CHAPTER ONE

Introduction and Preview

In recent years, the science of "networks" has become a very popular research topic and a growth area in many different disciplines. Two journals, *Social Networks* (first published in 1978) and *Network Science* (first published in 2013), have appeared that focus on network theory and applications. In its inaugural issue, the journal *Network Science* defined *network science* as the "study of the collection, management, analysis, interpretation, and presentation of relational data," and noted that the more one learns about networks, the more one sees networks everywhere. In this chapter, we describe why it is of interest to study relational data and networks, some background history, and the different types of network models that have been proposed.

1.1 What is a Network?

The focus of this book is on "relational data," which consist of relationships between pairs of entities (called *nodes* or *vertices*). These entities might refer to people, institutions, documents, webpages, genes, proteins, species of animals, transportation hubs, or even terrorist organizations. Interest usually focuses on understanding the nature of those relations (called *edges*) and formulating models that try to explain how those relations were created.

The Adjacency Matrix

How do we display relational data? If we have N nodes in the network, where the ith node is represented by the symbol v_i, $i = 1, 2, \ldots, N$,[1] the relational data are translated into a square $(N \times N)$-matrix, which we denote by

$$\mathbf{Y} = (Y_{ij}), \tag{1.1}$$

called an *adjacency matrix* (or *sociomatrix* in the social network literature),[2] in which the rows and columns each refer to the N nodes, and, in the binary case, each entry is either a zero or a one. In this matrix, if $Y_{ij} = 1$, this means that there is a relationship

[1]We use N as the size of the population network and n as the size of a sample network. In some situations, we may have access to the entire network of N nodes, but if the network is very large, we may have access only to a sample of $n \ll N$ nodes.

[2]In many books and articles on networks, the adjacency matrix is denoted by $\mathbf{A}$. Here, we prefer to use $\mathbf{Y}$ because later on we will use $\mathbf{X}$ to introduce covariates into the statistical model, and having $\mathbf{Y}$ as the elements to predict from $\mathbf{X}$ makes more pedagogical sense.

of a specified type between the pair of nodes v_i and v_j. On the other hand, if $Y_{ij} = 0$, this means that there is no such relationship between those two nodes. Usually, the diagonal entries, $\{Y_{ii}\}$, of the adjacency matrix are zero, indicating that v_i does not have a relation with itself; if $Y_{ii} = 1$, then we say that a *self-loop* is present at that node.

In certain instances, the elements of the adjacency matrix may be weights that are placed on the edges (rather than the binary 0 or 1). Such weights indicate the strength or importance of the relationships, or maybe the number of common nodes that a particular row and column have in common. For example, think of rows and columns as movies and the entries are the number of actors two different movies have in common.

Relationships between pairs of nodes can be either directed or undirected. With an *undirected network*, the adjacency matrix will be symmetric (the relation of node v_i to node v_j is the same as the relation of v_j to v_i), and so $Y_{ij} = Y_{ji}$. In the case of a *directed network*, the relation between two nodes will be directed (by an arrow $\rightarrow$) from one node to the other (i.e., $v_i \rightarrow v_j$), the adjacency matrix may not be symmetric (the relations between v_i and v_j may not be the same), and Y_{ij} may not be the same as Y_{ji}.

From the adjacency matrix, the pairwise relations are represented as a graph or network (called a *sociogram* in the social networks literature), which we denote as $\mathcal{G}$, and the connections (called *edges*) between pairs of nodes are modeled probabilistically. The set of nodes is represented as $\mathcal{V} = \{v_i\}$ and the set of edges is represented as $\mathcal{E} = \{E_{ij}\}$, where the edge joining nodes v_i and v_j is written as $E_{ij} = (v_i, v_j)$, so that the network is written as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. If there is no relation between nodes v_i and v_j, then that non-edge is not a member of the set $\mathcal{E}$. Some examples of networks are listed in **Table 1.1**.

1.2 Why Study Networks?

Different disciplines study networks for different reasons. Networks display interactions between various entities and it is of great interest to researchers to understand the nature of those interactions. The study of networks is truly a multidisciplinary topic with many research articles appearing in the journals of different fields. Here are a few of those motivations for studying networks.

