The Kuiper Belt

On January 10th, 1992, a freighter en route from Hong Kong to Tacoma, Washington got caught in a storm in the northern Pacific ocean and lost overboard shipping containers filled with 29,000 plastic bathtub toys. Ten months later these brightly colored ducks, turtles, beavers, and frogs began showing up on beaches up and down the coast of Alaska. Curtis Ebbesmeyer and James Ingraham, Seattle oceanographers who had been studying wind and sea circulation in the Northwest Pacific by systematically dropping small numbers of labeled floats from fixed locations and hoping for their recovery, quickly realized that they had a massive inadvertent experiment on their hands. While large computer models can predict the physics behind the winds and currents, the large numbers of different interactions that occur make precise prediction difficult. The bathtub toys provided a windfall of data that would allow them to chart out winds and currents and understand the complex interactions in more detail than was previously possible with their more limited data sets.

Like oceanographers with their limited numbers of data point to understand the complexities of currents, astronomers studying the formation and evolution of the outer solar system have for many years had few data points around which to spin their theories. There are 4 giant planets, each created and moved around by a complex suite of interactions. Trying to piece together all of these different interactions to discern the history was like trying to use just four large shipwrecks washed up on shore to understand the flows of all of the oceans' currents.

The breakthrough for astronomers came a few months before the first plastic ducks hit the beaches near Sitka. David Jewitt of the University of Hawaii and Jane Luu then of the University of California at Berkeley, after many nights of searching, found a single faint slowly moving object in the sky. After following it for a few days it became clear that they had found the first object beyond Neptune since Clyde Tombaugh had discovered Pluto in 1929.

Today more than 800 additional members of what is now known as the Kuiper belt have been discovered in the outer solar system, and their existence is changing the picture of the dynamical evolution of the outer solar system that held just a decade ago. These objects behave in many ways like test particles which trace the gravitational effects of the rearrangements and perturbations of the giant planets. The existence of hundreds of such objects strewn throughout the outer solar system gives the concrete data which allows us to attempt to trace the history of the outer solar system as surely as tracing the routes of hundreds of plastic bathtub toys washed on beaches shows the winds and currents of the northern Pacific.

Early expectations

The prediction of the existence of a belt of small bodies beyond the orbit of Neptune was made in 1950 by Gerard Kuiper using a seemingly weak but ultimately correct line of argument. Kuiper proposed a method to conceptually reconstruct the initial disk of gas and dust from which the entire solar system formed. He began by taking Jupiter, smashing it flat, and spreading that entire mass out into an annulus between the orbits of Jupiter and Saturn. This annulus now represents the region of the nebula that went into making Jupiter. We know, however, that some material that was in initially in this region of the nebula must have been lost, since Jupiter has a higher abundance of elements heavier than hydrogen and helium than does the sun. So we add a little more mass to the annulus to make up for these lost elements and bring that region of the theoretical nebula to solar composition. Now we do the same to all of the giant planets and then to the terrestrial planets (the terrestrial planets have lost almost all of their hydrogen and helium, so a large amount of extra material has to be added in). We now have an approximate reconstruction of the mass distribution of the initial nebula (Figure 1).

