This is another way some writers have construed the supposed evidential advantage of baryons over other candidates.) It suffices here to note that the observability of baryons may in fact be evidence *against* them, as discussed above: if the matter responsible for the dynamical discrepancy in galaxies and clusters *were* baryonic, we *ought* to be able to detect its presence. The only way to make sense of the idea that baryonic candidates have some epistemic advantage over other kinds of candidates is to note that some of the non-baryonic candidates are derived from more or less speculative

theories; the fact that the particles in question have not been observed just means that the probability of the theories from which they are predicted remains rather low. However, this apparent advantage quickly dissipates when the details of baryonic solutions are investigated: it seems very unlikely that it would be possible to sequester enough baryons to account for the dynamical discrepancies in a way that would also render them invisible to the investigative techniques now being brought to bear. In light of this, at present I see no evidential reason to prefer *any* matter candidate or class of candidates over any of the viable rivals.

The *totality* of what we know about stars, galaxies and clusters is evidence that any dark matter solution must explain, or at least be consistent with. In particular, in order to be viable a dark matter candidate must account for the dynamical discrepancy (the excess speed of rotation and the unexpected shape of the rotation curve, the increasing proportion of dark to light matter as scale increases, and so on), at the same time that it is consistent with our best theories of stellar evolution, galactic evolution, and with the fact that it cannot be directly detected. Candidates that meet this condition, at least on first glance, come in several kinds. Within each kind the evidence effectively RfP-measures (some) parameters of the solution (within better and worse margins of error depending on the case). Choosing amongst these candidates is then a judgement that amounts to preference for one set of background assumptions over another. At present, then, we do not have grounds to justify the choice of any dark matter solution that has been offered over its rivals. Note, however, that the evidence described in this chapter and the previous one has already been used to eliminate some proffered candidates. One can only hope that continued investigation will further narrow the class of viable candidates, and constrain ever farther the dynamically important characteristics of the dark matter, even if the evidence is never sufficient for a definitive choice of one particular solution.

CHAPTER 6

ALTERNATIVE THEORIES OF GRAVITATION AS SOLUTIONS TO THE ASTROPHYSICAL DYNAMICAL DISCREPANCY: A PROBLEM IN THEORY CHOICE

> *Of all the laws of physics, the one best verified by its innumerable consequences is surely the law of universal gravity; the most precise observations on the movements of the stars have not been able up to now to show it to be faulty. Is it, for all that, a definitive law? It is not, but a provisional law which has to be modified and completed unceasingly to make it accord with experience.*
>
> —*Pierre Duhem (A passage accidentally omitted from the English edition of* The Aim and Structure of Physical Theory, *first edition 1906, translated from the French by Gillies (1998, 308).)*

6.0 INTRODUCTION

As the quotation at the head of this chapter shows, Pierre Duhem, writing even before the advent of the Special let alone the General Theory of Relativity, was cognisant of the fact that the then-accepted Newtonian theory of gravity had not been definitively established despite its detailed agreement with all the dynamical evidence then available (Mercury aside). The quotation shows Duhem's awareness that the validity of the law of gravity depends on its success in saving the detailed motions of distant stars.[1] Duhem

---

[1] At the time Duhem wrote this, it had still not been settled that there were other galaxies external to the Milky Way, and the available observations of the motions of stars were still too primitive to allow the detection of the dynamical discrepancies that came to light in the studies of Babcock, Smith and Oort just a

here reminds us that no hypothesis is ever beyond test, revision or rejection, no matter how strong its apparent evidential support might be.

In the same way that Newton's theory was overturned in light of new theories and new evidence, it is possible that General Relativity (**GR**), though well-supported by its available evidence, will have to be revised. In fact, as I will describe in this chapter, GR's present evidential situation with regard to distant dynamical systems is not very different from the evidential situation of Newton's Universal Gravitation (**UG**) at the time Duhem wrote. Just as Duhem had no good tests of UG for distant stars, we have no informative tests of GR at galactic and greater scales. This is significant because the dynamical discrepancies at these scales possibly indicate the need to develop a new theory of gravity, and without good tests for these very large and distant systems the problem of choosing between the various rival gravitation theories is more than difficult.

Some writers have objected to the fact that matter solutions require introducing so much exotic matter when there are no independent empirical grounds for asserting its existence. Despite such complaints, only a few attempts have been made to revise or replace the theory of gravitation in order to account for the dynamics of large-scale astrophysical systems without the need for dark matter. (Other alternatives to GR have of course been proposed for other reasons: see Will 1993.) But, despite the paucity of articulated alternatives, the dispute between General Relativity and its rivals with regard to explaining the dynamical discrepancies is a classic illustration of the problem of theory choice. This chapter will explore the various factors involved in trying to choose among possible theories of gravitation that could be included as parts of solutions to these discrepancies. Since a minimum criterion for a theory to be an alternative in a problem of theory choice is that it be empirically adequate, I spend a fair bit of time discussing the basic evidence for and against the competing theories. One important philosophical conclusion that comes to light in this discussion is that the choice between dynamical theories will depend crucially on *non-dynamical* evidence, and on higher-order evidence of the sort mentioned in Chapter 2.

---

few decades later. In any case it seems nearly certain that when he spoke of "the movements of the stars" Duhem had in mind the orbits of binary stars around one another.

In section 6.1, I discuss two alternative gravitational theories that have been proposed as solutions to the dynamical discrepancies which avoid asserting the existence of dark matter. Although it seems that neither of these theories is in the end viable, the general lessons about what any successor to GR must achieve evidentially, stand. The discussion in 6.1 also sets up the more philosophical discussion in 6.2 of the problem of theory choice in light of the underdetermination of theory by evidence. I argue, by developing an idea of Duhem's, that it is possible to make reasonable, evidentially based theory choices despite the ambiguity of falsification, and despite the deductive circularity of confirmation pointed out by Hume. Besides providing a philosophical context for the discussion of the specific case of theories of competing gravitation offered as potential solutions to the dynamical discrepancy, my account here shows that an evidential choice between the rivals *is* possible in general, provided that evidence of the right sort becomes available. In section 6.3, I discuss the problem of curve-fitting as a model of the Humean underdetermination problem. This is appropriate because at least one of the rival gravitational solutions (Milgrom's; see below) was originally constructed as a curve-fitting problem; further, it allows me to introduce the idea that the simplicity of theories can be a factor in theory choices, and to briefly discuss the impact of observational error on the problem of theory choice. In section 6.4, I discuss three methodological critiques of GR and dark matter theories proposed by Mannheim, author of a rival theory of gravity. Mannheim argues that his theory is superior because its (more popular) rival suffers from being less simple, *ad hoc* and unfalsifiable, and less unified. I shall argue that his methodological critiques turn out to be unfounded, and that there is therefore no reason (in the present evidential situation) to reject GR as applied to galaxies or larger systems, and no reason to reject the idea of a dark matter solution to the discrepancies. In fact, in section 6.5, I argue that there is a methodological argument originating with Newton's Rules of Reasoning which would support the provisional acceptance of GR at galactic and greater scales, despite the fact (as I argue throughout the chapter) that there is (and perhaps can be) no dynamical evidence at those scales which could support GR over any rival theory that is empirically equivalent on the stellar-system scale tests.

Although everyone must admit, as a matter of logic, that gravitational solutions to the dynamical discrepancy are *possible*, many physicists seem to take such solutions to be

implausible—or even illegitimate—*because of the supposed evidential support for GR.* One of the things I am concerned to make clear here is that it is *not* illegitimate or unreasonable to suppose that alternative theories of gravity are plausible candidate solutions for the dynamical discrepancy. For one thing, the evidence for thinking that GR is the correct theory of gravity at *all* astronomical distance scales is extremely weak. (This claim will be substantiated in detail below.) It requires a complex philosophical or methodological argument (as opposed to a direct appeal to observational evidence) in order to find any (good) reason to prefer GR over its rivals at galactic and greater scales in the present evidential context. (This argument is sketched near the end of this chapter.) For another thing, there is a long and respectable history of proposing revisions to the dominant dynamical theory in order to try to overcome recalcitrant discrepancies between predictions and observations. In fact, the history of suggestions that the law of gravitation could be "non-Newtonian" goes back as far as Newton himself, insofar as Newton recognised that alternatives to the phenomena measure alternative values of his theoretical parameters: Newton could not unify the lunar precession under the inverse square framework, and he saw that one in-principle possible solution to the discrepancy was to have the power law for the Moon be slightly different than inverse square. Although Newton did not actually develop this suggestion in detail, Clairaut did.[2] Similarly, other persistent discrepancies in celestial mechanics inspired other attempts to reconcile theory with observation by altering the theory.[3] These examples show that

---

[2] The lunar motions continued to vex the best minds in celestial mechanics until Clairaut, after first proposing a change in the force law, finally showed that the lunar motions are in fact consistent with the inverse square law of Newtonian gravitation. By 1787 Laplace had constructed a theory whose predictions were good to within half a minute of arc. Toward the end of the nineteenth century George William Hill discovered a method that considerably simplified the calculation of orbits in a special case of the three-body problem, of which the Sun-Earth-Moon system is an instance. This was the foundation of the Hill-Brown lunar theory developed by Ernest W. Brown between 1897 and 1908, which is still used (with appropriate relativistic corrections) to calculate lunar ephemeredes. See Peterson (1993) and Smart (1953, Chapter 18).

[3] As mentioned in Chapter 2, for a time G.B. Airy, an Astronomer Royal, advocated a gravitational solution to the Uranus discrepancy. Simon Newcomb in 1895 showed that the motion of Mercury's perihelion measured the power law of a Newtonian force of gravity to have $n = -2.000000157$, although de

proposing *ad hoc* modifications to dynamical theories in order to try to account for discrepant observations is not unusual—and sometimes such proposals even turn out to be correct. More to the point, the merit of such proposals, like any empirical hypothesis, can only be evaluated by detailed observations and careful evidential reasoning.

With regard to the dark matter issue itself, van den Bergh (1961) and Finzi (1963) were among early proponents of the possibility that non-Newtonian forces could be responsible for the observed motions in clusters—van den Bergh considered but rejected the action of non-gravitational forces, whereas Finzi explicitly suggested that non-Newtonian gravitation could explain the cluster motions. (See Trimble 1990, 359.) Finzi points out, among other things, *that the solar system tests of GR confirm the theory only for distances of interaction eight orders of magnitude smaller than 1 kpc*. (Recall that a typical radius for a spiral galaxy is a few tens of kiloparsecs, and that the distances between galaxies in clusters are typically of the order of megaparsecs.)[4] Thus flat rotation curves "may indicate that the gravitational attraction of a galaxy decreases more slowly than $1/r^2$ at large r" (Finzi 1963, 22).

Sitter showed in 1913 that this "ugly" power law is inconsistent with the motion of the Moon's perigee (Earman and Janssen 1993, 133-4). (Recall, too, that the Hill-Brown theory had earlier shown the motion of the Moon to be consistent up to the precision available in the observations with Newton's inverse square law.) And, of course, Einstein's overthrow of Newtonian gravitation was ultimately required in order to give a satisfactory account of Mercury's motions. See Grosser 1979 [1962] on the history of the Uranus discrepancy and its resolution in the discovery of Neptune; see Roseveare 1982, and Earman and Janssen 1993, on the Mercury episode.

The power law describes how the gravitational attraction varies with distance; it is an interesting question why no one has proposed that the power law itself is dependent on distance or field-strength. We may also ask, under what conditions *could* we justify choosing a power law which varies with distance or field strength?

[4] Binary star systems have also provided very strong confirmation of GR (changes in the periods of binary pulsars exactly match the expectation derived from GR's predictions about the energy that should be radiated from such systems in the form of gravitational waves). But again, the interactions in question take place on scales *much* less than a single parsec. Thus this evidence is "short scale" or "stellar system" evidence; in this respect it does not matter than binary systems are distant from us.

It was not until about 60 years after the initial discovery of the dark matter problem that more or less fully-fledged proposals for gravitational solutions were finally developed.[5] Milgrom (1986) proposed a non-Newtonian gravitation theory as a solution to the dynamical discrepancies, a theory which he calls the "Modification of Newtonian Dynamics" (**MOND**).[6] This theory, which can be interpreted as describing either a modification to the Newtonian law of gravitation or to the Newtonian concept of inertia (Sanders 1999, L23), is proposed as the weak-field, low-velocity limit of whatever relativistic gravitation theory turns out to be the correct one. MOND differs from the Newtonian limit of GR in that it supposes that gravity is not accurately described by the Newtonian action when the gravitational field strength is below a certain empirically determined threshold. Thus MOND is an entirely non-relativistic theory. Philip Mannheim has however developed a *relativistic* gravitation theory which he claims fully explains galactic and cluster dynamics without the need for dark matter (see, for example, Mannheim and Kazanas 1989, Mannheim 1994, 1993, and 1992). I will here refer to Mannheim's theory as the "Conformal Theory of Gravity" (**CTG**) since his original motivation for developing it was to make gravitation conformally invariant. (He wanted to do this, in turn, in order to give gravity a theoretical description more like that of the other fundamental forces, so as to facilitate the project of unifying gravity with the other forces.) Both MOND and CTG will be considered in more detail below.

---

[5] This is probably due to two factors: one, discussed in previous chapters, is the fact that astronomers as a group did not take the dark matter problem seriously until the mid-1970s; the other is the bias, still present amongst physical scientists, in favour of GR. This bias seems to me to in fact have a sociological origin, although I will also later outline a methodological argument which could support a preference for GR.

[6] Milgrom on alternative theories of gravity: "Newton's law fails when objects approach the speed of light. For that we need Einstein's theory of relativity. What I am suggesting is that Newton's law must also be amended when the gravitational accelerations are very, very small, as they are in a galaxy's outer fringes" (as quoted in Bartusiak 1993, 213-14). In the context of galactic dynamics, Mannheim's suggestion about very small gravitational accelerations can be rephrased as being about the way the acceleration varies with large *distances*. Another way Milgrom could have put the point is this: Einstein's theory fails when we approach the realm of the very small. For that we need a theory of quantum gravity. What Milgrom is suggesting is that the low-velocity, weak-field limit of GR must also be amended when the distances between gravitating bodies are very great, as they are in the case of stars in the outer fringes of a galaxy.

The specific choice between GR and CTG, as an example of the generic problem of the choice between matter and gravity solutions to the dynamical discrepancies, both illuminates and is illuminated by a discussion of the general problem of theory choice. It may at first seem as if the choice between GR and CTG is simply a choice between gravitation theories, not between matter and gravity solutions to the dynamical discrepancy. But since opting for matter solutions amounts to accepting GR, the choice between matter and gravity solutions is really a choice between GR and other theories of gravity. Of course, no gravitation theory can make predictions, let alone be tested, without a theory of the distribution of matter in the systems in question. It turns out, then, that in order to test any dynamical law at gal.. tic or greater scales one needs to assume a matter distribution. The most convenient way to express this assumption is to grant that all theories assume the existence of the distribution of *visible* (baryonic) matter indicated by the light curves for galaxies and clusters, while GR needs to assume large amounts of dark matter and CTG assumes there is no dark matter.

6.1 GRAVITATIONAL SOLUTIONS TO THE DYNAMICAL DISCREPANCY

Imagine a quasi-Newtonian theory in which there is superimposed on the standard Newtonian action of gravity an additional component whose strength varies with distance so that it is too weak to be detectable in "short" distance interactions (say, up to the size of our solar system) but which grows to have a significant effect for interactions over greater distances (say, the size of a galaxy or cluster). (In other words, the quasi-Newtonian theory makes predictions which are observationally indistinguishable from those of the Newtonian theory for interactions taking place over distances smaller than a stellar system, but noticeably different for interactions taking place over larger distances.) If this additional component has the right form, it could account for the galactic rotation curves without the need for vast quantities of dark matter, while at the same time being observationally indistinguishable from the solar system observations (and therefore from the predictions of GR for the solar system). It is easy to see that one could construct such a theory *whatever* the galactic dynamical phenomena turned out to be: just start with Newton's theory, and add in an effect which is undetectable until one gets to galactic scales, and which has the form required to save the galactic phenomena to an adequate

degree of precision. If we had such a theory, would the fact that it saves the dynamical phenomena be reason enough for us to accept the theory?

In discussions of theory choice it is sometimes pointed out that one could have a "morce – gorce" gravitation theory, where *morce – gorce = force* (Glymour 1980, 356-7), that is, where the components morce and gorce combine in such a way as to exactly reproduce the predictions of the accepted dynamical theory in every respect and to the same degree of precision. Put in these terms, what MOND and CTG effectively present us with are theories according to which the dynamical law (or rather, the non-relativistic limit of the correct dynamical law) is a Newtonian "morce" plus a non-Newtonian "gorce", where this extra non-Newtonian component is vanishingly small until one gets to scales larger than stellar systems, so that one does not notice a difference between the Newtonian predictions and MOND or CTG until one gets to galactic scales.

Milgrom's original papers start from the idea that postulating large quantities of invisible matter is probably undesirable in itself, and from the recognition that an alternative to doing so is to modify the dynamical law using which we try to account for the observed motions of large scale systems like galaxies. There are two ways to interpret Milgrom's theory, either as a modification to the dynamical law, or as a revision of the concept of inertia. The second possible interpretation drops out of even Milgrom's later discussions, so I will ignore it here. (In any case, the two interpretations have the same observable consequences.) Milgrom constructs his new force law by doing curve-fitting on the visible masses and observed rotation curves of a sample of well-studied spiral galaxies. In other words, instead of inferring the mass from the observed rotation on the assumption that the Newtonian limit applies, he infers the dynamical law on the assumption that the visible mass is all the mass that is present. The result of this is as follows. "[T]he central idea is that the law of gravity or inertia assumes a specific nonstandard form below a fixed, universal value of the acceleration, a_0, the one parameter of the theory" (Sanders 1996, 117). Then,

> [T]he basic MOND relation between the acceleration g and the Newtonian acceleration g_N [is]: $\mu(g / a_0)g = g_N$, with $\mu(x)$ being the interpolating function of MOND. . . . MOND may be viewed as either a modification of gravity or as a modification of inertia. "Modified" gravity is described by the generalized Poisson equation discussed in Berkenstein and Milgrom (1984),

> which is of the form $\nabla\cdot[\mu(|\nabla\varphi|/a_0)\nabla\varphi] = 4\pi G\rho$, where φ is the (MOND) potential produced by the mass distribution ρ [and a_0 is the acceleration constant of MOND]. For systems with [spherical symmetry, this equation] is exact in this theory. It was also shown to be a good approximation for the acceleration in the midplane of disk galaxies. (Brada and Milgrom 1999, L17-8, and references therein.)

The equation $\nabla\cdot[\mu(|\nabla\varphi|/a_0)\nabla\varphi] = 4\pi G\rho$ here is the "field equation for the nonrelativistic gravitational potential produced by a density distribution $\rho(r)$. . . . [where] $\mu(x)$ satisfies $\mu(x << 1) \approx x$, $\mu(x >> 1) \approx 1$" (Milgrom 1986, 617). "The expression for the force between two masses $m_1 \geq m_2$ ([where] $m_1 + m_2 = M$) must be of the form $h(m_2/m_1)M^{3/2}R^{-1}(Ga_0)^{1/2}$, when the distance R between the two masses becomes very large" (Milgrom 1986, 618).

The empirically derived acceleration limit of MOND also offers possible dynamical explanations of several unexplained phenomenological laws. (For example, MOND offers potential explanations of "the Fish law, by which the distribution of the central surface brightnesses in elliptical galaxies is sharply cut off above a certain value," and "the Freeman law in its revised form, whereby the distribution of central surface brightnesses of galactic disks is cut off above a certain value" (Brada and Milgrom 1999, L18, and references therein).) Since MOND implies an upper limit on the maximum possible acceleration producible by dark halos, it is possible to compare MOND with numerical simulations of structure formation. The acceleration limit arises naturally in MOND, and Brada and Milgrom claim that since MOND itself is simply the result of fitting the dynamical law to galactic rotation curves (which themselves may be the result of a non-Newtonian dynamical law or of dark halos), the limit must hold independently of whether or not MOND is the correct quasi-Newtonian limit of the ultimate relativistic gravitation theory (1999, L17). They note, however, that for at least some well-known (and apparently successful) numerical simulations of structure formation, there seems *not* to be an upper limit on the maximum halo mass or acceleration contribution (Brada and Milgrom 1999, L18). Thus if the limit result truly is correct independently of the correctness of MOND, then these models of structure formation would be ruled out. It is not known, so far as I am aware, whether there are Newtonian models of structure formation which somehow manage to satisfy the acceleration limit on dark halos; if there

are none, and the acceleration limit is correct. Newtonian explanations of galactic dynamics would be ruled out. However, it is not clear to me that Brada and Milgrom have proved their contention that the acceleration limit is correct even if MOND is not. One thing that *is* certain is that if apparent galactic halos are found to cause accelerations exceeding the MOND limit, that fact would falsify MOND. Brada and Milgrom (1999, L17) report that for at least one sample of Newtonian best-fits of disk and halo matter distributions to the observed galactic rotation curves, the rotation curves do seem to imply a maximum halo acceleration below the MOND limit—they do not supply details of the calculation, so the claim is hard to verify. Note, however, that were it found that galaxies do have maximum halo accelerations consistent with the MOND result, this would not necessarily provide support for MOND, since there could be many reasons why galactic halos happen to have this maximum acceleration. For example, some fact about the formation and evolution of structure could end up in a *de facto* maximum halo size, without this being a "principled" limit of the kind MOND would require.

As Sanders (1999) notes, the dynamical discrepancy for clusters has in the last twenty years been significantly reduced by the discovery of visible matter not previously detectable, in particular by the X-ray detection of hot intra-cluster gas. Despite this new evidence, the mass discrepancy remains: although the hot intra-cluster gas may contain as much as 4 or 5 times the stellar contribution in rich clusters (Sanders 1999, L23), the visible matter still "fails by at least a factor of 3—more typically a factor of 10—to account for the Newtonian dynamical mass of clusters" (L23). Newly acquired data from the *Einstein* X-ray satellite observatory provides a sample of 207 clusters (Sanders 1999, L23, and references therein), of which 93 provide data reliable for Sanders' study (L24). Sanders shows that "the MOND dynamical mass within [some arbitrary radius] r_{out} is related to the Newtonian dynamical mass as $M_m = M_N [1 - (a_0/a)^2]^{1/2}$". Here, a is the acceleration resulting from the mass within r_{out}, and "a_0 is the MOND acceleration parameter found to be 0.8×10^{-8} cm s^{-2} from galaxy rotation curves" (L24).

Sanders then compares the fit of the dynamical mass as determined in both theories to the visible mass of the clusters as determined from a virial calculation based on the X-ray studies. The mean ratio in the sample for the Newtonian calculation is

"$<M_V \quad M_{obs}> = 4.4 = 1.6$" (L24); "MOND reduces the virial discrepancy in clusters: $<M_m \quad M_{obs}> = 2.1 = 1.2$" (L25). Using MOND to analyse clusters thus decreases the mass discrepancy by a factor of 2. Even so, MOND still seems to require around twice as much mass in clusters than is visible. On the face of things, this would seem to be inconsistent with the original motivation of MOND to explain astrophysical motions without the need for dark matter. Perhaps we should therefore embrace dark matter, or at least abandon MOND in favour of some other modification to dynamics which will successfully recover the dynamical evidence without the need for any dark matter.

But these results for clusters are in fact inconsistent with the motivation for MOND *only* if it can be shown that other solutions which would allow MOND to be retained are less probable than not. "Strictly speaking, the fact that MOND predicts more matter than is currently observed is not a falsification of the idea; more matter may be present and possibly observable" (Sanders 1999, L25)—this is just to say that the catalogue of visible matter may not yet be complete. "That the tally of ordinary baryonic matter may not yet be complete is suggested by several observations, in several clusters, of diffuse star light...and ultraviolet emission apparently from warm clouds"; cool gas is also not presently detectable, but must exist to some extent at least (Sanders 1999, L26). It is doubtful that any of these sources could entirely resolve the Newtonian discrepancy, but they could reduce the MOND discrepancy still further. "[I]f MOND were to require a component of nonbaryonic dark matter with significant cosmological density, then this certainly would be inconsistent with the spirit of the idea. There is nothing in this analysis that demands the presence of such a component" (Sanders 1999, L25).

Furthermore, "In many astrophysical contexts, a factor of 2 discrepancy can often be accommodated by reconsidering the effects of the several idealized assumptions that are not realized in every case" (Sanders 1999, L25). This is an interesting point, one which raises the question of how good the agreement between the two measures of mass has to be before we would consider there to be no discrepancy. The errors are indeed difficult to calculate and compare, but a factor of two uncertainty is definitely close to the best we could hope for given the difficulty of the observing situation and the available evidence. (Recall that until recently there was a factor of 2 uncertainty in the Hubble constant, and that this was the largest single source of error in dynamical mass

measurements: the error in the Hubble constant is now around 10%.) This said, the present discrepancy between the MOND dynamical mass and the visible mass of clusters could be reduced because of other factors: "It should be recalled that the MOND parameter a_0 is determined from the rotation curves of nearby galaxies. If the cluster distance scale differed from that of local galaxies (i.e., $H_0 = 75$ locally, but $H_0 = 50$ at the distance of clusters) then the remaining MOND discrepancy would disappear" (Sanders 1999, L25). Recent observations (mentioned in Chapter 1 and the Appendix) indicating that the Hubble expansion is accelerating lend some credibility to this idea, although the measured differences between the local and the distant Hubble constant are much smaller, and the distances much greater, than what Sanders mentions here.

So the fact that the MOND analysis of clusters results in a dynamical discrepancy of about a factor of two does not necessarily count against the theory. Clearly, more research is needed in order to settle the issue. However, there is another class of observations which may in fact rule out MOND. These are the observations of large scale gravitational lensing of background galaxies by foreground clusters. "Strong lensing observed in clusters typically requires a total projected mass in the inner 100-200kpc between 10^{13} and 10^{14} [solar masses], which is evidently not present in the form of hot gas or a normal stellar population" (Sanders 1999, L25). This is well above the mass density of cluster cores allowed by MOND: lensing *never* occurs when MOND applies since the critical density of lensing of this type is about 5 times greater than the maximum density at which MOND could apply (Sanders 1999, L25). It is also well above the visible mass density in cluster cores. In other words, significant dark matter *is* required in the cores of clusters in order to make the observed gravitational lensing possible, even if MOND is otherwise correct. To be forced to a modified law of gravitation and *still* have to introduce large amounts of dark matter seems the least desirable option. As we will see, gravitational lensing is also a significant stumbling block for Mannheim's CTG.

Since MOND is explicitly non-relativistic, it is not meant to be a serious contender as a *replacement* for GR. We use Newtonian dynamics to study galaxies and clusters because we think that they should satisfy the weak-field, low-velocity limit of General Relativity. MOND (if shown to be empirically adequate) could be the true weak-field,

low-velocity limit of some future relativistic gravitation theory. If that is the case, then Newtonian gravitation is the "short-scale" limit of MOND.

The fact that every alternative to GR that has so far been proposed has been shown to be evidentially inferior to GR (see Will 1993) does not mean that GR is the best of all possible relativistic gravitation theories, just the best so far described (as judged on the basis of the available evidence); the successor to GR could be lurking in the next issue of *Physical Review*. Independently of the astrophysical dynamical discrepancies, we have reason to expect that GR must be replaced, namely we have strong reason to think that GR is inadequate at the very short length scales at which quantum mechanical effects become important. The hoped-for theory of quantum gravity has not yet come on the scene; still farther off is a so-called Theory of Everything, which is supposed to unify all the forces of nature (electromagnetism, the weak and strong nuclear forces, and gravity) under a single theory. Nevertheless, we do have quite a bit of information about what these successor theories must be like, if they truly exist.

