4-2016

Growing Networks with Positive and Negative Links

Corynne Smith Dech
College of William and Mary

Recommended Citation
Dech, Corynne Smith, "Growing Networks with Positive and Negative Links" (2016). *Undergraduate Honors Theses.* Paper 922.
https://scholarworks.wm.edu/honorstheses/922

Growing Networks with Positive and Negative Links

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Mathematics from The College of William and Mary

by

Corynne Smith Dech

Accepted for Honors

Leah B. Shaw
Leah B. Shaw, advisor

Ross J. Iaci

Rex K. Kin
Rex K. Kincaid

Anke van Zuylen

Williamsburg, VA
April 20, 2016

Growing Networks with Positive and Negative Links

Corynne S. Dech

Department of Mathematics,
College of William and Mary,
Williamsburg, VA 23187-8795, USA

Email: csdech@email.wm.edu

Acknowledgements

I would like to express deep gratitude to my advisor Dr. Leah B. Shaw for her endless support, guidance, and commitment to this project. I also wish to thank the members of my committee, Dr. Ross J. Iaci, Dr. Rex K. Kincaid, and Dr. Anke van Zuylen. I am very thankful for the opportunity to collaborate with Shadrack A. Antwi, without whom much of this project would not be possible. Finally I wish to thank the Monroe Scholar Program for supporting my research.

Abstract

Scale-free networks grown via preferential attachment have been used to model real-world networks such as the Internet, citation networks, and social networks. Here we investigate signed scale-free networks where a link represents a positive or negative connection. We present analytic results and simulations for a growing signed network model and compare the signed network to an unsigned scale-free network. We discuss several options for preferential attachment in a signed network model. Lastly we measure preferential attachment in a real-world network and discuss the advantages and disadvantages of data fitting methods.

Contents

1 Introduction **1**

1.1 Networks . . . 1

1.2 Scale-Free Networks . . . 2

1.2.1 Barabási-Albert Model . . . 4

1.2.2 Barabási-Albert Model Degree Distribution . . . 7

1.3 Outline . . . 10

2 Signed Network Model **12**

2.1 Notation and Definitions . . . 12

2.1.1 Simulation Methods . . . 15

2.2 Separate Preferential Attachment . . . 16

2.2.1 Analysis of Separate Preferential Attachment . . . 17

2.2.2 Simulation Results . . . 20

2.2.3 Cross Correlation . . . 21

2.3 Other Attachment Methods . . . 22

2.3.1 Ratio Preferential Attachment . . . 23

2.3.2 Weighted Product Preferential Attachment . . . 24

3 Data Fitting **27**

3.1 Background and Assumptions . . . 27

3.1.1 Nonlinear Preferential Attachment . . . 28

3.2 Least Squares Fitting . . . 29

3.2.1 Logarithmic Least Squares Fitting . . . 30

3.2.2 Basic Least Squares Fitting . . . 32

3.2.3 Weighting of Data Points . . . 34

3.3 PAFit Maximum Likelihood Estimation . . . 36

3.3.1 Signed Maximum Likelihood Estimation . . . 40

4 Conclusions **43**

4.1 Future Work . . . 44

Chapter 1

Introduction

1.1 Networks

A *network* is a graph representation of a system where members are connected to other members. We define a network by a set of nodes (or vertices) and a set of links (or edges) connecting those nodes. Networks are used in a diverse range of disciplines to model complex systems such as metabolic networks, epidemiological networks, and the World Wide Web [2, 10, 17]. *Growing networks* are complex networks that evolve as a function of time, where new nodes and links are added or removed throughout the lifespan of the network [5]. Since many real-world networks, such as social networks and collaboration networks, expand and change over time it is important to study the processes by which networks grow so that we may understand and predict the behavior of these networks.

While directed networks specify the direction of a link between two nodes, we focus on undirected networks. We further simplify by only including networks without self-connections or multiple edges between the same two nodes. The *degree* of a node is the number of links incident to that node. The *degree distribution* of a network is the probability distribution of degrees that occur in the network. The degree distribution function $P(k)$ is the likelihood that a randomly selected node will have a degree of k.