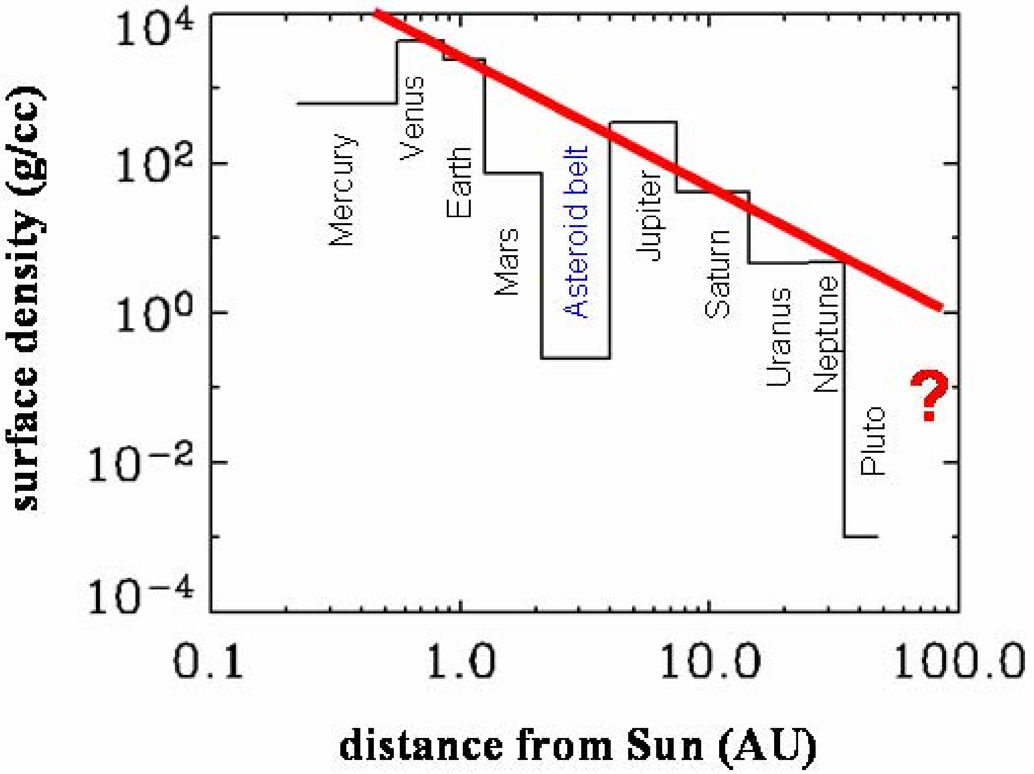


Figure 1 *An estimate of the surface density of the nebular disk from which the planets formed. Distance from the sun is in astronomical units (AU), where 1 AU is the distance from the earth to the sun. The density estimate is made by spreading the total mass in each planet into an annulus stretching to the next nearest planets and adding a little extra to account for what was thought to have been lost. In 1950 Gerard Kuiper argued that the solar system shouldn't abruptly end beyond Neptune and proposed the existence of a belt of small unseen bodies which we now know as the Kuiper belt***.**

Kuiper noted that the surface density of the nebula smoothly dropped from the inside to the outside until, beyond Neptune, the density plummeted. He reasoned that the nebula should not have an abrupt edge and that beyond Neptune was a realm where densities were never high enough to form large bodies, but where small icy objects exist instead. Furthermore, he suggested, this region could be the source for the comets which periodically come blazing through the inner solar system.

Almost 4 decades later, Martin Duncan of Queen's College and colleagues, taking advantage of the growing power of computers for dynamical simulations of the long-term gravitational influence of planets, showed that one class of comets—the Jupiter-family comets—were best explained by the existence of a band of small bodies just beyond the orbit of Neptune, which they named the Kuiper belt. Jewitt and Luu found the first object in this hypothesized Kuiper belt just 6 years later.

The edge of the solar system?

One of the first surprises after the discovery of the Kuiper belt was that it appears to contain only about 1% of the mass needed to make up for the deficit noted by Kuiper. Even more interesting, early studies by Alan Stern at the Southwest Research Institute suggested that the Kuiper belt did not even contain enough mass to have formed itself. Stern pointed out that for the largest Kuiper belt objects build up from the gradual accumulation of the smaller objects—the typical way in which solid bodies in the solar system are thought to form— would have taken longer than the age of the solar system.

An apparent resolution to this discrepancy can be seen from examining Kuiper's hypothetical initial nebula and noting that the asteroid belt also appears to be about100 times under dense. The loss of the vast majority of the original asteroids is an expected consequence of the interaction with the nearby giant Jupiter. A similar type of process is likely to have taken place in the outer solar system, where an initially much more massive Kuiper belt was severely depleted by interaction with nearby Neptune.

An interesting potential consequence of this local loss in the Kuiper belt is that somewhere beyond the influence of Neptune, the Kuiper belt might increase in density by a factor of 100 or more. Interestingly, however as more and more objects began to be detected, it became clear that the numbers of detections of Kuiper belt objects decreases dramatically beyond about 43 AU (an astronomical unit, or AU, is the distance from the sun to the earth; in AU, the giant planets have semimajor axes of 5.2, 9.5, 19.5, and 30.0 while Pluto is at 39.5). For many years arguments circulated that this drop off in detections was a simple consequence of more distant Kuiper belt objects being much fainter (the $1/r^2$ decrease in the intensity of light happens twice for Kuiper belt objects, once on the way from the sun to the object and then on the way back to the earth, so an object at 50 AU is only $(40/50)^4 = 0.41$ times the brightness of one at 40 AU).