A social network analyst would try to understand the edges of a network and how to interpret them. An edge may indicate that one individual "likes" another individual (i.e., a directed edge), that there is a "close friendship" between two individuals, that two individuals are members of the same organization, that an actor has appeared in the same movie as another actor, that two authors have collaborated in scientific research, that two nations trade with each other or have a strategic alliance, and so on. Social scientists are also interested in developing search algorithms for massive networks such as the World Wide Web that would identify sociological information and "cyber communities," where nodes are `html` pages and directed edges are hyperlinks between webpages.

A statistical physicist would seek to understand how networks are created, how they grow, and their topological properties. An engineer would be interested in simulating the failure of a national power grid, where the nodes represent generator substations (providing the source of electricity), transmission substations (providing connections), and distribution substations (providing power) that span the country

Table 1.1 Examples of random networks

Social networks: Examples include friendships between individuals (e.g., *Facebook*, *MySpace*, *Google Plus*, *LinkedIn*), or alliances between firms. The nodes represent people (e.g., students, company directors, film actors, or even just e-mail addresses), organizations, firms, or nations, and the edges reflect social relationships between pairs of nodes, such as e-mail messages. Social scientists refer to nodes as "actors" and edges as "ties." Social relationships do not have to be symmetric, and so the edges tend to be directed. Most social networks are sparse, but characterized by high local clustering of nodes and a low average diameter (exemplified by the "six degrees of separation" phenomenon).

Document networks: The nodes in a document network represent documents from some corpus of documents (e.g., scholarly manuscripts, webpages, text documents, medical records, or images) and the edges represent links from a given document to another document through some user interface.

Information networks: The most well-known example is the World Wide Web, with billions of webpages (nodes) and hyperlinks (directed edges) from one page to another. Hyperlinks are small bits of highlighted text or pushbuttons that, when clicked on, will take you to the address of another related webpage. There is usually no reciprocal arrangement here between webpages, and so edges are directed. A special type of information network is the citation network of academic papers.

Biological networks: There are many different kinds of biological networks. For example, in *gene regulatory networks*, the nodes represent genes, proteins, their corresponding mRNAs, and protein–protein complexes, and the edges represent individual molecular reactions, such as regulation, through which the products of one gene affect those of a target gene. Most gene regulatory networks are large and complicated. In *protein–protein interaction networks*, the nodes represent proteins and the edges represent pairs of proteins that have been shown to bind together. When proteins bind with each other, they form protein complexes that perform many of the functions in a cell. Large protein–protein interaction networks are especially of interest.

Ecological networks: Nodes represent species and edges are the number of shared interactions between different species. One type of interaction present in a forest ecosystem is a host–parasitic interaction. Another example is a *food web*, which represents the "who-eats-whom" system of hierarchical relationships between species; a directed edge between nodes represents members of one species consuming members of the other species.

Transportation networks: The nodes may be geographical locations and the edges joining pairs of nodes may represent roads, railways, or airport routes that join the locations.

and the edges represent high-voltage transmission lines. A computer scientist would be concerned about the spread of a computer virus, where the edges of a computer network represent the propagation of the virus.

A computational biologist would be interested in classifying functions of proteins and, therefore, studies protein–protein interaction networks to identify groupings of proteins. A neuroscientist would be interested in understanding how neural circuits and systems function in the human brain, how these circuits relate to one another, how they vary between individuals, and how the brain activates changes from memory to attention to movement; in this network, nodes represent individual brain regions (i.e., subnetworks of a large-scale brain network) and the directed edges represent interactions between these components. A health worker may be interested in understanding the spread of a certain disease (e.g., HIV, Ebola, Zika, SARS,

COVID-19), and the directed edges represent the transmission of the disease from one individual to another.

Lawyers would be interested in the e-mail correspondence between company executives (directed edges) following a bankruptcy filing and federal investigation of possible accounting fraud (e.g., the Enron case). A financial analyst would be interested in the links between financial institutions (e.g., banks, insurance companies, stock exchanges) that would explain periods of optimism and pessimism in financial markets. Law-enforcement personnel need to track terrorist cells, and the edges of a terrorist network represent contacts made between two members of such a cell.

1.3 A Little Bit of History

The study of network analysis originated from a number of different sources. Probably the earliest instance of a graph is the riddle called the *Seven Bridges of Königsberg* that the famous mathematician Leonard Euler studied in 1736 and published in 1841.[3] The problem involved the city of Königsberg, which was built on both sides of the Pregel River. The city included two large islands connected to each other and the mainland through seven bridges. See **Figure 1.1** The problem was to map out a path through the city that would cross each bridge only once, from end to end.