Consider the simple and commonly studied random network proposed by Erdős and Rényi [21]. The Erdős-Rényi (ER) random network model has a fixed set of n nodes and randomly places a given number of links between those nodes such that every possible link has the same probability of being created. The nodes in this network will all have a similar degree, approximately the average degree $\langle k \rangle$ of the network. We observe that the degree distribution for an ER random network is a Poisson distribution where $\langle k \rangle$ has the greatest probability of occurring in the network [1]. Networks that appear in the real world, however, are built according to principles more complex than ER random networks.

1.2 Scale-Free Networks

The discovery of degree distributions that follow a power law in several real-world networks has led to growing interest in scale-free networks [2, 20]. A *scale-free network* is a network with a degree distribution that follows a power law. We use the "$\sim$" symbol to denote values that are proportional to each other in the limit of large k. Thus the degree distribution of a scale-free network is

$$P(k) \sim k^{-\gamma} \tag{1.1}$$

where $\gamma > 0$ is the *degree exponent.* Since

$$\log P(k) \sim -\gamma \log k \tag{1.2}$$

we can expect a logarithmic plot of $P(k)$ versus k to be linear with slope $-\gamma$. Most scale-free networks have a degree exponent in the range $2 < \gamma < 3$ [1, 20]. Scale-free networks observed in the real world include citation networks, social networks, movie actor collaboration networks, the Internet, and the World Wide Web [5].

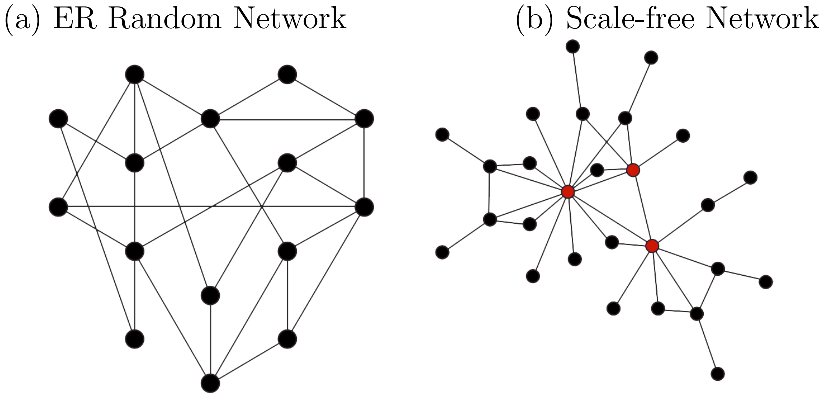


Figure 1.1: (a) An ER random network does not contain hubs and most nodes have degree close to the mean. (b) A scale-free network has many low-degree nodes and a few hubs (red nodes). Figure from [6].

The World Wide Web was first suggested to be a scale-free network in 1999 by Albert, Jeong, and Barabasi, who constructed a directed network of over 300,000 web pages where a link represents a hyperlink pointing from one web page to another [2]. Because these links are directed, each web page has an in-degree k_{in} of hyperlinks leading to that web page and an out-degree k_{out} of hyperlinks from that web page to another web page. The degree distributions were found to follow a power law such that $P(k_{in}) \sim k_{in}^{-2.1}$ and $P(k_{out}) \sim k_{out}^{-2.45}$. This significant deviation from the Poisson distribution of a random network suggests a need to define the mechanism that creates scale-free networks.

A visual inspection of a scale-free network reveals the appearance of highly-connected nodes, called *hubs*, that are not present in random networks (see Figure 1.1). A power law degree distribution falls off much more slowly than a Poisson distribution which is approximately Gaussian. Therefore a scale-free network has many low-degree nodes and a few very high-degree nodes, compared to the degree distribution of an ER random network where most nodes have degree close to the mean. This explains the absence of hubs in a random network.

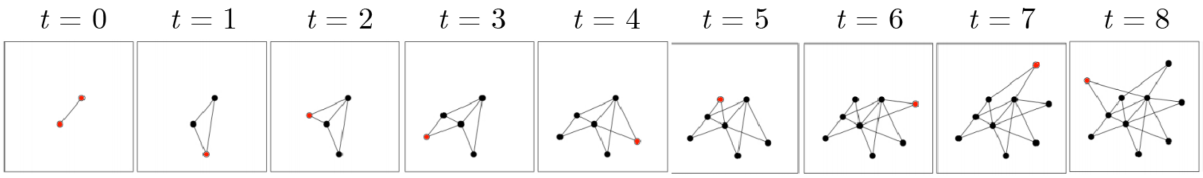


Figure 1.2: Growth of BA model from $t = 0$ to $t = 8$ where 1 node and 2 links are added at each time step. The red node is the newest node at each time step. Figure from [4].