In recent years separate analyses by Lynne Allen, then at the University of Michigan, and colleagues, and by Chad Trujillo and I showed that this lack of detections was indeed statistically significant even when taking into account the expected dimming of distant objects. The true distribution of Kuiper belt objects has a strong peak at about 43AU and then decreases quickly beyond that. My most recent analysis shows that we can rule out the existence of a resumption of a Kuiper belt at a density 100 times that of the known Kuiper belt out to a distance of more than 100 AU. It appears that we have truly found the edge (or at least *an* edge) of the solar system at about 50 AU.

The existence of an abrupt edge was what led Kuiper to suggest the presence of an unseen band of planetesimals to begin with. Some suggested that perhaps the Kuiper belt did indeed continue outward but that the bodies became significantly smaller or darker, but no physically plausible reason for such an abrupt change could be found. Shigeru Ida of the Tokyo Institute of Technology and colleagues proposed that a close encounter with a passing star could have stripped material from the outer edge of the solar system. Such an encounter would leave a very clear signature in the inclinations of the remaining Kuiper belt obects, but my analysis showed that this signature is absent. I have long speculated that a possible explanation for the edge is that one or more moderately large planets remain to be discovered in the outer solar system and that the missing mass is tied up in these bodies, but a large scale survey for the largest bodies in the outer solar system being performed by Trujillo and I has covered most of the likely area where such planets would be found and has yet to uncover anything larger than our discovery of Quaoar (see box), which is about half the size of Pluto (though we're continuing to look).

The structure of the Kuiper belt

In addition to an expectation of a much higher mass to the Kuiper belt, initial expectations of the Kuiper belt were that it would be a quiescent, smoothly varying band of objects slowly decaying in density with distance from the sun. The objects were expected to be in relatively circular (low eccentricity) orbits slowly dropping off in number at increasing semimajor axis. The objects closest to Neptune might be expected to have slightly higher eccentricities owing to the gravitational influence of the giant planet. The true picture of the Kuiper belt that has emerged from the first decade of observations is dramatically different than these expectations and is best shown by examining a plot of semimajor axis versus eccentricity for the known objects in the Kuiper belt (Figure 2). The Kuiper belt contains multiple complex dynamical structures. These structures are the clues which will allow us to untangle the complex interactions and begin to understand the early evolution of the outer solar system.

The Plutinos

The first unusual structure noted in the Kuiper belt was the pile up of objects with semimajor axes of approximately 39 AU and moderate to large eccentricities. Pluto, with a semimajor axis of 39.5 AU and an eccentricity of 0.25, is the largest member of this population, which have come to be known as the Plutinos. An important clue to the origin of Pluto and the Plutinos is that they are situated in 3:2 mean motion resonance with Neptune, meaning that every time Neptune circles 3 times around the sun, a Plutino circles precisely twice.

For many years this curious arrangement of the orbits of Neptune and Pluto was considered a coincidence that fortuitously allowed Pluto to survive by preventing close encounters with Neptune even though their orbits cross. It was hypothesized that many other such objects might have existed at one time but that one by one these other objects had close encounters with Neptune and were scattered away. In 1993, however, Renu Malhotra of the University of Arizona realized that Pluto's orbital resonance and high eccentricity, though unusual, would be expected if at some point in the past Neptune's orbit had slowly expanded outward and Pluto had been captured and pushed outward into its 3:2 resonance (see box). She calculated that Neptune would have to have moved by as much as 10 AU to explain the orbit of Pluto. Another important constraint is that the movement of Neptune must have been quite smooth or Pluto would have escaped the resonance.

