THE ORIGIN AND EVOLUTION OF PLANETARY NEBULAE

SUN KWOK

University of Calgary, Canada

Contents

1

History and overview

The first planetary nebula was observed by Charles Messier in 1764 and was given the number 27 in his catalog of nebulous objects. The final version of the Messier catalog of 1784 included four planetary nebulae (PN) together with other nonstarlike objects such as galaxies and star clusters. The name planetary nebulae was given by William Herschel, who found that their appearances resembled the greenish disk of a planet. With better telescope resolution, nebulae that are made up of stars (e.g., galaxies) were separated from those made up of gaseous material. PN were further distinguished from other galactic diffuse nebulae by that fact that PN have definite structures and are often associated with a central star. This distinction became even clearer with spectroscopy. The first spectrum of a PN (NGC 6543) was taken by William Huggins on August 29, 1864. The spectra of PN are dominated by emission lines, and not a continuous spectrum as in the case of stars. The first emission line identified was a Balmer line of hydrogen (Hβ), although stronger unidentified lines could be seen in the spectrum. Since the spectra of PN are entirely different from those of stars, their luminosity cannot be due to reflected starlight.

The idea that PN derive their energy from a nearby star was first considered by Herschel (1791). However, no further progress was made for another century. Hubble (1922), using data obtained with the Mount Wilson 60- and 100-in. telescopes, found a correlation between the magnitude of the central star and the size of the nebula. He therefore argued that the emission-line spectrum of PN is the result of the nebula absorbing the continuous radiation from the central star. In order to explain the strength of the Hβ line, Menzel (1926) suggested that all the stellar output beyond the Lyman limit (912 Å) must be utilized to ionize the hydrogen (H) atom. The mechanism that the lines of hydrogen and helium (He) are emitted as the result of recombination between the nucleus and electron after the nebula is photoionized was quantitatively developed by Zanstra (1927). Most importantly, Zanstra was able to determine the number of Lyman continuum photons emitted from the observed ratio of the Balmer line to stellar continuum flux, and was therefore able to deduce the temperature of the central star (see Section 7.1.1). The central stars of PN were found to have very high temperatures, which were much hotter than those of any other known stars at the time.

However, a number of strong nebular lines remained unidentified by laboratory spectroscopy and were suggested to be due to some unknown element "nebulium." The

strength of the lines led to the conclusion that these lines must originate from known elements of high abundance but are emitted under unusual conditions. One such condition is the low density of the interstellar medium. Russell *et al.* (1927) speculated that certain atoms with metastable states, which do not have the time to emit radiation because of collisional deexcitation in the high-density terrestrial environment, will radiate under interstellar conditions. Bowen, in 1928, identified eight of the strongest nebular lines as being due to metastable states of N^+ (singly ionized nitrogen), O^+, and O^{++}. These metastable states lie a few electron volts above the ground state and can be collisionally excited by electrons freed by the photoionization of hydrogen. The presence of highly excited, strong optical lines of oxygen was explained by Bowen (1935) as being the result of a fluorescence mechanism.

Since the forbidden lines are collisionally excited, and therefore remove energy from the kinetic energy pool of the electrons, they represent a major source of cooling of the nebula. Menzel and Aller (1941) were able to show that, no matter how hot the central star, cooling by the forbidden lines limits the electron temperature to <20,000 K.

Observations with better spectral resolution led to the discovery that the emission lines in PN are broad, or even split. This was correctly interpreted as expansion, and not rotation, of the nebula (Perrine, 1929). With the adoption of a size of 0.3 pc and an expansion velocity of 30 km s^{-1}, the dynamical lifetime of the PN can be estimated to be $\sim 10^4$ yr.

1.1   Planetary nebulae as a phase of stellar evolution

At the beginning of the 20th Century, when stars were believed to evolve from high temperatures to low temperatures, PN were thought to be very young stars because of their high temperatures. From his studies of the velocity distribution of PN, Curtis (1918) found that PN are more similar to late-type stars and are unlikely to be young objects. Theoretical understanding of the origin of PN began with Shklovsky (1956b), who suggested that PN are progenitors of white dwarfs (WDs) and descendants of red giants. By tying PN to red giants and white dwarfs, Shklovsky recognized that these stars must be evolving rapidly. This view was supported by Abell and Goldreich (1966) who used the expansion velocities of PN and the escape velocities of red giants to argue that PN are the ejected atmospheres of red giants. Using the total number of galactic PN of 6×10^4 (as estimated by Shklovsky) and a lifetime of 2×10^4 yr, Abell and Goldreich showed that PN must be forming at a rate of 3 per year. Since this is of the same order as the number of stars leaving the main sequence, they suggested that practically all low-mass stars will go through the PN stage. This established the importance of PN in the scheme of stellar evolution.