Any theory that will eventually replace General Relativity must satisfy a high standard of empirical success, across a very broad range of phenomena. In the same way that Einstein's theory had to include Newton's successes as weak-field and low-velocity limits of the relativistic equations of motion, any successor to GR must (minimally) be able to predict and explain all of the phenomena that GR is capable of predicting and explaining, and it must do so to at least the same degree of precision. General Relativity explains the motions in the solar system (including Mercury's perihelion shift), the gravitational redshift of light, the Shapiro time delay for signals passing near the Sun, the gravitational bending of light rays by massive bodies, and the formation of black holes, and it predicts other effects which have not yet been adequately tested (for example the frame-dragging effect for bodies orbiting nearby massive bodies). It does all of this to an extremely high degree of precision. But note that being able to predict and explain all this is merely a minimum standard for a successor theory, since empirical success is more than just the agreement of prediction with fact. Newton's standard of empirical success, discussed especially in Chapter 2, demands that a theory's fundamental parameters be measured from phenomena. As Harper (1997a; Harper and DiSalle 1996) has argued, GR

satisfies Newton's ideal of empirical success: this means that any successor to GR must satisfy this same ideal at least as well.[7]

The implication of this for successors to GR which are inspired by the desire to explain the dynamics of galaxies and clusters without the need for vast amounts of dark matter is that those theories must also meet Newton's ideal of empirical success to at least the same degree. This is no mean feat. And in the same way that General Relativity includes Newton's Universal Gravitation as an approximation in the limiting case of low velocity and weak gravitational fields, a successor to GR which is supposed to explain the dynamical discrepancies must include GR as an approximation at the distance scale of roughly stellar-system sized interactions. This is because observational tests place strong constraints on the behaviour of gravitational theories at those scales: the observations agree with the predictions of GR to high precision in this domain of distances.

As it turns out, however, tests on stellar system scales are really the *only* available tests for theories of gravitation. The success of GR with regard to these tests is usually taken to provide warrant for the applicability of GR to larger dynamical systems and even to the evolution and structure of the universe as a whole. But the applicability of GR at larger scales can be no more than an *assumption* in the present evidential context. (See Mannheim 1994b, and Ellis 1975, 1980, 1985, 1999.) As with Newton's argument to Universal Gravitation, we perform an inductive generalisation upon a set of "locally-derived" pieces of evidence, where this evidence is consistent with itself and ideally involves independent measures of theoretical parameters from several phenomena. Newton's extension of the principle of mutual gravitation to *all* bodies is the step of the argument on which the inductive risk is focused (Smith 1999). As we know, Newton's bet failed: extending UG to all phenomena on the basis of short-scale, weak-field tests.

---

[7] The Parameterized Post-Newtonia formalism discussed in Will (1993, especially Chapter 4) provides a way of testing and comparing metrical gravitation theories. Harper and DiSalle (1996) argue that this method of testing exemplifies Newton's ideal of empirical success.

[8] An FRW universe is a solution to the Einstein field equations for the whole universe, one that assumes a zero intrinsic curvature and a perfectly homogeneous and isotropic matter distribution. Among the interdependent properties of FRW universes identified by Ellis and homogeneity, isotropy, distortion effects and global curvature.

turned out to be incorrect. In the same way, it is evidentially risky to extend GR to all dynamical systems regardless of scale on the basis of stellar system or shorter scale tests.

Let me take the two relevant cases in turn, starting with the evidence for GR's applicability to the universe as a whole, and then the evidence for GR's applicability to large scale dynamical systems. G.F.R. Ellis (1985) has argued persuasively that cosmology not only *in fact* relies on unverified assumptions (for example, the Cosmological Principle and the Copernican Principle, which state respectively that the universe is homogeneous and isotropic, and that we do not occupy a privileged or unusual region of the universe), but that cosmology must *necessarily* rely on untestable assumptions. Ellis argues, among other things, that the very nature of the project of cosmology is such that we cannot determine by means of any purely observational test what the large scale structure of spacetime is. This in turn means that it is also impossible to test which law of gravitation holds for the universe as a whole: such a law can only be tested provided that we know well the very factors Ellis argues are unavailable to us.

The details of this, while interesting in themselves, are beyond the scope of the present discussion. It suffices to note that there are two factors which lead to this result. First, in order to verify any of the main properties of Friedmann-Robertson-Walker (FRW) universes[8], one has to assume some or all of the others (no independent check of each of the properties is possible). Second, alternative non-FRW cosmological models can always be found which would account for the observations just as well as any FRW model does. The clearest case (Ellis also gives several others) which shows that the characteristics of FRW universes cannot be tested independently is the case of distortion effects. A distortion effect is a change in the appearance of a distant object as compared to what it would look like across a completely flat spacetime, a change induced by the passage of the light from that object through a region of spacetime which is distorted (that is, not flat). Distortion effects can be induced by gravitational lensing—we know this happens in the case of gravitationally lensed background galaxies, which end up looking like extended luminous arcs—or the distortion can be caused by large-scale properties of the spacetime itself. In an FRW universe there would be no distortion effects (the global spacetime curvature is zero, the matter distribution perfectly homogeneous and isotropic). We would therefore need to confirm that there are no distortion effects in order to

confirm that our universe is actually FRW. But we cannot check for distortions in the images of distant objects unless we know what shape they have originally, and obviously we do not have this information. We cannot, for example, rule out the possibility that the objects that we call elliptical galaxies are really spherical objects seen through a distorting spacetime. Even if we were to treat every source as a point source, finding a perfectly homogeneous and isotropic distribution of sources across the entire sky would not prove that there is no distortion effect. This evidence would be just as consistent with an actually homogeneous and isotropic source distribution as it would be with any number of inhomogeneous and anisotropic distributions taken together with suitably selected global distortion effects. A homogeneous and isotropic apparent distribution of sources is indeed consistent with the assumption that the universe is FRW—but observing such a distribution obviously does not prove that the assumptions of the model are right. (On the other hand, if we assume that the universe is FRW and we find that sources are *not* distributed homogeneously and isotropically, that would be strong evidence that the universe is not actually FRW: such a distribution of sources can only be accounted for by either a really inhomogeneous and anisotropic distribution, or by a metrically-induced distortion effect, both of which are inconsistent with the universe really being FRW.)

Ellis' argument casts into doubt Will's claim (1993, 310-19) that cosmology has been a testing ground for gravitation theory since the 1920s. It is true that various cosmological observations (for example, the Hubble recession and the cosmic microwave background) have been *taken* as confirming that the universe satisfies the Big Bang model, and therefore the FRW spacetime model (which the Big Bang assumes), and therefore General Relativity (because the FRW model is a solution to the GR field equations). But if Ellis' arguments are correct, this supposed confirmation is illusory or extremely weak, amounting to no more than showing that GR is consistent with the available cosmological observations *given* some intuitively plausible but rather strong and evidentially unsupported assumptions. Relative to other (equally unsupported) assumptions, the cosmological evidence is equally consistent with universe models very different from the FRW model (for example, ones that are not isotropic or have regions of intrinsic spacetime curvature). If we had some way of confirming one set of assumptions over the others, we could perhaps make some progress toward deciding what the true

large scale structure of the universe is, and this would allow us to confirm a theory of gravitation at cosmic scales. But if, as Ellis argues, there is no way to determine empirically whether or not the universe is in fact homogeneous and isotropic, it is necessarily the case that there is no way to determine with confidence which spacetime model fits the universe best.

With the argument for cosmological scales behind us, let me now turn to the issue of the evidential basis for applying GR to the scales of galaxies and clusters. The fact that GR is *not tested* at scales larger than a stellar system can be demonstrated by considering the evidence that is taken to support GR. Gravitational redshift is known from terrestrial experiments, and from studies of the light emission of individual stars. The motions of planets within our solar system are now constrained by extremely precise laser and radar ranging evidence as well as by optical geocentric observations. Similarly, binary star systems are well studied, but although these systems are very distant from us the gravitational interactions in question take place on relatively "short" scales (what I call "stellar system" scales because binary star systems are never much bigger, astronomically speaking, than solar systems). The gravitational deflection of radiation by the Sun is well studied, and the observations are consistent with GR to a high degree of precision. (Will 1993, 332.) (The gravitational lensing of background galaxies by foreground clusters is a more difficult case that will be discussed later.)

What all these tests (and others) have in common is that they involve interactions happening over relatively short scales, at most about the size of a stellar system. These successful tests of General Relativity at short scales are consistent with GR being the correct gravitational theory at very large distances, but they do not give us direct evidence that this is indeed the case—the evidence is also consistent with a very different gravitational action at large scales. Newton's argument to "Universal" Gravitation did not establish empirically that other stars gravitate but rather used the available evidence of gravitation among nearby bodies plus some principles of theory choice (Newton's Rules of Reasoning) as the foundation for making an inductive extension of the local law to *all* bodies. In the same way, the hypothesis that GR applies at large scales is not supported empirically by direct evidence. One could even argue that the dynamical evidence from galaxies and clusters *contradicts* the hypothesis that GR holds at those

scales (relative to the hypothesis that the visible matter is the only matter present). Which horn of the dark matter dilemma to embrace thus depends on which hypothesis (either that GR should apply, or that there should be no other kinds of matter except those we already know) is more plausible.

What I have shown so far is that GR is *in fact* not tested at galactic and greater scales; I will now argue that purely dynamical tests of GR—or of *any* gravitational theory—are *impossible* at these scales, given the kinds of dynamical information available and likely to become available to us.[9] This is important, if correct, since it suggests principled limits to scientific knowledge. Moreover, it has serious negative implications for the prospects of solving the dark matter problem. The argument is based on what I shall call "the dark matter double bind", which goes as follows. Note that in order to test a dynamical theory, one must show (minimally) that it correctly predicts the motions of a system of bodies, given some initial configuration of that system. To specify the initial configuration, one must specify the distribution of bodies as well as their masses and velocities. Thus in order to test GR using the motions of a given spiral galaxy, say, one needs to know in advance what the mass distribution is. But the very existence of the dynamical discrepancy calls into doubt the assumption that the visible matter is all the matter that exists in galaxies. In fact, we have no warranted idea about what the matter distribution might be. Conversely, if we had reason to think that some particular dynamical law was true of the galaxy, we could use this dynamical law in concert with observations of the motions of visible matter to reliably infer the overall mass distribution. But, as I have suggested above, we have no tests to confirm that GR (or any other dynamical law) applies to galaxies. Thus, we cannot obtain a dynamical test of a gravitational theory at galactic and greater scales, or even compare rivals, unless we assume the matter distribution in advance, and we cannot empirically determine the matter distribution unless we first know which dynamical law applies. Purely dynamical

---

[9] If we had several billion years, we could perhaps study the perturbations of galaxies within a cluster on one another, and from these perturbations obtain a dynamical mass for them independent of the dynamical masses obtained by rotation curves. The kinds of dynamical evidence required in order to get around the dark matter double bind are very likely to remain permanently out of our reach.

tests of alternative theories of gravitation at galactic and greater scales are therefore impossible (again, relative to the kinds of information potentially available to us). The very fact of the dynamical discrepancy introduces doubt about the form of the matter distribution and/or the dynamical law that applies at those scales.

If we could find external and independent (for example, non-dynamical) reasons to think that the dynamical law or the matter distribution has a certain form, we could use that information to construct a probable inference to the form of the other parameter of interest. It is difficult to imagine a non-circular reason of this type for the matter distribution, however: a theory of the evolution of structure is perhaps the best hope for independent information about the matter distribution in galaxies and clusters,[10] but the evolution of structure depends on the dynamical law which governs it. All the tests which support GR over its rivals are at the scale of individual stellar systems, and it is hard to imagine how there could be dynamical evidence which would by itself provide adequate epistemic warrant for preferring GR over its rivals at large scales. This result applies equally GR and to any rival gravitation theory proposed to cover large scale phenomena.

The upshot is that whatever reasons we have for thinking that GR applies to large scale systems, they are not based on *dynamical* evidence that this is so. There is no direct evidence that GR is the correct theory of gravitation for the universe as a whole,[11] and similarly there is no direct evidence that GR applies to large scale astrophysical systems. Many alternative combinations of matter distributions and laws of gravitation could reproduce the observed dynamics. GR is not confirmed at galactic and cluster scales relative to its rivals because there is no way to distinguish them on the basis of the dynamical evidence, since we have no independent information about the matter distribution in these systems. There are, however, some other kinds of possible reasons

---

[10] Electromagnetic or particle detectors at best would allow us to infer a *minimum* mass and distribution: if dark matter is truly dark, no detection of it will be possible at all, and so we must try to find independent theoretical reasons to think the total matter distribution is such-and-such.

[11] If GR is not the correct law, replacing it with the correct law will cause huge upheaval in cosmology, where very many of the basic results—from theories of the evolution of the universe and its ultimate fate, the Hubble expansion, the formation and evolution of large-scale structure, and so on—depend on GR or its Newtonian limit in some way.

or indirect evidence that could be relevant to choices amongst rival gravitational theories. Among these are criteria for theory choice such as simplicity, non-*ad hoc*ery and explanatory unification. Perhaps most interestingly, it turns out that the key evidence available for deciding amongst *dynamical* theories at galactic and greater scales (indeed, perhaps the only possible evidence given the other epistemic constraints of the situation) is *non-dynamical* evidence.

We can compare the dark matter double bind for galaxies to the situation for the solar system. In the solar system, in contrast to galaxies and clusters, the available dynamical evidence, through a complex network of interdependent evidence relations, provides support for the matter distribution hypothesis and for the dynamical law governing the motions, and the confirmation of each hypothesis in turn provides further support for the other. That is, the two hypotheses are *mutually supporting*. The success of perturbation analyses in particular shows at one and the same time that GR is satisfied by the solar system, and that there are no unknown bodies of dynamical significance in any part of the solar system. Galaxies are not fundamentally different in kind *qua* dynamical systems, so the difference has do with our epistemic access to them. In the solar system, dynamical discrepancies (for example those involving Mercury and Uranus) were turned into successes for the dynamical law, which contributed to the mutual support of the matter and gravitation hypotheses (they are now really considered a single theory of the solar system). The confirmation of the gravitation theory, especially the strong confirmation which the theory earns from these strong tests, in itself provides grounds for thinking that the visible matter is all the matter there is. This is of course a fallible result. More exactly, because alternative hypotheses are logically possible (for example, it could be the case that the solar system behaves exactly *as if* GR were operating on the visible matter, even though the actual cause is some fortuitous combination of some other law and overall matter distribution), the epistemic warrant of the theory provides probabilistic grounds for believing that any excess matter that exists in the solar system is outside certain ranges of mass and location. Of course, the gravitation hypothesis could be wrong, in which case the matter hypothesis would be mistaken as well. But notice that the available evidence puts very strong constraints (relative to very weak and plausible background assumptions which we are not likely to

be willing to give up—things like the laws of motion for bodies moving at velocities much slower than the speed of light) on the possible ways in which and amounts by which the dynamical theory could be wrong in the solar system. These constraints in turn mean that the actual matter distribution can be only so different from what we would infer given the correctness of the solar system limit of GR. Thus the smaller the margins of error in the measurements used to confirm our theory of the solar system, the smaller the amount by which the predictions of any replacement theory can diverge in the tested realms from those of our current theory. This is true in principle also for galaxies and clusters, it is just that the evidence available from these systems is so much weaker and less detailed. The unfortunate fact is that this evidential disparity could be a permanent situation, because our distance from these systems and their intrinsic size puts practical limits on our ability to study the details of their dynamics.

The only real hope for confirming the details of the overall matter distribution that is inferred from the visible matter in conjunction with a dynamical theory is that we will acquire observations of some radiation or particle signature that indicates the existence and distribution of the dark matter in that system. If we can acquire independent information that a particular dark matter distribution exists, this would obviously redound to the credit of the dynamical theory which predicted that dark matter distribution. Thus the confirmation of a dynamical theory at galactic and greater scales depends crucially on *non-dynamical* evidence. This evidence may or may not become available, depending on which (if any) dark matter candidate is correct. The possibility of knowing the law of gravitation and the matter distribution in systems as large as or larger than galaxies thus depends on a state of the world over which we have no control and about which at present we have no information. We therefore cannot now predict whether it will even be possible for us to eventually acquire reliable and detailed information about large scale dynamical systems that will be comparable in quality to the analogous information we have about the solar system.

Even if relevant non-dynamical evidence *does* become available, the prospects for acquiring knowledge of the sort we would like to have are hardly encouraging. First, the confirmatory force of non-dynamical evidence about the overall matter distribution is mitigated by the fact that many different overall theories (including dynamical laws, dark

matter candidates and matter distributions) could result in this same non-dynamical evidence. Further, the matter distribution will not be inferable *exactly* from the non-dynamical evidence: the margins of error in this evidence mean that theories predicting a whole range of matter distributions will be consistent with the evidence. Of course, as the margins of error in this evidence decrease, the range of possible theories, and the differences between their predictions, becomes smaller and smaller. The non-dynamical evidence *could* eventually reduce the class of rival dynamical theories to a state where we consider the differences between them to be insignificant.

Philip Mannheim (with Kazanas 1989; Mannheim 1994, 1993, and 1992) claims that his Conformal Theory of Gravity (CTG) takes over all the solar system successes of GR, in the sense that the predictions of CTG are observationally indistinguishable from those of GR up to all distances for which we have tests of GR. But the predictions of CTG diverge dramatically from those of GR at the scale of typical spiral galaxies, where CTG predicts flat galactic rotation curves of the right sort while assuming the presence of only the normal matter whose light is visible and no additional dark matter. Thus GR (plus a some specific model of a dark matter distribution) is claimed to be observationally indistinguishable from CTG. We have to choose, then, between CTG versus GR+DM as our explanation of galactic and cluster dynamics. (Note that when I write "GR+DM" I mean some specific, though unspecified, dark matter candidate.)

Note that whether or not CTG itself is viable, most of the issues raised here remain. The relative epistemic support for GR as against some empirically adequate rival which does not require dark matter is a difficult matter to decide.

Mannheim and Kazanas (1989) claim to have come up with CTG while considering the problem of unifying gravity with the other fundamental forces. The other forces are conformally invariant, so the project of unification seemed to them to require making gravity conformally invariant.[12] (This brings them back to a formulation briefly

---

[12] Let me give a brief statement of what conformal symmetry is. Members of a set of operations on the points of spacetime are conformally invariant or symmetric just in case those operations leave the origin fixed and preserve angles between vectors. The conformal symmetry group includes rotations, dilations and contractions, and some other transformations. Scale invariance implies Weyl symmetry, or the higher conformal symmetry. Readers interested in the technical details of CTG are referred to Mannheim and

considered by Weyl, and hence they sometimes refer to their theory as "Weyl gravity".) They then happened to notice the facts mentioned above, namely that CTG agrees with GR at stellar system scales and predicts flat galactic rotation curves without the need for hidden matter.

The important thing about the origin story of CTG is that it shows that the theory was not designed specifically to solve the dynamical discrepancy. Scientists and others often take this sort of unexpected concurrence for a theoretical virtue: it is supposed to show that the theory is not *ad hoc* with respect to the phenomena it "accidentally" accounts for. Mannheim and Kazanas play this as their strongest card:

> While there have been several attempts made in the literature to account for the galactic rotation curves without the introduction of any so far undetected dark matter. . . . the present approach differs from them in that it was not at all introduced for the purpose of resolving a specific problem in astrophysics; rather our proposed solution is a by-product of a theory of gravity which is itself based on a principle, namely that of local conformal invariance of the world geometry. (Mannheim and Kazanas 1989, 638)

The practical outcome of Mannheim's introduction of the conformal symmetry to the gravitational action is that we have essentially two components interacting. One component of the action falls off with distance as one would expect; the unexpected thing is a second component of the action (Mannheim says resulting from a new linear potential term in his equations) the strength of which *increases* with distance. At short distances, the contribution of this linear term is unmeasurably small, and thus the predictions of CTG are observationally indistinguishable from those of GR for the solar system. CTG can thus take over all the empirical successes of GR in this sphere. This is something we see GR doing in the transition from Newtonian gravity, namely the successor theory takes over the successes of its predecessor. GR also successfully predicts several new phenomena, and is also RfP confirmed by some of them.

Mannheim explains the possibility of his theory replacing GR, which most people think of as solidly established, by pointing out that

---

Kazanas (1989) in the first instance, and then to Mannheim's later papers (for example 1994, 494-95). "What mainly distinguishes the conformal gravity program. . .from other alternative approaches is that it sets out to generalise not the Newtonian potential, but rather the Swarzschild solution, so that from the outset the theory is fully covariant and fully relativistic" (Mannheim 1995, 1-2).

> Since the conformal gravity potential reduces to Newton on short enough distance scales and then first deviates from it galactically. . .. the viewpoint espoused by Mannheim and Kazanas is that from the study of the solar system we only measure the first few terms in a perturbation series and that at larger distances the series may simply differ from that inferred from Newton-Einstein. i.e. that it may depart from the standard model in precisely the kinematic region where the conventional wisdom is currently having problems. with the apparent need for dark matter then simply stemming from having guessed the wrong series. (Mannheim. no date. 4)

As Mannheim says elsewhere. "Not only does most of our information regarding gravity derive from a study of the solar system.. . .so does most of our intuition" (Mannheim 1994. 488).

If the succession of CTG over GR is to go forward. we should expect to see the same pattern as we saw in the case of GR's succession over Newtonian gravity. It is not clear that we do. First. it is not clear that CTG really takes over *all* of the empirical successes of GR. CTG is not yet well enough developed or explored for us to be able say for sure; gravitational lensing is one problematic area (see below). CTG *does* make some new predictions. for example about galactic rotation curves and cluster velocity dispersions. and Mannheim has tried to find ways to tie the CTG action to the Hubble constant (which would yield another testable prediction). The trouble is that the present state of the evidence is such that CTG's prediction of flat rotation curves without dark matter cannot be evidentially distinguished from GR's prediction of similar rotation curves but with vast quantities of dark matter. There is. therefore. no differential confirmation from the present large scale dynamical evidence. and since the two theories agree about shorter-scale predictions. on the present evidence the two radically different options are evidentially indistinguishable. Mannheim attempts to adduce methodological arguments for thinking GR inferior to CTG. but as I argue below this attempt fails. Had it succeeded. it would have provided reason to prefer CTG over GR. and thus reason to prefer a gravity solution over a matter solution to the dynamical discrepancy in the present evidential situation. But as I argue in the final section of Chapter 6. methodological considerations can be adduced which favour provisionally retaining GR as the gravitational theory at galactic and greater scales. These reasons are fairly weak, however. and I also describe some possible evidence that. were it to become available.

would tip the balance in favour of CTG or some other gravitational explanation of the dynamical discrepancy.

6.2 THEORY CHOICE AND THE UNDERDETERMINATION PROBLEM

The dark matter discrepancy is a classic example of the failure of a prediction of a well-entrenched theory. I use the phrase "well-entrenched" advisedly, as implying that workers in the field *take* the dynamical theory involved to be worthy of acceptance. Whether the theory really *is* worthy of acceptance can only be judged relative to a body of evidence and a theory of confirmation. The logic of the confirmation of theory by evidence is an interesting and involved topic, one that needs adequate treatment in order to give a complete epistemic account of scientific reasoning. I will restrict my remarks on confirmation here to the minimum necessary to set up the problem of underdetermination and its epistemic consequences, which I will discuss in some detail because it is instantiated in an especially interesting way in the dark matter case.

The "empirical adequacy" of a theory, although it can be described in different ways, depends minimally on whether the theory in question (together with its background assumptions) makes predictions that are *logically consistent* with the known observations.[13] (Hereafter, when I say that a theory makes a prediction I should be read as saying that a theory plus the appropriate background assumptions and initial conditions make the prediction.) The basic ideal of confirmation is that when a theory T has as a logical consequence a statement P that is also entailed by another statement O, where O itself is arrived at by empirical investigation, T receives a boost of epistemic warrant thereby. More succinctly, if T entails P, and O entails P (for example, when O is equivalent to P), then the discovery of O confirms T (to some degree).[14] I must here

---

[13] Since judgements of logical consistency apply only to sets of sentences, we compare the theory (a set of sentences) and its predictions (another set of sentences—logical consequences of the first set in conjunction with another set of auxiliary hypotheses) against evidence *statements* which are warranted as the result of some observation process.

[14] Most accounts of instance confirmation (for example Hempel 1965) skip a step here, saying merely that if T entails E, then observing E confirms T. But an observation process involves a physical interaction plus

ignore a rather large amount of detail concerning how, under what conditions and to what extent a successful prediction confirms its parent theory, but everyone (at least everyone who believes that the confirmation of scientific theories by evidence is possible at all) agrees that, in most kinds of cases, a successful prediction does redound to the epistemic credit of the theory in question. (See Hempel 1965 and Achinstein 1983 for an *entrée* to the literature on confirmation theory.)

I will here briefly discuss two famous problems related to confirmation in order to get at the problem of "the underdetermination of theory by evidence" as it is relevant to the dark matter issue, and especially to the choice of a dynamical theory at galactic and greater scales. The first is Hume's problem of induction, which applies generally to all types of ampliative inference, and casts doubt on the possibility of confirming theories by their positive instances. The second is Duhem's problem of the ambiguity of falsification, which casts doubt on the possibility of disconfirming theories by their negative instances. Let me discuss each of these in turn.

Hume's problem may be set up with the example of enumerative induction, although the problem applies to every kind of induction. Positive instances, no matter how many are to hand, are never sufficient to establish with certainty (that is, by deduction) that a theory expressed as a universal statement is true. Any enumerative induction, from a set of observations all with a uniform character, to the universal generalisation that all objects of the observed type have those characteristics, is therefore deductively invalid. (No matter how many white swans I observe, I will never be licensed to draw the conclusion that all swans are white as a deductive certainty.) Hume's problem is that there is no way to provide a non-circular justification of our inductive practices. Probabilistic (which is to say ampliative or non-deductive) accounts of confirmation have it that positive instances or correct predictions confirm theories by

---

a theory-mediated ampliative inference, the outcome of which might need to be put into yet another inference in order to yield a statement in the same terms as the prediction made by the theory being tested. For example, a typical astronomical observation might yield information *O* about the spectrum of some star; a further inferential step is required in order to turn this observation into information *P* about the star's surface temperature and chemical composition, which could then be used to confirm, say, a theory of stellar evolution which has *P* as a consequence.

degrees, so that given some favourable evidence a theory has some degree of probability (always less than 1). So a quantity of observations of white swans confirms to some degree the hypothesis that all swans are white, but the hypothesis remains fallible (and corrigible) in light of future observations. Hume's problem remains even for these probabilistic accounts, however, since we now have to justify our belief that the evidence shows the conclusion of our inductive argument to be probable. (Again, I must neglect the details, and a huge body of literature.)

Duhem's problem has to do with the evidential import of observations that contradict predictions. Now, in most or many cases, and certainly in all scientifically interesting cases, theories make no predictions except in concert with a set of background (or "auxiliary") hypotheses and information. This, it turns out, means that the inference pattern of *modus tollens* (the valid deductive inference: if P then Q, not-Q, therefore not-P) does not provide the definitive disconfirmation one might have expected when predictions turn out to be wrong. As Duhem points out,

> [T]he physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses that constitute this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (Duhem 1982 [1914], 187)

Finding that some prediction entailed by a set of hypotheses and background assumptions is not the case logically requires that at least one member of that original set is false. The trouble is, the failure of the prediction itself does not tell us where the "arrow of *modus tollens*" should point among the sentences involved in the prediction, whether on the main hypothesis or on one of the background assumptions. Because there are many possible responses to such a failure of prediction, each of which is equally possible given the evidence, we may speak of the "ambiguity" of falsification.