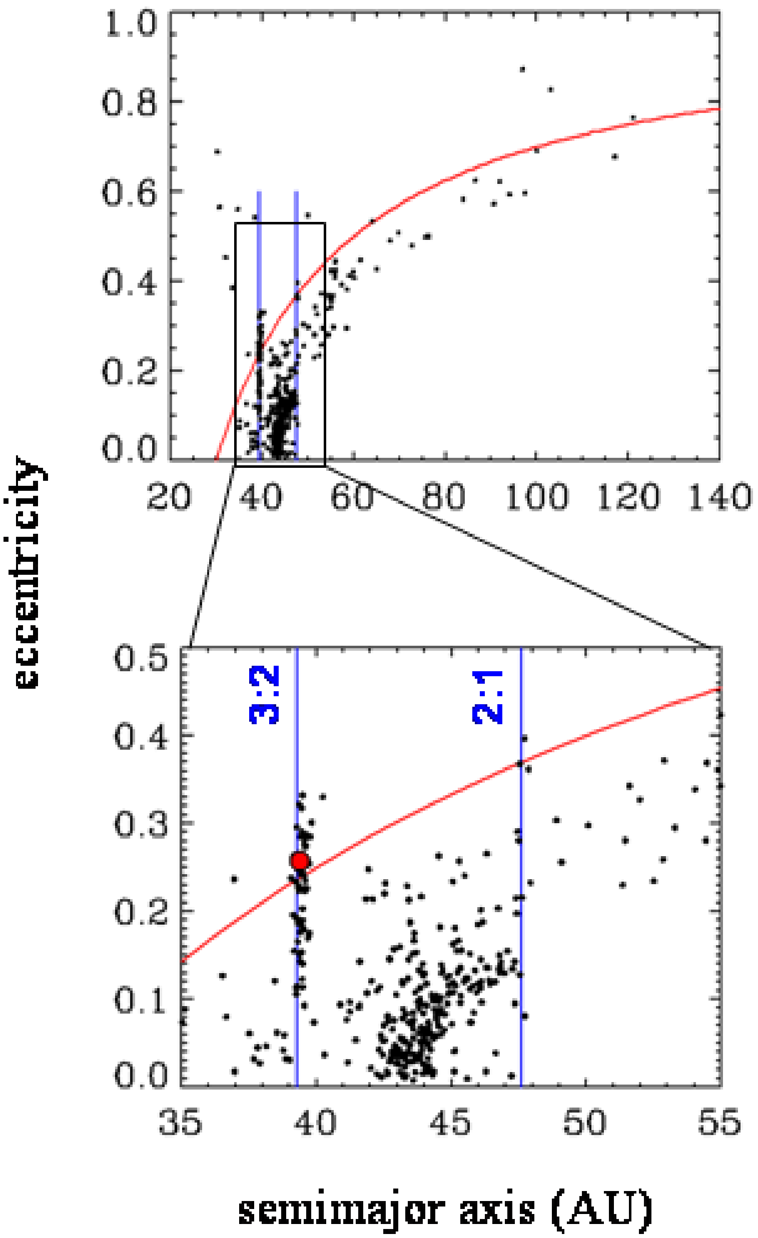


Figure 2 The structure of the Kuiper belt. A plot of the semimajor axis versus the eccentricity of the known objects in the Kuiper belt shows many unexpected dynamical structures. The red line shows the location of objects which have sufficiently high eccentricities to cross the orbit of Neptune. The blue lines show the locations of objects which are in orbital resonances with Neptune. In the 3:2 resonance, the Kuiper belt object orbits the sun twice for every three times Neptune orbits. In the 2:1 resonance the Kuiper belt orbits once for every two Neptune orbits.

While this process could naturally explain the orbit of Pluto and also correspondingly of the Plutinos, which are all on similar resonant high eccentricity orbits, why would Neptune's orbit have expanded? The answer was explained almost a decade before the discovery of the Kuiper belt by Juan Fernandez and Wing Ip at the Max Planck Institutes, who examined the interaction between