Although Shklovsky successfully drafted a qualitative scenario for PN evolution, the details of the transition from red giants to PN to WDs remained very poorly known for another 20 years. For example, in the 1960s it was commonly believed that the horizontal branch was an essential phase of the evolution of low mass stars. However, the way that PN are related to horizontal branch stars was not at all clear.

In this book, we present a modern view of the origin and evolution of PN, tracing their origins to the mass loss on the asymptotic giant branch (AGB). The circumstellar envelopes that are created by the mass-loss process over a period of 10^6 yr are swept up by a new fast stellar wind into the shell-like structure that we observe in PN. The
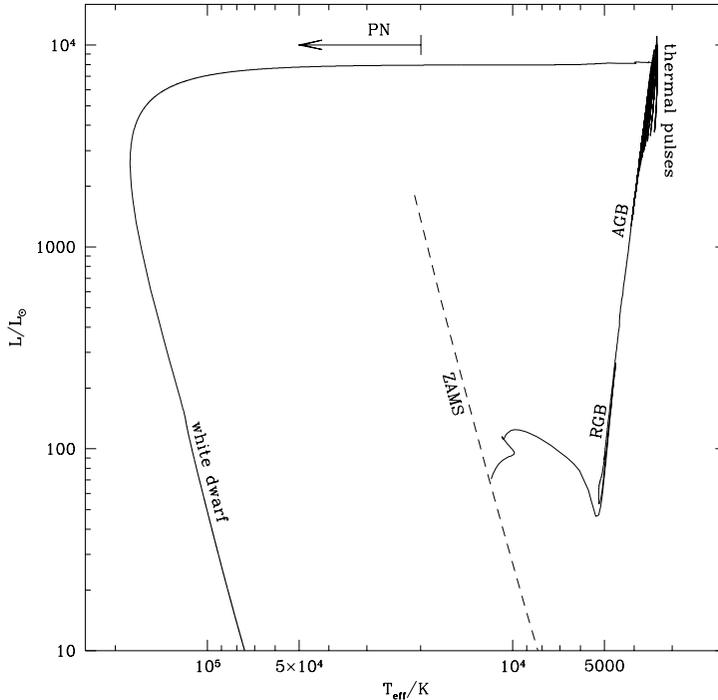
Fig. 1.1. The evolutionary track of a $3M_\odot$ star on the H-R diagram beginning from the zero age main sequence (ZAMS), through the red giant branch (RGB) and AGB to PN and ending as a white dwarf (Figure from T. Blöcker).

interaction of the two winds creates a high temperature-bubble that exerts pressure on the shell and causes it to expand. The core of the AGB star, having lost its envelope, evolves to the higher temperature as its thin remaining H envelope is burnt up by nuclear processes. The increasing output of UV photons will gradually ionize the shell, which at the same time is growing in mass as more AGB wind material is swept up. When the stellar H envelope is used up by nuclear burning, the core will cool down and decrease in luminosity to become a white dwarf (Fig. 1.1).

1.2   Discovery and identification

Based on their diffuse appearances, PN were first cataloged together with galaxies and clusters as part of the New General catalog of Clusters and Nebulae (NGC) in 1887. Many PN carry their NGC designations to this day. In the 20th Century, new PN were discovered either by their appearances on photographic plates or their emission-line spectrum. For example, PN were identified by Abell, using photographs obtained with the Palomar 48-in. Schmidt telescope, and by Minkowski, using objective prism plates taken with the Mt. Wilson 10-in. telescope. Examinations of the Palomar Atlas by Abell (1966), Kohoutek, and others have led to the identification of large numbers of PN. Through objective prism surveys, hundreds more PN were found by Minkowski (1964), Henize (1967), and Thé (1962). More recent discoveries of PN have used a number of methods:

Table 1.1. *Number of known planetary nebulae*

Year	Objects Called PN	True & Probable PN	Possible PN
1967 PKCGPN	1063	(846)	
1992 SECGPN	1820	1143	347
1996 *1st suppl.*	+385	+243	+142
1998 AAO/UKST Hα survey	+>300	+>150	

Table adapted from Acker (1997).