Hypothetico-deductive (H-D) accounts of scientific evidence and theory choice provide no adequate response to the Duhem problem. There are many logically possible ways to account for any discrepancy between prediction and observation, and H-D has no resources that provide grounds for preferring one among the many incompatible theoretical structures that are consistent with all the evidence available at any one time. To put the point in the usual way, given any body of evidence, there exists an indefinitely

large set of conjunctions of hypotheses and background assumptions, where each member of this set entails the same observations. Each of these is therefore equally well-supported by the evidence (according to the H-D and the standard instance confirmation accounts). The basic H-D idea provides no grounds for distinguishing evidentially between members of the set of predictively adequate hypotheses; if we wish to make epistemic distinctions between predictively equivalent theories we must invoke considerations that go beyond mere predictive adequacy. (We must find grounds for saying, for example, that although predictively equivalent, the hypotheses are nevertheless not all on a par, that some of them enjoy greater evidential support than others.)

One consequence of the ambiguity of falsification is the impossibility of so-called crucial experiments. A crucial experiment is supposed to provide definitive proof of one hypotheses by providing definitive disproof of (all) its rivals. Duhem argues that there is no analogue in the physical sciences of the method of *reductio ad absurdum* in mathematics. That is, one cannot prove a hypothesis true by proving its contrary (or contraries) false. This is partly a practical, partly a principled limitation. The practical part is that it is impossible to construct an exhaustive disjunction of all possible theoretical systems capable of saving the phenomena; the principled part is that since falsification is ambiguous, contrary evidence never provides definitive disconfirmation of the contraries of a theory.

Both Hume's problem and Duhem's problem, at their most basic levels, pose the following choice problem: given a certain body of evidence, how do we decide in favour of one of the possible theories over its rivals (where the "possible" theories are the internally consistent ones also consistent with all the available evidence, and where the rivals are all the other theories which meet the empirical standards to the same extent)? What these two problems demonstrate is the inadequacy of deductive logic by itself for scientific reasoning. This is really no surprise, since it is obvious that scientific reasoning is and must be ampliative. In short, what Hume's problem and Duhem's problem show is that evidence plus deductive reasoning always underdetermines theory choice.

Hume's problem and Duhem's problem do not, by themselves, say anything about the possibility or impossibility of ampliative principles of theory choice. It is, however,

very difficult to articulate acceptable principles of ampliative theory choice. Some philosophers have in fact argued that rational, evidentially justified theory choice is impossible. But the history of science shows that scientists *do* make theory choices, and that evidence matters to them in those choices.[15] Various principles of ampliative reasoning have been suggested as justifications for those choices (or as normative guides for theory choice), but we must acknowledge that this is an open field and the final answers are not to hand. Among the possible principles of theory choice that come up in the problem of deciding upon the dynamical law at galactic and greater scales are simplicity, non-ad-hocness, explanatory unification; these will be discussed below in response to Mannheim's charges against dark matter.

Let me first mention some aspects of underdetermination in more detail. Larry Laudan (1996) distinguishes two different underdetermination theses, neither of which entails the other. The first thesis, "Humean Underdetermination" (HUD) is clearly correct, but also quite weak.

> HUD: for any finite body of evidence, there are indefinitely many mutually contrary theories, each of which logically entails that evidence. (Laudan 1996, 31)

This thesis merely asserts that the fact that a theory makes correct predictions is no guarantee of its truth. To assert the argument, "If theory *T* is true, then observation *O* holds; observation *O* holds; therefore, *T* is true," is to commit the deductive fallacy of affirming the consequent. *O* could hold for some reason besides the truth of *T*. (This is "Humean" in the minimal sense that it follows more or less directly from Hume's remarks on induction. The prediction "The next swan will be white." may be correct even if it is derived from the false hypothesis "All swans are white.") Note that the fact that this logical possibility exists does not tell us much about the epistemology of science. As Laudan (1996) and others have taken pains to emphasise, the deductive

---

[15] In the Neptune case, for example, we see that the prior evidential support for Newton's theory was a significant factor in Adam's and Le Verrier's choice to opt for matter solutions, and that the matter hypotheses they settled on were guided as much as possible by evidence from the known perturbations of Uranus. Likewise, in the Mercury case, the successful prediction of the perihelion precession was an important factor in the choice of General Relativity over rival gravitation theories.

underdetermination embodied in the HUD thesis in no way implies that evidence *absolutely* underdetermines theory: *ampliative* rules of evidential reasoning could still provide adequate warrant for rational theory choices. So there could well be evidential or methodological grounds for preferring one or some of the predictively adequate theories over the others. Thus those who would assert radical underdetermination are bound to provide additional arguments.

Quine famously asserts the thesis that, "Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system [of belief]" (Quine 1998 [1951], 297). Quine mentions four possible strategies for revising one's belief set in order to retain a given theory in the face of contrary evidence: change the meanings of terms, alter the laws of logic, plead hallucination, and modify background or auxiliary assumptions in the web of belief (Laudan 1996, 35; cf. Quine 1998 [1951]). Quine shows that it is logically possible to reconcile one's beliefs with any available evidence by these means. Of the strategies Quine mentions, the first three seem obviously bad: the first one merely changes the subject, and the second and the third, like the first would result in a huge loss of empirical support for the theoretical structure as a whole (what Quine calls the "web of belief"). If we must treat every single piece of evidence as defeasible by an unfalsifiable assumption about hallucination, that would call into doubt the very possibility of knowledge.

Quine's fourth strategy is more interesting. It surely is possible, in many cases at least, to modify background assumptions so that any evidence becomes logically compatible with any theory. We may recall Duhem's point that only a whole group of theories (Quine says the whole of science) faces the tribunal of experience, and thus that contrary evidence falsifies the theoretical structure as a whole but does not specify which part is responsible for the predictive failure. (The fact that we are to take the whole of science as the unit of analysis earns Quine's view and similar views the label "holism".) Since we can imagine many different logically possible modifications of the theoretical structure, each of which would eliminate or otherwise explain away the failed prediction, Quine's fourth strategy for reconciling theory with contrary experience in effect proposes a second underdetermination thesis. Laudan phrases it this way:

> QUD: any theory can be reconciled with any recalcitrant evidence by making suitable adjustments in our other assumptions about nature. (Laudan 1996, 36)

If we read the QUD thesis as claiming that all such adjustments are epistemically of a piece, then the QUD thesis makes a much stronger claim than the HUD thesis. (Note, however, that Quine himself would *not* say that all possible adjustments of the web of belief are epistemically of a piece. Laudan's accusation that Quine is a radical epistemic relativist is rather unfair given that Quine devotes a lot of space in his writings to trying to articulate and defend principles of ampliative theory choice.) On this strong (or "radical holist") reading QUD is an assertion not only of deductive but also of *ampliative* underdetermination (on the strong reading QUD implies that *no* theory can be shown to be epistemically superior to any other, *whatever* the available evidence, and *whatever* our best rules for ampliative reasoning might be).[16]

But the logical possibility of retaining any belief "come what may" does not by itself support a claim that all theoretical structures that save the phenomena are equally rational or equally evidentially supported (not every possible modification to the web of belief will be equally evidentially prudent). The strong reading of the QUD thesis is probably not supportable; it is not, for example, Quine's reading. I will argue against it here by giving an account that does distinguish epistemically among the various possible modifications to a theoretical structure that could be made in response to recalcitrant evidence. In so doing I am trying to sketch an evidential or epistemic solution to the

---

[16] Glymour bemoans, "Holism is the predominant epistemological theme nowadays, even though much of what is said in elaboration of it seems to defy both good sense and historical fact" (Glymour 1980, 148). Glymour diagnoses the rise of radical holism through a semi-historical chain of causes and inferences (Glymour 1980, 148-9); the main factor seems to be that some philosophers have taken the apparent failure of H-D confirmation or instance confirmation, as tried by logical empiricists of various stripes, as indicating the *general* failure of attempts to establish a theory of evidential relevance. But such an inference commits the fallacy of denying the antecedent: If H-D instance confirmation is correct, then it is possible for evidence to differentially support one theory over its rivals; H-D and instance confirmation are not correct; therefore, it is not possible for evidence to differentially support one theory over its rivals. In short, radical holists seem to require the assumption that either H-D or instance confirmation establishes the link between evidence and theory, or nothing does; clearly, this assumption relies on an incomplete survey of the logical space of possible accounts of evidential relevance.

problem of theory choice, in light of the fact that a moderate version of the QUD thesis seems to be correct.

Glymour[17] cites at least five reasons why the predictive equivalence of two theories need not imply their epistemic parity. (i) One theory but not the other might contain hypotheses not tested by the phenomena. (ii) The two theories might share hypotheses, but those hypotheses be tested by the evidence only with respect to one of the theories. (iii) The phenomena might repeatedly test a single hypothesis of one theory, but with regard to the other independently test several of the hypotheses. (iv) The two theories might postulate common properties, but only one permit the determination of the values of those properties for various systems. (v) The hypotheses tested in one theory may be important and central ones, whereas only peripheral hypotheses are tested in the other. (See Glymour 1980, 182; compare 153.)

One thing that is implied here is that consistency with the facts is by itself no indication of the degree of empirical support of a theory. And as Newton's methodology illustrates, empirical support is—in some significant cases at least, and at least in part—a matter of a meta-induction on a consilience of multiple measurements of theoretical parameters from several independent phenomena. The distinction between predictive success and empirical support upon which I wish to insist here is crucial for the evidential analysis of the case of competing gravitational solutions to the astrophysical dynamical discrepancy. The distinction requires that *even if* we had rival solutions which were predictively equivalent, we would still have to perform a close investigation of the details of the theories' respective evidential credentials in order to evaluate their relative empirical support. The distinction between predictive success and empirical support

---

[17] Addressing the Duhem problem and answering its supposed implications for the possibility of rational theory choice was a large part of Clark Glymour's motivation in his *Theory and Evidence* (1980). As Glymour suggests, an adequate model of scientific reasoning ought to be able to account for the fact that in actual situations involving the failure of prediction scientists often *do* lay the blame on *specific* parts of the theoretical structure rather than on others, and seem to do so in a rational manner that is based on evidence. Glymour's account of "bootstrap confirmation" was only a partial success; see Earman and Glymour (1988), where Glymour admits that he has to give up on significant parts of his account. But the points about evidential relations which I make use of here stand whether or not the bootstrapping account fails.

entails, furthermore, that we draw a divide between predictive adequacy and empirical adequacy; thus even a predictively adequate theory (or, the most predictively successful theory available at any given time) may fail to be empirically adequate, may fail to have a *high enough* degree of empirical support. Being able to draw *this* distinction, in turn, then, enables us to formulate justifications—once we have gone beyond the present account and have described in detail the criteria of evidential adequacy which go beyond mere predictive success—for many deep-rooted theoretical judgements, for example, against overly *ad hoc* theories, or against theories which are mere conjunctions of accurate observation statements.

If Glymour, Laudan and others are right, even predictive equivalence does not imply epistemic parity. The choice problem in that case is one of trying to find or invoke principles of ampliative reasoning so as to establish the epistemic superiority of one of the predictively equivalent theories over the others. But yet another kind of choice problem bears mentioning here. It is the problem of choosing between theories that are predictively equivalent with regard to all the phenomena *for which we have detailed evidence*, but whose predictions differ (or may differ) with regard to other phenomena for which the relevant data are less well (or not at all) established. The best response in such a situation may be simply to try to improve our knowledge of the phenomena with regard to which the hypotheses make differing predictions, and to suspend judgement until evidence which forces the choice becomes known. But there may arise cases in which this is not an option—because it is impossible (either *simpliciter* or within some reasonable time frame), whether for practical or for principled reasons, to acquire the necessary information—and with regard to which we may yet have a strong desire or need to choose some hypothesis or other. Such a situation arises, I claim, in the case of the dynamical discrepancy for galaxies and clusters: we must choose *some* theory of gravitation, at least provisionally, even just to be able to investigate the issue, and finding a solution is vitally important for our understanding of the universe. Not only must we decide whether to pursue the class of matter or the class of gravitation solutions, we must also decide which candidate from among the chosen class of solutions is the best option. The philosophically interesting thing about the choice problem presented by the debate between CTG and GR+DM is that in this case we seem to have a situation involving

predictively *inequivalent* theories that are apparently evidentially indistinguishable on any dynamical evidence we might hope to be able to acquire.

The radical holist interpretation of the QUD thesis, according to which any one or more of the statements in the web of belief can be revised or replaced with epistemic impunity, seems to make the assumption that all parts of the theoretical structure have the same degree of prior epistemic warrant. But this is manifestly not the case. Some beliefs in the web have *much* higher prior evidential warrant than others. We *do* have, therefore, *prima facie* reasons—albeit ampliative and therefore defeasible ones—for thinking that some revisions to the theoretical structure are evidentially better than others. Of course it is true that, in the absence of further information about prior evidence and degrees of epistemic confidence, the arrow of disconfirmation loosed by a failed prediction cannot be aimed at any particular part of the theoretical structure. But we *do* usually also have additional information, for example large parts of the structure will have been used successfully to make other predictions. This does not mean that all members of that group are true (that would be to affirm the consequent) but they thereby acquire increased epistemic warrant, and our first attempts at least to reconcile our theory with the new facts should try to preserve them. The fact that other evidence matters is something that Duhem, for example, recognizes. Although Duhem argues for the deductive ambiguity of falsification, and even recognises that many different revisions to the theoretical structure could in principle be made which would be compatible with the observations, he does *not* conclude from this that it is impossible to distinguish epistemically between the various logically possible responses to a failure of prediction.

Likewise, although Duhem denies the possibility of so-called "crucial experiments" he allows that the evidential judgement of a scientist—Duhem calls it "good sense" (Duhem 1982 [1914], 216-18)—will nevertheless indicate how to assess the relative weight of evidence in a case of theory choice: "Good sense is the judge of hypotheses which ought to be abandoned" (Duhem 1982 [1914], 216). Thus, though Foucault's experiment on the velocity of light in water as compared with air was not (because nothing is) a crucial experiment in the sense that it proved with absolute certainty the falsity of the particle theory of light and the truth of the wave theory, nevertheless scientists of good sense understood that the experiment established with high probability

the relative superiority of the empirical warrant for the wave theory. Though some scientists and historians of science *took* this experiment to be "crucial" in the strong sense, what it really did was to provide relatively stronger (perhaps much stronger) empirical support for the wave theory.[18] So Duhem certainly does not think that the impossibility of performing crucial experiments, or the ambiguity of falsification from which this impossibility follows, entails that evidential reasoning is impotent, or that *any* theory can *rationally* be retained in the face of any evidence.[19] Duhem recognises the *logical possibility* of maintaining any theory in the face of any evidence, by making suitable changes to the theoretical structure, but his theory of "good sense"—to be honest, he gives no more than a nod towards a sketch of such a theory—is meant to indicate *that* and *how* theory choice remains possible despite deductive underdetermination and the ambiguity of falsification. Clearly the (tacit) principles backing up a scientist's good sense are ampliative rules of theory evaluation and theory choice.

Let me now briefly outline an argument about the role of evidence in directing the arrow of disconfirmation. This is to fill in, in a plausible if preliminary way, Duhem's

---

[18] If I may invoke a Cartesian distinction, Foucault's experiment does not provide absolute certainty, but it does provide *moral* certainty, that is, grounds for practical certainty sufficient for action and belief. (Descartes, in his *Principles of Philosophy*, advocates the view that scientific knowledge can at best be morally certain; this seems to get much less attention in discussions of Descartes than does his view about the possibility of absolute certainty with regard to some foundational metaphysical precepts. Descartes' views about the nature and possibility of scientific knowledge are actually quite compatible with Duhem's very strong instrumentalist tendencies.) We might say, with this Cartesian distinction in mind, that Foucault's light experiment is "morally crucial". It leaves no reasonable doubt remaining about the definitive superiority of the wave theory, although this judgement could be overturned by future evidence (and eventually was), and even though we *could* have retained the particle theory by continuing to make suitable adjustments to the theoretical structure. Gillies (1998, 310) makes similar remarks, but interprets Duhem's "good sense" as a *community* standard which can result in an experiment being *in practice* crucial.

[19] The apparent contradiction of Duhem's strong instrumentalist leanings—that the goal of science is nothing more than to save the phenomena, and that any theory that achieves this goal is of equal moment (see Duhem 1969 [1908])—with his remarks on good sense and crucial experiments, is merely apparent: once we have Foucault's experiment, the particle theory no longer saves the phenomena, so it does not meet the minimum standard for theories, and this is why it is reasonably rejected given the evidence.

sketch of his theory of good sense. My attempt here is still no more than a sketch, since a complete theory would require an extensive analysis of the probability dynamics of evidence and confirmation, an analysis which would go far beyond the scope of the present project. It is, nevertheless, important to say *something* about this issue here in order to explain my views about the prospects for an evidential solution to the astrophysical dynamical discrepancy.

Note first that the claim that predictively equivalent theories are rarely if ever epistemically equivalent becomes more obvious once one considers Duhem's stricture that in order to be acceptable a theory must be consistent with *all* the available evidence at any given state of science (Duhem 1969 [1908], 117), and when one pays close attention to the details of historical examples. The standard example, the debate between the Ptolemaic and Copernican theories, is in fact *not* an example of predictive equivalence (and it is certainly not an example of epistemic parity): even in their original forms, the two systems made quite different predictions about the phases of Venus, the distances of the planets, and so on.[20]

Most of the hypotheses in a theoretical system falsified by recalcitrant evidence are involved in the failed prediction only in the sense of what Duhem calls "experiments of application" (Duhem 1982 [1914], 183). In the first instance at least, the scientist will take the arrow of falsification *not* to fall on these auxiliaries. What makes logical and evidential sense of this, in the face of the deductive ambiguity of falsification, is the fact that the auxiliaries involved in predictions merely "by application" have *already* been involved in "experiments of testing" and have therefore acquired some (usually substantial) degree of confirmation thereby. (For example, optical theory is involved by application in any telescopic observation; and optics is highly confirmed by laboratory experiments.) So, where the auxiliary hypotheses have independent support from other tests, we have grounds (up to some degree of confirmation) for taking them as provisionally established, and this helps to direct the arrow of disconfirmation. So sometimes refuting evidence really refutes.

---

[20] The Tychonic system is another matter: it does eventually take Newton's physics to resolve the debate in favour of Copernicus over Tycho.

Clearly it is still *logically possible* that *any* hypothesis involved in a theoretical structure making a false prediction could be at fault, even a hypothesis with a high degree of previous independent confirmation. But removing or revising a well-tested auxiliary would require re-interpreting the experiments of testing—and *all* the experiments of application—of which that auxiliary is a part and which originally were taken to confirm it. This will in turn require new hypotheses and auxiliaries: it is easy to see that performing this kind of revision will have far-reaching ramifications within the theoretical system. It is hard to imagine that the new auxiliaries will (immediately in any case) provide equally good epistemic foundations for the facts and theories we take to be known. Thus, rejecting or revising a well-confirmed auxiliary will not necessarily lead to an evidentially good theoretical system, even when it leads to correct predictions of the formerly falsifying fact.[21] The mere fact of the availability of an indefinite number of

---

[21] Klee (1992, 488) mentions an argument due to J.D. Greenwood that is related to the claim that some modifications to the theory structure would lead to an overall *decrease* of epistemic warrant for the theory structure as a whole.

> Greenwood calls Quine's bluff on this issue of the alleged limitless capacity for accommodating test theories to recalcitrant observations. What the Quine-Duhem thesis [Laudan's QUD] ignores, according to Greenwood, is that *some* adjustments to the auxiliary hypotheses and ceteris paribus clauses have "ripple effects" throughout the rest of one's total theory which serve to undermine the *prior* observational evidence for the theory under test. Greenwood calls such adjustments "degenerating", and he considers them epistemologically self-defeating. If this argument is to have any epistemic force, then it should apply to the classic case of the underdetermination of theory by data: the attempt to "save" Newtonian theory from recalcitrant observational evidence by postulating the existence of undetectable "universal forces" which shrink measuring rods, bend light rays, and slow down clocks. . . . A Greenwoodian, we can surmise, would argue that such universal forces would undermine almost all prior observational evidence for classical Newtonian theory—so the salvation would be self-defeating. (Klee 1992, 488)

Greenwood's point that some modifications to the theoretical structure would undermine previous epistemic support is a good one, although Klee's treatment of it is not exactly correct. Contrary to what Klee says here, it would be perfectly possible to construct a theory of universal forces such that all the pre-relativistic observations were untouched, that is, where the universal forces caused observable changes in bodies only in the realm in which the recalcitrant observations were found. In the same way, Milgrom and Mannheim propose modifications to (the weak-field, low-velocity limit of) GR that leave the solar system evidence untouched. If such a theory is correct, it would transform support for GR into support for the successor theory. All I mean to say here is that some significant attempts to modify a background theory in order to save a higher level theory in the face of recalcitrant evidence *will not* be self-defeating in Greenwood's

logically possible alternative theoretical systems does not mean that all such systems will have equal epistemic warrant.

Let me quote a passage from Lakatos which seems tailor-made for this discussion. Lakatos modifies Quine's idea of a web of belief whose boundary conditions are experience by talking about a "hard core" of theory surrounded by a "protective belt" of auxiliary hypotheses. Lakatos holds that a commitment to a particular research programme just means that we will not be willing to modify theories in the "hard core", and that we will try to protect them in the face of contrary evidence by modifying auxiliaries in the protective belt.

> The story is about an imaginary case of planetary misbehaviour. A physicist of the pre-Einsteinian era takes Newton's mechanics and his law of gravitation, (N), the accepted initial conditions, *I*, and calculates, with their help, the path of a newly discovered small planet, *p*. But the planet deviates from the calculated path. Does our Newtonian physicist consider that the deviation was forbidden by Newton's theory and therefore that, once established, it refutes the theory N? No. He suggests that there must be a hitherto unknown planet *p'* which perturbs the path of *p*. He calculates the mass, orbit, etc., of this hypothetical planet and then asks an experimental astronomer to test his hypothesis. The planet *p'* is so small that even the biggest available telescopes cannot possibly observe it: the experimental astronomer applies for a research grant to build yet a bigger one. In three years' time the new telescope is ready. Were the unknown planet *p'* to be discovered, it would be hailed as a new victory of Newtonian science. But it is not [discovered]. Does our scientist abandon Newton's theory and his idea of the perturbing planet? No. He suggests that a cloud of cosmic dust hides the planet from us. He calculates the location and properties of this cloud and asks for a research grant to send up a satellite to test his calculations. Were the satellite's instruments. . .to record the existence of the conjectural cloud, the result would be hailed as an outstanding victory for Newtonian science, but the cloud is not found. Does our scientist abandon Newton's theory, together with the idea of the perturbing planet and the idea of the cloud that hides it? No. He suggests that there is some magnetic field in that region of the universe which disturbed the instruments of the satellite. A new satellite is sent up. Were the magnetic field to be found, Newtonians would celebrate a sensational victory. But it is not. Is this regarded as a refutation of Newtonian science? No. Either yet another hypothesis is proposed or. . .the whole story is buried in the dusty volumes of periodicals and the story never

---

sense. But it is easy to see that *some* such modifications *would* be self-defeating, so that Greenwood's argument successfully defeats radical relativist versions of QUD.

> mentioned again. [Lakatos's footnote:] At least not until a new research programme supersedes Newton's programme which happens to explain this previously recalcitrant phenomenon. In this case the phenomenon will be unearthed and enthroned as a 'crucial experiment'. (Lakatos 1970, 101-02)

It is unfortunate that Lakatos focuses here on an imaginary example since, except for the satellites, there were two very similar historical cases available to him which would have been more enlightening, namely the Uranus and Mercury cases. The messy details of the real reasoning situations, though more difficult to analyse, are more informative about general principles scientific reasoning. Lakatos' target, in any case, is Popperian falsificationism; Lakatos essentially just wants to show that the simplistic-seeming account of falsification given by Popper does not work. As Lakatos points out immediately following this passage, a theory never contradicts a singular observational statement unless taken in conjunction with a "non-existence statement" (almost always conjectural) asserting that no additional factors are operating. A failed prediction therefore refutes not some theory alone but the theory *plus its "ceteris paribus" clause*. The failed prediction can then be made inconsequential for the main theory, simply by modifying the *ceteris paribus* clause (that is, some of the untested or weakly supported auxiliary assumptions necessary in order to make any prediction).

My account turns out to be similar to Lakatos' except in that I maintain that there are evidential relations (not just a prior commitment to a research programme) that determine which theories will be protected by modification of the auxiliaries, and which auxiliaries will be modified. There can arise *evidential* situations in which a theory, even one in the "hard core", will be rejected (thus contrary to Kuhn I think that theory change is or can be guided purely by evidence). Since according to Lakatos making the choice to direct the arrow of falsification at the hard core is to abandon one's research programme, he seems to admit only two possible outcomes of discovery contrary evidence, namely successful *ad hoc* protection of the theory and ignoring the anomaly until one research program replaces another.[22]

---

[22] For Lakatos, the rationality of the succession of one research program over another can only be judged retrospectively, that is, with the perspective of sufficient temporal distance. In contrast, I claim that in some cases at least we can judge the evidential warrant of a theory change, even a major one, while it is happening. Such judgements are relative to a given body of evidence and a given body of theory, of course,

It will be useful to mention the explicit connection between the debate between CTG and GR, and the conclusions reached here about crucial experiments. Crucial experiments in the strict sense are impossible—to do a *reductio ad absurdum* would require us first to give an exhaustive list of all the possible theories that could save the phenomena under consideration, and then to definitively disconfirm all but one of these, which is clearly an impossible task. Nevertheless, evidence could conceivably become available which would allow us to decide more or less definitively in favour of some particular solution to the dynamical discrepancy. Some instances of possible evidential conditions of this type are described elsewhere in this chapter.

The choice between CTG and GR plus a theory of dark matter is a case of HUD, with a twist: the two theories, almost by definition, agree on all the available and projected dynamical data. Where they differ is with regard to predictions that at first blush seem merely tangential, that is, not really the point of either dynamical theory. These are, for example, predictions about what kinds and amounts of particles and radiation we should be able to detect with devices of various designs here on Earth. This non-dynamical and especially higher order evidence is extremely important for the dark matter problem, and represents the best hope for finding a solution to it soon. The account of theory choice sketched in this section shows that evidentially based theory choice is possible despite problems of theory choice both real and imagined.

6.3 CURVE-FITTING: THE ROLE OF SIMPLICITY IN THEORY CHOICE

Because I wish to make a point later about how the error in the observations relevant to the dark matter issue impacts on the specific choice problem under consideration here, I wish to make some remarks about how the HUD problem is compounded by the problem of measurement error. I will do so by considering the curve-fitting problem, which I take to be an interesting model or instance of the HUD underdetermination problem. Besides the heuristic value of the curve-fitting problem for discussing theory choice, and for introducing the role of simplicity therein, the present analysis bears directly on MOND-type theories, which are constructed as curve-fitting

---

and are fallible (because ampliative). But so are Lakatos's retrospective judgements, so this does not mean judgements made now are more likely to be wrong than ones made later.

problems: a gravitational action is fitted to the rotation velocities for a sample of galaxies (see, for example, Sanders 1996).

The curve-fitting problem is well-known, of course: how do we determine which of all the possible lines or curves on a graph best fits the available data points? To begin to answer this question we need to say what "fit" and "best" mean in this context. A curve "fits" a set of data points provided that (or to the extent that) it hits or comes sufficiently close to them. Since observations unavoidably contain some error, "hitting" a data point means passing through its error bars. (Figure 1 in Chapter 2, the rotation curve for our solar system, is an example of a theoretical curve that hits its data points with extreme precision.[23]) Finding a curve that hits all of the data points is (or is analogous to) finding a law that explains all of the available data. But we construct scientific laws in order to be able to predict future data as well as to explain past data, which means that a curve's (or a law's) expectation of fitting *future* data is something we should be concerned with in deciding which of the possible curves is the best one.