Neptune and all of the small bodies left over after the formation of the giant planets. Neptune moves from simple conservation of momentum. As small bodies approach Neptune and are scattered inward toward the other planets or outward towards the edge of the solar system, Neptune must move a tiny amount in the opposite direction. In general, Neptune has an equal probability of scattering an object inward or outward and thus moving outward or inward. But objects scattered outward are still gravitationally bound to the sun and will eventually make their way back to have another close encounter with Neptune again. The net effect is that objects scattered outward do nothing to Neptune's orbit. Objects scattered inward, however, have a much lower chance of coming back, as they are likely to hit or gravitationally interact with one of the other giant planets. If the object gets close to the giant Jupiter it is likely to be thrown out of the inner solar system completely (which, incidentally, leads to the formation of the Oort cloud, the source of the second type of comets). The net effect is that Neptune scatters more objects inward than outward and thus Neptune itself must move outward. The other giant planets move, too, as they scattered objects, Uranus and Saturn, like Neptune tend to scatter more objects inward then outward, and so move further from the sun, while Jupiter, with no interior giant planets, tends to scatter more objects outward than inward, and thus moves closer to the sun.

The smoothness of the planetary migration, which plays an important part in the capture of the resonant objects, is a function of the size and numbers of objects in the disk. Large numbers of small objects will lead to a smooth migration, while a smaller number of larger objects will make the migration jumpy and cause objects to be lost from the resonances.

Though Fernandez and Ip had proposed this migration of the outer planets years earlier, the idea of large scale motion of giant planets did not really catch on with astronomers until Malhotra's suggestion that the pile-up of Plutinos in an inescapable consequence of such migration, at which point the idea that the giant planets migrated instantly became part of the standard lore of the early evolution of the outer solar system.

The scattered belt

The next most strikingl aspect of the structure of the Kuiper belt is the long tail of objects at high semimajor axis and high eccentricity (Figure 2). The fact that the majority of these objects have closest approaches to the sun similar to Neptune's semimajor axis is a give away as to their origin. These are the remnants of the scattering and migration process described above. These objects must have had a close approach with Neptune sometime in the past and are now on large eccentric orbits which will eventually lead them to have additional close encounters with Neptune. Some of these objects have close encounters with Neptune will be scattered inward and will eventually make their way into the inner solar system to become the Jupiter family comets that we see today. These scattered objects in the Kuiper belt thus preserve a small remnant signature of the process of scattering and migration that once moved Neptune long distances through the outer solar system and also provide a direct look at the objects that are likely to be the source of the Jupiter-family comets.

The classical Kuiper belt

The final population visible in the Kuiper belt consists of the bodies in relatively circular orbits between about 41 and 48 AU (Figure 2). These bodies most resemble the earliest expectation of what Kuiper belt objects should be and thus have become known as the classical Kuiper belt objects. But a close look at the classical Kuiper belt objects shows that even these are not the pristine undisturbed Kuiper belt envisioned. My work has shown that the classical objects appear to be two separate superimposed populations, one a low-inclination dynamically unperturbed population, and the second a much higher inclination dynamically stirred population.

The existence of separate dynamically hot and cold populations existing in the same place is almost as odd as taking a pot of water off of the stove and finding that it is half scalding and half tepid. No simple dynamical process is capable of starting with a dynamically unexcited population and exciting half of it while leaving the other half unperturbed. The only conclusion that seems possible is that the two separate populations were made either at different times or in different places or both and are now fortuitously superimposed. Interesting evidence supporting the separate formation of the two populations was found by Hal Levison and Alan Stern, who showed that the largest objects in the Kuiper belt are all part of the dynamically hot population. Chad Trujillo and I also found that the unexcited cold population is distinctly different in color than the other objects (though no one knows what colors actually mean, so they are difficult to interpret). Both of these findings suggest that the different dynamical populations of the classical belt are physically distinct and arose separately. The natural conclusion is that the cold classical population, which consists only of small red bodies, is the only pristine part of the Kuiper belt, and everything else formed elsewhere (presumably in the more dense regions closer to the sun, where the larger bodies could more easily have formed) and was somehow transported to where we see it today.

A nice explanation for these two separate populations has recently been proposed by Rodney Gomes of the Observaotrio Nacional in Brazil who has used ever more powerful computer simulations to suggest that the populations are also a (somewhat) natural consequence of the migration of Neptune. While previous work had by necessity considered the processes of scattering and migration separately, steady increases in computer speed allowed Gomes to combine the two processes and see their interplay. He found that sometimes objects that were scattered into high inclination orbits that intersected the region of the classical Kuiper belt. If Neptune were not migrating, these objects would eventually have another close encounter with Neptune and scatter elsewhere. But with Neptune's migration, occasionally objects remained trapped in these seemingly odd orbits. Interestingly, the objects that were trapped in these orbits all came from regions closer to the sun, while the colder population remained unperturbed. Thus the dilemma above seems resolved, with the small red dynamically cold objects in the outer solar system being the only objects which actually formed in place and the larger high inclination population being an interloper from closer in in the solar system.

