Modeling Human Networks Using Random Graphs

Lee Bernick

May 4, 2018

Abstract

One natural way to model human networks, such as social networks, transportation networks, or the internet, is with random graphs. This paper summarizes the foundations of random graph theory, developed by Paul Erdös and Alfred Rényi in 1958, and some common techniques used to analyze random graphs. Three more generalized random graph models are also explored: the configuration model, the small-world model, and the preferential attachment model. The similarity of these models to human networks is evaluated based on four criteria: average path length, degree distribution, clustering coefficient, and static or dynamic nature of the graph.

1 Introduction

What do MIT's online course catalog, Boston's MBTA transit system, and the set of friendships among the class of 2018 have in common? All are examples of networks created by humans. On the surface, it may seem like the process by which friendships develop would not be related to the decisions the city of Boston has made about where to place bus and subway stops. However, we will see that these networks do in fact have several mathematical properties in common.

A natural way to model networks is by using graphs where the nodes represent elements in the network and the edges represent connections between those elements; for example, links between web pages or instances of contact between people afflicted with a communicable disease. The creation of these networks are commonly modeled as random processes. This paper will begin with the earliest literature on such randomly created graphs, demonstrating some common methods of analyzing properties of large random graphs. The earliest random graph model is known as the Erdös-Rényi model, created by Paul Erdös and Alfred Rényi in 1958. The bulk of Erdös and Rényi's work focused on investigating connectivity and component structure characteristics of an infinitely large Erdös-Rényi graph.

Erdös and Rényi's work has a number of useful applications, most notably bond percolation, an important field of physics. However, it has several shortcomings when it comes to modeling networks created by human interaction, simply because the algorithm used to create Erdös-Rényi graphs is not representative of the way people interact with each other. Since Erdös and Rényi's foundational work, several mathematicians have created more generalized random graph models that are more suited to practical applications of human network modeling. Some important generalized random graph models include the small-world model, the configuration model, and the preferential attachment model, and each has different strengths for network modeling.

2 Foundations of Random Graph Theory: Erdös-Rényi Random Graphs

Erdös-Rényi random graphs are one of the simplest types of random graphs, and can be constructed in three different ways. In the model $G(n,p)$, introduced by Edgar Gilbert in 1959, a graph with n vertices is constructed by connecting each pair of vertices (i,j) independently with probability p [6]. The model $G(n,m)$ was introduced by Paul Erdös and Alfred Rényi in the same year, and is constructed by choosing uniformly at random from all graphs with n vertices and m edges [3]. One can also think of a stochastic process $G_n(m)$, where m ranges from 0 to $\binom{n}{2}$, and a random edge is added at each time step. In the limit of large n, the models are equivalent if $m = p\binom{n}{2}$. This paper will focus primarily on the model $G(n,p)$, as it is often the easiest to analyze.

2.0.1 Degree Distribution

Definition 2.1. ***degree distribution****: The probability distribution of the fraction of nodes with a given degree; or equivalently, the probability that a* *random node has a given degree.*

Because the existence of each edge in an Erdös-Rényi random graph is a Bernoulli random variable with parameter p, the degree of a given vertex is a Binomial random variable with parameters $n-1$ and p. The degree distribution $P(k)$ can therefore be written as

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

As the number of vertices becomes large, the degree distribution converges to a Poisson distribution with parameter $(n-1)p$:

$$P(k) = \frac{((n-1)p)^k e^{-(n-1)p}}{k!}$$

Erdös-Rényi graphs are therefore said to have Poisson degree distribution.

2.1 Phase Transitions

In many situations, it can be instructive to examine what happens to large random networks as the edge existence probability p varies. Erdös and Rényi introduce the term **phase transition** to describe a significant change in a given property of a random graph as p crosses a threshold. Generally, analysis of statistical properties of Erdös-Rényi graphs is performed in the limit as n approaches ∞, and the threshold value of p is calculated as a function of n.