As Forster and Sober (1994, 8, *et passim*) argue, since we know that observational evidence is error-prone, we should expect that any curve that fits the present evidence *exactly* will turn out to fit future data poorly: such a curve is said to "over-fit" the present data. There is then a kind of trade-off between the goodness of fit to present data and the likelihood of future fit. The "best" fit curve is thus not merely the one that most closely approaches all the known data points. The best fit curve is the one constructed on the present data which is most likely to continue to successfully fit all future data. The question then is to decide which of the many logically possible statistical methods for choosing this curve is most reliable: that is, which method leads to correct theory choices most often, and when wrong leads to choices as little different from the truth as possible. Of course, it is difficult to give a definitive answer to this question, in part because we cannot know the future. The Akaike criterion discussed by Forster and Sober (see below) merely provides a probable answer to the question of best fit: theory choices made on

---

[23] This example may cause confusion in connection with the later discussion of over-fit. The solution is to notice that the curve for the inverse square law is the *simplest* curve that hits these data points well.

these grounds can turn out to be mistaken, but in the statistical long run the criterion (or some other like it the relevant respects) should be a reliable guide.

Another (more usual) way of describing the choice problem in curve fitting is this: we seek the *simplest* curve that fits the data well. The reason we ought to prefer simpler curves[24] can now be understood as arising from the problem of over-fit. In a given evidential context, more complex curves can always be found which will better fit any set of available data than simpler curves, but these more complex curves are likely to be poor predictors of future data because of their tendency to over-fit. Forster and Sober (1994, 5) propose an analogy to information theory: if the true curve is the signal, the error in the observations is the noise; overly complex curves tend to fit more to the noise rather than to the signal. Thus future data, even if (*per impossible*) free of error, are very unlikely to fall on the too-complex curve, and thus more complex curves tend to be poorer predictors of future data.

The procedure involved in solving the curve-fitting problem, according to Forster and Sober, is first to choose the family of curves which best balances the prospects of fit against over-fit, and only then to pick the individual curve from within the chosen family that best fits the available data. A formula developed by Akaike[25] proposes a precise rate

---

[24] Note that Forster and Sober interpret simplicity in the context of curve-fitting as being measured by the relative paucity of the adjustable parameters in the equations describing the family of curves to which the curve belongs. In particular, the *powers* of the equations representing families of curves partition curves into classes according to (one measure of) their relative simplicity. Thus a linear curve (of form $x = a + cy$) is simpler in this sense than a parabolic curve ($x = a + by + cy^2$). Forster and Sober (1994, 11) take pains to emphasise that on this approach the simplicity is a property of the families of curves rather than being a property of individual curves independently. Note that other approaches to simplicity in curve fitting consider the parameters of individual curves, not their families.

[25] See Forster and Sober 1994, especially page 10. Akaike's Theorem provides a solution to the curve-fitting problem in virtue of the fact that it allows one to choose the curve that, given the available data, has the highest estimated probability of future success ("closeness to truth"). In words, the theorem says that the estimated accuracy or closeness to the truth of a family of curves is equal to the inverse of the number of data points times the difference of the logarithm of the "likelihood" (that is, the probability of the data given the curve) of the best-fitting member of the family minus the number of adjustable parameters of that family. That is, *Estimated [A (family F)] = (1/N) [log-likelihood (L(F)) - k]*. This

of exchange between goodness of present fit and simplicity. One in effect compares the best fitting curve from each family, and finally settles (in a fallible but empirically corrigible judgement) on the curve that achieves the best compromise between fit and simplicity relative to the available evidence. The intimate details of this, while interesting in themselves, are not important here since I merely want to use the curve-fitting problem as a model for choice problems when the evidence underdetermines the theory, and as a way to introduce simplicity as a factor in these choices.[26]

Now we have the resources to address the point about observational error and underdetermination. Clearly, if the available observations were error-free, the true curve would be one that passes through all the data points. But given any finite set of data, an

---

> explains why fitting the data at hand is *not* the only consideration that should affect our judgement about what is true. The quantity k [measuring the number of adjustable parameters of the family of curves] is also relevant; it represents the epistemic bearing of simplicity. A family F with a large number of adjustable parameters will have a best member L(F) whose likelihood is high; however, such a family will also have a high value for k. Symmetrically, a simpler family will have a lower likelihood associated with its best case, but will have a low value for k. . . . A simpler family is preferable if it fits the data about as well as a more complex family. . . . *A slight improvement in goodness-of-fit will not be enough to justify the move to a more complex family. The improvement must be large enough to overcome the penalty for complexity* (represented by k). (Forster and Sober 1994, 11; emphasis added)

If this general strategy is correct—and it may be correct even if, as some critics have suggested, Akaike's Theorem itself is not the best equation for formalising the strategy; see Bandyopadhyay and Boik 1998—one very important thing it shows is that simplicity is not an *extra*-empirical criterion for theory choice.

[26] Rather than the Akaike criterion propounded by Forster and Sober, most often scientists use the "least squares" method of choosing a curve, which involves choosing that curve which minimizes the sum of the squares of the differences between the curve and the data points. But the curve that truly minimizes this sum will be the curve that hits all the data points exactly; that is, it will over-fit the data. Thus the least squares method is usually used in combination with a tacit or incompletely specified preference for "simpler" curves. The Akaike criterion shows why and how this works. All statistical methods for solving curve-fitting problems attempt to provide curves likely to fit future data while explaining as well as possible the available evidence. The "best fit" curve is the one constructed on the present data which is mostly likely to successfully predict the future evidence. The question then is to decide which of the many logically possible statistical rules actually best achieves this aim. Forster and Sober argue that Akaike's criterion is better for this purpose than the least-squares method and that it provides solutions to a number of philosophical and other problems related to theory choice, but they do not argue for it by showing its superiority to all possible rules for choosing curves.

infinite number of curves will pass through all the points exactly (this follows simply from the fact that the real number line is continuous; see Figure 7). We would have a choice problem in curve-fitting even if there were no error in the data. Error in the data, which is unavoidable in actual observing situations, compounds the choice problem because now fitting the data (roughly) means passing through the error bars: there is clearly yet another infinity of curves which fit the data in this sense. In one sense this does not matter, since by definition it is impossible to empirically distinguish curves that differ from one another by less than the margin of error in the data. But where this does have an impact is precisely with regard to the question of (the probability of) future fit: some of the presently empirically indistinguishable curves will be better predictors of future data than others. This is especially true where the future data falls in a region of the data-space where no other data was previously available (see below). Again simplicity in the sense discussed by Forster and Sober is a plausible candidate criterion for choosing among these curves. If we can establish some such criterion, we will have a powerful tool for *empirically* deciding between "empirically equivalent" theories.

Note that there are two kinds of "future" data, which I will call interpolations and extrapolations (relative to some previously existing data set). Both are relevant to the issue of "best fit". Future data points that fall between data points already at hand test the law in that realm to a higher degree, and thus can be used to improve the precision and accuracy with which the theoretical curve approximates the true curve in that realm. Future data that fall beyond the end points of previous data sets are most useful for constraining the class of hypotheses which are consistent with the previous data but which diverge from one another elsewhere; extrapolations thus also constrain the class of hypotheses considered as candidates for explaining the previous data. Until we have such extrapolated data, there are no strictly *evidential* constraints on the form the law takes beyond the end points of the data (see Figure 8).

To take examples of each kind of new data: (1) Further observations of planetary motions in our solar system confirm or make more exact our knowledge of the short-scale, low-velocity, weak-field limit of whatever relativistic gravitation theory is the correct one (and now that our observations are becoming increasingly accurate, planetary observations even give some information about the relativistic corrections to the limit

results). (2) In contrast, new data about the dynamics of galaxies, were it to become available, would rule out some of the previously viable candidates for the non-relativistic limit of whatever gravitation theory is the correct one. We might, for example, find some dynamical evidence that would rule out MOND at galactic scales, and this would remove it altogether as a candidate even though its predictions are observationally indistinguishable from the Newtonian limit of GR (and from the observations) at stellar system scales. So judgements of "best fit" must try to maximise the likelihood of future fit to both interpolated and extrapolated future data.

Figure 7.

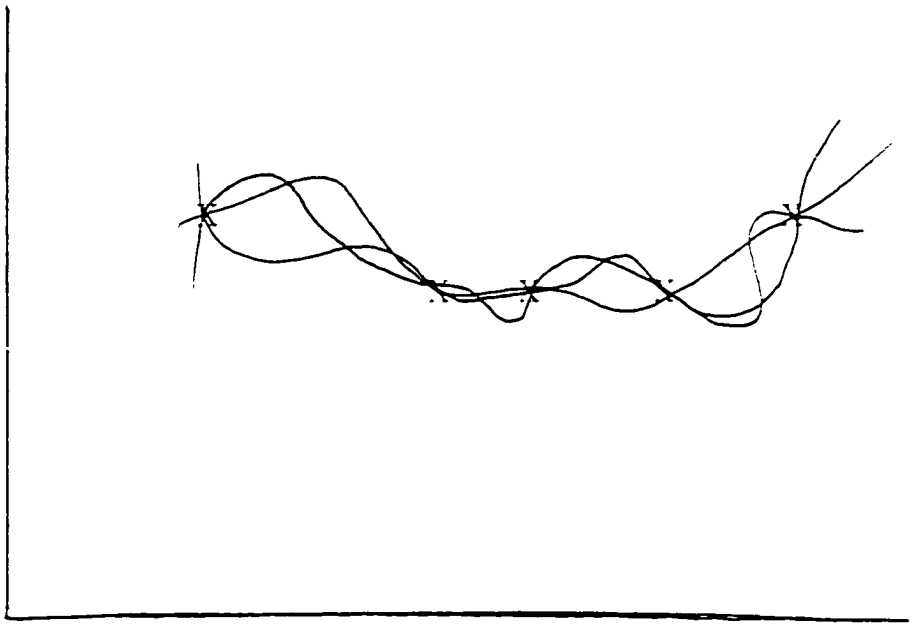


Figure 8.

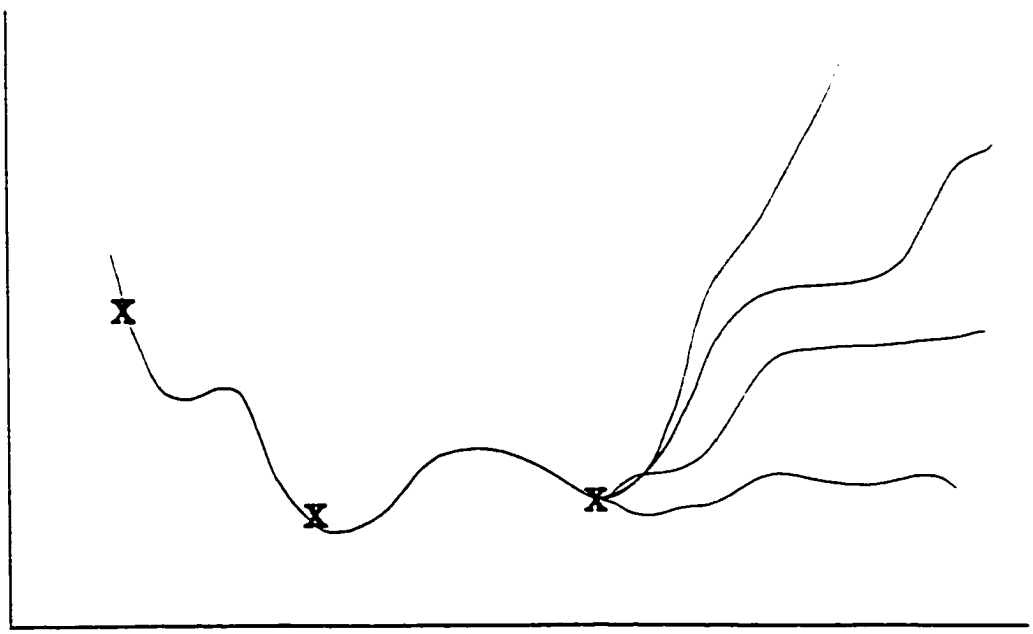


FIGURE 9.

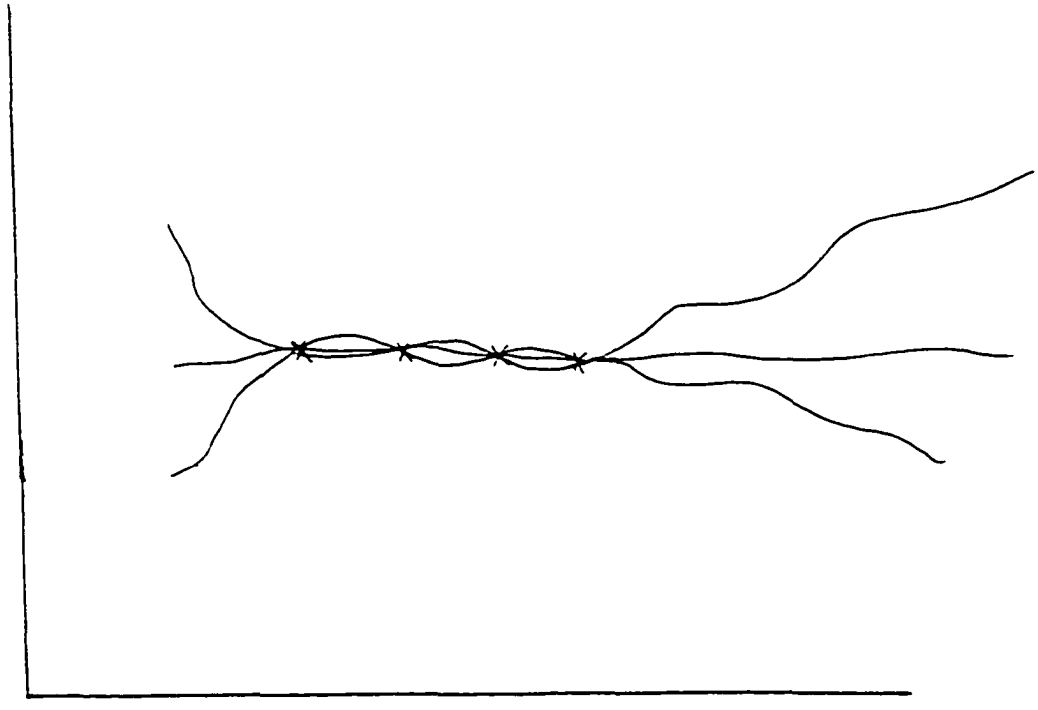
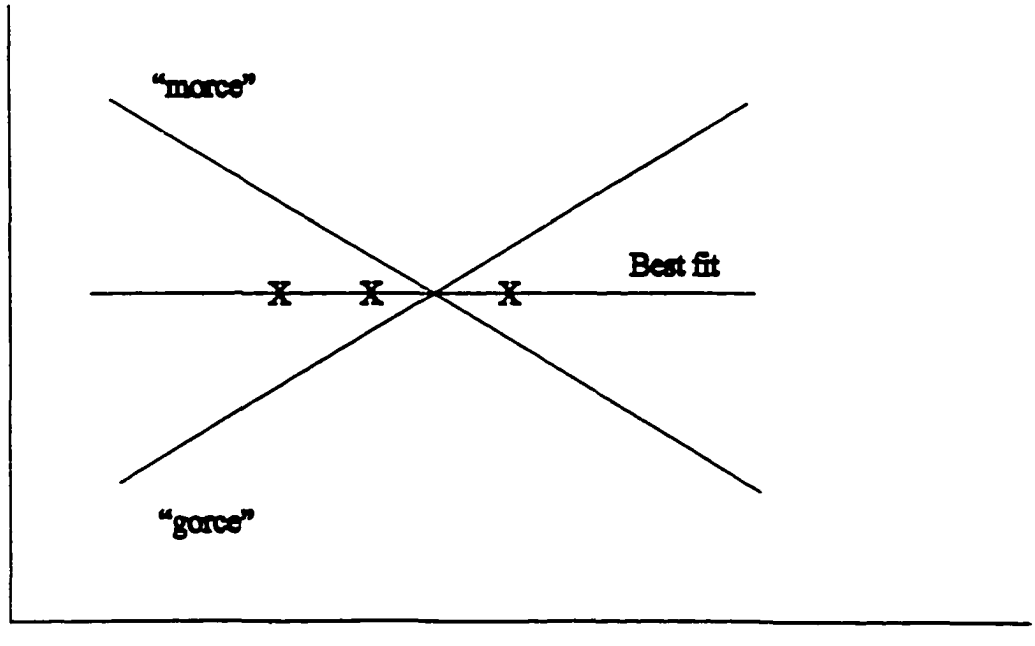


Figure 10.



In the present evidential context with regard to choosing the dynamical law operative at galactic and greater scales, we have no strong evidence in favour of any of the rivals. This means that the set of hypotheses consistent with the short and long scale evidence contains hypotheses whose predictions at long scales diverge significantly from one another (as in Figure 9). Ideally we would like to have dynamical evidence which would constrain the candidate hypotheses at large scales, but for reasons discussed above such evidence may in principle be unavailable to us. So, on what grounds, or by what rule, shall we adopt a hypothesis about large scale dynamics?

The analogy of the curve-fitting problem to the underdetermination of theory by evidence now needs to be made explicit. The data points obviously correspond to the available evidence, and the curves to the candidate hypotheses: the question of the best fit curve corresponds to the question of which of the many hypotheses consistent with the available evidence is the one we ought to pick. The question of what rule to use in picking the best curve (which is equivalent to deciding what features—for example closeness of fit to present data and simplicity—are reliable indicators of future success or closeness to truth) has an obvious correlate in the question of what rule to use to pick one from among the candidate hypotheses as the best one.[27] Unfortunately, it is not clear what the choice rule(s) should be for either problem. Determining which such rules achieve our epistemic aims is one of the great open questions in the philosophy of science. The most I can show here (see below) is that simplicity is not a useful criterion in the present evidential context with regard to the choice between GR and CTG.

One respect in which HUD underdetermination differs from the curve-fitting problem is that in HUD there are an indefinite number of ways to produce *each* set of predictions (as encapsulated in a given curve). Thus if the "best fit" curve for some data set is, for example, a straight line (as in Figure 10), there will be a multitude of different ways to produce that curve with "morce–gorce" type theories. This is relevant to the current discussion because both Milgrom and Mannheim present their theories as

---

[27] Some rules for theory choice might be more reliable than others at picking hypotheses more likely to succeed in the long run. With such a rule in hand, then, we might be able to reliably decide between rival

involving an additional action which sums together with the Newtonian action to accurately predict the galactic rotation curves without the need for dark matter, while at the same time making only an unmeasurably small addition to the predictions at shorter scales. But any number of combinations of two or more actions could be concocted so as to add up in the right way at the right scales. So even once we have settled on the "best fit" form of the *phenomenological* law summarising the data or predictions at all scales, we still have to pick which of all the possible *theoretical mechanisms* for producing those predictions is the best one. Here again we are forced to invoke criteria such as simplicity or theoretical unity if we are to make such a choice.

6.4 ASSESSING MANNHEIM'S COMPLAINTS AGAINST DARK MATTER

The choice problem with regard to CTG and GR–DM does not involve complete replacement of the entire theoretical system.[28] The two theories agree about the background information, and even about most of the auxiliary hypotheses, and they save the presently available dynamical evidence indistinguishably well.[29] The scope of the possible differences of the ultimate gravitational theory from GR, note too, is highly constrained by the available evidence. The CTG challenge hits GR precisely where it is evidentially weakest, namely, where it is forced to make untested assumptions about matter distributions, and in realms where (as I will argue) no independent tests of GR's

---

theories which are predictively equivalent on the available evidence but divergent elsewhere *even though* we have no evidence to constrain the form of the law in the regions where the rivals diverge.

[28] Note that when I write "GR–DM" I mean the conjunction of General Relativity with some specific dark matter candidate; I do not specify which one because we do not know at this point which dark matter candidates are really contenders for saving the dynamical evidence when considered in combination with GR. Nevertheless, "GR–DM" is supposed to designate (albeit abstractly) a hypothesis of the same specificity as CTG.

[29] I discuss this in more detail below, but let me here note that in the category "dynamical evidence" I include rotation curves, velocity dispersions, and potentially even arguments from the evolutions of large scale structure, but not (for example) evidence from gravitation lensing. I take dynamical evidence to be evidence based on observations of motions from which we can make the kinds of inferences discussed in Chapter 2.

applicability seem possible. Mind you, CTG suffers the same epistemic defects in that realm; its potential success cannot depend merely on GR's weaknesses in these respects. Rather, as Mannheim claims, if CTG proves to be rationally or methodologically superior to a GR–DM explanation, it will be because of *other* features of the theory than its mere predictive success with regard to the dynamical evidence it is trying to explain.

I will consider three main points of Mannheim's methodological critique of dark matter theories. (This critique is developed in several of his papers, but the nicest discussion is in his 1994.) The three points are these: (1) GR is not tested at galactic scales, which means we have no reason to reject out of hand or to fail to consider alternatives to GR. (2) The Conformal Theory of Gravity gives a *simpler* or *more unified* explanation of the observed phenomena, and so it is to be preferred over GR plus a dark matter hypothesis. (3) Dark matter theories are *ad hoc* and/or unfalsifiable, and are therefore to be rejected. I will consider each of these complaints in what follows.

Mannheim points out that our gravitational theories have only been tested at solar system distances, and in the weak field, low velocity limit. There is therefore no proof that the Newtonian limit of GR is the correct law to use in dynamical mass measurements of galaxies and clusters. Unlike in the case of the solar system, we do not have independent measures of the power law of the force of gravity at galactic distances. The idea that gravity is the same at galactic scales is not directly evidentially supported by the argument to "Universal" Gravitation (whether in the Newtonian or the Einsteinian form), because that argument only involves tests up to stellar system scales. This means that the assumption that GR applies to galaxies does not have empirical support in the sense defined by Reasoning from Phenomena: galactic rotation is not a phenomenon unified under Universal Gravitation by having been used independently to measure the power law for the force of gravity. Note that I have only said that the assumption of Universal Gravitation is in fact unjustified at the scale of galaxies, not that it is unjustifiable or definitely wrong. It is an empirical question, yet to be answered, whether the Newtonian limit or some alternative has better empirical support from galactic rotation phenomena. In the following three sub-sections I consider in more detail some of Mannheim's methodological complaints against the dark matter hypothesis, and find them wanting.

6.4.1 Principles of Theory Choice: Simplicity

Cosmology is a discipline in which explicit attention is (and must be) paid to principles of evidential reasoning and methodology. Herman Bondi makes some remarks in his *Cosmology* (1960) that are useful here in the discussion of competing theories.

> The overriding principle must be that of the economy of hypotheses, but in comparing different theories according to this principle, one must take account of *all hypotheses* involved in them, whether originally tacitly assumed or not. Replacing an old assumption with which many are acquainted by a new one (even if strikingly novel) does *not increase* the number of hypotheses required, and it is quite wrong to consider such a change a disadvantage. The value of a hypothesis depends primarily on its fruitfulness, i.e., on the number and the significance of the deductions that can be made from it, and not on whether it requires a change in outlook and is considered "upsetting". (Bondi 1960, 6; italics as original)

Bondi's point here is important. He is arguing that it is a mistake to think of newly introduced hypotheses as necessarily increasing the complexity of a theoretical framework, simply because they differ from accepted hypotheses. He expands on this in his next paragraph:

> There seems to be a widespread tendency to consider any extrapolation from observational data, however great, to be "self-evident" and therefore not as a special hypothesis to be counted in computing the number of hypotheses required by the theory. This attitude, understandable though it is, is clearly utterly mistaken. When, for example, Milne (1935) suggested that the ratio of the speeds of dynamical and atomic phenomena varied with time in a certain slow but significant manner, he was accused of having introduced an additional hypothesis. This is plainly not so: he merely *replaced* the customary hypothesis that the ratio of the speeds was the same at all times by another hypothesis, viz. that the ratio varied in a certain manner. His assumption, though new, was no more speculative than the old one. Therefore the old and the new have to be treated on the same footing. Hence both of them have to undergo equally the test as to whether they are fruitful, i.e. whether conclusions can be drawn without additional assumptions, whether these conclusions form a self-consistent scheme, and whether this scheme agrees with observation. (Bondi 1960, 6)

We may consider the MOND and CTG frameworks in this light: it is a mere assumption that the power law of the gravitational attraction is constant to arbitrarily high radius (and therefore accurately described by the Einsteinian or the Newtonian theory). As discussed above, no gravitational theory has been *tested* at distances much larger than

the solar system. The only evidential constraint on the form of a gravitational law, then, is that the law make predictions consistent with the phenomena in the range of distances that have been tested. What happens beyond that distance scale is, evidentially speaking, up for grabs. Any number of theories will make predictions that are in agreement with the short-scale tests, but these theories may differ quite significantly at other scales. The very limited information we have about the motions of bodies acting under the influence of gravity at galactic distances, say, does not provide an evidential basis for distinguishing between the alternative theories. This is because of the fact that while it is perfectly possible to use a gravitational theory to make a dynamical determination of the mass interior to a given orbiting body, it is not possible to do so *without* making some assumption about the form of the gravitational action. In the ideal case where we know a galaxy's overall mass distribution in advance, we could use the observed orbits of stars around the galactic centre to distinguish between gravitational laws that make different predictions from one another at large distances (while agreeing with each other with regard to the short-scale predictions for which we do have constraining evidence). The procedure would be rather straightforward: specify the (gravitating) matter distribution, apply each candidate gravitation theory to produce a predicted rotation curve or velocity dispersion, and compare this prediction against the observed motions. But since we have no way of knowing the mass distribution in advance—and especially since in this circumstance making *any* assumption about the mass-to-light ratio amounts to assuming exactly what needs to be empirically determined—this choice procedure is not an option. We must, therefore, find some other way to adjudicate amongst all the possible gravitational hypotheses (explicitly formulated or not) that save the phenomena at the scales where detailed tests are available and yet differ with regard to their predictions at other scales.

If it were the case that all the theories agreeing with regard to phenomena at scales for which we have good tests *also* agreed (to within the margins of difference we are able to discern) about predictions at *all* scales, then the choice problem would be somewhat less interesting than it is. The case of the alternative theories of gravity designed to solve the dynamical discrepancies without the need for dark matter are *not* like this, however, since their predictions differ significantly from GR's at galactic and greater distances.

One issue obscured by Bondi's mode of presenting the issue of theoretical simplicity is the fact that where an assumption is present in a theory, that assumption may itself be of greater or lesser complexity. Thus, while Bondi is right that the mere fact that one such assumption is replaced by another is by itself no sign of an increase in the overall complexity of a theory, this is true in general only where the two assumptions are each of the same degree of complexity. The simplicity of a theory depends not just on the *number* of hypotheses it needs to assume in order to save the phenomena, but also on the *character* of those assumptions. A linear law for the change of the ratio of the speeds of dynamical and atomic phenomena, to take Milne's theory mentioned by Bondi, will be simpler than if the new assumption is something to the effect that the ratio oscillates over long periods according to a law with higher power. It is particularly important to consider this aspect of theoretical simplicity in the case of the alternative theories of gravity proposed as solutions to the dynamical mass discrepancy. While competing alternatives may not differ in terms of the sheer number of hypotheses they each include, they may well differ with regard to their overall complexity (say, in virtue of the number of free parameters of each theory for which there is no theory to predict their values).

Questions about the simplicity of a theory do not meaningfully arise with regard to a single theory by itself. That is to say, whether or not a theory is simple is a *comparative* question, one that arises when we find ourselves forced to choose between two theories that save the known phenomena (or perhaps some restricted domain of the phenomena) equally well. According to the underdetermination thesis, we are *always* in such a situation. In practice, though, it is rare to have *even* one empirically adequate theory to hand, let alone more than one, so the practical choice problem becomes somewhat academic (although the question of the epistemic status of the accepted theory remains).