Euler reworded the problem in abstract terms, which later became the foundation of graph theory. He represented the four land masses as nodes of a graph and the seven bridges as edges. See **Figure 1.2**. He then showed that the key to the problem depended upon the degrees of the nodes.[4] Euler showed that for there to be a path

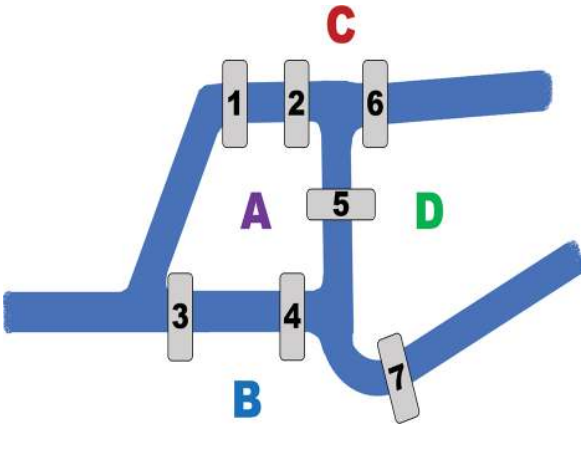

Figure 1.1 The Seven Bridges of Königsberg problem (freehand sketch). The four land masses are A: Big Island; B: Southern Bank; C: Northern Bank; and D: Small Island, and the edges are the seven bridges 1: Krämerbrückenfest (Merchant's Bridge); 2: Schmiedebrücke (Forge Bridge); 3: Grüne Brücke (Green Bridge); 4: Kottelbrücke (Connecting Bridge); 5: Honigbrücke (Honey Bridge); 6: Holzbrücke (Wooden Bridge); 7: Hohe Brücke (High Bridge).

[3]Königsberg, which was then in Prussia, was later renamed Kaliningrad, Russia.
[4]The "degree" of a node is the number of direct links (or edges) a node has to other nodes.

that satisfied the conditions of the problem, the graph would have to be connected,[5] and there would have to be zero or two nodes of odd degree. The map of the bridges showed that the latter condition did not hold (there were three nodes of degree 3 and one node of degree 5), and so no such solution exists to the problem.[6]

It was not until the 1930s that a systematic study of social networks (also called *sociometry*) was initiated by Jacob L. Moreno, a psychiatrist, and Helen Jennings, a social psychologist, who looked specifically at (1) the relations between prison inmates and (2) relations between residents in a girls' reform school (Moreno, 1932, 1934). At about the same time, W. Lloyd Warner and his business-school colleagues independently studied a social network component of the Western Electric research on industrial productivity. It was Moreno who introduced the *sociogram*, which was an early representation of a network with individuals as nodes and social relations between individuals as edges. Some development of these ideas using matrix representations of networks took place during the late 1940s and 1950s, and mathematicians introduced graph theory to social networks (Cartwright and Harary, 1958).

Interest in social networks was sporadic over the next decade, but some development of the field was carried out over the next 30 years, during which time 16 centers of social network research were set up in various countries. An additional such center was started in the early 1970s by Harrison C. White, who changed the way social networks were studied by developing mathematical models of social structure, including patterns of social relationships, and important applications to economics and sociology. During the 1970s, he and his students also introduced the concepts of "structural equivalence" and "blockmodel," which impacted the way that social network analysis was viewed by the social science community.

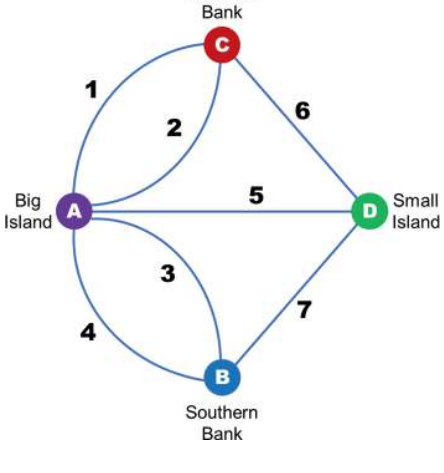

Figure 1.2 Euler's graph representation of the Seven Bridges of Königsberg problem. The nodes of the graph are the four land masses (A, B, C, D) and the edges are the seven bridges (1–7).