Definition 2.2. ***threshold function****: For a given property P, $t(n)$ is a threshold function if*

- $\mathcal{P}[G(n,p)$ *exhibits property* $P] \overset{n\to\infty}{\to}$ *0 if* $\frac{p(n)}{t(n)} \overset{n\to\infty}{\to} 0$.
- $\mathcal{P}[G(n,p)$ *exhibits property* $P] \overset{n\to\infty}{\to}$ *1 if* $\lim_{n\to\infty} \frac{p(n)}{t(n)}$ *diverges.*

In the example below, we will determine the threshold function $p(n)$ above which an Erdös-Rényi graph is connected and below which it is not connected. However, connectivity is not the only property that exhibits an abrupt phase transition. A graph property P is **monotone** if, for some graph G having property P, any graph on the same set of vertices containing G as a subgraph also has property P. In 1996, Ehud Friedgut and Gil Kalai showed that any monotone graph property has a sharp threshold function [4]. Erdös Rényi also analyzed threshold functions for the emergence of a large ($\mathcal{O}(\log n)$ size) connected component, an important property in applications such as bond percolation theory and epidemic modeling.

2.1.1 Example: Connectivity

In this section, we will determine a threshold function $p^*(n)$ for connectivity by showing that for $p(n) = \omega(p^*(n))$, the probability that $G(n,p)$ is completely connected converges to 1 as n approaches ∞. Likewise, for $p(n) = o(p^*(n))$, the probability that $G(n,p)$ is disconnected converges to 0 as n approaches ∞.

Theorem 1. *Let* $p(n) = \frac{\lambda \log n}{n}$*.* $G(n,p)$ *is connected with high probability for* $\lambda > 1$ *and disconnected with high probability for* $\lambda < 1$*.*

Note that this proof is stronger than the definition for a threshold function.

Proof.

Lemma 2. $G(n,p)$ *is almost surely disconnected for* $\lambda < 1$*.*

Proof. The probability of $G(n,p)$ being disconnected can be bounded from below by the probability that at least one node is isolated.

Let $I_i = 1_{\text{node i is isolated}}$, and $X = \sum_{i=1}^{n} I_i$. First, we will calculate $\mathbb{E}[X]$ and $Var(X)$.

$$\mathbb{P}[I_i = 1] = (1-p)^{n-1} = (1 - \frac{np}{n})^{n-1}$$
$$\lim_{n\to\infty} \mathcal{P}[I_i = 1] = e^{-pn}$$

Plugging in $p(n) = \lambda \frac{\log n}{n}$, we have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}[I_i = 1] &= e^{-\lambda \log n} = n^{-\lambda} \\ \mathbb{E}[X] &= n * n^{-\lambda} \\ Var(X) &= \sum_{i=1}^{n} Var(I_i) + \sum_{i\neq j} Cov(I_i, I_j) \\ &= n * Var(I_i) + n(n-1)Cov(I_i, I_j)\end{aligned}$$

Define $r = (1-p)^{n-1}$.

$$\begin{aligned}
Var(X) &= nr(1-r) + n(n-1)(\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]) \\
\mathbb{E}[I_1 I_2] &= \mathbb{P}[I_1 = 1] * \mathbb{P}[I_2 = 1 | I_1 = 1] \\
&= (1-p)^{n-1} * (1-p)^{n-2} = \frac{r^2}{1-p} \\
Var(X) &= nr(1-r) + n(n-1)(\frac{r^2}{1-p} - r^2) = nr(1-r) + n(n-1)\frac{r^2 p}{1-p}
\end{aligned}$$

$\lim\limits_{n\to\infty} Var(X) = nr + n^2 r^2 p$

Again, plugging in $p(n) = \lambda \frac{\log n}{n}$ gives us

$$\begin{aligned}
Var(X) &= n * n^{-\lambda} + \lambda n \log n * n^{-2\lambda} \\
&= \mathcal{O}(n^{-\lambda+1} = \mathcal{O}(\mathbb{E}[X])
\end{aligned}$$

Next, we will calculate an upper bound bound on the probability that there are no isolated nodes.