Furthermore, what degree of complexity a hypothesis has is of little moment by itself. One thing that is sometimes implied in imperatives to make theories simple is Ockham's Razor. But this principle of parsimony is just designed to eliminate from theory *superfluous elements*, to avoid *unnecessary* complication (we are not to introduce more principles or entities when fewer will do). Ockham's Razor does not advocate simplicity *per se* or for its own sake, but rather enjoins us to choose the *simplest possible theory from among those that save the phenomena.*

In part what the foregoing implies is that we cannot perform a straightforward comparison of the dynamical predictions of GR and CTG at galactic scales in order to decide between them, nor can we compare their relative complexity. For one thing, GR without a dark matter hypothesis is not capable of saving the phenomena at that scale; this much we know from the evidence, cited in previous chapters, about galactic rotation curves and so on. If (the Newtonian limit of) GR is to be involved at all in the saving of galactic rotation curves, it is only in conjunction with some specific theory of dark matter. So the conjunctive theory "GR+DM" is what has to be compared with CTG. And assessing the relative simplicity of theories that make very different assumptions (about basic physics as well as about the matter content of the universe) will be extremely difficult to say the least. Moreover, we do not know whether CTG can eventually be shown to save *all* the available data (dynamical and other), or whether some particular GR+DM model can; this means that we do not know whether the choice between these two theories *is* a choice from among the theories that save all the relevant phenomena. Thus in our present stage of knowledge it is premature to try to judge the relative simplicity of these theories, and it is therefore impossible to invoke simplicity as a principle of theory choice at the present time.

According to the Akaike framework, "a simpler family [of curves] is preferable if it fits the data about as well as a more complex family" (Forster and Sober, 11). The problem then is to determine whether the law describing the CTG action is really simpler than GR's, and whether it actually does fit the data with as much success. From one angle, it seems as if CTG might be simpler than GR, since CTG imposes an additional symmetry on gravitational interactions, thus reducing the number of degrees of freedom involved. This may, however, be the wrong way to look at the question, since the move from the second-order to the fourth-order Poisson equation represents an increase in the complexity (by the Akaike measure) of the law describing the gravitational action. And, what perhaps comes to the same thing, the introduction of the linear potential term is an increase in the number of adjustable parameters that go into the gravitational equation. Then again, CTG account reduces the number, and the number of kinds, of entities required in order to explain the dynamical phenomena. Clearly, making the judgement of

relative simplicity would require finding some way to negotiate between the various factors mentioned here.

I cannot give a good argument about whether CTG is in fact more complex in the relevant sense than GR, but I can say what our attitude toward accepting the theory should be *if* it is. If solar system tests were our only consideration, and CTG and GR are really predictively equivalent at the solar system scale (as Mannheim claims they are), then we should reject CTG in favour of GR. At this distance scale, CTG fails to satisfy the criterion of empirical success derived from Akaike's Theorem: a slight improvement in goodness-of-fit (and there is *no* improvement in this case) "will not be enough to justify the move to a more complex [theory]. The improvement must be large enough to overcome the penalty for complexity" (Forster and Sober, 11).

This is really not the right level of comparison, however, because we have more than the solar system tests to consider. Since we want to take galactic rotation into account, and GR can only do so by assuming the existence of some dark matter distribution, we have to compare CTG against GR+DM. But if CTG by itself can account for the rotation curve, it is more unified than GR+DM; CTG is probably also simpler, since the dark matter particle theories have at least several additional free parameters (including number, mass and density distribution). On those grounds, we should be inclined to favour CTG. But it may be possible to use other phenomena to constrain the nature of the dark matter, and perhaps even to measure its parameters precisely.[30] If that turns out to be the case, then once the dark matter theory has had its parameters determined, the GR+DM theory may not be much more complex than CTG, in which case we would have to turn to other considerations to choose between them. (Note, again, that we must compare CTG against a *specific* dark matter theory, and not just against the general idea of dark matter, since we need the detailed predictions of a specific theory in order to evaluate empirical success.)

---

[30] Dynamical evidence in concert with observations of extragalatic background light, particle detection schemes, and fundamental physics could together provide us with reliable RfP estimates of the total mass of dark matter, individual particle mass, number and detailed distribution.

The possibility of carrying out this simplicity analysis is contingent on the two options being empirically successful. But we do not know yet whether CTG really is empirically adequate—its predictions have not been fully explored in all domains where we have data. Nor do we know whether any dark matter theory *can* be empirically adequate. (The candidates that have not been ruled out so far *can* be used in models sufficient to explain the dynamical data, but we do not know for sure yet whether those models are consistent with *all* the available data, for example data about radiation and particle backgrounds.) For this reason, we are unable to perform the simplicity comparison in our present epistemic situation.

One important piece of information necessary for determining whether or not GR and CTG are in fact empirically equivalent, and adequate, that was lacking from Mannheim's own development of the conformal theory has recently been supplied by others. I am referring to the CTG predictions regarding "gravitational lensing", the deflection of light passing near massive bodies. We know that General Relativity has been highly successful at predicting the amount of deflection by the Sun of the light of background stars. GR has even been applied on cosmic distance scales (apparently successfully), thereby allowing us to explain the observations of background quasars lensed by foreground galaxies, among other phenomena. As it turns out, gravitational lensing poses a rather serious challenge for CTG, one that is perhaps fatal to the idea of using that theory in order to avoid dark matter.

Briefly, gravitational lensing occurs whenever a background light source lies close enough to the line of sight of an intervening massive object. The amount of the observed deflection of the image(s), as well as what counts as "close enough" alignment to the line of sight, depends on the mass of the so-called lensing body or lens and on the distance between the lens and the observer. (See Figure 6, Chapter 5.) In Chapter 5 we came across the "micro-lensing" of individual background stars by foreground stellar or sub-stellar objects within the Milky Way. Here we are considering cases where background objects such as quasars and galaxies are lensed by foreground galaxies or clusters. Depending on the precise alignments involved and the mass distribution within the lens, lensing configurations can produce luminous arcs, arclets, rings and multiple images of the background object. (Unlike in the case of microlensing (Chapter 5) since the motions

of source and lens across the line of sight are undetectably small, the images produced are for all practical purposes permanent.)

We can solve the gravitational lensing equations (of a given gravitational theory) to determine the mass of the lens—essentially, using the phenomenon of the deflection of light in a particular configuration as a measure of the mass of the lensing object. Note that I consider this to be *non-dynamical* evidence because it does not depend on the motions of bodies in the system whose mass is being determined.) In the case of background quasars lensed by spiral galaxies, the result of this measurement seems to confirm the order of magnitude of mass that is measured dynamically for typical spirals. Similarly, the observed lensing by clusters seems to require copious amounts of dark matter. It seems, then, that we have two independent measures of mass—from dynamics and from lensing—both of which give us roughly the same values for systems of the same types. The agreement of these two apparently independent measures sounds like excellent grounds for thinking we can trust what the measurements seem to be telling us, namely that there is a high proportion of dark matter in galaxies and clusters.

The details of the gravitational lensing case do, however, lead to some misgivings about whether the two measures really arrive at the same value for galactic mass, but even so it does not seem that the two could be different enough to not count as confirming each other to some degree. That they should only accidentally arrive at roughly the same (incorrect) value for the galactic mass seems rather implausible: that they could have a common systematic error is difficult to conceive, since the two methods seem to be quite independent of each other (one relies only on the Newtonian limit, while the other relies on the specifically relativistic parts of GR).

Unfortunately, there is not yet (so far as I am aware) a case where the mass of a single galaxy is measured by both methods. (Because of the geometry required for lensing, the galaxies which "lens" quasars are typically very far away, and hence too dim to allow the detailed spectroscopic work required in order to obtain rotation curves.) There is, therefore, no double measurement of the mass of any individual galaxy. However, the dynamical measurements of the masses of spiral galaxies arrive at roughly the same total mass in each case. (Within less than an order of magnitude difference; spirals seem to come in only a rather narrow range of possible masses.) Gravitational

lensing measures arrive at the same order of magnitude of mass for spirals. If CTG's gravitational lensing predictions are significantly different than GR's, then CTG will have to explain in detail how it can be that this agreement of measures can be merely coincidental. If there were a double measurement for a single galaxy, its result would weigh decisively (in the sense of a "morally crucial" experiment) in favour dark matter or against it, depending on whether the two measures agreed or not.[31]

In the case of the gravitational lensing predictions, the observations are consistent with GR provided that the visible mass in the lensing bodies is not all the mass present. (In other words, there is a discrepancy, similar to the dynamical discrepancy, between the observed amount of lensing and the visible matter distributions, given GR.) By assuming GR one can infer facts about the overall matter distribution of the lens from the characteristics of the lensed images. The best fits seem to be more or less homogeneous spherically symmetric haloes extending to several times the visible radius of the galaxy or cluster in question—which is just what the dynamical measures also indicate.

GR is consistent with the lensing observations; but it is surely not the only possible gravitational theory that is. Other theories, such as CTG, could also be consistent with the lensing observations in the same way, namely they would permit us to deduce a different overall matter distribution in these systems from the observed images. Given this situation, the only way to turn gravitational lensing into a meaningful test of GR(+DM) as compared to its rivals would be to acquire detailed independent information about the mass distribution in a lensing galaxy or cluster, and then to compare this to the predictions made from the observed lensing characteristics by assuming GR and its rivals. But because of the dark matter double bind, it seems unlikely that this independent

---

[31] Mannheim (1995a) quipped in response to a question I posed to him about the evidential force of this agreement of measures that the gravitational lensing and the dynamical measures must simply be wrong by the same amount. This is clearly an inadequate response, especially since the dynamical measures can be carried out using just the Newtonian limit whereas the lensing measures rely on the distinctly non-Newtonian parts of GR. The successor theory would have to explain this coincidence, if it is one. It is nevertheless still possible that the appearance of the independence of these coincident measures is misleading. But in any case the lensing calculations of Edery and Paranjape (1998) seem to make CTG's prospects look dim: see below.

information about the matter distribution could ever become available. (As mentioned elsewhere, the best hope for independent evidence about the mass distributions in large scale astrophysical structures, since it cannot come from dynamics, would be from theories of structure formation. But theories of structure formation have to assume a total mass content of the universe *and* the gravitational law governing the evolution.) We will need to find other, indirect means of trying to decide the question of which gravitational theory and matter distribution is correct of which gravitational theory and matter distribution is correct.

Edery and Paranjape (1998) mention that it is possible to use galactic rotation curves to set the parameters of CTG and it is also possible to use gravitational lensing by galaxies and clusters to do so. They show, however, that the results of these two attempts to allow phenomena to measure the parameters of the theory arrive at incompatible results (on the assumption that the visible matter is all the matter there is). This suggests that CTG without dark matter cannot account for all the relevant galactic scale phenomena: using the parameter values determined from fits to the galactic rotation curves, CTG *would* require large amounts of dark matter at the scale of galaxies and clusters in order to be able to save the gravitational lensing observations, just as GR does. This is an important result. Nothing in CTG rules out the possibility of having dark matter too, but the need for copious amounts of dark matter certainly goes against the motivation for seeking a gravitational solution to the dynamical discrepancy in the first place.

The details are as follows. Edery and Paranjape (1998) undertake to test the viability of Mannheim's theory by checking its predictions regarding gravitational lensing. What they find is that in CTG

> besides the usual (Einstein) deflection of $4GM/r_0$ we obtain an extra deflection term of $-\gamma r_0$, where γ is a constant and r_0 is the radius of closest approach [of the light beam]. With a negative γ, the extra term can increase the deflection on large distance scales (galactic or greater) and therefore imitate the effect of dark matter. Notably, the negative sign required for γ is opposite to the sign of γ used to fit galactic rotation curves. (Edery and Paranjape 1998, 1)

They note elsewhere that the value of γ obtained from gravitational lensing is also about equal in absolute magnitude to the value obtained from the fits to galactic rotation curves.

The values of γ obtained in both ways are consistent with the solar system data, that is to say, its value as determined by both methods is very small, so that the extra effect CTG would add to gravitational lensing results would only be noticeable over extremely large distances such as those involved in galactic and cluster lensing interactions.[32]

Edery and Paranjape (1998, 6) remind us that there is a discrepancy between the General Relativistic gravitational lensing predictions and the observations, insofar as the visible mass in galaxies and clusters is insufficient, by between one to two orders of magnitude, to account for the amount of light deflection observed. Thus considering alternatives to GR for the description of the geodesics of light rays is perhaps a good idea. The deflection angle for a light ray passing near a massive body is given by $\Delta\varphi = 4GM/r_0$ in GR, and in CTG by $\Delta\varphi = 4\beta r_0 - \gamma r_0$. Since $\beta = GM$ in the Einstein version of the calculation, the light bending predictions of the two theories differ only in γ. The magnitude of γ is constrained by the fact that GR's predictions for light bending by the Sun as seen from Earth are in very close agreement with the observations, given the mass of the Sun as determined by dynamical measures. This means that γ must be small, so that CTG does not disagree with the lensing observations in the solar system.[33] This is similar to the constraints on γ we have from solar system dynamics.

---

[32] Mannheim and Kazanas (1989) derive the vacuum solution $B(r) = A^{-1}(r) = 1 - 2\beta/r - \gamma r - kr^2$. The γ mentioned above is in both cases the constant multiplying the distance in the linear term of this equation.

[33] Edery and Paranjape write:

> To date, the best measurements of the deflection of light from the sun were obtained using radio-interferometric methods and verified Einstein's prediction to within 1%. The measured deflection at the solar limb was 1.761 ± 0.016 arc sec compared with Einstein's prediction of $4GM_\odot$ $R_\odot$ = 1.75 arc sec. Using the Weyl deflection angle [$\Delta\omega = 4\beta r_0 - \gamma r_0$] these measurements constrain the constant γ to the range $3.45 \times 10^{-19}\ cm^{-1} \geq \gamma \geq -1.8^{-} \times 10^{-18}\ cm^{-1}$. Clearly, the solar gravitational deflection experiments constrain strongly the order of magnitude of γ but leave open the possibility for a positive or negative γ. (Edery and Paranjape 1998, 13)

Measuring γ from large scale lensing to the highest possible precision would require detailed parameterised lens models because the matter distribution in galaxies and clusters is unknown and (because of the dynamical discrepancy) these systems cannot be assumed to act like point masses or to be spherically symmetric (Edery and Paranjape 1998, 13). But assuming that these systems are spherically symmetric

Edery and Paranjape conclude that "there is a glaring incompatibility between these two analyses [the one from galactic dynamics, the other from gravitational lensing]. This means that Weyl gravity [a.k.a. CTG] does not seem to solve the dark matter problem, although this does not signal any inconsistency of Weyl gravity itself" (1998, 7). Perhaps worse, if one uses the non-negative γ obtained from Mannheim's fits of CTG to galactic rotation curves in gravitational lensing calculations, the deflection angle expected from the visible mass of a galaxy or cluster will be *less* than that of GR; in other words, CTG requires *even more* dark matter than GR does in order to account for the observed gravitational lensing! My point here is twofold. First, it now seems that CTG plus just the visible matter *is not* predictively equivalent to GR+DM, which means that criteria of theory choice such as simplicity should definitely not come into play. Second, the original motivation for suggesting CTG as a rival to GR is eliminated by the fact that CTG would require even more dark matter than GR in order to account for the observed lensing. Note, however, that even if this result is correct it does not mean that *no* possible rival to GR exists which does not require dark matter. In fact, the underdetermination thesis tells us that such rivals should exist. Thus the lessons of the CTG case about the evidential status of GR and the problems of theory choice in the case of searching for a solution to the dynamical discrepancy still stand whether or not CTG itself is viable.

6.4.2 Principles of Theory Choice: *Ad Hocery and Unfalsifiability*

The very idea of postulating unseen (and perhaps unseeable) matter in order to account for galactic rotation curves has been called *ad hoc* and unfalsifiable by Philip Mannheim (1994, 493ff.), especially because it seems that dark matter theorists can alter the details of the solution at will when new observations become available (they abandon brown dwarfs in favour of some nearly-massless particle, neutrinos in favour of axions, and so on), and still retain the assumption that the problem is a *mass* problem. Yet proposing the existence of some distribution of dark mass in itself seems very much like proposing the existence of Neptune. What is the difference, if there is one, that makes the

---

yields an "order of magnitude" approximation to γ that agrees with Mannheim's value for this parameter as determined from galactic rotation curves, but which has opposite sign (Edery and Paranjape 1998, 14).

Neptune case the standard example which shows that "*post hoc* revision is not always bad" (Forster and Sober, 17), while the dark matter case is *ad hoc* in some pejorative sense? Mannheim (1994, 489) calls the prediction of Neptune "the only successful prediction of dark matter theory."[34] A supposed problem with dark matter theories closely related to *ad hocness* is the fact that that some dark matter theories require the existence of matter which may be in principle not directly observable, and this would seem to violate a Popper-style requirement for falsifiability: the non-observation of this kind of dark matter would not falsify the hypothesis.

Specifically, Mannheim argues of the various attempts to fit a dark matter distribution to the galactic rotation curves that,

> while the fits are certainly phenomenologically acceptable, they nonetheless possess certain shortcomings. Far and away their most serious shortcoming is their *ad hoc* nature, with any found Newtonian shortfall being retroactively fitted by an appropriately chosen dark matter distribution. In this sense dark matter is not a predictive theory at all but only a parameterization of the difference between observation and the luminous Newtonian expectation. As to possible dark matter distributions, none has convincingly been derived from first principles as a consequence of, say, galactic dynamics or formation theory. (Mannheim 1994, 493.)

Mannheim's main point here is that a new dark matter theory can be invented to account for almost any new data we might encounter: that sort of manoeuvre seems to him to be *ad hoc*. It is a general feature of the various attempted computer simulations of cosmic evolution and the formation of structure that roughly spherical, roughly correctly sized mass distributions form, and that they are quite similar to what is expected for dark matter halos. (Dubinski 2000; see Blanford 1997.) As discussed in Chapter 5, the particles considered in numerical simulations come in two kinds, dark matter and "ordinary" matter. The dark matter acts only by gravity, while the ordinary matter can dissipate energy through electromagnetic radiation. Thus by assuming that dark matter has this character one can calculate that haloes of the sort required (given that GR is correct) in

---

[34] However dramatic, the claim is in fact false: as Chapter 3 describes, there were successful predictions of dark binary companions to several stars in the late 1800s, and these predictions were confirmed when observational capabilities improved sufficiently. This is just more evidence that *some* cases (at least) of introducing dark matter can be highly successful.

order to explain galactic rotation curves will evolve naturally from the initial distribution. If this is right, Mannheim's complaint about deriving the distributions "from first principles" (in the physicist's sense) is less telling that one might have supposed. (Although it is worth noting again that these simulations *assume* GR as the main law governing the evolution of large scale structure.) Moreover, as I hope is clear from discussions in earlier chapters, it seems to me that the correct way to view this is to see that the galactic phenomena, when put into relation with the dynamical assumptions of the Newtonian limit of GR, can be turned into RfP-style *measurements* of the parameters of the dark matter distribution. That is, if we grant the Newtonian assumptions, the rotation curves force us to accept the existence of a spherical halo of matter ten to 100 times more massive than the total luminous matter, surrounding the visible disk and extending significantly beyond it. There is nothing *ad hoc* about this. (If there is anything epistemically dubious about the inference, it is whether there is epistemic warrant for retaining the assumption that the Newtonian limit of GR is satisfied by these systems.)

Mannheim's further complaint here that dark matter is not a predictive theory at all is very strange. It is similar to the complaint he makes when he says that the most popular model for the distribution, an isothermal gas sphere,

> is motivated by the very data that it is trying to explain. However, careful analysis of the explicit dark matter fits is instructive. Recalling that the inner region is already flat for Newtonian reasons, the dark matter parameters are then adjusted so as to join on to the Newtonian piece. . .to give a continuously flat curve in the observed region. This matching of the luminous and dark matter pieces is for the moment completely fortuitous. . .a conspiracy. (Mannheim 1994, 493.)

Of course it is the case that for a given galaxy from whose rotation curve we develop a theory of the dark matter distribution, that distribution only explains the rotation curve after the fact. The "conspiracy" of the smooth joint between the models for the inner and outer regions of the disk is merely the result of RfP measurement on the assumption that the Newtonian limit holds. The dark matter distribution arrived at by this method does indeed make novel predictions, for example about points at the extremity of the rotation curve not yet observed, and about other galaxies whose rotation curves were not part of the sample used to derive the general features of the model of the dark matter distribution.

(Specific dark matter candidates do also make novel predictions, for example about particle detections on Earth, or the intensity and spectrum of the UV extra-galactic background, and so on.)

Mannheim's complaint here is in turn related to his charge of unfalsifiability.

> Since dark matter only interacts gravitationally. . .. and since it can be freely reparameterized as galactic data change or as new data come on line, it hardly qualifies as even being a falsifiable idea, the *sine qua non* for a physical theory. . . . Since the great appeal of Einstein gravity is its elegance and beauty, using such a band aid solution for it essentially defeats the whole purpose. (Mannheim 1994, 493-4.)

There are several points to make with regard to the claim that the dark matter hypothesis is unfalsifiable. First, Mannheim's way of phrasing the accusation treats dark matter theories monolithically, but any *particular* dark matter theory *is* actually falsifiable. At a minimum, we can check to see whether its predictions succeed in capturing what is observed in dynamical systems of various kinds. If the candidate can be made to yield non-dynamical predictions—for example, about some flux of radiation or particles—we can also check for that. The hypothesis that all the missing mass is in a supermassive black hole at the centre of the galaxy, for example, can be proven false by the fact that such a thing would not affect the shape of the predicted rotation curve beyond the edge of the visible light. So, the charge of unfalsifiability against matter solutions is disproved by the fact that some proposed candidates from that class have already been ruled out on evidential grounds. Second, it is important to note that the appeal of GR is *not* its "elegance and beauty", whatever that might be, but its empirical success. It is the empirical success of GR in the realms at which it has been tested, including its survival of severe tests, and the fact that dynamical inferences to unknown matter have succeeded in the past, that invite us to make the inductive leap to the idea that GR is the right theory at all scales. This inference is natural, though perhaps not well-founded.

Perhaps Mannheim's complaint is that every time a dark matter candidate is ruled out there is always another waiting in the wings; on this interpretation what is "unfalsifiable" is the idea that the rotation curve discrepancy is due to missing *mass*, and not to having the wrong theory of gravity. We can never know *a priori* whether one or the other kind of solution will be more empirically successful in any particular case. But if we think instead of the situation as where the phenomena can be used to measure the

nature of the dark matter, we can see that the process Mannheim complains about is in fact empirically and methodologically sound: it is just the reduction in the range of viable rivals, the reduction of the error in the measurement of the nature of the dark matter.

Forster and Sober note that

> the Quine-Duhem thesis states that the core theory may always be shielded from refutation by making after-the-fact adjustments in the auxiliary hypotheses. . . . The classic example is Ptolemaic astronomy, where the model may always be amended in the face of potential refutation by adding another circle. (16-7)

We need to draw a distinction, then, between reasonable *post hoc* revision and unacceptable *ad hoc* revision, but there is no easy way to do this. "Leverrier's postulation of Neptune's existence to protect Newtonian mechanics from the anomalous wiggles in Uranus' orbit" (Forster and Sober, 17) is an example of *post hoc* theory revision which Lakatos lauds as an excellent move. For Lakatos what distinguishes good from bad *post hoc* revisions is that the good ones make novel predictions (Forster and Sober 1994, 17; Lakatos 1970). Among the novel predictions in the Neptune case are things like previously unnoticed perturbations of Saturn's orbit, and the existence of an observable body at a certain location in space.

In the dark matter case, once the parameters of a particular theory have been fixed we also have new predictions, for example, about what sorts of detectors would be required in order to observe the dark matter particle (or perhaps even that the dark matter particle is not observable by us). But more importantly, while a given dark matter hypothesis is constructed so as to be able to recover the data-so-far, there is an implicit prediction that it will be consistent with any *new* observations that might become available. Among the new situations the hypothesis must cover are things such as the extensions of known rotation curves to longer radii, or discoveries of new galaxies. So dark matter theories do make novel predictions, and therefore are acceptable *post hoc* protections of Newtonian theory according to this criterion.[35]

---

[35] There is a large literature on novel predictions, some of which requires that "novelty" cannot simply be "more of the same". It is important to note that the extension of a rotation curve to greater radii is not just more of the same: every additional parsec over which the rotation curve remains flat is a novel discovery in the strong sense, and therefore for a theory to predict that beforehand is for it to make a novel prediction in

Note that Lakatos' criterion is consistent with what Akaike's Theorem tells us, namely that *future fit* matters, and not just fit-so-far. If a dark matter theory fits the data-so-far as well or better than CTG, but fails to fit future data where CTG succeeds, its degree of empirical success goes down. Forster and Sober suggest that "a research programme is *degenerate* [in Lakatos' sense] just in case loss in simplicity is not compensated by a sufficient gain in fit to data" (17). The problem here, once again, is that neither CTG nor any specific GR–DM theory is sufficiently developed to allow us to make this sort of determination. Perhaps this means that Akaike's method is a good one in principle since it gives us a rigorous way of determining which of two theories is more empirically successful under the right conditions, but that in practice it is not very useful since in important cases (like the dynamical discrepancy) it cannot tell us anything about which theory we ought to prefer.

Mannheim's charge of *ad hocery* against individual candidate dark matter particles and against the very idea of dark matter is based on the notion that new dark matter particles are proposed just in order to address the dynamical discrepancy. There are several things that must be said in response to this.

(1) It simply is not true that dark matter particles are proposed just in order to address the dynamical discrepancy. Almost all of them have independent motivations (from astrophysics or particle theory). Admittedly, most of them do have their particular parameters adjusted so that they can exactly account for the dynamical discrepancy.

(2) Even so, some of these instances at least can be interpreted to involve the operation of Reasoning from Phenomena. The parameters of these candidates are measured by systematic discrepancies between the phenomena and the theoretical expectation (relative to plausible and widely accepted theoretical principles that in some cases even have independent warrant). That is, by assuming GR it becomes possible to allow the dynamical phenomena, for example, to measure the total mass and the shape of the distribution of dark matter. From these constraints one can construct a theory what the dark matter is, or decide between competitors. If this is "*ad hoc*" then so is Newton's

---

the strong sense, that is a prediction of a phenomenon never before observed, one that is surprising given our other theories, background knowledge, and previous experience.

argument to UG, and so is all the dynamical evidence (at various scales up to stellar system-sized interactions, including binary stars) that we have in support of GR.

(3) It is never the case that a dark matter particle is considered definitively or even moderately well supported in virtue of merely having been fitted into a scheme that appears to successfully account for a certain body of dynamical evidence. As the discussions in Chapters 4 and 5 show, once a candidate has been proposed, it must be shown to pass a plethora of independent tests. For instance, the candidate must be shown to be consistent with dynamical evidence at all scales (from the historical stability of the solar system to the motions in our own and other galaxies of various types, to clusters and superclusters). It must not emit noticeable amounts of radiation at any wavelength. Its quantity, decay rate and decay signature must also be consistent with the extra-galactic background light. And so on. Thus individual hypotheses about dark matter particles *are* subject to independent test and *are* falsifiable, even if Mannheim is correct that they are proposed specifically to save a certain body of dynamical evidence. Besides this, astrophysicists often cite the fact that some candidate particle had its first origin in particle physics or some other realm not related to the dark matter problem, and this shows that astrophysicists are aware of the problem of *ad hocery* and are concerned to avoid the appearance of it at least.