[5]A graph is *conected* if any two nodes can be linked by a sequence of nodes and edges.

[6]In 1875, the city of Königsberg built an eighth bridge between nodes B and C, which provided a simple solution to the problem. However, during the Second World War, most of the city was bombed and two of the bridges (Merchant's Bridge and Green Bridge) were destroyed.

During the 1950s, social network researchers started to consider the notion of "cohesive groups," which had been initiated in sociology many decades earlier, without actually defining what was meant by a "group." As social networks became a popular research area, the concept of a cohesive group became more formalized as a structural component of a network. In this scenario, social links within a group would be dense, whilst those between groups would be sparse. Since then, dozens of mathematical, probabilistic, statistical, and computational models for identifying cohesive groups have been proposed in the social networks literature.

During the late 1950s, mathematicians Paul Erdős, Alfréd Rényi, and Edgar N. Gilbert introduced probability models for the study of random graphs (Erdős and Rényi, 1959, 1960; Gilbert, 1959). The model studied by Erdős and Rényi had a fixed number of nodes and a given probability that a pair of nodes would be joined by an edge, and then they looked at what would happen when the number of edges increased. Although these probability models were much too simple, they did produce some astonishing theoretical results. One of the most surprising results was the discovery that a giant connected component would emerge when the edge probability exceeds a certain threshold.

Because of the state of mainframe computers and remote, batch-oriented computation during the 1960s, and the paucity of publicly available network data, the theoretical development of random graphs was pretty much the only game in town. The Internet was still in its infancy and could not yet play a role in social networking. E-mail was introduced in 1971 and, in 1979, *CompuServe* was the first Internet service that allowed e-mail exchanges. Compared with what we know today, theoretical models of social structure were not able to produce a reasonable representation of real-data networks.

In the 1980s, several important steps forward were made to help model social networks. First, parametric statistical models were proposed to model the structure of social relationships. The first such model was the p_1 model (Holland and Leinhardt, 1981). The p_1 model was proposed for the analysis of directed networks and was based upon the exponential family of distributions with unknown parameters to be estimated from network data called "dyads" (i.e., elements $D_{ij} = (Y_{ij}, Y_{ji})$, $i < j$, of the adjacency matrix $\mathbf{Y}$). This was followed by the ERGM (exponential random graph model) or p^* model (Frank and Strauss, 1986) and later by the p_2 model that incorporated covariates into the p_1 model (van Duijn, 1995). Second, the introduction of *stochastic blockmodels* was proposed as a tool for community detection (Holland, Laskey, and Leinhardt, 1983). The theory underlying stochastic blockmodels involved a concept of *stochastic equivalence*, which was a generalized version of "structural equivalence" (by adding a probability component) that defined the notion of similarities amongst the nodes and edges.

Starting in the 1990s, the world began to shrink when the following dramatic developments occurred: the World Wide Web was introduced in 1990; the first major web browser *Mosaic* appeared in 1993, followed by *Netscape* in 1994 and Microsoft *Internet Explorer* in 1995; high-speed computational facilities became more readily available; high-density data storage became faster and more efficient, which led to the emergence of large-scale databases; more people were connected to the Internet, which could be used to transport all types of data for download; and computer software could be used for data manipulation, analysis, and graphics of networks with millions of nodes. Social networking sites, such as *Six Degrees* (1996, closed down

in 2002), *Friendster* (2002, closed down in 2015), *LinkedIn* (2003), *MySpace* (2003), *Facebook* (2004), *Twitter* (2006), *Pinterest* (2010), and *Instagram* (2010), sprang up and changed the way people related to each other. With all these advances, more complicated network models were made possible. Researchers tried to make network models more realistic so that they could reproduce the complex structures of real networks. As a result, this desire to be more realistic led to research programs being focused on building statistical models to try to explain how networks were generated.

It was also during the late 1990s that physicists, who had not shown any previous interest in social networks, began to contribute significantly to the research work in this area. In efforts to build models for graph partitioning, they refocused attention from nodes to edges of the network. This development, however, was not taken kindly to at the time by those who had worked in the area for many decades and were irritated by these physicists, whom, they felt, had "crashed the world of social networks" (Bonacich, 2004). They claimed that physicists ignored the social networks literature, that they took the research topics of those working in social networks and claimed them as topics in physics, and that they encouraged many other physicists (viewed as "alien invaders") to get involved, who then "completely overwhelm[ed] the traditional social network analysts" (Freeman, 2011).