1.2.1 Barabási-Albert Model

The Barabási-Albert (BA) model is an algorithm for constructing a scale-free network utilizing two mechanisms: *growth* and *preferential attachment* [2].

While the ER random network model has a fixed number of nodes, most real-world networks grow over time as new nodes are added. For example, citation networks grow with each new publication, movie actor networks expand with each new movie, and thousands of new web pages are added to the World Wide Web every day. The BA model incorporates *growth* by adding one node to the network at every time step (Figure 1.2).

In the ER random network model every node is equally likely to obtain a new link. *Preferential attachment* is a mechanism where new nodes prefer to connect to existing high-degree nodes. For example, a new web page is more likely to include hyperlinks to other web pages that are already well-known. Similarly, a newly published paper is most likely to cite existing well-known (i.e., highly cited) papers. In the BA model the probability that a node will receive a new link is proportional to that node's degree. We define k_i to be the degree of node i and $\Pi(i)$, called the *attachment kernel*, to be the probability that a new link will attach to node i such that

$$\Pi(i) \propto k_i. \tag{1.3}$$

To build a network using the BA model we begin with a small network of m_0 nodes. This base network is complete, meaning that each node is connected to every other node

Term	Meaning
t	current time step
$N(t)$	number of nodes at time t
$M(t)$	number of links at time t
m	number of links connected to new node
m_0	starting size of network
k_i	degree of node i
t_i	time that node i is added
$\Pi(i)$	probability a new link will attach to node i

Table 1.1: BA model definitions

in the network. At every time step a new node will be added to the network. Then m links will be added to the network connecting the new node to m existing nodes. At $t = 1$, the new node must have at least m existing nodes with which to connect; thus, we set our base network size to $m_0 = 2m$. See Table 1.1 for a full listing of the variables in the BA model.

We define $N(t)$ to be the number of nodes in the network at time t and $M(t)$ to be the number of links in the network at time t. Therefore,

$$N(t) = m_0 + t \tag{1.4}$$

$$M(t) = \frac{m_0(m_0 - 1)}{2} + mt. \tag{1.5}$$

For large t we can ignore the m_0 term, resulting in the following approximations:

$$N(t) \approx t \tag{1.6}$$

$$M(t) \approx mt. \tag{1.7}$$

When a new node is added to the network, m existing nodes are chosen to connect to the new node using preferential attachment. The probability that a given node i is chosen as the target for a given link is determined by its attachment kernel, $\Pi(i)$:

$$\Pi(i) = \frac{k_i}{\sum\limits_{j=1}^{N(t)} k_j}. \tag{1.8}$$

The attachment kernel is normalized by dividing by the sum of the degrees of all the nodes in the network so that the sum of the attachment kernels of all nodes is 1. For simplicity, we will drop the limits of summation in the normalization term; a summation of k_j in this thesis is over all nodes in the network.

We observe an important feature of the BA model by studying the relationship between a node's age t_i and its degree k_i. By assuming continuous degrees across the network, we can obtain an approximate time evolution of an individual node's degree. At time t, a given node i has m chances, each with a probability of $\Pi(i)$, to be chosen to connect to the new node. Thus its change in degree with respect to time is as follows [1]:

$$\frac{\partial k_i}{\partial t} = m\Pi(i) = m\frac{k_i}{\sum\limits_j k_j}. \tag{1.9}$$

Because each new link increases the total degree sum by 2, the sum of all degrees is equal to twice the total number of links. Thus $\sum\limits_j k_j = 2M(t)$. By Equation 1.7 for large t we can use the substitution $\sum\limits_j k_j = 2mt$ resulting in

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \tag{1.10}$$

The degree of node i when it first enters the network at t_i is m, resulting in the initial condition $k_i(t_i) = m$. By integrating Equation 1.10 using this initial condition we obtain

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}. \tag{1.11}$$

This definition exposes an important feature of the BA model: older nodes become hubs because they have more chances to gain links. The factor $\frac{1}{t_i}$ implies that the earlier a node is added the higher its degree. This observation explains why the network contains a range of degrees further from the mean degree than in a random network.