- Search for nebulosity by using photographic plates or CCD frames
- Comparison of red and infrared National Geographic-Palomar Observatory Sky Survey (POSS) plates
- Search for radio emission from *Infrared Astronomical Satellite (IRAS)* sources with appropriate colors
- Systematic survey of globular clusters and the galactic bulge
- Hα survey of the galactic plane

The first catalog devoted exclusively to PN was made by Curtis (1918), which contained photographs of 78 PN. This number was increased to 134 in the catalog of Vorontsov-Velyaminov (1934). The catalog of Galactic Planetary Nebulae by Perek and Kohoutek in 1967 has over 1000 PN included. In the Strasbourg-ESO PN catalog (Acker *et al.*, 1992), 1,143 objects are listed as true PN, 347 as possible PN, and another 330 as mis-classified PN. Since the publication of the Strasbourg-ESO Catalog, a number of PN candidates, selected based on their *IRAS* colors, have been confirmed by optical and radio observations (van der Steene *et al.*, 1995, 1996). These and other new PN are included in the First Supplement to the Strasbourg-ESO Catalog of Galactic PN (Acker *et al.*, 1996). In the Anglo-Australian Observatory/UK Schmidt Telescope Hα survey of the southern galactic plane, many new faint and extended PN were discovered (Parker and Phillipps, 1998). When this survey is completed, there is a potential of nearly doubling the number of PN cataloged. A summary of the numbers of known PN is given in Table 1.1 and a plot of the galactic distribution of PN is shown in Fig. 1.2.

The numbers in Table 1.1, however, do not represent the total population of PN in the Galaxy. Many PN are hidden by interstellar extinction in the galactic plane, and most of the PN on the other side of the galactic center are not seen. Old PN have a very low surface brightness and are difficult to identify. Distant PN are stellar in appearance and cannot be easily distinguished from stars. It is estimated that the total number of PN in the Galaxy can be 10 times higher (see Chapter 18).

1.3   Confusion with other galactic objects

The identification of PN is based on a combination of morphology (shell plus central star) and spectroscopy (strong emission line spectrum with little or no continuum, see Fig. 1.3). The most common confusing sources are emission-line galaxies, reflection nebulae, HII regions, symbiotic stars, M stars, and other emission-line stars. For example,
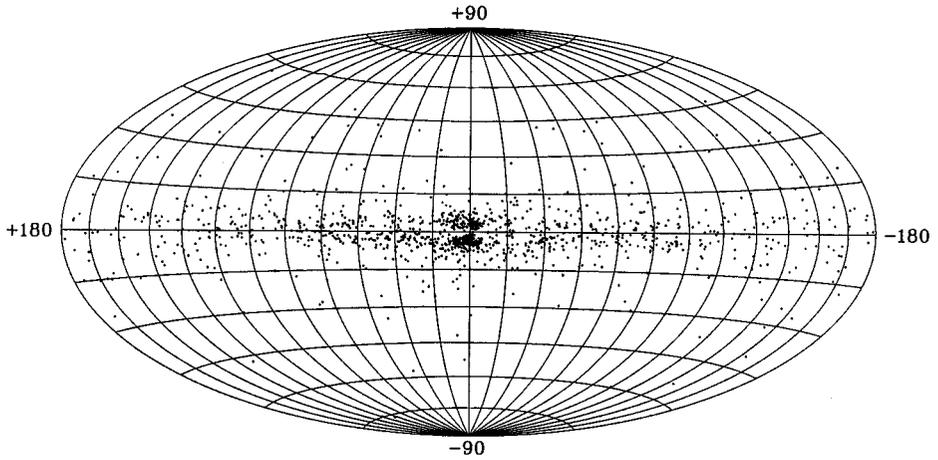
Fig. 1.2. The galactic distribution of PN in the Strasbourg-ESO catalog of Galactic PN.
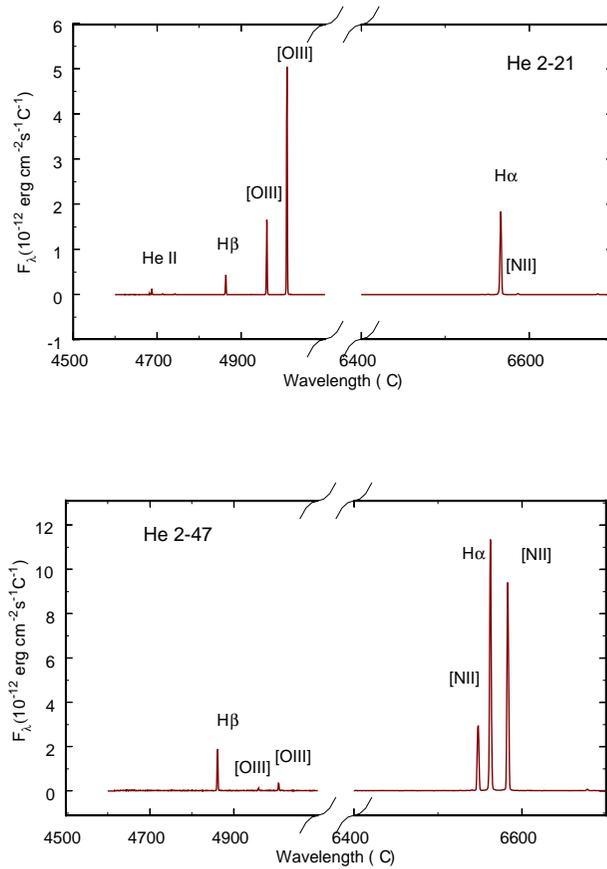