Mannheim's charge may be better founded with regard to cosmological dark matter. The succession of hot, cold, mixed and other cosmological dark matter models proposed in order to fit the preconceived notion that the universe must be at the critical $\Omega = 1$ mass density to the age and the observed level of structure in the universe often seems purely *ad hoc*. There are two things to say about this. First, the numerical modelling of the cosmological dark matter in computer simulations is epistemically problematic not just because of the *ad hocery* problem, but also because it completely ignores the underdetermination problem: in principle we should be able to construct an indefinite number of cosmological models which will evolve to final mass distributions that happen to be qualitatively similar to the structure we observe the universe to have.[36] But second,

---

[36] For example, models with different initial mass densities will evolve to have qualitatively similar kinds of structure at the present epoch provided that the relevant dynamical laws and cosmological parameters are

if Mannheim was thinking of cosmological dark matter when he made the charge of *ad hocery*, even if that charge held up it would not establish that *dynamical* dark matter is epistemically or methodologically suspect in the same way. Mannheim may not have paid enough attention to the significant differences that exist between dynamical and cosmological dark matter.

Most scientific theorising is *post hoc*: we have to know what the phenomena are before we can begin to try to give a theoretical account of them. One might, however, find reason to impose a stricture to the effect that no theory should be considered *confirmed* simply in virtue of saving the phenomena it was designed to explain. Thus saving the phenomena is a kind of minimum standard that a hypothesis must meet in order to be a viable contender: the hypothesis does not get confirmed unless it makes successful novel predictions. But note that dark matter candidates meet this standard as well. As Chapter 5 illustrates, dark matter particles are formulated (or their parameters specified) in response to empirical information, and they are always subjected to rigorous tests of the kind described above.

6.4.3 Principles of Theory Choice: Unification

Many discussions of the problem of theory choice mention the role and importance of theoretical or explanatory unification. To unify a phenomenon with some others is to incorporate that phenomenon into a scheme originally constructed as an account of other (seemingly independent) phenomena. What successful unification shows is that the apparently disparate phenomena are in fact of the same kind in so far as they are governed by the same set of laws. The value of unification has its roots in a kind of simplicity: it is more economical to use one theory instead of two to account for two phenomena. But

---

chosen appropriately. Thus, we get simulations yielding more or less the visible structure if we assume a critical mass density in matter, a high value for the Hubble constant and no cosmological constant, or if we assume a non-critical matter density, a lower value for the Hubble constant, and a "dark energy" field which dominates cosmic space curvature and contributes an effective cosmological constant. (We can find evidence for or against models that are predictively equivalent to each other with regard to their predictions about the evolution of large scale structure (and which agree with the observations) only by appealing to things other than the observed large scale structure, such as the recent observations mentioned in the appendix about the fact that the Hubble constant seems to be accelerating.)

unification is also supposed to do more than this. Friedman (1983, 244ff.) argues that a unified theory receives a bigger boost of confirmation from its successful prediction of two different phenomena than would two separate theories making the same predictions. (This differential confirmation, Friedman also points out, is one reason why theoretical laws are to be preferred over purely phenomenological laws making the same predictions.) This is an important point, one which indicates that a methodological preference for unified theories could be based on the *evidential superiority* of unified theories. Unification would in that case contribute to solving problems of underdetermination by narrowing down the class of acceptable theories: it would take us from the set of theories which make a certain set of successful predictions, down to the set of theories which make those predictions *and* have a higher degree of epistemic warrant as a result. Unification thus possibly gives us epistemic grounds for preferring a sub-class of the class of predictively equivalent theories.

Harper (1997a) also talks about an even stronger sort of unification, one operative in Newton's methodology. The diverse phenomena in this case are brought under the same set of (fairly low level) laws (for example, the Laws of Motion). Given these background assumptions, systematic dependency theorems can be proved that make such phenomena as the harmonic law for a system of planetary orbits measure the power law for the centripetal forces maintaining the bodies in their orbits to be inverse square (see Chapter 2). Then, by constructing or choosing the phenomena so that "systematic dependencies" are present, the phenomena are turned into measurements of a parameter of a higher level theory purporting to explain those phenomena. (Thus both zero orbital precession and harmonic law ratios for the solar system measure the power law of the centripetal force producing those phenomena to be inverse square: $F \propto r^n$, where $n = -2$.) By "systematic dependencies" here we mean that alternatives to the phenomena would measure (systematically correlated) alternative values of the parameter. For example forward orbital precession would yield $n < -2$, and backwards $n > -2$, by the formula $n = [(360/(360 - p))^2 - 3]$, where p is the number of degrees of precession per revolution. (See Harper 1997, and Chapter 2 above.) The margins of error in measurements of the parameter depend on the margins of error in the empirical inputs, on the degree of

probability attached to the background assumptions, and on the degree of sensitivity of the measurement inference.

Thus "unification" in the Newtonian sense means that with the same set of minimal background assumptions in place, *diverse phenomena* yield *accurate, agreeing measurements* of the *same* theoretical parameter(s) of a *single* theory that explains all the phenomena in question. The agreement is of an especially strong sort since alternatives to the phenomena would measure the parameters to have different values: clearly, then, this kind of unification and agreeing measurements provides an unusually strong degree of confirmation of the theory as well. So the fact that all the phenomena agree on this measurement provides strong epistemic warrant for the belief that, within the margins of error, the parameter really has this value universally—that is, the confirmation of the theory provides warrant for the inductive generalisation of the measured parameter value to further cases. Of course this step of inductive generalisation remains, as always, fallible—and the more dissimilar the further cases are from the paradigmatic cases, the riskier the generalisation. With the advent of a new phenomenon, we might falsify the universalised claim. But the fact that we have *detailed* agreement, and the fact that we let the phenomena decide what values theoretical parameters are to take, rather than inventing and universalising them arbitrarily or *ad hoc*, seems to increase the chances of the success of the generalisation. Moreover, on this method, and in contradistinction to hypothetico-deductive methods, if the RfP-derived theory is ultimately falsified, we know with certainty that it is nevertheless *approximately* predictively correct (at least for phenomena sufficiently similar to the ones used to perform the unificatory argument), so that any successor theory must reproduce and explain all of the original theory's empirical successes, and with at least the same level of empirical success. (See the discussion of Newton's fourth Rule of Reasoning in Chapter 2, and more below).

This sort of unification is considerably stronger than the usual (H-D) sort:

> On this view the sort of unification we have been discussing is an empirical virtue, not merely a pragmatic desideratum to be applied after the empirical demands have been met. This is one respect in which the methodology of actual scientific practice differs from an hypothetico-deductive model which would measure empirical success only by global fit with the data. (Harper, Bennett and Valluri 1994, 132)

What this sort of unification does is enable us to give *empirical determinations* of theoretical parameters, relative to fairly innocuous background assumptions (but see Smith 1999b: Newton's Laws of Motion *do* have independent confirmation, so they are not unjustified assumptions). This is clearly better than H-D, where we would simply propose some theory and arbitrarily fix the values of its parameters, and check to see whether by using this theory we are able to predict the observations. H-D is susceptible to well-known difficulties of confirmation by positive instances, not the least of which has its source in the Humean principle of underdetermination. The truth of HUD shows that merely saving the phenomena is an insufficient standard for rational theory acceptance. Recognition of HUD and the apparent failure to develop detailed, adequate and convincing standards for ampliative theory choices, has led various philosophers of science to adopt strong versions of conventionalism, instrumentalism and epistemic relativism, because they believe that we cannot do better than H-D given HUD. But RfP shows that in at least some evidential situations a higher standard of empirical success *is* achievable. The question then is whether the dark matter debate is an evidential situation in which Newton's higher standard of empirical success can be fulfilled in the service of theory choice.

If we take Newton's argument for the "universality" of his law of gravity as the ideal way to establish such a theory, then it would seem that the standard ways of investigating large scale structure and the motions of clusters and galaxies proceed in the wrong direction. Newton starts with a minimal and plausible set of background assumptions in his Laws of Motion, some (relatively unobjectionable) Rules of Reasoning, and from there finds or constructs phenomena of various types and at several distance scales to measure the parameters of the gravitational action. When he finds through this process that each phenomenon measures the power law to be inverse square, he makes the inference (an inductive generalisation relying on his Rules of Reasoning) that since all these phenomena are found to have or to be effects satisfying a single theoretical description, we should attribute to them the same cause. Thus the universality of the gravitational attraction is (supposed to be) established.

Of course, Newton's universe was quite different from our own: his argument proves that his law of gravity applies to terrestrial phenomena and to motions in our solar

system. Newton speculated that stars also mutually gravitate with one another. Newton's cosmology did not recognise any dynamical structures at larger scales. However, ours does, so in principle we ought to try to use RfP to check whether phenomena in those larger structures also measure the parameters of GR to be the same as solar system tests do. Only if they do should we make the induction to the true universality of the gravitational action. To be more precise, *if* we can find such phenomena, then we have to determine whether the parameter measurements from phenomena at different scales agree. If they do agree, we can perform the next step in an argument analogous to Newton's argument for Universal Gravitation. If the measurements do not agree then we can reject the theory. However, it might not be possible to find the kinds of phenomena required in order to perform this comparison. This depends on how the world turns out to be, as well as on our state of knowledge (both observational and theoretical). In a case where we are unable to perform the parameter measurement for phenomena at the next scale (or level of dynamical structure), then we should generalise the theory whose parameters are best RfP measured by shorter scale phenomena, so that we do not allow "mere hypotheses" to defeat an empirically well tested theory.

The method of constructing models of the universe which assume a law of gravity (Einstein's) and a matter distribution (homogeneous, isotropic and $\Omega = 1$, for example), and then checking to see whether the model that results "looks like" our universe, was appropriate—because we had no other way of proceeding—when our state of knowledge about galactic dynamics and large scale structure was essentially non-existent. (Before we knew of the existence of galaxies external to the Milky Way, and before Hubble's observations challenged the assumption that the universe is static, there was no reason to try to look for dynamical effects anyway.) Thus, for the original developers of the standard (FRW) cosmological models, the hypothetico-deductive approach was acceptable, if not very enlightening. But given the detailed observations that have become available to us, especially in the last two decades, regarding the dynamics of galaxies and clusters, and given fairly detailed models of structure formation now constrained by the cosmic background radiation, the Hubble-age of the universe and observations of present-day structure, an alternative and methodologically superior approach is in principle open to us: namely Newton's.

Newton's approach would enjoin us to try, if possible, to use dynamical phenomena at galactic and greater scales to measure the parameters of a gravitational theory; only if this can be done, and only if the results come out right, will we have dynamical evidence for taking gravity to apply to all bodies in the universe and at all times, according to the same law. This is a pattern of reasoning that could in principle be employed in support of alternative laws of gravity whose actions differ from those of GR with regard to predictions beyond the scale of the solar system. But to what extent can we hope to bring such a program to completion? The main stumbling block to this approach is that in order to apply the RfP framework to phenomena at galactic and greater scales, one would need already to know the masses and motions involved. But the nature, amount and distribution of unknown matter in large scale astrophysical systems is just what is in question. One could try the RfP approach by *assuming* some value for the dark matter content of galaxies (for example) and then letting the observed motions measure the parameters of the law of gravitational interaction. But unless there are strong constraints on the amount and distribution of mass present in the systems studied, many very different results could be obtained in this way. Unfortunately the constraints are rather weak. The visible mass (which, recall, is independent of any theory of large scale dynamics) does provide a firm lower limit for the mass content of galaxies and clusters. One upper limit on the mass is obtained by assuming the Newtonian action, but that begs the question with regard to the approach suggested here.[37] The dark matter double bind described above thus suggests the principled impossibility of carrying out for galaxies and other large scale structures the sort of Newtonian unification described here.

We would have similar reason to complain against Newton's argument from solar system motions except that there are several mutual consistencies that act as checks on the masses of the bodies under consideration (from perturbations of one planet on several others, and on comets, and from the motions of moons around primaries, not to mention artificial satellites and probes). There is an urgent need, then, for an independent check

---

[37] In principle one could adopt a non-Newtonian low-velocity limit which required even more dark matter than the Newtonian assumption requires, but so far no one has proposed that we ought to abandon GR *and* introduce huge amounts of dark matter!

on the masses of galaxies, which can be used to constrain the masses assumed in an RfP attempt to measure the parameters of the gravitational interaction.

As mentioned above, gravitational lensing may provide just such a check on the masses of galaxies. One worry, however, is that the agreement of the gravitational lensing and dynamical masses of galaxies is merely an artifact. In order to rule this out, we would have to show that the two methods are truly independent of each other, that is, that the assumptions of one methods do not automatically force it to agree with the outcome of the second method. On the one hand it is plausible that the two methods are indeed independent, since dynamical measures rely only on the Newtonian limit of GR while lensing calculations depend on the specifically relativistic parts of GR. But on the other hand the Newtonian limit is the limit *of GR*, so perhaps the two ways of measuring mass are not independent.

Barring the possibility of proving the clear independence of these two methods of measuring mass (which, if independent, would provide strong warrant in favour of GR and thus in favour of a matter solution to the dynamical discrepancy), what other possible paths could lead to an evidential decision in favour of one class of candidates, or one of the rivals within a class? One recourse available to us would be to try to gather non-dynamical evidence for a particular overall mass distribution in astrophysical systems. One way to do this would be to detect a dark matter particle in an Earth-bound detector, and to then make statistical arguments about its distribution in our galaxy, and by analogy in others. If this distribution plus the visible matter allowed us to use the Newtonian limit of GR to deduce the observed dynamics, we could be fairly confident that GR applies to galaxies. This would not be definitive proof, however, because the statistical arguments from the particle detections to the distributions could be mistaken, or because some other DM particle is contributing even though we have not yet detected it.

Because of the dark matter double bind, dynamical evidence (of the sort available to us) at galactic and greater scales cannot by itself weigh in favour of any gravitational theory (and therefore any theory of the overall matter distribution). And it could turn out not to be possible to use gravitational lensing to provide and independent check on the masses of galaxies. Even so, we are not lost. Other sorts of non-dynamical evidence could possibly become available which would allow a strictly evidential (although of

course ampliative) decision in favour of some particular solution to the dynamical discrepancy.

If this evidence is not forthcoming, and until we have it, we need some basis according to which to make provisional judgements about what lines of research to pursue. Below I sketch an argument in favour of retaining GR at galactic and greater scales. This is to privilege matter solutions to the dynamical discrepancy, but since this privileging is not based on direct evidence we should not altogether rule out or stop pursuing to some extent gravitational solutions as well.

6.5 Conclusions

One "loose end" in the argument to Universal Gravitation is the part of the "universalisation" that extends the law to other, ever more distant bodies and their minutest parts, bodies *not shown* to obey or be consistent with Newtonian gravity. This part of the inference Newton supports by an appeal to his third Rule of Reasoning: "*The qualities of bodies, which admit neither intension nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever.*" In the discussion of this rule he writes:

> [I]f it universally appears, by experiments and astronomical observations, that all bodies about the earth gravitate towards the earth, and that in proportion to the quantity of matter which they severally contain; that the moon likewise, according to the quantity of its matter, gravitates toward the earth; that, on the other hand, our sea gravitates towards the moon; and that all planets mutually gravitate one towards another; and the comets in like manner towards the sun; we must, in consequence of this rule, universally allow that all bodies whatsoever are endowed with a principle of mutual gravitation. (Newton 1995 [1726], 321)

Since the Rules were set out *as* Rules only in the Second and Third editions of the *Principia*, there is a certain air here of Newton invoking, *ad hoc*, just those principles of reasoning which will justify the conclusion about UG he has already reached. Nevertheless, the Rules are probably sound methodological principles; they are, at least, fruitful guidelines for pursuing empirical research in some fields. As Smith (1999a, 1999b) puts it, following the methodology of the Rules allows Newton to put all of the epistemic risk of the hypothesis of universal gravitation into the step of its inductive

generalisation: Newton showed in detail that all the bodies in the solar system known to him (the Moon excepted) obey his law of gravitation up to the margins of error in the positional data available to him. (As mentioned above, later attempts beginning with Clairaut and culminating in the Hill-Brown lunar theory did succeed in bringing the Moon under the rubric of Newtonian theory.) In fact, the solar system phenomena measure the parameters of the gravitational theory, and thus achieve a very high degree of empirical success. In particular, reasoning from phenomena shows that departures from inertial motion in the solar system result from mutual gravitational attractions between the planets obeying the law $F = GmM/r^2$. And thus the existence of perturbations, which at first seemed to threaten Newton's theory, in the end became very strong support for it. Likewise if the predictions of the existence of dark matter are found to be correct, that fact would provide extremely strong confirmation of GR.

Rule 3 gives grounds for extending by inductive generalisation what we know by experiment about some bodies, to all other bodies. As with any inductive rule, the inductive principle involved in Rule 3 will give strong support to the extension of the law to cases that are similar to the cases with which we are familiar. The more things the new cases have in common with the known cases, the more likely it will be that the generalisation is close to correct for those cases. However, where the unknown cases are very unlike the known cases, the grounds for the generalisation will obviously be weaker; the greater the differences between new and known cases, the less likely it is that the generalisation will be correct. This point posed no particular problem in Newton's cosmology, since he was unaware of any bodies or dynamical systems that were very different from those on which he had used Reasoning from Phenomena. Stars, he conjectured, are bodies like our Sun; the universe, he thought, is infinite in extent and filled with stellar systems like our own. But in fact there are levels of dynamical structure in the universe of which Newton had no inkling. Since galaxies, clusters, superclusters, domain walls, and other structures are very unlike any of the bodies on which Newton's argument to Universal Gravitation is founded, we have grounds for doubt about whether the laws of gravitation that hold in the regions and for the types of bodies which we are able to study will hold truly universally.

In addition to this sort of inductive skepticism based on a lack of relevant similarity between known and new cases—which suggests not that universalising the properties of gravitation will be false, but that more evidence is needed in order to adequately establish just what gravitational laws do govern higher levels of structure—there is another (by now familiar) worry that can be raised. It is one based on the notion that the law governing gravitational interaction might differ with the distance scale involved.

One way we might have tried to justify retaining GR (or rejecting it in favour of some other hypothesis) would have been to attempt a unification of the sort mentioned above: coming up with an empirically adequate unified explanation of gravitational phenomena at all distance scales, had it been possible, would have led to evidential support for that unified theory over other rival explanations. However it seems very hard to judge, especially in our restricted evidential position, whether GR plus some theory of exotic matter is more "unified" than an alternative gravitation theory plus a hypothesis to the effect that the only matter is the known matter. Certainly, in the present state of things, neither of these options comes close to satisfying Newton's stronger form of unification. We lack sufficient detailed information about the dynamics of large systems, and this could be a permanent predicament. (If we *did* have further reliable information about the overall matter distribution, then we might be able to do an RfP unification for some law of gravity using the astrophysical phenomena.)

It seems then that none of the popular principles of theory choice considered here (simplicity, *ad hocness*, falsifiability, unification) are of much use for constraining possible solutions to the dark matter problem in the present evidential situation. Attempts to invoke such principles as grounds for preferring one candidate solution over the rivals are premature. The prospects for future evidence make higher order and non-dynamical evidence the most likely sources of improved knowledge in this area.

Are there any methodological principles which would allow us to make provisional theoretical decisions despite the lack of this desired further evidence? Consideration of Newton's Rule Four leads to one possible answer:

> In experimental philosophy, propositions gathered from phenomena by induction should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions. [Newton

> comments:] This rule should be followed so that arguments based on induction may not be nullified by hypotheses. (Newton 1999 [1726], 796)

This is a brave pronouncement: it seems to say that success in the RfP methodology within some limited realm makes the universalisation of a hypothesis empirically better supported than any rival, unless very strong counter-evidence comes along. Really, though, Newton is not claiming that Universal Gravitation *is* the correct theory at all scales for all phenomena: rather, he is proposing to use UG until some rival can be *shown* to do *at least as well* with regard to the standard of RfP empirical success. More exactly, Newton wants to preclude the possibility of utterly untried hypotheses being thought of as legitimate rivals to RfP-tested theories. So Newton lays it down that a theory's having been successfully RfP-tested in a limited realm gives that theory pride of place in application to broader realms. Furthermore, *if* the old theory *is* successful in the new realm, it will receive a huge boost of confirmation, making it much better confirmed than a "pretender" theory would have been even if predictively adequate. And if there is a discrepancy between the predictions of the old theory and the observations in the new realm then, as Smith argues, that discrepancy itself can become higher order evidence for a successor theory which *is* RfP confirmed. So the kind of theoretical conservatism embodied in Rule Four has as its motivation the best chances of increasing the epistemic warrant of our empirical knowledge. Rule Four is a kind of regulative ideal for scientific investigation.

By Newton's Rule 4, on Harper's interpretation (1997, 1999, and elsewhere), a legitimate rival (something more than a "mere hypothesis") is a competing hypothesis that meets or exceeds the level of empirical success (of the sophisticated kind) that the entrenched theory has. A theory must show potential to meet this high ideal in order to be considered seriously as a rival to a well entrenched theory, and the higher the degree to which it achieves this ideal, the more seriously the alternative hypothesis should be taken.[38] Clearly, then, the theory that meets this ideal to the fullest extent in its realm of application is the one to be preferred over its rivals.

---

[38] This requirement does not limit the possibility of developing new theories, it just sets down a standard that new theories must meet in order to eventually be considered legitimate rivals to theories whose evidence makes them satisfy Newton's ideal of empirical success already. This requirement is designed to

This is persuasive in realms within which the dominant theory *has* had its parameters measured in the RfP way. But it is not clear to me what force this way of arguing has in realms in which the theory has *not* been tested or had its parameters measured in this particular way. The argument outlined in the discussion of Newton's Rules in Chapter 2 gives a reason for supposing that systems beyond our ability to test will have properties like those of systems within the reach of our experiments. Newton's Rule 4 enjoins us to not let theories be defeated by mere hypotheses, and to accept the laws obeyed by all bodies subject to experiment to be the laws of all bodies whatsoever. But surely this supposition has a *much* weaker epistemic status than does the argument from similarity of effects to the similarity of causes with regard to those systems whose parameters *have* been measured from the phenomena in the relevant way.

As I have argued here, GR+DM and CTG have comparable empirical warrant on the available dynamical evidence. Because of the dark matter double bind, no simple dynamical evidence can distinguish these rivals epistemically. Other kinds of indirect, higher-order and non-dynamical evidence could possibly become available which would bear on the issue. The prospects for gravitational lensing providing the sort of non-dynamical evidence sufficient for selecting a theory of gravity at galactic and greater scales were discussed above. (I also discussed the only extant account of lensing in CTG, and mentioned results which seem to remove the motivations for considering CTG as an alternative to dark matter, but which certainly do not remove CTG from the competition altogether.) To take another sort of example, the discovery of a hitherto unnoticed radiation background could possibly provide support for some particular dark matter candidate, and once that model is properly specified we can deduce from the observed dynamics which gravitational law is correct.

This said, in our present epistemic situation no such non-dynamical evidence is available. Given the state of the present evidence, one possible reason to prefer GR over

---

prevent entrenched theories with strong evidential bases from being overturned by what Newton calls "mere hypotheses". Because of the problems related to the ampliative nature of all of our scientific theories, we can never be sure that even a theory which does meet Newton's ideal of empirical success is correct, and thus we should always be prepared to consider and develop rivals to entrenched theories, since they could turn out to be even more empirically successful than the theories we have now.

rivals for phenomena taking place over galactic and greater scales is a k' id of theory-conservatism, and this is perhaps licensed by Newton's Rule Four.

If Rule Four is correct, the best chance for improving our knowledge of the dynamics of large scale systems is to provisionally employ GR to analyse those systems, even though GR in fact has no confirmation in that realm, and despite the evidential underdetermination this implies for our choice of a gravitation theory in that realm. If the theoretical and observational processes of developing and testing dark matter candidates continues to come up negative, we may reach a point where all the likely matter solutions have been exhausted, in which event we should start to take gravitational solutions more seriously.

Despite the argument for theoretical conservatism developed here, and even though we have not yet exhausted the matter candidates, we should still make room for new gravitational theories. There is a practical question about how to divide scarce research resources. How much should we put into theories that seem like extreme long shots from our present epistemic situation? I cannot give an explicit answer to this question—there probably cannot be any hard and fast rules—but the following considerations bear on the issue, and specifically on how we should treat candidate gravitational solutions to the astrophysical dynamical discrepancy. As I have interpreted it, RfP and Newton's Rule 4 give new theories a standard of success to aim at. A theory which does not meet this ideal of empirical success is not a viable contender to replace a theory that does. Nevertheless, we should not rule out the possibility of the entrenched theory being superseded by preventing possible rivals from being developed. No matter what the state of evidence for the dominant theory, and especially when a major unresolved discrepancy exists, it will always be possible for the dominant theory to be replaced by a successor which satisfies the Newtonian ideal more fully or to a higher degree. For this reason, I think that some fraction of the research resources devoted to the dark matter problem *should* go into the development of alternative theories of gravity, even though none of the available alternatives to GR has strong warrant at this point. The possibility of a gravitational solution to the astrophysical dynamical discrepancy should not be dismissed out of hand. Moreover, given the present evidential context the astrophysics community should probably give more credence to the viability of alternatives to GR. In the present

evidential context, the *only* possible support for GR at galactic and greater scales must come from some methodological principle like Rule 4. Such a rule could certainly turn out (in general, or in a particular case) to provide unsound methodological advice. Note, too, that given the failure after an exhaustive search to identify *any* direct evidence for dark matter (and in the absence of some well-founded theoretical reason to expect to be unable to directly detect it), it *would* be possible to do an RfP inference to some alternative gravitation theory—we see the proponents of MOND and CTG attempting this already by letting rotation curves determine some of the parameters of those theories.

In this chapter I have discussed a cluster of issues surrounding the problem of theory choice, in particular as it applies to the choice of which dynamical theory to use in the solution to the astrophysical dynamical discrepancy. I described two gravitational theories proposed as alternatives to dark matter, and found them insufficiently supported to count as legitimate rivals to GR. But I also noted that GR itself does not have direct evidential support from interactions taking place over galactic and greater scales. I argued that some principle such as Newton's Fourth Rule of Reasoning is necessary in order to ground the provisional acceptance of the applicability of GR at these scales; this line of argument can in turn be taken to suggest that most but not all of the research resources devoted to this problem should be directed towards finding matter rather than gravity solutions. As an answer to the threat of strong ampliative underdetermination and radical holism I sketched an expansion of Duhem's theory of "good sense" on which it is not impossible to gather evidence and use principles of ampliative inference to make reasoned "morally definitive" (albeit fallible) theory choices. I also described evidence that, were it to become available, would enable us to make an evidential decision (according to these principles of ampliative theory choice) in favour of either GR or CTG (depending on what the evidence turns out to be). For example, if the observations of gravitational lensing can be developed more completely, they could lead to a strong reason to prefer GR over CTG or other gravitational solutions to the dynamical discrepancy. Such an evidentially based choice would still be fallible, but the inference might be probable enough to satisfy us epistemically, depending on just what the evidence turns out to be. As I pointed out, the existence of the dark matter double bind implies that the best hope for finding the sort of evidence required in order to make a

epistemically respectable decision between matter or gravity solutions is likely to be found in non-dynamical evidence. At present, though, this desired evidence is not available, and this is one reason for the failure of Mannheim's methodological critique of dark matter i. terms about simplicity, *ad hocery*, falsifiability and unification.

The short story is that evidential reasoning is more subtle than even some of its philosophical proponents (for example, hypothetico-deductivists) have supposed. And contrary to the claims of some of its opponents (Kuhnians, radical epistemic relativists of all stripes), not only is evidential reasoning *not* impossible or powerless in the face of the problem of underdetermination, but we have seen here both: (1) how evidential reasoning could—when (or if) more empirical information becomes available—provide a highly epistemically warranted solution to the astrophysical dynamical discrepancy; and (2) how methodological rules guide us in the present evidential situation to provisionally accept the applicability of GR at galactic and greater scales, and therefore to pursue matter solutions as the best hope for maximising the empirical support of our account of astrophysical systems.

