# 1

# Introduction

My purpose here is to discuss the past and present of the dark matter hypothesis: how it has developed that most astronomers and physicists now believe that the matter content of the Universe is dominated by an unseen, non-luminous substance that interacts with ordinary matter, protons, neutrons and electrons, primarily through the force of gravity. This description is personal and based largely upon my perspective as an interpretive astronomer. It is also necessarily biased. Throughout most of my career, for the past 40 years now, I have been involved – at times peripherally, often directly – in research on the discrepancy between the detectable mass of astronomical systems and the inferred Newtonian dynamical mass. Since my graduate student days, I have worked at institutes where consideration of this problem, both theoretical and observational, has been a dominant theme. My views on these developments are certainly colored by my experience at these particular institutes and, no doubt, by my own prejudices. But I do hope that the account that I will give here is reasonably honest and fair.

Forty years ago, I was a graduate student at Princeton University. In the Peyton Hall basement, every Wednesday, there was a lively lunch meeting attended by staff members and students. Theses projects would be described, new ideas would be tossed out and batted around, and often politics (in that lively rebellious period) would be discussed in a highly dialectical manner. One Wednesday – it must have been in 1969 – one of our young assistant professors, Jerry Ostriker, appeared at lunch with a radical new idea. Jerry was an expert on the stability of rotating fluid spheroids (and many other subjects as well). He had been following with interest the computer simulations of disk galaxies which, at that point, were becoming extremely sophisticated, involving large numbers of particles all interacting gravitationally. He had noticed that in these simulations disks of particles which were initially supported against gravity by rotation – let's say, centrifugal force – did not seem to remain that way. The round disks developed elongated shapes and heated

up – that is, they became more like hot pressure-supported systems rather than rotating systems.

This corresponded perfectly to what Jerry knew about rotating fluid spheroids: it is impossible to construct such an object supported entirely by rotation; Newtonian dynamical systems supported by rotation are unstable. But our galaxy, the Milky Way Galaxy, appears to be held up almost entirely by rotation; the stars near the Sun are moving on nearly circular orbits about the center of the Galaxy. How is it that the Galaxy can remain rotationally supported and yet stable? Jerry's brilliant leap was to suggest that the Galaxy, in fact, is not rotationally supported – that the rotationally supported disk is only one component of the Galaxy. There is another major component, a spheroidal component, at least equal in mass to the disk, and this system is primarily pressure supported. Because no such massive spheroidal component is seen, it must be dark – a dark halo.

On that Wednesday, this suggestion appeared radical; I recall that it caused a great stir and considerable argument, especially from some of the more senior staff members such as Martin Schwarzschild. He raised a number of questions, most of which concerned the composition of the dark halo (Schwarzschild was an astronomer after all). What is the dark halo made out of? Low-luminosity stars possibly – red dwarfs – remnants of dead stars – white dwarfs. How might it be detected by means other than its gravitational influence? An infrared glow around galaxies, perhaps; high-velocity, low-luminosity stars, maybe. No one could have supposed at that point that the halo might consist of weakly interacting, subatomic particles. This would have been far too radical. Not one of us would have dared to suggest, even if they had thought of it, that Newton's laws might need revision on the scale of galaxies and larger; that would have seemed insane.

In 1973, Ostriker, joined by his Princeton colleague, Jim Peebles, published this proposal which by that point had been bolstered by their own N-body calculations; the idea provoked even more controversy in the larger community than it had on that Wednesday afternoon in Princeton (Chapter 3). Although this was a radically new idea with an entirely theoretical basis, there had been considerable earlier evidence that astronomical systems contain large quantities of unseen matter. In 1933 the Swiss astronomer Fritz Zwicky had made the first systematic kinematic study of a cluster of galaxies and pointed out that in order to gravitationally bind the cluster the actual mass had to be several hundred times larger than the observed mass in stars (Chapter 2). Earlier, in 1932, the Dutch astronomer Jan Oort, by looking at the motion of the stars above the galactic plane, concluded that there must be about 50% more mass in the Galaxy disk than is evidenced by luminous stars. But Oort's dark matter was distributed in the plane of the Galaxy, like most of the observed stars; this would probably not solve Ostriker's stability problem. Moreover, Oort included the undetected component of the interstellar medium, dust and

gas, as part of the dark component so that it did not seem, at the time, particularly mysterious.