Fig. 1.3. Top: typical optical spectrum of PN. Bottom: optical spectrum of a low-excitation PN (data from T. Hua).
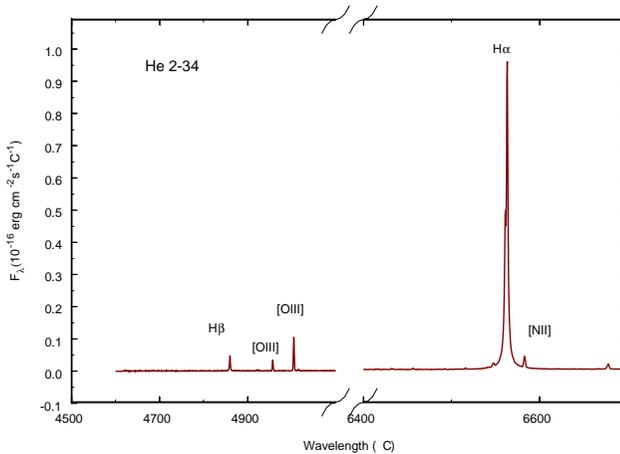
Fig. 1.4. Optical spectrum of a symbiotic star.

symbiotic stars show many similar nebular properties as PN but are classified separately because of the simultaneous presence of TiO absorption bands, and they are commonly believed to be interacting binary stars. Figure 1.4 shows the optical spectrum of the symbiotic star He2-34. From the optical spectrum alone, it is very difficult to distinguish it from a PN. Only the detection of water band and long-term near-infrared variability confirms the presence of a Mira in the system (Feast *et al.*, 1983).

Ring nebulae around massive stars also have morphologies similar to PN (Chu, 1993). They are classified differently only because of their sizes, or the luminosities of their central stars. Since these properties are distance dependent, the separation of these two classes of objects is not always easy. Many ring nebulae around massive stars were once cataloged as PN, for example, AG Car (PK $289 - 0°1$) and NGC 6164-5 (PK $336 - 0°1$). There are also nebulae classified alternately between PN and ring nebulae, e.g., M1-67 and We21. In the case of M1-67, it was first classified as a HII region by Sharpless (1959) but was included in the PN catalog of Perek and Kohoutek (1967) based on its high heliocentric velocity. Cohen and Barlow (1975) suggested that it is a ring nebula based on its infrared and radio properties, and as a result it was removed from the PN catalog by Kohoutek (1978). However, its PN classification was again suggested by van der Hucht *et al.* (1985). Finally, a detailed abundance analysis of the nebula confirms that it is ejected from a massive star (Esteban *et al.*, 1993). The example of M1-67 shows that misclassification can easily occur for less well-studied objects.

Unfortunately, there is not a universally accepted definition of PN. As an example, whereas Acker *et al.* (1992) excluded symbiotic stars, Kohoutek (1994) has continued to include them in his supplements to the catalog of Galactic Planetary Nebulae (Perek and Kohoutek, 1967). Kohoutek (1989) used a combination of observational properties of the nebula and the central star to define a PN. For example, Kohoutek places density, size, and expansion velocity ranges on the nebula, and temperature, luminosity, and gravity limits on the central star in order for an object to qualify as PN. These observational definitions not only reflect properties commonly observed in

PN, but also represent the imposition of our theoretical understanding of the phenomenon. Clearly observational criteria alone are not sufficient and a combined approach is necessary.