BIBLIOGRAPHY

Aaronson, Mark. (1983). "Accurate Radial Velocities for Carbon Stars in Draco and Ursa Minor: The First Hint of a Dwarf-Spheroidal Mass-to-Light Ratio." *Astrophysical Journal*, **266**, L11-15.

Achinstein, Peter. (ed.) (1983). *The Concept of Evidence*. Oxford: Oxford University Press.

Achinstein, Peter, and Owen Hannaway. (1985). *Observation, Experiment, and Hypothesis in Modern Physical Science*. Cambridge, MA: The MIT Press.

Alcaniz, J.S., and J.A.S. Lima. (1999). "New Limits on Ω_Λ and Ω_M from Old Galaxies at High Redshift." *Astrophysical Journal*, **520**, L87-L90.

Alcock, C., *et al.* (1997). "The MACHO Project Large Magellanic Cloud Microlensing Results from the First Two Years and the Nature of the Galactic Dark Halo." *Astrophysical Journal*, **486**: 697-726.

Alcock, C., *et al.* (1993). "Possible Gravitational Microlensing of a Star in the Large Magellenic Cloud." *Nature*, **365**, 14 Oct 93, 621-22.

Aubourg, E., *et al.* (1993). "Evidence for Gravitational Microlensing by Dark Objects in the Galactic Halo." *Nature*, **365**, 14 Oct 93, 623-25.

Audouze, J. and J. Tran Thanh Van. (eds.) (1988). *Dark Matter: Proceedings of the XXIIIrd Rencontre de Moriond*. Gif-sur-Yvette, France: Editions Frontieres.

Babcock, H.W. (1939). "The Rotation of the Andromeda Nebula." *Lick Observatory Bulletin*, 498, 41-51.

Babcock, H.W. (1938). *On the Rotation of the Andromeda Nebula*. Ph.D. dissertation, University of California at Berkeley.

Bahcall, John N. (1984a). "K Giants and the Total Amount of Matter Near the Sun." *Astrophysical Journal*, **287**, 926-44.

Bahcall, John N. (1984b). "The Distribution of Stars Perpendicular to a Galactic Disk." *Astrophysical Journal*, **276**, 156-68.

Bahcall, John N. (1984c). "Self-consistent Determinations of the Total Amount of Matter Near the Sun." *Astrophysical Journal*, **276**, 169-81.

Bahcall, John N. (1983). "The Ratio of the Unseen Halo Mass to the Luminous Disk Mass in NGC 891." *Astrophysical Journal*, **267**, 52-61.

Bahcall, John N., and Jeremiah P. Ostriker. (1997). *Unsolved Problems in Astrophysics*. Princeton, NJ: Princeton University Press.

Bamford, Greg. (1996). "Popper and his Commentators on the Discovery of Neptune: A Close Shave for the Law of Gravitation?" *Studies in the History and Philosophy of Science*, **27.2**, 207-232.

Bandyopadhyay, Prasanta S., and Robert J. Boik. (1998) "The Curve-Fitting Problem: A Bayesian Rejoinder." <http://scistud.umkc.edu/psa98/papers/bandyo.pdf>. Philosophy of Science Association 1998 Biennial Meeting.

Bartusiak, Marcia. (1993). *Through a Universe Darkly: A Cosmic Tale of Ancient Ethers, Dark Matter and the Fate of the Universe*. Avon Books: New York.

Benacerraf, Paul, and Hilary Putnam. (eds.) (1964). *Philosophy of Mathematics: Selected Readings*. Engelwood Cliffs, NJ: Prentice-Hall.

Bekenstein, Jacob, and Mordehai Milgrom. (1984). "Does the Missing Mass Problem Signal the Breakdown of Newtonian Gravity?" *Astrophysical Journal*, **286**, 7-14.

Berry, Arthur. (1961 [1898]). *A Short History of Astronomy: From the Earliest Times Through the Nineteenth Century*. New York: Dover Publications, Inc.

Bertoloni Meli, Domenico. (1993). *Equivalence and Priority: Newton versus Leibniz*. Oxford: Clarendon Press.

Bertotti, B., R. Balbinot and S. Bergia. (eds.) (1990). *Modern Cosmology in Retrospect*. Cambridge: Cambridge University Press.

Binney, James, and Scott Tremaine. (1987). *Galactic Dynamics*. Princeton, NJ: Princeton University Press.

Blaauw, Adriaan, and Maarten Schmidt. (1965). *Galactic Structure*. Chicago: University of Chicago Press.

Blanford, R.D. (1997). "Unsolved Problems in Gravitational Lensing," in Bahcall and Ostriker (1997), 93-108.

Blout, B.D., E.J. Daw, M.P. Decowski, Paul T.P. Ho, L.J Rosenberg, D.B.Yu. (2000). "A Radio Telescope Search for Axions," astro-ph/0006310.

Brada, R., and M. Milgrom. (1999). "The Modified Newtonian Dynamics Predicts an Absolute Maximum to the Acceleration Produced by 'Dark Halos'," *Astrophysical Journal*, **512**, L17-8.

Brackenridge, J. Bruce. (1995). *The Key to Newton's Dynamics*. Berkeley: University of California Press.

Brody, Baruch A. (1970). *Readings in the Philosophy of Science*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Brown, James Robert, and Jurgen Mittelstrass. (1989). *An Intimate Relation: Studies In The History And Philosophy Of Science: Presented To Robert E. Butts On His 60th Birthday*. Dordrecht: Kluwer Academic Publishers.

Börner, Gerhard. (1992). *The Early Universe: Facts and Fiction*. Second Enlarged Edition. Berlin: Springer-Verlag.

BOOMERanG experiment. (2000). http://oberon.roma1.infn.it/boomerang/

Burrows, Adam, and James Liebert. (1995). "Probing Dark Matter." *Nature*, **373**, 19 Jan 95, 191-2.

Carnap, Rudolf. (1995 [1966]). *An Introduction to the Philosophy of Science*. Edited by Martin Gardner. New York: Dover Publications, Inc.

Carnap, Rudolf. (1964 [1956]). "Empiricism, Semantics and Ontology." As reprinted in Benacerraf and Putnam (1964), 233-248.

Carnap, Rudolf. (1960 [1937]). "Elementary and Abstract Terms." As reprinted in Danto and Morgenbesser (1960), 150-158.

Carr, Bernard. (1993). "Dark Matter Strikes Again," *Nature*, **363**, 06 May 93, 16-7.

Chandrasekar, S. (1967). *An Introduction to the Study of Stellar Structure*. (Originally published 1939, University of Chicago Press.) New York: Dover Publications, Inc.

Churchland, Paul M., and Clifford A. Hooker. (eds.) (1985). *Images of Science: Essays on Realism and Empiricism, with a Reply from Bas C. van Fraassen*. Chicago: University of Chicago Press.

Clerke, Agnes M. (1903). *Problems in Astrophysics*. London.

Clerke, Agnes M. (1885). *A Popular History of Astronomy During the Nineteenth Century*. Edinburgh: Adam & Charles Black.

Cline, David. (1993). *Astroparticle Physics and Novel Gamma-ray Telescopes*. Bellingham, WA: SPIE.

Cohen, I. Bernard. (1985). *The Birth of a New Physics*. Revised and updated. New York: W.W. Norton and Company.

Colless, Matthew, and Andrew M. Dunn. (1996). "Structure and Dynamics of the Coma Cluster." *Astrophysical Journal*, **458**, 435-54.

Cornell, James. (ed.) (1989). *Bubbles, Voids and Bumps in Time: The New Cosmology*. Cambridge: Cambridge University Press.

Crowe, Michael J. (1994). *Modern Theories of the Universe: from Herschel to Hubble*. New York: Dover Publications, Inc.

Crowe, Michael J. (1990). *Theories of the World from Antiquity to the Copernican Revolution*. New York: Dover Publications, Inc.

Curd, Martin, and J.A. Cover. (eds.) (1998). *Philosophy of Science: The Central Issues*. New York, NY: W.W. Norton & Company.

Danto, Arthur, and Sidney Morgenbesser. (eds.) (1960). *Philosophy of Science*. New York: World Publishing Company.

Davis, D.S., and R.E. White, III. (1996). "ROSAT Temperatures and Abundances for a Complete Sample of Elliptical Galaxies." *Astrophysical Journal*, **470**, L35-L40.

de Bernardis, P. *et al.* (2000). "A Flat Universe from High-Resolution Maps of the Cosmic Microwave Background Radiation." *Nature*, **404**, 955-959. astro-ph/0004404

Dicke, R.H. (1970). *Gravitation and the Universe*. Philadelphia: American Philosophical Society.

DiSalle, R., W.L. Harper and S.R. Valluri. (no date) "General Relativity and Empirical Success." off-print.

DiSalle, R., W.L. Harper and S.R. Valluri. (no date) "Reasoning from Phenomena in General Relativity." manuscript.

Disney, M. (1986). *The Hidden Universe*. New York: MacMillan.

Dorling, Jon. (1988). "Reasoning from Phenomena: Lessons from Newton." *PSA 1988*, Vol. 2, 197-208.

Dorling, Jon. (1979). "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem." *Studies in the History and Philosophy of Science*, **10**, 177-187.

Dubinski, John. (2000). "Dynamical Evolution of Galaxies in Clusters." Department of Physics and Astronomy Colloquium, UWO, 02 February 2000.

Duhem, Pierre. (1982 [1914]). *The Aim and Structure of Physical Theory*. Translated by Philip P. Wiener from the second French edition. Princeton, NJ: Princeton University Press.

Duhem, Pierre. (1969 [1908]). *To Save the Phenomena: An Essay on the Idea of Physical Theory from Plato to Galileo*. Chicago: University of Chicago Press.

Earman, John, and Clark Glymour. (1988). "Discussion: What Revisions Does Bootstrap Testing Need? A Reply." *Philosophy of Science* 55, 260-264.

Earman, John, and Michel Janssen. (1993). "Einstein's Explanation of the Motion of Mercury's Perihelion," in Earman, Janssen and Norton (1993), 129-172.

Earman, John, Michel Janssen and John D. Norton. (eds.) (1993). *The Attraction of Gravitation*. Boston: Birkhauser.

Earman, John, and Jesus Mosterin. (1999). "A Critical Look at Inflationary Cosmology," *Philosophy of Science*, **66.1**, 1-49.

Earman, John, and John D. Norton. (eds.) (1997). *The Cosmos of Science: Essays of Exploration*. Pittsburgh, PA: Pittsburgh University Press.

Eddington, Arthur Stanley. (1994 [1917]). "The Motions of Spiral Nebulae," as reprinted in Crowe (1994), 264-68. Originally published in *Monthly Notices of the Royal Astronomical Society*, **77**, 375-77.

Eddington, Arthur Stanley. (1924). "On the Relation Between the Masses and Luminosities of the Stars," *Observatory*, **47**, 107-14.

Edelstein, Jerry, Stuart Bowyer, and Michael Lampton. (2000). "Reanalysis of *Voyager* Ultraviolet Spectrometer Limits to the Extreme-Ultraviolet and Far-Ultraviolet Diffuse Astronomical Flux," *Astrophysical Journal*, **539**, 187-190.

Edery, A., and M.B. Paranjape. (1998). "Classical tests for Weyl gravity: deflection of light and time delay," NASA pre-print achives, arXiv:astro-ph/9708233 v2, 21 Apr 1998. Published in *Physical Review D*, **58** (1998), p.024011.

Einasto, Jan, Ants Kaasik and Enn Saar. (1974). "Dynamic Evidence on Massive Coronas of Galaxies," *Nature*, **250** (July 26, 1974), 309-10.

Ellis, G.F.R. (1999). "The Different Nature of Cosmology," *Astronomy and Geophysics*, August 1999, 4.20-4.23.

Ellis, G.F.R. (1985). "Observational Cosmology After Kristian and Sachs," in Stoeger (1985), 475-86.

Ellis, G.F.R. (1980). "Limits to Verification in Cosmology." *Annals of the New York Academy of Sciences*, **336**, 130-60.

Ellis, G.F.R. (1975). "Cosmology and Verifiability." *Quarterly Journal of the Royal Astronomical Society*, **16**, 245-64.

Faber, S.M., and J.S. Gallagher. (1979). *Annual Review of Astronomy and Astrophysics*, **17**, 135.

Faber, S.M., and D.N.C. Lin. (1983). "Is There Non-luminous Matter in Dwarf Spheroidal Galaxies?" *Astrophysical Journal*, **266**, L17-20.

Fabricant, D., and P. Gorstein. (1983). "Further Evidence for M-86's Massive, Dark Halo." *Astrophysical Journal*, **267**, 535-46.

Felton, J.E. (1984). "Milgrom's Revision of Newton's Laws: Dynamical and Cosmological Consequences." *Astrophysical Journal*, **286**, 3.

Finzi, Arigo. (1963). "On the Validity of Newton's Law at Long Distances." *Monthly Notices of the Royal Astronomical Society*, **127**, 21-30.

Forster, Malcom. (1995). "Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation." *British Journal for the Philosophy of Science*, **46**, 399-424.

Forster, Malcolm. (1988b). "The Confirmation of Common Component Causes." *PSA 1988*, Vol. 1, 3-9.

Forster, Malcom, and Elliot Sober. (1994). "How to Tell when Simpler, More Unified or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *British Journal for the Philosophy of Science*, **45**, 1-35.

Franklin, Allan. (1986). *The Neglect of Experiment*. Cambridge: Cambridge University Press.

Freeman, K.C. (1970). "On the Disks of Spiral and S0 Galaxies." *Astrophysical Journal*, **160**, 811-30.

Friedman, Michael. (1983). *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science*. Princeton, NJ: Princeton University Press.

Friedman, Michael. (1979). "Truth and Confirmation." *Journal of Philosophy*, **LXXVI**, No. 7, 361-381.

Galeotti, P., and David N. Schramm. (eds.) (1990). *Dark Matter in the Universe*. NATO Advanced Study Institute. Dordrecht: Kluwer.

Galison, Peter. (1987). *How Experiments End.* Chicago: University of Chicago Press.

Geller, Margaret. (1989). "Mapping the Universe: Slices and Bubbles," in Cornell (1989), 50-72.

Geller, Margaret, and John Huchra. (1989). "Mapping the Universe," *Science*, **246**, 17 Nov 89, 897-903.

Gingerich, Owen. (ed.) (1984). *Astrophysics and Twentieth-Century Astronomy to 1950: Part A.* The General History of Astronomy, Vol. 4. Cambridge: Cambridge University Press.

Gillies, Donald. (1998). "The Duhem Thesis and the Quine Thesis," as reprinted in Curd and Cover 1998, 302-19.

Ghosh, Amitabha. (1995). "Determination of True Velocity Dispersion and the Dark Matter Problem in Clusters of Galaxies," *Astrophysics and Space Science*, **227**, 41-52.

Gawiser, Eric. (2000). "Limits on Neutrino Masses from Large-Scale Structure," astro-ph/0005475.

Glymour, Clark. (1980). *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Gnedin, Yu. N. (1997). "Astronomical Searches for Nonbaryonic Dark Matter," *Astrophysics and Space Science*, **252**: 95-106.

Govroglu, Kostas. (1989). "Simplicity and Observability: When Are Particles Elementary?" *Synthese*, **79.3**: 543-558.

Gray, David F. (1997). "Absence of a Planetary Signature in the Spectra of the Star 51 Pegasi," *Nature*, **385**, 795-96.

Gray, David F. and A.P. Hatzes. (1997). "Non-radial Oscillation in the Solar-Temperature Star Pegasus 51," *Astrophysical Journal*, **490**, 412-24.

Gribbon, John, and Martin Rees. (1989). *Cosmic Coincidences: Dark Matter, Mankind, and Anthropic Cosmology*. New York: Bantum Books.

Grosser, M. (1979 [1962]). *The Discovery of Neptune*. First published by Harvard University Press. New York: Dover Publications, Inc.

Grünbaum, Adolf. (1976). "*Ad Hoc* Auxiliary Hypotheses and Falsificationism," *British Journal for the Philosophy of Science*, **27**, 329-362.

Guth, Alan H. (1997). *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins*. Reading, MA: Addison-Wesley.

Guth, Alan H. (1981). "The Inflationary Universe: A Possible Solution to the Horizon and F[illegible]tness Problems." *Physical Review D*, **23**, 347-56.

Hacking, Ian. (1989). "Extra-galactic Reality: The Case of Gravitational Lensing." *Philosophy of Science*, **56.4**, 555-81.

Hanson, N.R. (1962). "Leverrier: The Zenith and Nadir of Newtonian Mechanics." *Isis*, **53**, 359-377.

Harper, William. (1999). "The First Six Propositions in Newton's Argument for Universal Gravitation." *The St. John's Review*, Vol. XLV, No. 2, 74-93.

Harper, William. (1997a). "Isaac Newton on Empirical Success and Scientific Method." in Earman and Norton (1997), 55-86.

Harper, William. (1997b). "Newton's Phenomena," manuscript.

Harper, William. (1991). "Newton's Classic Deductions from Phenomena." *PSA 1990*, Vol. 2, 183-196.

Harper, William. (1989). "Consilience and Natural Kind Reasoning," in Brown and Mittelstrass (1989), 115-72.

Harper, William, Bryce Hemsley Bennett and Sreeram Valluri. (1994). "Unification and Support: Harmonic Law Ratios Measure the Mass of the Sun," in Prawitz and Westerståhl (1994), 131-46.

Harper, William, and Robert DiSalle. (1996). "Inferences from Phenomena in Gravitational Physics," *Philosophy of Science Proceedings*, S46-S54.

Hawking, Stephen, and Werner Israel. (eds.) (1983). *Three Hundred Years of Gravitation*. Cambridge: Cambridge University Press.

Hawkins, Michael. (1997). *Hunting Down the Universe: The Missing Mass, Primordial Black Holes and Other Dark Matters*. Reading, MA: Addison-Wesley.

Hawkins, M.R.S. (1993). "Gravitational Microlensing, Quasar Variability and Missing Matter," *Nature*, **366**, 18 Nov 93, 242-45.

Hempel, Carl G. (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: The Free Press.

Herschel, John F.W. (1872). *Outlines of Astronomy*. Tenth Edition. New York: D. Appleton and Company.

Hetherington, Noriss S. (ed.) (1993a). *Encyclopedia of Cosmology: Historical, Philosophical and Scientific Foundations of Modern Cosmology*. New York: Garland.

Hetherington, Noriss S. (1988). *Science and Objectivity: Episodes in the History of Astronomy*. Ames, Iowa: Iowa State University Press.

Hinchcliffe, Ian. (1987). *Proceedings of the Theoretical Workshop on Cosmology and Particle Physics*. Singapore: World Scientific.

Hodge, Paul W. (ed.) (1984). *The Universe of Galaxies. Readings from Scientific American*. New York: W.H. Freeman and Company.

Hogan, Craig J. (1993). "In Search of the Halo Grail." *Nature*, **365**, 14 Oct 93, 602-03.

Hogg, David W., Gerald Quinlan, and Scott Tremaine. (1991). "Dynamical Limits on Dark Mass in the Outer Solar System." *The Astronomical Journal*, **101**, 2274-2286.

Holt, Stephen. (1991). *After The First Three Minutes*. New York: American Institute of Physics.

Horwich, Paul. (1982). "How to Choose between Empirically Indistinguishable Theories." *Journal of Philosophy*, **LXXIX**, No. 2, 61-77.

Hoskin, Michael, and Owen Gingerich. (1980). "On Writing the History of Modern Astronomy." *Journal for the History of Astronomy*, **xi**, 145-6.

Huchra, John P., Margaret J. Geller, Valerie de Lapparent and Harold G. Corwin. (1990). "The CfA Redshift Survey: Data for the NGR-30 Zone." *Astrophysical Journal Supplements*, **72**, 433-70.

Israel, Werner. (1987). "Dark Stars: The Evolution of an Idea." in Hawking and Israel (1987), 199-276.

Jammer, Max. (1997 [1961]). *Concepts of Mass in Classical and Modern Physics*. First published by Harvard University Press. New York: Dover Publications, Inc.

Jerusalem Winter School for Theoretical Physics. (1987). *Dark Matter in the Universe*. Vol. 4. Singapore: World Scientific.

Jones, Harold Spencer. (1956). "John Couch Adams and the Discovery of Neptune," in Newman (1956), Volume II, 822-39.

Kahn, F. and L. Woltjer. (1956). "Intergalactic Matter and the Galaxy," *Astrophysical Journal*, **130**, 705-17.

Kapteyn, J.C. (1922). "First Attempt at a Theory of the Arrangement and Motion of the Sidereal System." *Astrophysical Journal*, **50**, 302-328.

Klee, Robert. (1992). "In Defense of the Quine-Duhem Thesis: A Reply to Greenwood." *Philosophy of Science*, **59.3**, 487-91.

Knapp, Gillian. (1995). "The Stuff Between the Stars." *Sky and Telescope*, May 1995, 20-26.

Kolb, Edward W., and Michael S. Turner. (1990). *The Early Universe*. Redwood Cliffs, CA: Addison-Wesley.

Kolb, Edward W., and Michael S. Turner. (eds.) (1988). *The Early Universe: Reprints*. Redwood Cliffs, CA: Addison Wesley.

Kosso, Peter. (1988). "Dimensions of Observability," *British Journal for the Philosophy of Science*, **39**, 449-467.

Kosso, Peter, and Cynthia Kosso. "Central Place Theory and The Reciprocity Between Theory and Evidence." *Philosophy of Science*, **62**, 581.

Kormendy, J. (1987). *Dark Matter in the Universe*. IAU Symposium No. 117. Dordrecht: Reidel.

Koyré, Alexandre. (1965). *Newtonian Studies*. London: Chapman & Hall.

Krauss, Lawrence. (2000). *Quintessence: The Mystery of the Missing Mass in the Universe*. New York: Basic Books. First published (1989) as *The Fifth Essence*.

Kuhn, Thomas S. (1970). *The Structure of Scientific Revolutions*. Second edition, enlarged. Chicago: University of Chicago Press.

Kuijken, Konrad, and Gerhard Gilmore. (1991). "The Galactic Disk Surface Mass Density and Galactic Force Kz at z = 1.1 kiloparsecs." *Astrophysical Journal*, **367**, L9-L13.

Lakatos, Imre. (1970). "Falsification and the Methodology of Scientific Research Programmes," in Lakatos and Musgrave (1970), 91-196.

Lakatos, Imre, and Alan Musgrave. (eds.) (1970). *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Lankford, John. (1984). "The Impact of Photography on Astronomy," in Gingerich (1984), 16-39.

Laudan, Larry. (1996). *Beyond Positivism and Relativism: Theory, Method and Evidence*. Boulder, CO: Westview Press.

Lederman, Leon M., and David N. Schramm. (1995). *From Quarks to the Cosmos: Tools of Discovery*. New York: Scientific American Library.

Lerner, Eric J. (1995). "On the Problem of Big Bang Nucleosynthesis." *Astrophysics and Space Science*, **227**, 145-149.

Lerner, Eric J. (1991). *The Big Bang Never Happened: A Startling Refutation of the Dominant Theory of the Origin of the Universe*. New York: Times Books/Random House.

Leslie, John. (ed.) (1998). *Modern Cosmology & Philosophy*. 2nd edition. Amherst, NY: Prometheus Books. First published (1990) as: *Physical Cosmology and Philosophy*. New York: Macmillan.

Lightman, Alan and Roberta Brawer. (eds.) (1990). *Origins: The Lives and Worlds of Modern Cosmologists*. Cambridge, MA: Harvard University Press.

Lin, D.N.C., and S.M. Faber. (1983). "Some Implications of Non-Luminous Matter in Dwarf Spheroidal Galaxies." *Astrophysical Journal*, **266**, L21-L25.

Longair, M.S. (ed.) (1974). *Confrontation of Cosmological Theories with Observational Data*. IAU Symposium No. 63. Dordrecht: D. Reidel.

Loewenstein, Michael, and Raymond E. White, III. "Prevalence and Properties of Dark Matter in Elliptical Galaxies." *Astrophysical Journal*, **518**, 50-63.

Lynden-Bell, D., and G. Gilmore. (1990). *Baryonic Dark Matter*. Dordrecht: Kluwer.

Lynden-Bell, D., *et al.* (1988). "Spectroscopy and Photometry of Elliptical Galaxies. V—Galaxy Streaming Toward the New Super-galactic Cluster." *Astrophysical Journal*, **326**, 19-49.

Madsen, Jes, and Richard I. Epstein. (1984). "Firm Bounds on the Neutrino Mass from the Distribution of Dark Matter in Galaxies." *Astrophysical Journal*, **282**, 11-18.

Majernik, V. (1996). "Is Dark Matter Created in the Gravitation Field of Galaxies?" *Astrophysics and Space Science*, **240**, 133-139.

Mannheim, P.D. (1995a). Lecture to the Department of Physics, University of Western Ontario, 22 November 1995.

Mannheim, P.D. (1995b). "Cosmology and Galactic Rotation Curves," University of Connecticut Preprint #UCONN 95-07, November 1995.

Mannheim, P.D. (no date). "Microlensing, Newton-Einstein Gravity, and Conformal Gravity." [manuscript]

Mannheim, Philip D. (1994). "Open Questions in Classical Gravity." *Foundations of Physics*, **24.4**, 487-511.

Mannheim, Philip D. (1993). "Linear Potentials and Galactic Rotation Curves," *The Astrophysical Journal*, **419**, 150-154.

Mannheim, Philip D. (1992). "Conformal Gravity and the Flatness Problem." *The Astrophysical Journal*, **391**, 429-432.

Mannheim, Philip D., and Demosthenes Kazanas. (1994). "Newtonian Limit of Conformal Gravity and the Lack of Necessity of the Second Order Poisson Equation." *General Relativity and Gravitation*, **26.4**, 337-361.

Mannheim, Philip D., and Demosthenes Kazanas. (1989). "Exact Vacuum Solution to Conformal Weyl Gravity and Galactic Rotation Curves." *The Astrophysical Journal*, **342**, pp. 635-638.

Mateo, Mario. (1994a). "Hunting For Dark Matter." Department of Astronomy Colloquium, University of Western Ontario, 15 March 1994.

Mateo, Mario. (1994b). "Searching for Dark Matter," *Sky and Telescope*, Jan 94, 20-4.

Matthewson, D.S., V.L. Ford and M. Buchhorn. (1992). "No Back-Side In-Fall into the Great Attractor." *Astrophysical Journal*, **389**, L5-L8.

Meadows, A.J. (1984). "The New Astronomy," in Gingerich (1984), 59-72.

Michell, John. (1784). *Philosophical Transactions of the Royal Society*, **74**, 35-.

Michell, John. (1767). *Philosophical Transactions of the Royal Society*, **57**, 246-.

Milgrom, Mordehai. (1994). "Dynamics with a Non-Standard Inertia-Acceleration Relation: An Alternative to Dark Matter in Galactic Systems." *Annals of Physics*, **229**, 384-415.

Milgrom, Mordehai. (1989). "On the Stability of Galactic Disks in the Modified Dynamics and the Distribution of their Mean Surface Brightnesses." *Astrophysical Journal*, **338**, 121-27.

Milgrom, Mordehai. (1986). "Solutions for the Modified Newtonian Dynamics Field Equations." *Astrophysical Journal*, **302**, 617-25.

Milgrom, Mordehai. (1983). "A Modification of the Newtonian Dynamics as a Possible Alternative to the Hidden Mass Hypothesis." *Astronomical Journal*, **270**, 365-70.