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{n} (i - \mathbb{E}[X])^2 \mathbb{P}[X = i] \\
&\le (0 - \mathbb{E}[X])^2 \mathbb{P}[X = 0]
\end{aligned}$$

Since $Var(X) = \Theta(\mathbb{E}[X])$,

$$\begin{aligned}
\mathbb{E}[X] \le \mathbb{E}[X]^2 \mathbb{P}[X = 0] \\
\mathbb{P}[X = 0] \le \frac{\mathbb{E}[X]}{\mathbb{E}[X]^2} = \frac{1}{\mathbb{E}[X]}
\end{aligned}$$

For $\lambda < 1$, $\lim_{n\to\infty} \mathcal{P}[X = 0] \to 0$. Since in the limit of large n, the probability that no nodes are isolated is at most 0, the probability that $G(n,p)$ is disconnected approaches 1. □

Lemma 3. *$G(n,p)$ is almost surely connected for $\lambda > 1$.*

Proof. In order to show that $G(n,p)$ is connected, we must show that there is only one connected component by putting an upper bound on the probability that there exists a connected component which is separate from the rest of the graph.

$$\mathbb{P}[\text{nodes i through k separate from rest of graph}] = (1-p)^{k(n-k)}$$
$$\mathbb{P}[\text{any k nodes separate from rest of graph}] \binom{= n}{k(1-p)^{k(n-k)}}$$

Using the union bound formula,

$$\binom{\mathbb{P}[\text{graph is disconnected}] \le \sum_{k=1}^{n/2} n}{k(1-p)^{k(n-k)}}$$

Stirling's formula states that $k!\ (\frac{k}{e})^k$ and therefore $\binom{n}{k \le \frac{n^k}{(k/e)^k}}$.

$$\mathbb{P}[\text{graph is disconnected}] \le \sum_{k=1}^{n/2} (\frac{n * e}{k})^k (1-p)^{k(n-k)}$$

Plugging in $p(n) = \lambda \frac{\log n}{n}$ and working through some messy algebra shows that $lim_{n\to\infty}\mathbb{P}[\text{graph is disconnected}] \to 0$ for $\lambda > 1$. □

□

3 Characterization of Realistic Network Models

While Erdös and Rényi's work formed a basis for random graph theory, it is not a suitable model for analyzing many of the common networks discussed in the introduction of this paper. The Erdös-Rényi model does a poor job of representing human networks for a simple reason: the probability of links between any two nodes in a human network is far from independent. The next few sections will discuss some simple means of measuring similarity between theoretical models and observed networks, and evaluate the Erdös-Rényi model on each of these criteria. First, realistic networks often have few connections separating any given pair of nodes; second, they often contain hubs with large degree; third, their degree distribution follows a power law, and last, they grow in size.

4 Six Degrees of Separation: Modeling Networks with Short Path Lengths

Many people have heard the famous theory of ”six degrees of separation”: based on Stanley Milgram’s 1967 social science research, this theory posits that any two humans are separated by an average of six mutual acquaintance linkages [7]. While Milgram’s techniques have faced their fair share of criticism, networks ranging from the neurons of the C. elegans worm to wires in electrical power grids have similarly been shown to display short average distances between nodes [8].

This characteristic of networks can be quantified by calculating the average path length of a model.

4.1 Average Path Length

Definition 4.1. *average path length: The length of the shortest path between a pair of nodes, (i, j), averaged over all pairs of nodes.*

A realistic random graph model should have a short average path length, i.e. an average path length which is asymptotically smaller than the number of vertices. The Erdös-Rényi model has a short average path length, given by $l = \mathcal{O}(\log n)$ [5].

5 Constructing Hubs in Random Graph Models

Another property commonly observed in human networks is the degree to which people tend to cluster together. For example, mutual acquaintances may be based on attendance at the same school, church, or workplace, and connections in transportation or utility networks are often based on physical proximity. The quantitative measure of the tendency of nodes in a network to cluster together is known as the **clustering coefficient**.

5.1 Clustering Coefficient

Definition 5.1. *local clustering coefficient: The local clustering coefficient of a vertex is the number of edges in its neighborhood as a fraction of the max-*

imum possible number of edges. In an undirected graph, this is given by the formula $\frac{2*(\text{\# of edges})}{n(n-1)}$, where n is the number of vertices in the neighborhood.

There are multiple ways of measuring a clustering coefficient of an entire graph, but the simplest way is to simply average the local clustering coefficients of every vertex in the graph. This metric was introduced by Ducan Watts and Stephen Strogatz, the creators of the Small-World random graph model.