Pushing it all out

With the explanation of the origin of the dynamically hot population, the entire dynamical structure in the Kuiper belt appears to be explained, but to get his simulation to resemble the solar system Gomes had to make one critical *ad hoc* assumption. In his simulations, Neptune always migrates to the very edge of the initial disk of the Kuiper belt. To make Neptune stop its migration at 30 AU, Gomes had to end the main Kuiper belt at 30AU. To allow the presence of the cold classical Kuiper belt, Gomes had to artificially put a less massive belt of objects beyond this initial edge. No plausible explanation for this initial configuration could be found.

Hal Levison and Alessandro Morbidelli at the Observatoire de la Cote d'Azur devised a scheme to solve the mystery of the location of Neptune simultaneously with the mystery of the edge of the solar system near 50 AU. They noted that the edge of the classical Kuiper belt appears to coincide quite nicely with location of the 2:1 mean motion resonance of Neptune at 48 AU. They hypothesized that this location was not a coincidence and that perhaps the *entire* Kuiper belt – including the supposedly primordial dynamically cold classical objects – had been pushed out by the process of resonance capture. With the continued advances of computer power, they examined Neptune's migration through a massive disk of particles which ended at 30AU. The computer advances allowed them to resolve the disk into an ever increasing number of ever smaller objects. The results were surprising: not only did they get the expected capture into the 3:2 resonance and into the 2:1 resonance, but some objects that had been initially pushed out in the 2:1 resonance had dropped out of the resonance and ended up with low eccentricities in the region of the classical Kuiper belt. The appearance of the low eccentricity objects was a surprise, as the resonance capture and pushing out of objects had always previously been assumed to monotonically increase the eccentricity of the object (see box). So where did the low eccentricity objects come from? When a computer simulation gets sufficiently complicated, understanding the results are almost as difficult as understanding the real universe. But the big advantage in a simulation is that you can change things and see what happens. Levison and Morbidelli re-ran their simulation artificially forcing Neptune to migrate but ignoring the mass of the objects. They found that the eccentricities increased as expected. Further analysis found an effect that had not been previously noticed. The Kuiper belt objects which are captured into resonances are all individually small, but collectively they exert a torque on the orbit of Neptune which causes Neptune to precess differently than it would in the absence of the objects. This precession in turn causes a back-reaction on the Kuiper belt objects which causes the eccentricities of the objects to oscillate. Previously, it was though that any objects in resonances would have to have high eccentricities on account of migration, but with the oscillation causes by Neptune they now have a much wider range. When the migration is sufficiently jumpy some of the objects will fall out of the 1:2 resonance as they are being pushed outward and these objects will have a similar range of eccentricities as seen in the classical Kuiper belt.

In this story, *all* of the Kuiper belt objects formed inside of the present location of Neptune and were carried out as Neptune migrated. The edge at the location of the 2:1 location occurs because that is the most distant resonance that can capture and move objects. Intuitively, it appears difficult to reconcile this mechanism with the physical differences in the separate classical populations, but when the dynamics become sufficiently complicated intuition often does not serve very well. To date, the computer power does not exist to attempt to understand if this process will segregate objects from different regions of the initial nebula. Such a reconciliation of the physical observations and the computers simulations remains an important test of this new idea.

The current story

Putting all of the current ideas together we are left with a coherent and possibly even correct picture of the formation and evolution of the outer solar system. The current story goes like this. First, the entire disk was much smaller than ever previously expected. The nebula must have had an edge at 30 AU for Neptune to have ended up there now. Uranus and Neptune could have been well inside 20 AU, which relieves the long standing problem of the difficulty of the formation of these planets at their current. Neptune and the other giants began migrating through the disk of planetesimals, pushing out objects in resonances and scattering them into elliptical orbits. The force of the planetesimals on Neptune causes Neptune to precess, which in turn caused eccentricities of the planetesimals to oscillate rather than simply increase. As Neptune migrated some of the scattered objects became stranded in what we now call the dynamically hot classical belt and some of the objects fell out of the 2:1 resonance due to Neptune's occasional large jumps to become what we call the cold classical belt. When Neptune reached the edge of the disk at 30 AU everything came to an end and we are left with the planetary arrangement that we have today.