1.2.2 Barabási-Albert Model Degree Distribution

We solve for the exact degree distribution of a network grown according to the BA model to verify that the network is scale-free and to identify the degree exponent. We find the steady state of the degree distribution $P(k)$ by using a rate equation approach developed by Krapivsky, Redner, and Leyvraz [12]. We wish to find $N(k,t)$, the number of nodes with degree k at time t. We define $P(k,t) = \frac{N(k,t)}{N(t)}$ to be the proportion of nodes with degree k at time t. After one time step, $N(k,t)$ is decreased by the number of degree k nodes that gain a link (becoming degree $k+1$ nodes) and is increased by the number of degree $k-1$ nodes that gain a link (becoming degree k nodes).

The number of degree k nodes that gain a link is the product of the number of links being added, the number of degree k nodes, and the probability that a degree k node will be linked to by incoming link (i.e., its kernel). Thus the number of degree k nodes that gain a link is:

$$mN(k,t)\frac{k}{\sum\limits_j k_j} = mN(k,t)\frac{k}{2M(t)}. \tag{1.12}$$

By Equation 1.7 for large t we can use the substitution $\sum\limits_j k_j = 2mt$ resulting in

$$N(k,t)\frac{k}{2t} = N(t)P(k,t)\frac{k}{2t}. \tag{1.13}$$

It follows that the number of degree $k-1$ nodes that gain a link is $N(t)P(k-1,t)\frac{k-1}{2t}$.

Then together

$$N(t+1)P(k,t+1) = N(t)P(k,t) - N(t)P(k,t)\frac{k}{2t} + N(t)P(k-1,t)\frac{k-1}{2t}. \tag{1.14}$$

By Equation 1.6 for large t, $N(t) \approx t$ and $N(t+1) \approx t+1$ so this becomes

$$(t+1)P(k,t+1) = tP(k,t) - P(k,t)\frac{k}{2} + P(k-1,t)\frac{k-1}{2}. \tag{1.15}$$

This equation holds for $k > m$.

Because every node has a degree of at least m, $P(k,t) = 0$ for $k < m$. One new node of degree m is added to the network at each time step. Thus we can define the following boundary condition:

$$(t+1)P(m,t+1) = tP(m,t) - \frac{m}{2}P(m,t) + 1. \tag{1.16}$$

We solve for the steady state $P(k,\infty) = P(k)$. Using Equation 1.15 and 1.16 gives us

$$P(k) = \begin{cases} \frac{k-1}{k+2}P(k-1) & : k > m \\ \frac{2}{m+2} & : k = m. \end{cases} \tag{1.17}$$

We solve the recurrence relation by iteration:

$$P(m) = \frac{2}{m+2} \tag{1.18}$$

$$P(m+1) = \frac{m}{m+3}P(m) = \frac{2m}{(m+2)(m+3)} \tag{1.19}$$

$$P(m+2) = \frac{m+1}{m+4}P(m+1) = \frac{2m(m+1)}{(m+2)(m+3)(m+4)} \tag{1.20}$$

$$P(m+3) = \frac{m+2}{m+5}P(m+2) = \frac{2m(m+1)}{(m+3)(m+4)(m+5)}. \tag{1.21}$$

We observe that the iterations of $P(k)$ follow a simple recursive pattern, producing the