5.2 Small-World Model

The Small-World Model, also known as the Watts-Strogatz model after its creators, was developed in 1998 to address the limited amount of clustering present in Erdös-Rényi graphs [8]. (The model was refined in 1999 by Watts and Newman.) It is essentially a cross between ring lattices, which have a high degree of local connectivity, with Erdös-Rényi random graphs, with the desirable property of short average path lengths. A ring lattice in the context of the Watts-Strogatz model is a ring-shaped graph with n vertices, where each vertex is connected to the $\frac{K}{2}$ closest vertices to its left and the $\frac{K}{2}$ closest vertices to its right.

The Watts-Strogatz model starts with a ring lattice and a parameter $p \in [0, 1]$, representing the degree of similarity to a ring lattice or Erdös-Rényi graph. To create a Watts-Strogatz random graph, each of the $\frac{K}{2}$ edges to the left of a given node i are "re-wired" with probability p. An edge leaving i is re-wired by selecting another node j uniformly at random among the list of vertices that would avoid self-loops or duplicate edges, constructing edge (i, j), and deleting the original edge. Clearly, $p = 0$ corresponds to a ring-lattice (no re-wiring), and $p = 1$ corresponds to an Erdös-Rényi graph, as every edge is selected randomly.

The properties of this model can be analyzed by observing the behavior of the graph as p varies from 0 to 1, as shown in Figure 5.2. For example, the ring lattice has a high degree of local clustering, with a clustering coefficient of $\frac{3(K-2)}{4(K-1)}$, approaching the constant 3/4 for large K; while the Erdös-Rényi clustering coefficient $\frac{K}{N} \ll 1$ is small. Analysis of the average path length is similar; it is $\frac{N}{2K} \gg 1$ for a ring lattice and $\frac{\ln N}{\ln K}$ for an Erdös-Rényi graph. For values of p between 0 and 1, the average path length of the Watts-Strogatz graph is close to the short average path length of the Erdös-Rényi model, and the clustering coefficient is close to the large value for the ring lattice [8].

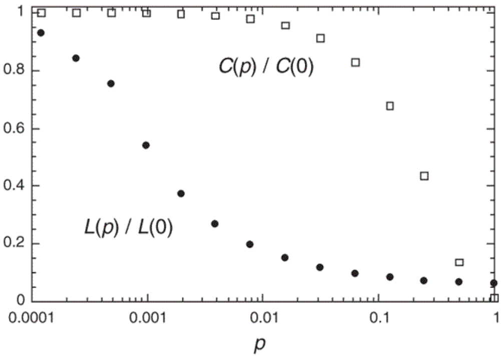

Figure 1: Path length (L) and clustering coefficient (C) of the small-world model as a function of the parameter p [8].

From these characteristics alone, this model would appear to be an excellent representation of human networks. However, its major shortcoming is its unrealistic degree distribution. The degree distribution of a ring lattice is simply the Direc delta function centered around K, while the Erdös-Rényi model has a Poisson degree distribution, as shown in section 2.0.1. The degree distribution of the Watts-Strogatz model is therefore too homogeneous (from the Dirac δ) and decays too quickly for large k to adequately represent scale-free networks.

6 Modeling Scale-Free Networks with Power-Law Degree Distributions

Another important graph property is the degree distribution, defined in section 2.0.1. Note that degree distribution is related to the clustering coefficient, but imparts more information about the structure of the graph. Many real networks have a power-law degree distribution, meaning the probability of a vertex having degree k is given by $P(k) \propto k^{-\gamma}$, where γ is some positive constant typically between 2 and 3. Such networks are called **scale-free**. Networks thought to be scale-free include social networks (such as the graph representing co-authorship of math papers), airline networks, and inter-bank payment networks.

6.1 Configuration Model

The Configuration random graph model, developed by Bender and Canfield in 1978, is a method of generating a random graph based on a desired degree distribution [2]. This construction can easily be used to model a scale-free network by simply generating a graph based on a power law distribution. The algorithm works as follows: First, choose a desired number of vertices, n, and a desired degree distribution, $P(k)$. Take a random (or pseudo-random) sample from $P(k)$ for each node i, and let this number d_i be the degree of node i. A list of nodes is generated in which each node i is listed d_i times, and edges are generated by randomly selecting pairs of nodes (i, j) from this sequence and constructing an edge between the nodes. The nodes selected are then removed from the list, and this procedure repeats until the list is empty. Note that the sum of degrees must be even, and that duplicate edges and self-loops are permissible; however, as n becomes large, these details have little impact on the overall structure of the graph.