Fortunately, many of the models proposed by physicists, based upon ideas that had previously appeared in the social science literature, came to be accepted by social network analysts and, thus, helped to accelerate a revolution in social networks. These ideas included:

1. The model for the "small world" effect (Watts and Strogatz, 1998) and the "six-degrees-of-separation" phenomenon in which individuals are connected to each other by a succession of a few individuals who know each other.
2. A "power-law" model for the degree distributions of large networks (Barabási and Albert, 1999) described networks in which a few nodes each had a large degree and many nodes each had a small degree, so that the degree distributions are skewed and deviate significantly from the Poisson-distributed degree distributions of random graphs.

The small-world effect had a long history, reaching back to the late 1970s, and skewed node-degree distributions went back to the 1930s with the work of Moreno and Jennings (1938), and later de Solla Price (1976). Apparently, Barabási and Albert were not aware of either of those two articles. These physicists also succeeded in showing that networks were present in many different areas and disciplines (other than the social sciences) and, in effect, they actually broadened the study of social networks.

A series of articles by Mark Newman and Michelle Girvan (Girvan and Newman, 2002; Newman and Girvan, 2004) introduced physicists and computer scientists to the problem of cohesive groups, now called *communities*. This led to a focus on the algorithmic study of graph partitioning and community detection. Although the Girvan–Newman algorithm was shown to have its problems, it did alter the way researchers thought about modeling networks and that it was important to design efficient algorithms for doing so. The next step in the development process was focused on creating fast and efficient algorithms that would partition a very large network into communities. These concerns also succeeded in bringing computational biologists, who have networks with millions of nodes, and computer scientists into the world of network research.

So, major advances in network science since the late 1930s came from the fundamental contributions of social scientists, mathematicians, physicists, computer scientists, probabilists and statisticians, including Nobel Laureates who worked on network-related problems. By the mid-1980s, however, very little research on social networks had appeared in statistical journals. But then, perhaps because of the developments due to physicists and the various computational advances, probabilists and statisticians became interested in the analysis of random networks and network structure. Articles on networks, which had mostly been published in physics and social science journals, now also started to appear with some regularity in the major journals in statistics and probability.

The focus of the articles appearing in the statistical literature has been on studying statistical models for the analysis of networks, especially parameter estimation of ERGMs using maximum likelihood, pseudo-likelihood, and approaches using MCMC (Markov chain Monte Carlo) algorithms, analyses of stochastic blockmodels, and other related issues such as the quality of network data (primarily social network data) and what to do when confronted with bias in parameter estimates.

As network sizes grew larger and larger, it became important to consider the following question: If networks were now too large and complex to analyze in their entirety, how can one sample from a network? The answer is not as obvious as one might think. A large literature has grown up on the many ways of sampling from a very large network. Sampling can be carried out by nodes or by edges, and variations on these themes have been proposed. There is also the question of sampling from a "difficult-to-access" network (such as individuals with HIV/AIDS, homeless persons, illegal drug users, or undocumented aliens) if neither the members of the network are known (perhaps they want to be "hidden" from public attention) nor the size of the network is known. The social science literature has produced some remarkable advances related to sampling difficult-to-access populations.

The most recent addition to the theoretical literature on networks is the work on large networks and their limiting properties (Lovász, 2012). This topic has become very popular, and the literature, which includes studies on exchangeable random arrays, graph coloring, property testing in networks, and graphons and graphon estimation, has grown quite rapidly.

1.4 Building Network Models

During the last couple of decades, we saw the rapid development of probabilistic, statistical, and computational tools for modeling, analyzing, and graphing network data. As large amounts of network data became readily available on the Internet, new probabilistic methods and new statistical models were proposed to try to understand the structure of complex networks.

1.4.1 How are Networks Generated?

Given a set of nodes that make up a network, a major research goal is to discover how the edges that link pairs of nodes are formed. Many ideas have been put forward:

- *Random graph models*. Pairs of nodes are randomly joined according to a given probability.