But observational evidence in support of the idea that spiral galaxies possessed a substantial, more extended unseen component was beginning to appear in the early 1970s. My first real position, in 1972, was at the National Radio Astronomy Observatory (NRAO) in Charlottesville, Virginia. This was primarily an observational institute and I was known as a "house theoretician". Radio astronomers at NRAO, such as Mort Roberts and Seth Shostak, had been observing the distribution and motion of neutral hydrogen in the outer parts of galaxies through the spectral line emitted by hydrogen at a wavelength of 21 cm (Chapter 4). They noticed that the rotational velocity of the gas does not seem to be declining with distance from the centers of galaxies as it should for a bounded mass distribution. The rotation velocity appeared to be constant well beyond the visible image of the galaxy. This was a very contentious result at the time, with heated debates about telescope side lobes and possible warping of the gas layers in spiral galaxies, but it was a clear early indication that there is a real discrepancy between the dynamical and visible mass in galaxies. And it was in complete accordance with the suggestion of Ostriker and Peebles.

Later in my career, in 1977, I accepted a position at the Kapteyn Astronomical Institute at the University of Groningen in the Netherlands, again, as a house theoretician at a primarily observational institute. A few years before that, the synthesis radio telescope at Westerbork, a one and one-half kilometer array of dishes used as a single telescope, had begun operating and was being applied to observe the distribution and motion of neutral hydrogen in spiral galaxies with relatively high spatial and velocity resolution. The radio astronomers at Groningen were making precise measurements of the "rotation curves" of spiral galaxies – how the gas rotates as a function of distance from the center well beyond the visible object. Consistent with the earlier observations, the rotation velocity was not seen to decline but remained constant with distance implying that the gas, although well beyond most of the light of the galaxies, is still immersed in the mass distribution of the galaxy – that the mass in the outer regions of the galaxies is dark. Coming from Princeton and from NRAO, with all my theoretical and observational prejudices, this was not a surprising result for me. I realize now that I was not as excited as I should have been. Westerbork was producing the most convincing and direct observational confirmation of an idea that was still quite tentative – the idea that the visible parts of galaxies were a tiny, shiny central component of a vast dark system.

Evidence from other sources had been mounting as well. High-resolution measurements of rotation curves from spectroscopic observations of optical emission lines by Vera Rubin and her collaborators were beginning to appear in the literature – these rotation curves were also flat out to the optical edges of the spiral

galaxies. Because the rotation velocity was not measured beyond the optical image, this did not constitute compelling evidence for dark matter, as I will discuss in Chapter 5; but that was not the perception at the time. These observations had an enormous impact on the growing realization that there was a substantial dark matter component in spiral galaxies. By the early 1980s this viewpoint was rapidly becoming the paradigm.

My own interest has been mostly centered on galaxies and the manifestations of the mass discrepancy on a galaxy scale. But evidence was mounting on other scales as well. In the 1970s satellites that could observe the sky at X-ray wavelengths (this radiation does not penetrate the atmosphere of the Earth) were launched into Earth orbit. It was discovered that distant clusters of galaxies were powerful sources of X-rays and that this emission is thermal radiation from vast pools of hot gas filling the clusters. In fact, the mass of gas generally exceeds that of the stars in galaxies by a factor of two or three. Could this be Zwicky's missing cluster mass? For such a gaseous object in equilibrium one can, by measuring the temperature and density distribution of the gas, determine the gravitational field and, hence, with Newton's law of gravity, the mass of the entire system. When this was done, it became apparent that most of the mass of clusters of galaxies was still unseen; that the clusters contained at least five or six times more mass than was detected in stars and gas. Was this dark matter the same as that in individual galaxies? It was, and is, generally assumed to be so.