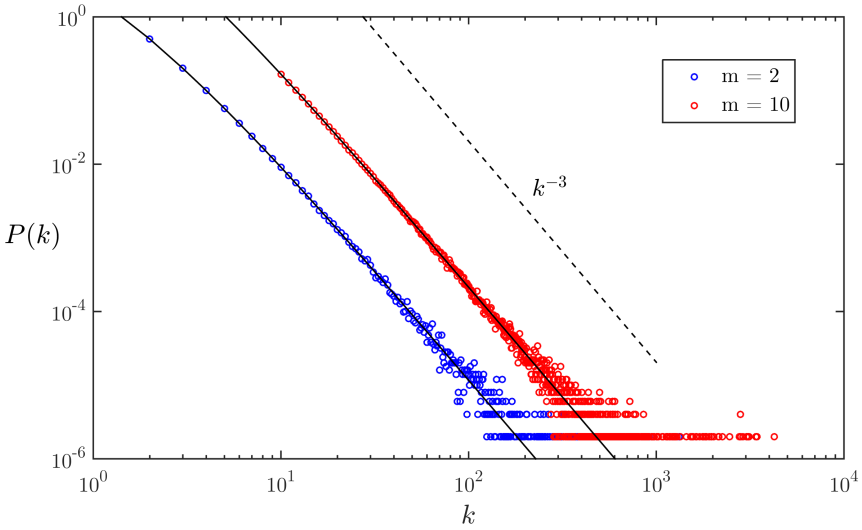


Figure 1.3: Degree distribution of a network simulation using the BA model at $t = 500,000$ for $m = 2$ and $m = 10$. Black curves are theoretical prediction from Equation 1.22. Dashed curve shows asymptotic slope of k^{-3}.

exact solution for the degree distribution of the BA model [12]:

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{1.22}$$

For large k, $P(k) \sim k^{-3}$, fulfilling the requirement in Equation 1.1 and confirming that the BA model produces a scale-free network.

We simulate the growth of a network using the BA model to verify the prediction for the degree distribution. The methods for this simulation are similar to what will be described in Section 2.1.1. Our simulation of the BA model matches the analytic prediction for the degree distribution given by Equation 1.22 (Figure 1.3). It is expected that the degree distribution plot exhibits substantial noise in the tail. This occurs because very high degrees with low probabilities would have an expected occurrence number between 0 and 1. Since the simulation obviously only allows for an integer number of occurrences, the observed probabilities for these high degrees will be noisy.

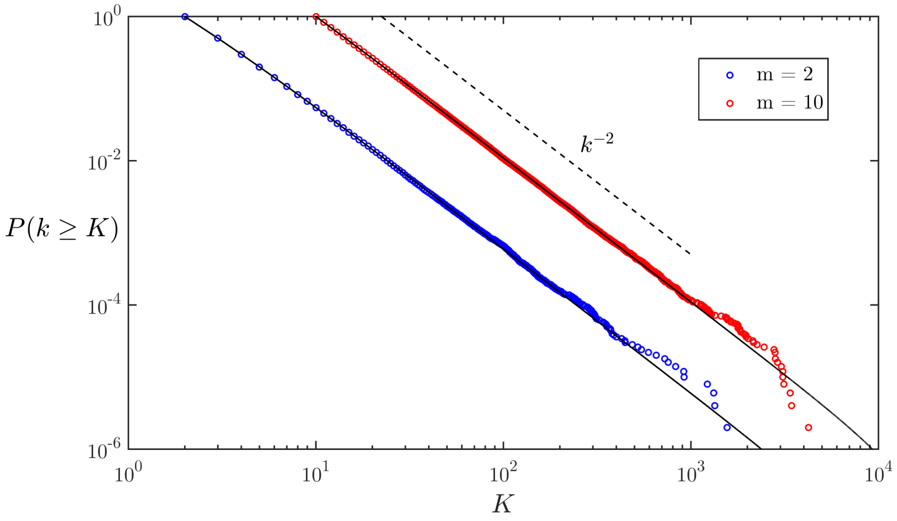


Figure 1.4: Reverse cumulative degree distribution of a network simulation using the BA model at $t = 500,000$ for $m = 2$ and $m = 10$. Black curves are theoretical prediction from Equation 1.22. Dashed curve shows asymptotic slope of k^{-2}.

One method for smoothing the degree distribution is to plot the reverse cumulative degree distribution, $P(k \geq K) = \sum_{k=K}^{\infty} P(k)$. By summing the probabilities of all degrees greater than or equal to the current degree we achieve a more accurate representation of the high degree probabilities. The reverse cumulative degree distribution is the integral of a degree distribution so we expect a network with a degree distribution of $P(k) \sim k^{-3}$ (as in Figure 1.3) to have a reverse cumulative degree distribution of $P(k \geq K) \sim k^{-2}$ (as in Figure 1.4).