- *Percolation models*. Bonds on a lattice are randomly opened or closed according to a given probability.
- *Small-world model*. Existing edges in a regular ring lattice are randomly rewired.
- *Configuration models*. Pairs of nodes are randomly joined subject to the condition that their degrees agree with a given degree distribution.
- *Expected-degree random graph models*. Edges are randomly selected so that the total number of edges is a Poisson random variable and the expected degree of a particular node is the weight attached to that node.

1.4.2 How do Networks Grow?

Attention has also been paid to the process of adding nodes and edges to an existing network with the goal of understanding how a network grows in size whilst still retaining its underlying properties:

- *Preferential attachment models*. Nodes and edges are randomly added in a sequential fashion to an existing network by linking, at each step, each new node to an existing node with probability proportional to its degree, so that the "rich get richer."
- *Random copying models*. A new node is added to an existing network, and then edges that emanate from that node are added by duplicating the edges of a randomly selected existing node.

Some of these ideas produce more realistic networks than others. When researchers were able to study real networks in detail, they found a number of common features that were not predicted by earlier models. They included power-law degree distributions, the presence of "hubs" (certain nodes that have very high degrees), and possible disjoint (or overlapping) communities in the network. So, any model worth its salt was expected to reproduce those features in its generated networks.

1.4.3 Statistical Models

One way of modeling network data is to use parametric statistical models, and several of these have been proposed for this purpose:

- The p_1 *model*, which uses the idea of independent "dyads" (pairs of entries in the adjacency matrix) for a binary network, has the form of a log-linear model whose likelihood is a member of an exponential family, where the model parameters are considered to be fixed effects, and maximum likelihood is used in a similar way as for contingency tables.
- The p_2 *model* is a logistic regression model that incorporates covariates into the p_1 model, where the network adjacency matrix is set as the dependent variable, and where the model parameters are considered to be random effects.
- The *exponential random graph model* or ERGM (also referred to as the p^* model in the social networks literature) is a member of an exponential family that incorporates variables based upon the number of edges (i.e., 1-stars), 2-stars, 3-stars, and triangles in the network, and a parameter that represents the average degree of the network.

- The *latent space model* is a conditional probability model that associates each node in the network with a point in some k-dimensional continuous latent space, where the presence of an edge between a pair of nodes is determined by the distance between the two points in the latent space (a small distance means likelihood of an edge, a large distance means likelihood of no edge).
- In the case of large networks, a theory of *graphons* has been proposed, and the problem of how to estimate a graphon is being considered and is now under serious development.

The p_1 and p_2 models have been used primarily within the social network community, although the ERGM has become generally more popular. Fitting an ERGM to network data by maximum likelihood is difficult, but workarounds have been proposed using logit models, pseudo-likelihood methods, and MCMC sampling methods.

1.5 Discovering Network Structure

Recognition of a number of common network features that have been observed in real-world networks has transformed the way in which network scientists now approach the study of network structure. Such features include power-law degree distributions, the presence of hubs, and disjoint or overlapping communities. These features led researchers to propose a variety of algorithms for discovering specific structures in networks. To accomplish these goals, they often adopted machine-learning techniques such as algorithms for graph partitioning and community detection.

1.5.1 Partitioning Algorithms

Any algorithm for partitioning a network into "communities" should be designed so that the communities each consist of groups of nodes that have the property that there are many edges between nodes in the same group, but only a few edges between nodes in different groups. The methods used include the following:

- *Graph cuts* employs a number of methods (binary cuts, normalized cuts, ratio cuts, multiway cuts) for dividing the nodes into two or more groups, where the objective function is different for each method.
- *Stochastic blockmodels* (SBMs) divide the nodes into disjoint communities or blocks where nodes are members of the same community if and only if they are stochastically equivalent, which means that nodes within the same community are exchangeable (with respect to some probability distribution).
- *Degree-corrected stochastic blockmodels* are stochastic blockmodels that include an additional parameter for dealing with heterogeneity in the degree of each node.
- *Spectral clustering* uses eigenvalues and eigenvectors of the graph Laplacians to divide the nodes of a network into two or more groups.
- *Overlapping-community models* relax the requirement of SBMs that networks be partitioned into disjoint communities, so that nodes can belong to more than one community.
- *Latent cluster models* extend the idea of latent space models, where communities in the network are formed, possibly using Bayesian model-based clustering of the points in a continuous latent space.