It was also becoming evident in the late 1970s that something is missing on a cosmological scale. The Universe is typically modeled as an expanding, isotropic, homogeneous fluid, and certainly on the largest scales it appears to be that way. The cosmic microwave background radiation, (the CMB) discovered in 1965 by Arno Penzias and Robert Wilson, should reflect density fluctuations in the cosmic fluid when the Universe was only 300 000 years old – when protons and electrons combined to make neutral hydrogen and the radiation decoupled from the matter. These fluctuations in the CMB were looked for and not found at the level of about one part in 10 000. This means that all of the structure that we observe in the Universe – from stars to galaxies to clusters of galaxies and to super clusters – has formed in the last 14 billion years or so by the gravitational growth of incredibly small fluctuations. This just did not seem possible in the context of the standard theory of gravitational instability. A solution to this problem is to add dark matter, but a special kind of dark matter: matter consisting of particles that interacts with light or ordinary (baryonic) matter primarily through gravity – "non-baryonic" dark matter. Because it is decoupled from the radiation, this dark matter fluid can begin to gravitationally collapse sooner than the normal baryonic matter – before the recombination of hydrogen. This gives the observed structure time to form from the very small density fluctuations. So dark matter on a cosmological scale appeared to be

a necessity as well. (The missing fluctuations were finally seen at a level of $10^{-5}$ by the COBE satellite in 1992. See Smoot *et al.*, 1992.)

But a completely new aspect of the dark matter problem emerged from these cosmological considerations. This cosmological dark matter is very different than what had originally been imagined for the dark halos surrounding galaxies. It is not small or dead stars, but subatomic particles – and not the ordinary subatomic particles like protons and neutrons, but something else which interacts very weakly – neutrinos perhaps, or something even more exotic, something not yet detected in terrestrial laboratories. At about the same time, particle physics theory was advancing beyond its so-called standard model. New ideas on the unification of forces were being proposed – grand unification and then, supersymmetry. These new theories provide a host of particle dark matter candidates in addition to the modest neutrino. Subatomic particles possess an attribute called "spin" that is quantized (it comes in distinct lumps). In supersymmetry every known standard-model particle is required to have a supersymmetric partner that differs by half-integer spin. So this theory, in effect, doubles the number of possible particles. Only one of these hypothetical particles – the lowest mass superpartner – is stable and long-lived and could be the dark matter. But because of this possibility, physicists became very excited about the prospect of dark matter – some even appeared to believe that they had invented dark matter. This union of astronomers, cosmologists and particle physicists led to the development of a new, interdisciplinary subject – astroparticle physics. Once again, astronomical observations had spawned not only a new paradigm, but a new field of study.

In the spring of 1982, I was taking a four-month sabbatical at NRAO and enjoying the Virginia spring while working on an absolutely unrelated topic – the jets observed to be emanating from some active galactic nuclei. In those days, preprints of scientific articles – pre-publication versions of papers which were usually in press already – were not placed on the Internet – there was no Internet – but were distributed in printed form between various scientific institutes. NRAO was definitely on this preprint circuit, and at some point, around April 1982, three preprints arrived from the Institute of Advanced Study in Princeton. These were preprints on the missing matter problem authored by an Israeli physicist, Mordehai Milgrom. I had actually encountered Milgrom before in a rather competitive way; he had independently developed a model that I had proposed some years before – a model for compact radio sources with apparent faster-than-light motion. But here, in these articles, Milgrom was proposing a very radical new idea – and not one that I could claim to have thought of. He was suggesting there is no dark matter but that the usual Newtonian dynamics or gravity was not applicable on these extragalactic scales. His hypothesis was called "modified Newtonian dynamics" or MOND for short. These preprints first brought home to me the realization that, after all,

dark matter is a sort of ether – a medium that is necessary to make observations consistent with the expectations of existing theory. If the theory is inappropriate on these scales, then perhaps there is no ether.