If Kuiper recreated his initial nebula using these current ideas, the density in this disk would now be higher closer to the sun, with all of the mass of the giant planets inside about 20 AU. One thing would remain the similar, though. At 30 AU the disk would end. And Kuiper might think to himself "it does seem natural that the solar system should have such an abrupt edge," and begin to look for a new answer to his dilemma.

The full story is not yet know. Many of the advances will come with the always increasing power of computer modeling to add more and more particles to a simulation that interact in more and more realistic ways. But in the end, while we might know all of the relevant forces, computer modeling of the full complexities of the real universe is never enough. The final answers are likely to come from finding one single plastic duck, unexpectedly washed ashore up on the shores of Ireland, and using it to trace the complex paths of all of the many forces that move plastic ducks and Kuiper belt objects to where they are found today.

BOX 1: How to find a Kuiper belt object

Kuiper belt objects are found by their motion with respect to the stars, the same way that Pluto was found 74 years ago. Astronomers point their telescope to an anonymous area of the sky (usually one that is close to the ecliptic, the plane of the planets in the solar system), typically take three image spread over the course of a few hours, and compare the images. The majority of objects moving in the images turn out to be the nearby asteroids (which are a minor component of the solar system compared to the Kuiper belt, but are much closer and therefore brighter). Over the course of a few hours, the motions that are seen for objects in the solar system are almost entirely due to parallax caused by earth's motion, thus the distant Kuiper belt objects are distinguished from the nearby asteroids by their much slower motion. In the images shown from our wide-field survey shown in the Figure below we have to search an area approximately 10000 times that shown for each object that we find. Computer aided image processing greatly simplifies the search process compared to the visual search for moving objects during which Clyde Tombaugh found Pluto.

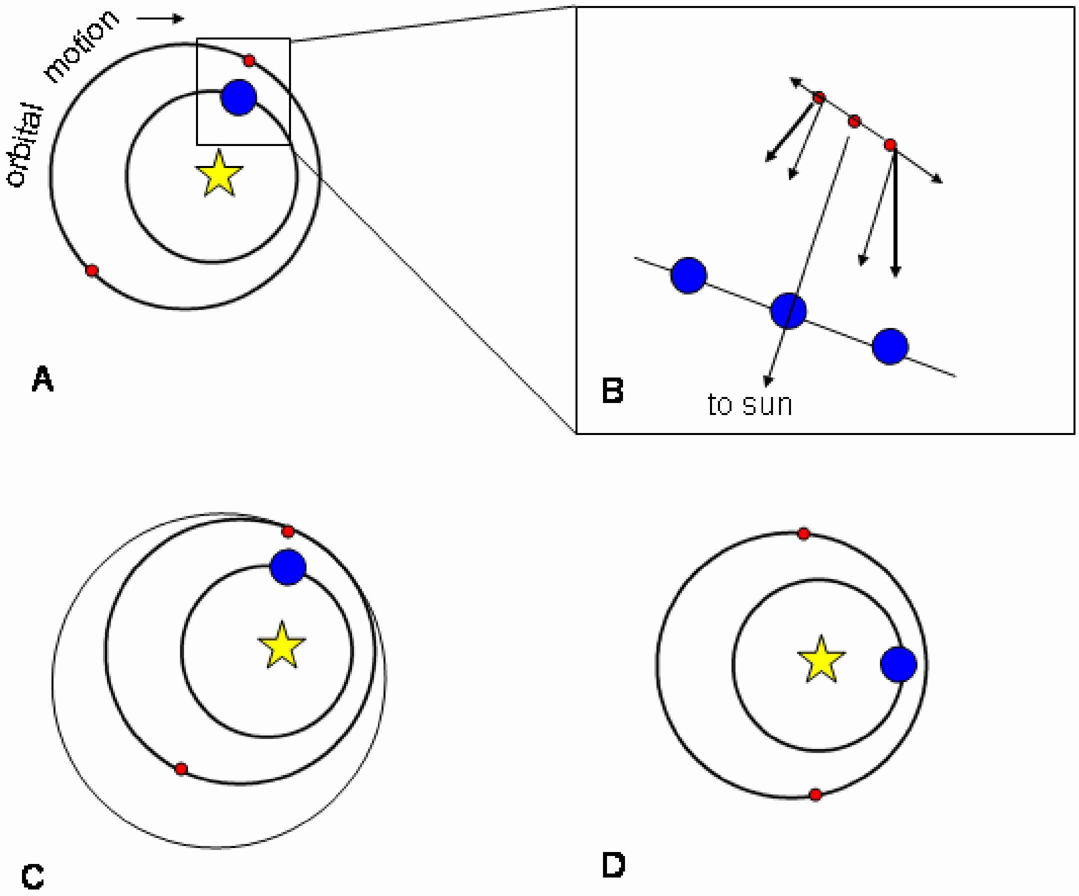
The discovery of Quaoar, a Kuiper belt object about half the size of Pluto, on 3 images from 4 June 2002. The images were taken 90 minutes apart from the Oschin Schmidt telescope at Palomar Observatory and show one object moving slowly with respect to the background stars. Quaoar is the largest known Kuiper belt object after Pluto and is on a remarkably circular orbit 43 AU from the sun.