The Configuration Model is a useful tool for generating graphs with a desired degree distribution, but it fails to adequately capture the high degree of clustering in human networks. We will next turn to a model that is more suited to this characteristic.

7 Dynamic Networks

All the random graph models discussed so far have been limited by their static nature: the number of vertices cannot change with time. However, there are many situations where it would be desirable to allow the number of nodes to change, for example, modeling a new user joining Facebook or an airline choosing to service a new city. A good network model should be **dynamic**, meaning the model allows the number of nodes to change over time.

7.1 Preferential Attachment Model

One of the most realistic network models is known as the preferential attachment model. This process was studied as early as 1925 but was popularized in 1999 by Barabási and Albert, in their studies of the model's applications to the internet [1]. The phrase ”preferential attachment” can be used to describe a number of processes in which wealth is distributed among individuals according to how much they already have; in this model, edges are constructed in proportion to nodes' existing degrees. The model's most important feature is the capability of network growth, as it describes a process for adding nodes over time. It also has a power law degree distribution with short average path length and high clustering, demonstrating its usefulness in modeling human networks [1, 5].

The Barabási-Albert algorithm for constructing a random graph with n vertices and expected degree m begins with a clique of size $m-1$, and adds one node at every time step. A newly born node will randomly attach to m other nodes, with probability proportional to the other node's current degree. That is, a node i born at time t_i will attach to node j with probability

$$m\frac{d_j}{\sum_{k=1}^{i-1} d_k}$$

.

Theorem 4. *Barabási-Albert graphs have power law degree distributions.*

Proof. We will first find an expression for the expected degree of a given node i at time t. When a new node j is added to the network, the probability that it attaches to node i is given by $m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)}$, since attachment probability is proportional to the current degree of the node in question. At time t, $m*t$ edges have been created, so this probability is equal to $\frac{m*d_i(t)}{2tm} = \frac{d_i(t)}{2t}$. We can then write the differential equation for the expected degree of node i at time t:

$$\frac{d}{dt}d_i(t) = \frac{d_i(t)}{2t}$$

with initial condition $d_i(t_i) = m$. Separating and integrating, we have

$$\int \frac{1}{d_i(t)}d(d_i(t)) = \int \frac{dt}{2t}$$

$$d_i(t) = m(\frac{t}{t_i})^{1/2}$$

Now, we can find the cumulative distribution function for the degree.

$$\begin{aligned}\mathbb{P}[d_i \leq k] &= \mathbb{P}[m(\frac{t}{t_i})^{1/2} \leq k] \\ &= \mathbb{P}[t_i \geq (\frac{m}{k})^2 t]\end{aligned}$$

After a large number of time steps, this equation reduces to $F = 1(-\frac{m}{k})^2$. Differentiating to find the probability mass function, the degree distribution $P(k) = \frac{2m}{k^3}$, proving that the preferential attachment model has a power law degree distribution with $\gamma = 3$. □

8 Conclusions

Random graph theory is a useful tool for modeling many real-life phenomena, including human-constructed networks such as transportation systems and the internet. The simplest model, Erdös-Rényi random graphs, provides a useful framework for analyzing phase transitions of random graphs as the likelihood of two nodes being connected varies. More generalized random graph models have been developed to address the Erdös-Rényi model's shortcomings in modeling human networks, most notably the false assumption that all connections between pairs of vertices are independent. These generalized models include Bender and Canfield's configuration model, Watts and Strogatz's small-world model, and Barabási and Albert's preferential attachment model. Of the three, the preferential attachment model is the best fit for human networks due to its power law degree distribution, short average path length, high clustering, and dynamic nature.

References

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] Edward A Bender and E.Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296 – 307, 1978.

[3] Paul Erdös and Alfred Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1958.

[4] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society*, 124, 1996.

[5] Agata Fronczak, Piotr Fronczak, and Janusz Holyst. Average path length in random networks. *eprint arXiv:cond-mat/0212230*, 2002.

[6] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 12 1959.

[7] Stanley Milgram. The small-world experiment. *Psychology Today*, 1:61–67, 1967.

[8] Duncan Watts and Steven Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.