Now Milgrom's idea is basically very simple: Newtonian dynamics is modified at low accelerations – that the familiar old formula $F = ma$ becomes more like $F = ma^2/a_0$ at accelerations below a critical value $a_0$. This simple modification appears to accomplish a great deal. It yields flat galaxy rotation curves in the limit of large radius (low acceleration), and provides a relation between the mass of a galaxy and its rotation velocity, or if mass is proportional to luminosity, a luminosity–rotation velocity relation. In fact, such a relation had been observed years before by Brent Tully and Rick Fisher – the Tully–Fisher relation – and Milgrom's acceleration-based modification provided a simple explanation of this correlation as resulting from existent physical law, as opposed to dark matter which attributed such scaling relations to the contingencies of galaxy formation. Moreover, MOND predicts that high-surface-brightness systems, like globular star clusters for example, should have no apparent dark matter problem within the visible object, and that low-surface-brightness systems, such as the dwarf spheroidal satellites of our own Galaxy, should have a large discrepancy.

I was fascinated by this idea, but I thought that it was probably not correct. Such a drastic modification would surely have other consequences – consequences for cosmology and large-scale structure in the Universe. It seemed to me that it was not just sufficient to explain a few facts about galaxies, the idea had to fit into a much larger picture. There is much more to explain than galaxies.

I let this go for a while, but then, a couple of years later, back in Groningen, I had my own idea. I read a paper by a French physicist, Joel Sherck, who proposed that, consistent with supersymmetry or its follower, supergravity, additional fields might exist in the Universe; fields which couple to matter with gravitational strength. One possibility is a vector field, but vector fields, like electromagnetism, produce a repulsive force between similar particles – an anti-gravity. The force would be carried by a particle, a so-called vector boson. Sherck wanted this vector boson to have a finite mass and therefore a limited range, but a range so small that it would have no actual macroscopic effect on scales of one meter or so where the inverse-square law of gravity had been carefully measured (the larger the mass of the field, the smaller its range). I picked up on this suggestion and warped it to my own purpose.

How could a repulsive force yield flat rotation curves? I thought – perhaps gravity, locally, is a mixture of repulsion and attraction, but slightly more attraction. Suppose also that the vector boson which mediates the repulsive force has such a small mass that its range would be on the scale of galaxies? This would mean that on a scale larger than a galaxy the repulsive force would die away leaving pure

attraction. It would be possible to have a larger effective gravitational attraction on extragalactic scales than on the sub-galactic scale. Adjusting the mass of the vector boson correctly and the ratio of repulsion to attraction correctly, one could produce flat rotation curves for spiral galaxies over a range of about a factor of 10 in radius. This, I thought, led to a more cosmologically acceptable model, because on the largest scale, there was a return to inverse-square attraction, and the Universe behaved as it would in the standard picture with 10 times more dark than visible matter. I might add here that I didn't know very much about general relativity in those days and didn't realize that my proposal would violate the local universality of free fall (first tested by Galileo in his famous, but probably fictional, Tower of Pisa experiment) in a very blatant and detectable way.

I immediately submitted a short paper to *Astronomy and Astrophysics* (the European journal) and waited to see what would happen. There were two reviewers of the paper, one of whom was Milgrom. He was very negative in his report. He pointed out that such a modification would, indeed, lead to a Tully–Fisher law, but the wrong Tully–Fisher law: $L \propto V^2$ instead of $L \propto V^4$, as is, so he claimed, more consistent with observations. I protested. I thought that the form of the Tully–Fisher law was not so evident at that point; it seems to depend upon the color in which the luminosity is measured, and in blue light it is more like $L \propto V^2$. I was so attracted by my idea that I thought that it must be published, and after much pleading with the editor (who occupied an office a few doors from my own), it was.