BOX 2: How an orbital resonance can move a Kuiper belt object

To understand how a Kuiper belt object can be pushed outward in its orbit by the process of resonance capture, one first needs to understand the dynamics of the resonance. A mean-motion resonance occurs any time one body orbits the sun once while another body orbits the sun in an integer fraction of that time. The simplest resonance is the 2:1 resonance, where one object (Neptune, in our example) orbits the sun twice for every time the other object (a Kuiper belt object) orbits the sun once. A consequence of this resonance is that every time the Kuiper belt object returns to one spot in its orbit, Neptune is always in the same spot in its orbit.

The important part of the interaction between Neptune and the Kuiper belt object comes at the moment the two have their closest approach (see Figure). As long as the Kuiper belt object has even a slight eccentricity, the forces are asymmetric around closest approach; the Kuiper belt object will always get a slight tug in the direction towards the perihelion (closest approach to the sun) of the orbit. Such a tug, if in the direction of motion of the Kuiper belt object orbit, will add angular momentum to the orbit which in turn causes an increase in semimajor axis (and eccentricity) and a decrease in orbital velocity. A tug in the direction opposite the orbital motion causes a corresponding decrease in semimajor axis (and eccentricity) and increase in orbital velocity. The overall effect is that the orbit of the Kuiper belt object in resonance will always oscillate about an equilibrium position, where the positive and negative tugs balance, by slow oscillations of its semimajor axis, eccentricity, and orbital velocity.

If Neptune’s semimajor axis expands by a small amount, Neptune’s orbital velocity slows, so the closest approach occurs a little past the equilibrium position. As before, the Kuiper belt object orbit and eccentricity slightly expand as the object works back toward the equilibrium position. Continued smooth outward movement of Neptune will cause continued movement of the Kuiper belt object and a continued increase in the object’s eccentricity. If Neptune’s orbit ever jumps by too large of an amount, however, the object may not be able to reestablish its equilibrium and it will be lost from the resonance and its orbit will quit expanding.



Resonant dynamics in the solar system. (A) Neptune (blue) is in a 2:1 resonance with a Kuiper belt object (red). Every time Neptune orbits the sun twice, the object orbits once. As a consequence, the closest approach of Neptune and the object always appear at the same spot. (B) The forces at closest approach are asymmetric; the object is slightly closer immediately post-encounter than pre-encounter, thus is gets a net kick in the direction of motion. (C) This kick adds angular momentum, which increases the semimajor axis and eccentricity of the orbit (greatly exaggerated here!). The next encounter between the object and Neptune will therefore come slightly later. (D) The net effect of these perturbations is for the object to oscillate about the equilibrium position show, where kicks in the direction of motion and opposite that are equal If Neptune's orbit is expanding, this process leads to the capture and expansion of the orbit of the Kuiper belt object.