One could go a step further and define PN as ionized circumstellar shells showing some degree of symmetry surrounding a hot, compact star evolving between the AGB and WD phases. Even in such a restrictive definition, it is still not clear how binary stars fit in. For example, PN with binary nuclei can go through mass transfer followed by thermonuclear ignition, which makes them very similar to symbiotic stars or novae. One or more mass transfer phases can occur in an interacting binary system, leading to many different evolutionary scenarios. Mass loss can occur under some of these scenarios, resulting in a PN-like object. For example, the nebular spectra of the symbiotic stars V1016 Cygni and HM Sge are believed to be due to the ionization of the Mira stellar wind by the companion white dwarf. In this book, I will avoid these complications and concentrate the discussion on single star evolution.

1.4   Plantary nebulae as a physics laboratory

PN present an ideal laboratory for the study of the interaction between radiation and matter. The system is simple. All the energy of a nebula is derived from a single source, the central star. Radiation emitted by the star is absorbed and processed by the nebula, which contains matter in ionized, atomic, molecular, and solid-state forms. Because early (pre-1970s) observations of PN were limited to the visible region, our knowledge was restricted to the ionized gas component. Through active interactions between atomic physics and nebular observations, considerable progress has been made. For example, nebular densities and temperatures can be measured by comparing the strengths of forbidden lines (see Section 3.5). However, such determinations depend on accurate values for the spontaneous decay rates and the collisional cross sections. Since forbidden lines include magnetic dipole and electric quadrupole transitions, the observations of these nebular lines stimulated the calculations of the wave functions of multielectron atoms and ions and the corresponding transition probabilities (Shortley *et al.*, 1941; Aller *et al.*, 1949). Applying the techniques of quantum mechanical scattering theory, Seaton (1954b) calculated the collisional cross sections of many ions.

With the use of the assumption of Zanstra, that all the Lyman photons are absorbed in the nebula, the relative strengths of the Balmer lines can be determined by solving the equation of statistical equilibrium if the spontaneous decay and recombination rates are known. The early work of Plaskett (1928) contained only seven levels. This was later improved by Menzel and Baker (1937), who set up an exact algebraic solution to the equations. Consequently, the relative intensities of the Balmer lines can be calculated by approximation techniques and can be compared with observations. With greatly improved computing capabilities in the 1960s, the theory of Balmer decrement was developed to a high degree of accuracy (Brocklehurst, 1970).

The confrontation between theory and observations continues as the quality of both nebular spectroscopy and computational methods improve. Early photographic spectrophotometric measurements were improved by photoelectric calibrations. This was followed by the use of the electronic camera, the image tube, the image-tube scanner, and more recently, the charged-coupled device (CCD). Since the early spectroscopic observations of PN by Wright (1918), extensive databanks on emission lines were built

up by Aller *et al.* (1955; 1963) and by Kaler *et al.* (1976). Recent advances in CCD technology have made possible high-resolution spectroscopy with high accuracy, and a substantial increase in the quantity of spectral information has been generated. For example, the number of detected and identified lines from the PN NGC 7027 has increased from the ∼250 tabulated in Aller (1956) to more than 1,000 (Péquignot, 1997). This advance has created the need for determinations of the energy levels of many new atomic of ions of common elements and their transition rates. The analysis of nebular spectra is now performed by computer codes known as photoionization models. These models store a large amount of atomic data in the code and use certain elemental abundance and stellar and nebular parameters as inputs. The calculated emergent line spectrum is then used for comparison with observations. The continued refinements of the atomic data have led to a reasonable agreement between the photoionization models and the optical spectra of PN.

This book is roughly organized into three parts. In Chapters 2-6 we describe the physics of the nebula. The physical processes in the ionized component are discussed in Chapters 2-4. The physics of neutral gas and dust components, which were discovered as the result of millimeter-wave and infrared observations, is treated in Chapters 5 and 6. The properties of the central stars of PN are summarized in Chapter 7 and the morphologies of PN are described in Chapter 8.

The PN phenomenon has its origin in the preceding stellar evolutionary phase, the AGB. The structure of AGB stars, and in particular the mass loss that occurs in that phase, is described in Chapter 10. The theory of evolution of central stars of PN is summarized in Chapter 11. The effects of AGB mass loss on PN formation and the subsequent dynamical evolution are discussed in Chapters 12 and 13.

The immediate progenitors and descendants of PN, the proto-PN and WDs, are discussed in Chapters 14 and 15, respectively. The formation rate and galactic distribution of PN (Chapter 18) and the testing of the evolutionary models of PN (Chapter 17) are dependent on an accurate knowledge of the distance scale (Chapter 16). The contributions of PN to the chemical structure of galaxies are presented in Chapter 19. The applications of PN as a tool to study the large structure of the universe are discussed in Chapter 20.