I cherished this idea for several years more, but then, the reality of galaxy phenomenology caught up with me in the form of two facts. The first fact is that Milgrom was right about the form of the Tully–Fisher law – when measured in the near-infrared emission from stars (the radiation from the old, low-mass stars that are the dominant component of the stellar disk), the relation really is more like $L \propto V^4$, as he said. The second is this: larger galaxies do not exhibit a larger discrepancy – big galaxies do not need more "dark matter". I had proposed a modification of gravity attached to a definite fixed-length scale. This means that galaxies which are larger than this length scale should have a larger discrepancy and smaller galaxies a smaller discrepancy or even no discrepancy at all. Being at an institute that was primarily observational and producing new rotation curves every day, I realized that this was not true. There are very small galaxies with a large discrepancy, and very large galaxies with a small discrepancy. The discrepancy seems to be more dependent upon surface brightness (the energy of radiation emerging per second per square meter at the source) than size, and surface brightness, in so far as it reflects surface density, is proportional to acceleration.

My idea seemed pitiful and lonely without any observational support, so even I had to abandon it. I think, actually, that many scientists have trouble with this. We become too deeply attached to ideas because they are ours – but confronted

by the facts, painful though it is, we are forced to forsake our pet theories. It must have been around 1985 when I realized that Milgrom was right. The only sort of modification of gravity or dynamics that could possibly replace dark matter was a modification attached to an acceleration scale. Then began for me a long period, still continuing, of work on MOND – observational and theoretical. I corresponded with and met another Israeli colleague of Milgrom's – the physicist Jacob Bekenstein. Jacob was a relativist – an expert in general relativity well known for his work on black holes – and he believed that MOND should be viewed as a modification of the theory of gravity. Jacob thought, and I agreed, that if MOND is to ever be acceptable it must connect to more familiar physics – it must be an aspect of a more general theory of gravity or inertia. I still think that this is true, but it is also true that what is "familiar" changes as well.

But what of dark matter? If MOND is right, is dark matter wrong? Simply defined, MOND is an algorithm for calculating the gravitational force in an astronomical object, from the observed distribution of ordinary baryonic (detectable) matter. And it works – at least on the scale of galaxies. Because it works, this is very problematic for dark matter – at least on the scale of galaxies. It would seem to imply a very precise coupling between dark matter and baryonic matter – a coupling that is not comprehensible in the context of standard or "cold" dark matter. On the other hand, cold dark matter is quite successful on cosmological scales; it predicts the formation of observed large-scale structure and the magnitude and distribution of fluctuations in the primordial cosmic microwave background. How could these two be reconciled?

But another interesting twist, which no one really imagined 20 years ago, emerged in the late 1990s: dark matter alone is not sufficient; it appears that, on a cosmological scale, "dark energy" is also required. This is a mysterious fluid with a negative pressure that does not dilute as the Universe expands and leads to the accelerated expansion of the Universe. In Einstein's theory of gravity, general relativity, the dark energy is embodied by the so-called cosmological constant. It may also be identified with the energy density of the vacuum, a concept of modern quantum field theory in which "empty" space is actually filled with virtual particles popping into and out of existence – virtual but gravitating. In this case, the vacuum energy density should be many orders of magnitude larger than it is observed to be; in fact, so large that the Universe as we observe it would be impossible. The observation of a tiny value for the vacuum energy density, tiny in terms of the expectations of quantum field theory, is one of the greatest puzzles in modern physics.

Now that we "know" the composition of the Universe, some cosmologists have become quite triumphal. There certainly has been enormous progress, but given this very strange composition – a mysterious and unnatural dark energy as well

as a dark matter fluid which has not been detected by any means other than its gravitational influence – triumphalism seems to be premature. To me, it appears presumptuous to assume that we human beings at this point in our development understand either the material content of the Universe or all of its physical laws.