Moore, Ben. (1994). "Evidence Against Dissipationless Dark Matter from Observations of Galaxy Haloes." *Nature*, **370**, 25 August 94, 629-31.

Musgrave, Alan. (1978). "Evidential Support, Falsification, Heuristics and Anarchism," in Radnitsky and Andresson (1978).

Ne'eman, Yuval, and Yoram Kirsh. (1986). *The Particle Hunters*. Originally published (1983) in Hebrew. Cambridge: Cambridge UP.

Newman, James R. (1956). *The World of Mathematics*. Four Volumes. New York: Simon and Schuster.

Newton, Isaac. (1999 [1726]). *The Principia: Mathematical Principles of Natural Philosophy*. Trans. I. Bernard Cohen and Anne Whitman. Berkeley, CA: University of California Press.

Newton, Isaac. (1995 [1726]). *The Principia*. Trans. Andrew Motte. Amherst, NY: Prometheus Books.

Nishinomiya-Yukawa Memorial Symposium. (1990). *Dark Matter in the Universe: Proceedings of the Third N-Y Memorial Symposium*. Berlin: Springer-Verlag.

Oepik, E. (1922). "An Estimate of the Distance of the Andromeda Nebula." *Astrophysical Journal*, **55**, 406-10.

Oldershaw, Robert L. (1998). "The Galactic Dark Matter: Predictions and Observations." *Astrophysics and Space Science*, **257**, 271-278.

Oort, Jan. (1965). "Stellar Dynamics," in Blaauw and Schmidt (1965), 455- 511.

Oort, Jan. (1960). "Note on the Determination of K_z and on the Mass Density Near the Sun." *Bulletin of the Astronomical Institute of The Netherlands*, **15**, 45-53.

Oort, Jan. (1932). "The Force Exerted by the Stellar System in the Direction Perpendicular to the Galactic Plane and Some Related Problems." *Bulletin of the Astronomical Institute of The Netherlands*, **6**, 249-87.

Ostriker, J.P. (1997). "What can be Learned from Numerical Simulations of Cosmology," in Bahcall and Ostriker, 1997, 115-35.

Ostriker, J.P., and P.J.E. Peebles. (1973). "A Numerical Study of the Stability of Flattened Galaxies: Or, Can Cold Galaxies Survive?" *Astrophysical Journal*, **186**, 467-80.

Ostriker, J.P., P.J.E. Peebles and A. Yahil. (1974). "The Size and Mass of Galaxies, and the Mass of the Universe," *Astrophysical Journal*, **193**, L1-L4.

Overduin, J.M., S.S. Seahra, W.W. Wesley, and P.S. Wessen. (1999). "Could Intergalactic Dust Obscure a Neutrino Decay Signature?" *Astronomy and Astrophysics*, **349**, 317-22.

Overduin, J.M., and P.S. Wessen. (1997). "Decaying Neutrinos and the Extragalactic Background Light." *Astrophysical Journal*, **483**, 77-86.

Parker, Barry. (1993). *The Vindication of the Big Bang: Breakthroughs and Barriers*. New York: Plenum Press.

Peebles, P.J.E. (1993). *Principles of Physical Cosmology*. Princeton: Princeton University Press.

Peebles, P.J.E. (1984). "Dark Matter and the Origins of Galaxies and Globular Clusters." *Astrophysical Journal*, **277**, 470-77.

Perlmutter, S., *et al.* (1998). "Discovery of a Supernova Explosion at Half the Age of the Universe." *Nature*, **391**, 01 Jan 1998, 51-54

Peterson, Ivars. (1993). *Newton's Clock: Chaos in the Solar System*. New York: W.H. Freeeman and Co.

Persic, Massimo, and Paolo Salucci. (eds.) (1997). *Dark and Visible Matter in Galaxies*. San Francisco, CA: Astronomical Society of the Pacific (Conference Series Volume 117).

Persic, M., and P. Salucci. (1988). "Dark and Visible Matter in Spiral Galaxies." *Astrophysics and Space Science*, **234**, 131-154.

Philip, A.G. Davis, and David DeVorkin. (eds.) (1977). *In Memory of Henry Norris Russell*. Dudley Observatory Reports, No. 13 (December 1977). Proceedings of Sessions I and II of IAU Symposium No. 80 (held in Washington, D.C., 02 November 1977) and other material (including some original papers by Russell).

Ponman, T.J., and D. Bertram. (1993). "Hot Gas and Dark Matter in a Compact Group." *Nature*, **363**, 06 May 93, 51-4.

Popper, Karl Raimund. (1972). *Logic of Scientific Discovery*. 6th impression revised. London: Hutchinson.

Prawitz, D., and D. Westerståhl. (eds.) (1994). *Logic and Philosophy of Science in Uppsala*. Dordrecht: Kluwer Academic Publishers.

Presley, C.F. (1960 [1954]). "Laws and Theories in the Physical Sciences." *Australasian Journal of Philosophy*, **XXXII**, No. 2, 79-103. As reprinted in Danto and Morgenbesser (1960), 205-225.

Quine, W.V.O. (1998 [1951]). "Two Dogmas of Empiricism," as reprinted in Curd and Cover 1998, 280-319. Originally published in *Philosophical Review*, **60**, 20-43.

Quine, W.V.O. (1992). *The Pursuit of Truth*. Revised edition. Harvard University Press: Cambridge, MA.

Quine, W.V.O. (1975). "Empirically Equivalent Systems of the World," *Erkenntnis*, **IX**, No. 3, 313-28.

Quine, W.V.O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W.V., and J.S. Ullian. (1970). *The Web of Belief*. Random House: New York.

Radnitzky, G., and G. Andersson. (eds.) (1978). *Progress and Rationality in Science*. Volume 58, Boston Studies in the Philosophy of Science. Dordrecht: D. Reidel.

Rees, M. J. (1987). "The Emergence of Structure in the Universe: Galaxy Formation and Dark Matter," in Hawking and Israel (1987), 459-98.

Reid, *et al.* (1999). "The Proper Motions of Sgr A*: I. First VLBA Results." Preprint astro-ph/9905075, 6 May 1999.

Riordan, Michael, and David N. Schramm. (1991). *The Shadows of Creation: Dark Matter and the Structure of the Universe*. New York: Freeman and Company.

Rockman, Joseph. (1998). "Gravitational Lensing and Hacking's Extragalactic Reality," *International Studies in the Philosophy of Science*, **12.2**, 151-163.

Roseveare, N. T. (1982). *Mercury's Perihelion from Le Verrier to Einstein*. Oxford: Clarendon Press.

Rubin, Vera. (1989). "Weighing the Universe: Dark Matter and Missing Mass," in Cornell (1989), 73-104.

Rubin, Vera C. (1983a). "Dark Matter in Spiral Galaxies," in Hodge (1984), 31-43. First appeared in *Scientific American*, June 1983.

Rubin, Vera C. (1983b). "The Rotation of Spiral Galaxies," *Science*, **220**, 24 June 83, p.1139.

Rubin, Vera C., and W. Kent Ford. (1970). "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions," *Astrophysical Journal*, **159**, 379-403.

Sacket, Penny D., *et al.* (1994). "A Faint Luminous Halo That May Trace the Dark Matter Around Spiral Galaxy NGC5907," *Nature*, **370**, 11 August 94, 441-43.

Salmon, Merrilee H., *et al.* (1992). *Introduction to the Philosophy of Science.* Englewood Cliffs, NJ: Prentice-Hall.

Salmon, Wesley C. (1967). *The Foundations of Scientific Inference.* Pittsburgh: Pittsburgh University Press.

Sanders, R.H. (1999). "The Virial Discrepancy in Clusters of Galaxies in the Context of Modified Newtonian Dynamics." *Astrophysical Journal*, **512**, L23-L26.

Sanders, R.H. (1996). "The Published Extended Rotation Curves of Spiral Galaxies: Confrontation with Modified Dynamics." *Astrophysical Journal*, **473**, 117-29.

Sanders, R.H. (1991). "Mass Discrepancies in Galaxies: Dark Matter and Alternatives." *Astronomy and Astrophysics Review*, 2, 1.

Sanders, R.H. (1986). "Alternatives to Dark Matter." *Monthly Notices of the Royal Astronomical Society*, **223**, 539-55.

Sanders, R.H. and M.A.W. Verheijen. (1998). "Rotation Curves of Ursa Major Galaxies in the Context of Modified Newtonian Dynamics." *Astrophysical Journal*, **503**, 97-108.

Schaeberle, J.M. (1896). "Discovery of the Companion to Procyon." *Astronomical Journal*, **17**, 37.

Sciama, D.W. (1998). "Updated Parameters for the Decaying Neutrino Theory and EURD Observations of the Difffuse UV Background." *Astronomy and Astrophysics*, **335**, 12-8.

Sciama, D.W. (1993). *Modern Cosmology and The Dark Matter.* New York: Cambridge University Press.

Seeds, Michael A. (1989). *Horizons: Exploring the Universe.* 3rd ed. Belmont, CA: Wadsworth Publishing Company.

Seidelmann, P. Kenneth. (1992). *Explanatory supplement to the Astronomical almanac.* Prepared by the Nautical Almanac Office, U.S. Naval Observatory; with contributions from H.M. Nautical Almanac Office, Royal Greenwich Observatory, *et al.* Rev. ed. Mill Valley, CA: University Science Books.

Shapere, Dudley. (1982). "The Concept of Observation in Science and Philosophy." *Philosophy of Science*, **49**, 485-525.

Sky & Telescope News Bulletin. (August 11, 2000). <space.sci.news>

Sky & Telescope News Bulletin. (March 31, 2000). <space.sci.news>

Smart, W.M. (1953). *Celestial Mechanics*. London: Longmans, Green and Co. Ltd.

Smart, W.M. (1947). "John Couch Adams and the Discovery of Neptune." *Occasional Notes of the Royal Astronomical Society*, **2**, August 1947, 1-56.

Smith, George. (1999a). "From the Phenomenon of the Ellipse to an Inverse-Square Force: Why Not?" Philosophy Department Colloquium, University of Western Ontario, 24 September 1999.

Smith, George. (1999b). "Newton's Laws of Motion." Philosophy Department Colloquium, University of Western Ontario, 25 September 1999.

Smith, George. (1999c). "How Did Newton Discover Universal Gravity?." *The St. John's Review*, Vol. XLV No. 2, 32-59

Smith, Sinclair. (1936). "The Mass of the Virgo Cluster." *Astrophysical Journal*, **83**, 23-30.

Smoot, George, and Keay Davidson. (1993). *Wrinkles in Time*. New York: W. Morrow.

Sober, Elliott. (1999). "Testability." Presidential Address to the Central Division of the American Philosophical Association, May 1999. In *Proceedings and Addresses of the APA*, **73:2**, 47-76.

Srednicki, Mark (ed). (1990). *Particle Physics and Cosmology: Dark Matter*. Amsterdam: Elsevier Science.

Standish, . (1992). In Seidelmann (1992).

Stein, Howard. (1991). "From the Phenomena of Motion to the Forces of Nature: Hypothesis or Deduction?" in *PSA 1990*, vol. 2, pp. 209-22.

Stoeger, W.R. (ed.) (1985). *Theory and Observational Limits in Cosmology: Proceedings of the Vatican Observatory Conference*. Vatican City: Specola Vaticana.

Swarzschild, Martin. (1965). *Structure and Evolution of the Stars*. (Originally published 1958, Princeton University Press.) New York: Dover Publications, Inc.

Tayler, Roger J. (1991). *The Hidden Universe*. Chirchester, Great Britain: Ellis Horwood Limited.

Torretti, Roberto. (1990). *Creative Understanding: Philosophical Reflections on Physics*. Chicago: University of Chicago Press.

Trimble, Virginia. (1993). "Dark Matter," in Hetherington (1993a), 148-58.

Trimble, Virginia. (1990). "History of Dark Matter in the Universe (1922-1974)," in Bertotti, *et al.* (1990), 355-362.

Trimble, Virginia. (1988). "Dark Matter in the Universe: Where, What, Why?" *Contemporary Physics*, **29**, 373-92.

Trimble, Virginia. (1987). "Existence and Nature of Dark Matter in the Universe," *Annual Review of Astronomy and Astrophysics*, **25**, 425-472.

Trimble, Virginia, and Andreas Reisenegger, (eds.) (1996). *Clusters, Lensing and the Future of the Universe*. San Francisco: Astronomical Society of the Pacific (Conference Series, vol. 88).

Tucker, Wallace H. (1988). *The Dark Matter: Contemporary Science's Quest for the Hidden Mass in Our Universe*. New York: Morrow.

Tyson, J.A., R.A. Wenk, and F. Valdes (1990). "Detection of Systematic Gravitational Lens Image Alignments—Mapping Dark Matter in Clusters," *Astrophysical Journal*, **349**, L1-L4.

Valluri, Sreeram, Curtis Wilson and William Harper. (no date) "Newton's Apsidal Precession Theorem and Eccentric Orbits." Manuscript.

Valtonen, M.J., and G.G. Byrd. (1986). "Redshift Asymmetries in Systems of Galaxies and the Missing Mass," *Astrophysical Journal*, **303**, 523-34.

van den Bergh, Sidney. (2000). "A Short History of the Missing Mass and Dark Energy Paradigms," arXiv:astro-ph/0005314, 15 May 2000. Forthcoming in V.J. Martinez and V. Trimble, eds., *The Development of Modern Cosmology*, ASP Conference Series.

van den Bergh, S. (1961). "Stability of Clusters of Galaxies," *Astronomical Journal*, **66**, 566-71.

Vanderburgh, William L. (1997). "Empirical Equivalence and Approximative Methods in the New Astronomy: A Defence of Kepler Against the Charge of Fraud," *Journal for the History of Astronomy*, **xxviii**, 317-336.

van Maanen, Adriaan. (1916). "Preliminary Evidence of Internal Motions in the Spiral Nebula Messier 101," *Astrophysical Journal*, **44**, 210-28.

Van Waerbeke, L., Y. Mellier, *et al.* (2000). "Detection of Correlated Galaxy Ellipticities on CFHT Data: First Evidence for Gravitational Lensing by Large-Scale Structures," <http:/xxx.lanl.gov/astro-ph/0002500>.

Walsh, D., R.F. Carswell and R.J. Weymann. (1979). "0957+461 A, B: Twin Quasistellar Objects or Gravitational Lenses?" *Nature*, **279**, 31 May 1979, 381-84.

Weinberg, Steven. (1993). *The First Three Minutes*. Second paperback edition, with new afterward. New York: Basic Books.

Weinberg, Steven. (1992). *Dreams of a Final Theory*. New York: Pantheon.

White, R. Stephen. (1996). "Can Baryonic Dark Matter be Solid Hydrogen?" *Astrophysics and Space Science*, **240**, 75-87.

Whitney, Cynthia Kolb. (1995). "How Can Spirals Persist?" *Astrophysics and Space Science*, **227**, 175-186.

Will, Clifford M. (1993). *Theory and Experiment in Gravitation Physics*. Revised edition (first edition 1981). Cambridge: Cambridge University Press.

Wittman, D.M., J.A. Tyson, *et al.* (2000). "Detection of Weak Gravitational Lensing Distortions of Distant Galaxies by Cosmic Shear at Large Scales." *Nature*, **vol.** 11 May 2000, 143-48.

Wszolek, Bogdan. (1995). "Observational Limits on Intergalactic Matter." *Astrophysics and Space Science*, **227**, 151-155.

Woodward, Jim. (1989). "Data and Phenomena." *Synthese*, **79.3**, 393-472.

Youden, W.J. (1998 [1962]). *Experimentation and Measurement*. New York: Dover Publications, Inc.

Zwicky, Fritz. (1957). *Morphological Astronomy*. Berlin: Springer Verlag.

Zwicky, Fritz. (1933). *Helvetica Physica Acta*, **6**, 110-.

APPENDICES

EVIDENCE FOR THE VALUE OF Ω_{Matter}

A.1 The Age of the Universe and Constraints on the Matter Content

The question of the age of the universe has gone back and forth several times in recent years—in its present incarnation the issue arises from an apparent conflict between the ages of globular star clusters (structures within the Milky Way) and the age of the universe as computed from the time required for the universe to expand to its present size (roughly the inverse of H_0). On several occasions it has seemed on the best available evidence that the universe as a whole was younger than some of the objects it contains—obviously an undesirable result. Alcaniz and Lima (1999, L87) quote some recent globular cluster ages (t_{gc}) in the range $t_{gc} \sim 13\text{-}15\ Gyr$ (or perhaps 2 Gyr less according to one study), whereas

> Recent measurements of the Hubble parameter from a variety of techniques are now converging into the range (1σ) $h = (H_0/100\ kms^{-1}\ Mpc^{-1}) = 0.7 \pm 0.1$.... This means that the expansion age for a FRW [Friedman-Robertson-Walker[1]] flat matter-dominated [$\Lambda = 0$] universe ($t_0 = 2/3H_0^{-1}$) falls within the interval $8.1\ Gyr \leq t_0 \leq 10.8\ Gyr$. (Alcaniz and Lima 1999, L87)

According to Alcaniz and Lima,

> the unique possible conclusion is that the 'age crisis' continues for closed and, at least moderately, for flat FRW models. . . . Actually, from the original matter-dominated FRW class with no cosmological constant, only extremely open universes may be old enough to solve (beyond doubt) the expanding age problem. (1999, L87)

Their own and other work on high redshift galaxies that appear to be old even though they are seen such a large fraction of the age of the universe ago makes the age crisis even more acute; more to the point for present purposes, this work provides an empirical determination of the total mass density, and of the relative contributions of matter and "dark energy" to that total.

---

[1] An FRW universe is a solution to the Einstein field equations for the universe as a whole, one which assumes a zero intrinsic global curvature and a perfectly homogeneous and isotropic matter distribution.

Standard closed and flat FRW models without a cosmological constant thus seem to be ruled out. A popular remaining option, one which solves the age crisis as well as having other theoretical motivations, is to introduce a non-zero cosmological constant. A cosmological constant has the effect of accelerating the Hubble expansion (or at least, it causes the gravitational deceleration of the Hubble expansion to be effectively lower), which means that at earlier epochs the Hubble parameter H had a lower value than it has now, and this in turn means that the universe had more time to evolve to its present state than it would on $\Lambda = 0$ models. With the right choice of Ω_M and Ω_Λ, one can find a universe with cosmological properties consistent with all available observations *including* the ages of globular clusters. Although "the possibility of a non-zero cosmological constant has not been proved beyond doubt and remains an essentially open question," the only serious competitors are extremely open universes (Alcaniz and Lima 1999, L87).

Studies of Type 1a supernovae (for example, Riess, *et al.*, 1998) have improved the known value of H_0 to $65 \pm 7\ kms^{-1}\ Mpc^{-1}$; Alcaniz and Lima attempt to constrain Ω_M (and Ω_Λ) through study of two very high redshift, but *old*, galaxies (at $z \sim 1.43$ and $z \sim 1.55$ respectively).[2] They are able to show that the ages of these galaxies require, respectively, $[\Omega_M \leq 0.37, \Omega_\Lambda \geq 0.5]$ and $[\Omega_M \leq 0.45, \Omega_\Lambda \geq 0.42]$. Combining their results with values calculated through studies of Type 1a supernovae, field galaxies, statistics of gravitational lensing and the cosmic microwave background radiation,

---

[2] Doppler shifts for light are cited in astronomy by using the parameter z, where $z = \Delta\lambda / \lambda_0$, that is, the change in wavelength divided by the wavelength of emission. (One knows the wavelength of emission by looking for the spectrographic signature of some known element such as hydrogen, and then calculates the change of wavelength due to relative motion of source and observer by measuring the distance this characteristic pattern has been shifted through the spectrum.) One can convert a redshift quoted this way to a recessional velocity by using the relativistic redshift equation: $v_r / c = [(z+1)^2 - 1] / [(z+1)^2 + 1]$. (Seeds 1989, 304) (Note that $z = 1$ corresponds to a recessional velocity of about half the speed of light. As v_r approaches c, z approaches infinity. Measurements of z are convertible to absolute distance measures assuming no significant proper motion and some value for the Hubble constant; that is, regions of a given recessional velocity correspond to a given radial distance from us.)

Alcaniz and Lima give best-fit overall limits on Ω_Λ as follows, given the best available data: $0.42 \leq \Omega_\Lambda \leq 0.7$. I quote part of a table from Alcaniz and Lima (1999, L89):

Table of Observational Limits on Ω_M and Ω_Λ.

Method	Ω_M	Ω_Λ
Type Ia supernovae	0.2 ± 0.4	0.4 ± 0.2
	$0.24^{+0.56}_{-0.24}$	$0.72^{+0.72}_{-0.48}$
Field galaxies	~ 0.5	
Statistics of GL	> 0.15	< 0.66
CBR	0.24 ± 0.1	0.62 ± 0.16
Old high-z galaxies	≤ 0.37	≥ 0.5
	≤ 0.45	≥ 0.42

What is important for my purposes here is that all these studies agree that Ω_M is much less than the critical density, and that *if* the universe is nevertheless flat (a fact that many cosmologists still take for granted or give theoretical arguments for—see Krauss (2000)—but which so far has no unambiguous observational evidence), then the main part of the total energy (which is equivalent to mass by Einstein's famous formula, $E = mc^2$) density of the universe is in the form of a cosmological constant. Currently the most highly favoured candidate is the energy of the vacuum. (Krauss 2000 enthusiastically supports the hypothesis of "dark energy", but unfortunately without giving very much detail about it.) And whatever this cosmological constant is, it is a form of *energy* rather than matter. In other words, since the overall matter contribution to Ω_{total} has now been measured to be about the same as the total dynamical mass, there seems to be little reason to accept the existence of additional, purely cosmological dark matter. Admittedly, a homogeneous universal distribution of cosmological dark matter would have no dynamical effect on objects embedded in it (Tayler 1991, 18), but such a cosmological dark matter distribution *would* have cosmologically important effects which so far our best evidence strongly indicates are not to be found in our universe.[3] Note that

---

[3] The methods by which the relative contributions to the overall mass density of matter and of the cosmological constant are established, are extremely interesting examples of evidential reasoning, in particular of using observed phenomena to empirically measure theoretical parameters—but unfortunately

this result with regard to the cosmological dark matter has no effect on the dynamical dark matter problem—that discrepancy is entirely independent, and none of its potential solutions are ruled out by the cosmological result.

At the August 2000 meeting of the International Astronomical Union, Wendy Freeman of Carnegie Mellon University and the Hubble Space Telescope Institute reported that the Hubble Space Telescope has completed its project on determining the Hubble constant: $H_0 = \text{~}4 \pm \text{~}\ kms^{-1}Mpc^{-1}$ (Sky and Telescope News Bulletin, August 11, 2000). The former factor of two uncertainty in the Hubble constant has now been replaced by an error of only about ten percent. As one can see, this more or less definitive result is consistent with the other results mentioned above. Again, unless the dark energy is causing the Hubble parameter to accelerate there would be an age crisis.

A.2 Supernovae Constraints on the Matter Contribution to the Total Mass Density

It turns out that for a Type 1a supernova (an explosion triggered when mass accretion onto a white dwarf star exceeds a critical value) there exists an empirical relation between peak absolute luminosity and the period over which the supernova brightens. This relation is established through observations of supernovae whose distances can be reliably known by independent means (for example, from Cepheid variables in their host galaxies). The relation is then extrapolated to more distant supernovae, so that from their observed periods one can calculate their absolute luminosities. Comparing the period and corresponding absolute luminosity of a supernova with its observed luminosity allows one to find its distance. (This is a version of the "standard candle" method of determining astronomical distances, as is the Cepheid method, except that since supernovae are so intrinsically bright, they can be detected over much greater distances.) Now, for truly "cosmologically" distant supernovae (for which the Hubble motion swamps any contribution of peculiar motion to the redshift), one can compare a supernova's redshift distance as calculated from the Hubble relation with its distance as calculated by the luminosity method. In effect, this allows one to calibrate the

---

describing these methods is beyond the scope of the present endeavour. See Alcaniz and Lima (1999) and Perlmutter, *et al.* (1998) for information about these methods.

Hubble constant. By repeating this method for more and more distant supernovae, one finds the value of H at earlier and earlier times.

The Supernova Cosmology Project has carried out this procedure, and has found that the so called "deceleration parameter", q_0 (a measure of the amount by which gravity is slowing the Hubble expansion, and therefore a function of the cosmic mass density), has a *negative* value: $q < 0$ means that the expansion of the universe is actually *accelerating* over time. This is only possible if there exists some sort of cosmological constant, Λ, that acts like negative gravity. Astronomers had hoped to measure Ω from q—the degree of deceleration is a measure of the total gravitational attraction of matter, and therefore of the total mass of the universe—but the Supernova Cosmology Project result shows that the contribution of some cosmological constant to the effective mass density of the universe cannot be ignored, as had been supposed for so long.[4] The question then is to try to determine the relative contributions of Ω_M and Ω_Λ, to H and q; of course, many different combinations of values of these parameters are initially plausible, since they can reproduce the observed rates of expansion.

The best fit of earlier data (involving supernovae at redshift $z \sim 0.4$) with the data from SN 1997ap ($z \sim 0.83$) "corresponds to a value of $\Omega_M = 0.6 \pm 0.2$ if we constrain the result to a flat universe ($\Omega_\Lambda + \Omega_M = 1$), or $\Omega_M = 0.2 \pm 0.4$ if we constrain the result to a $\Lambda = 0$ universe. These results are preliminary evidence for a relatively low-mass-density universe" (Perlmutter, *et al.*, 1998, 53). Obtaining more data at high redshift will enable us to distinguish more exactly between competing cosmological models, including the relative contributions of matter and the cosmological constant to the overall mass density. The result discussed here is based on a sample of one, so further examples of high redshift supernovae are required in order to improve the epistemic warrant of the result—note the very large error bounds on the results quoted above. But also note that even at the extremes of those error bounds, the matter contribution is well below the critical

---

[4] By the mass-energy equivalence, since Λ contributes to the energy density of the universe, it contributes to the overall mass density, Ω. Scientists often speak simply of the contribution of Λ to the universal energy density, but this is equivalent to mass density once an appropriate conversion of units is made. Note that there must be a huge amount of energy acting in order to produce a cosmological acceleration!

value. If correct, this result removes the motivation for thinking that the difference between the observed mass density and the critical value (about a factor of 10) is to be made up by extra strictly cosmological dark matter. Now it seems that if the universe is flat, then Ω_Λ, that is, dark energy, not dark matter, is the main contributor to the total mass density.

CURRICULUM VITAE

William L. Vanderburgh

Born 14 July 1970, Montreal, Quebec, Canada

EDUCATION

Ph.D., Philosophy, University of Western Ontario, London, Canada
September 1994 to January 2001

M.A., Philosophy, University of Western Ontario
September 1993 to August 1994

B.A. (Honors), Philosophy, University of Western Ontario
September 1989 to April 1993

AWARDS AND SCHOLARSHIPS

University of Western Ontario Graduate Research Fellowship, summer 1998
Nominated for Graduate Student Teaching Award, 1997-98
Social Sciences and Humanities Research Council of Canada Doctoral Fellowship, 1996-97 through 1997-98
Tuition Waiver, 1996-97 through 1997-98
Ontario Graduate Scholarship (declined), 1996-97
Special University Scholarship, 1993-94 through 1997-98
Graduate Studies Entrance Scholarship, 1993-94
Dean's Honor List, 1991-92, 1992-93
Rio Algom Education Award, 1989-90 through 1992-93

PUBLICATION

"Empirical Equivalence and Approximative Methods in the *New Astronomy*: A Defence of Kepler Against the Charge of Fraud," *Journal for the History of Astronomy*, **xxvii**, (November 1997), 317-336.