Here I want to describe the process of discovery over the past 40 years that has led to the development of the dark matter paradigm as well as the now standard cosmological model. Of course, these developments have spawned not only the paradigm but also its alternative, as I will discuss in Chapter 10. I will discuss the dark matter vs. MOND controversy as a conflict of paradigms, but my primary purpose is to provide the reader with a reasonably objective view of the major developments in the emergence of the dark matter–dark energy view of the world. Most of my own experience is in the field of galactic astronomy. So in this discussion I will emphasize galaxy-scale phenomenology which provides, after all, the primary observational evidence for dark matter that clusters on a small scale and is, possibly, directly detectable locally.

I will not discuss one very interesting aspect of the dark matter problem: the development of the astronomy of gravitational micro-lensing with the goal of detecting "massive compact halo objects" or MACHOs. This was a brilliant observational technique that spawned a new arena of astronomical research and provided the direct observational evidence that normal "baryonic" matter in the form of stellar and sub-stellar mass objects could not be the principal constituent of dark matter halos about galaxies. I refer the reader to the book on dark matter by Freeman and McNamara (2006) for a highly readable account of this development.

The level of this discussion should be appropriate for professionals as well as beginning students and interested readers with some scientific background. Therefore, the presentation is essentially non-mathematical. However, I include a pedagogic appendix that is primarily for those who are less familiar with astronomical concepts and terminology. Here I provide the most relevant formulae and definitions. This can safely be skipped by professionals or more advanced students, but the scientifically literate reader may find this survey to be useful as an introduction to the jargon as well as the more quantitative aspects of the problem. In particular, I focus on the following points:

(1) Electromagnetic radiation is the primary (but not the only) medium for observing objects in the distant Universe. What is the nature of electromagnetic radiation? How is it emitted and how does it propagate? What are spectral lines and how are they formed? How can we measure the velocity of an astronomical object toward or away from us by using spectral lines?
(2) It is important to be acquainted with aspects of scale in astronomy. What are the units of distance appropriate to galactic and extragalactic problems? How do we measure distance? What do we mean by apparent brightness and intrinsic luminosity of a star

or galaxy, and what are the appropriate physical units? What is meant by the surface brightness of astronomical objects? How do we measure the color and composition of stars and galaxies? What are the characteristics and morphological types of galaxies? What is the mass scale, luminosity, star and gas content of extragalactic objects?

(3) Familiarity with a few basic physical concepts is necessary – Newton's laws and classical mechanics – because this is how we measure the mass of gravitating systems.

(4) Dark matter is thought to be a substantial component of the entire Universe and required for the formation of observed structures such as galaxies and clusters; therefore I consider a few basic concepts of cosmology, which is the study of the structure and evolution of the Universe as a whole. I define the fundamental density parameter of cosmology and describe the known constituents of the Universe – visible matter and electromagnetic radiation. What is baryonic or non-baryonic matter? What do we mean by "dark energy"? I discuss the thermal history of the Universe and take up the question of how structure – stars, galaxies, clusters of galaxies – can form from an originally hot, highly homogeneous expanding Universe.

I assume throughout that the reader is familiar with scientific notation; that is, instead of writing 1 000 000 000, I write $10^9$, or $10^{-3}$ instead of 0.001. In the text, I write only the most basic equations, often without derivation, because of my generally qualitative and historical approach to this subject.

The style of this discussion is essentially narrative and personal. I have not only witnessed, but in some cases, been involved in these developments, so I do have a very direct interest. I have been privileged to work at institutes where much of the initial work on the dark matter problem, especially with respect to galaxies, has been carried out, and I know a number of the principal players who have shared their thoughts and enthusiasm. I have learned a great deal from the dark matter problem, not only about dark matter but also about the way in which science progresses and how scientists work. I will conclude with some general remarks on these sociological aspects of science as exemplified by the dark matter problem.

This work is a personal and by no means a complete or encyclopedic history of the subject. So I will not cite everyone who has made significant contributions to the study of the dark matter problem; I apologize in advance to those who may feel slighted. I do think that I have included reference to most of the major contributors in this field.

Finally, I hope that I can convey to the more general reader a sense of the excitement in this ongoing adventure of discovery and at least make the case that the adventure is far from complete.