1.3 Outline

This chapter presented an overview of growing scale-free networks. It is important to study the mechanisms by which scale-free networks grow so that we may predict how networks will behave in the future. Given the attachment kernel for a network, it is possible to predict which nodes are likely to obtain new links.

Chapter 2 discusses a special class of networks, called signed networks, where links have a label of positive or negative. The growth of signed networks is a new field and an accepted mechanism for growing signed networks is yet to be determined. In Section 2.2 we propose a basic method for signed network growth called separate preferential attachment and study it analytically. In Section 2.3 we propose several more complicated signed growth methods and discuss their motivation and consequences.

Chapter 3 describes existing techniques for fitting an attachment kernel to data from real-world networks. We test these methods on a data set from a signed online social network and discuss the advantages and disadvantages of each fitting method. We propose adaptations to these fitting methods that incorporate the sign of a link.

Chapter 4 summarizes our findings about growing signed networks and suggests further work in the study of growing networks with positive and negative links.

Chapter 2

Signed Network Model

2.1 Notation and Definitions

A *signed network* is a special class of networks in which each edge is labeled positive or negative. Most research on online networks focuses on the positive connections of friendship, trust, or collaboration; however, it is important to consider the negative connections of distrust, animosity, and controversy. Signed networks have been used to measure status and centrality in a social network of monks [7], to identify top users and unpopular "troll" users in the online social network Slashdot [14], and to observe structural balance in online networks [15]. Signed networks could also represent positive and negative reviews on product review sites like Amazon and eBay or upvotes and downvotes on voting sites such as Reddit or YouTube. A simple example of a signed network can be seen in Figure 2.1 which shows the alliances of countries leading up to World War I [8]. Here a green link (a positive link) means that the two countries are allies and a red link (a negative link) means the two countries are enemies.

Most existing research on signed networks has focused on structural balance theory [9,14,15]. We observe structural balance by looking at the signs of links in a cycle between three nodes. If all three links are positive, this cycle supports the principle that "the friend

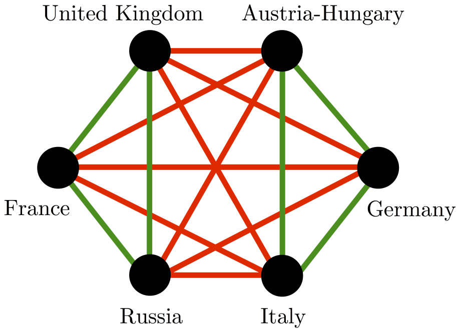


Figure 2.1: World War I alliance network [8]

of my friend is my friend." If one link is positive and two are negative, that cycle suggests that "the enemy of my enemy is my friend." According to structural balance theory these two types of cycles occur more often than cycles with all negative links or with only one negative link.

Our project, to the best of our knowledge, is the first attempt to model the growth of signed networks. We propose attachment kernels for adding positive and negative links and run simulations for the growth of a signed network where positive and negative links are added to the network at each time step.

With two kinds of links, we must introduce new terminology. The *positive degree* of a node is the number of positive links a given node participates in. The *negative degree* of a node is the number of negative links a given node participates in. The *total degree* of a node is the sum of its positive and negative degrees. The *positive degree distribution*, $P(k_+)$, of a network is the probability distribution of positive degrees that occur in the network, and likewise for the *negative degree distribution*, $P(k_-)$. The *joint probability distribution*, $P(k_+, k_-)$, gives the probability that a randomly selected node will have a positive degree of k_+ and a negative degree of k_-.

Because the types of signed networks we wish to model, specifically social networks, typically have a power law degree distribution we adapt the BA model for growing unsigned scale-free networks in order to grow a signed scale-free network. See Table 1.1 for

Term	**Meaning**
m_+	number of positive links added to each new node
m_-	number of negative links added to each new node
m_0	starting size of network
k_{+i}	positive degree of node i
k_{-i}	negative degree of node i
$\Pi_+(i)$	probability a new positive link will attach to node i
$\Pi_-(i)$	probability a new negative link will attach to node i
p	fraction of positive links in the network

Table 2.1: Growing Signed Network Model Definitions

the definitions of variables in the BA model, Table 2.1 for the definitions of new variables in the signed model, and Section 2.1.1 for a description of the initial network. At each time step one node is added to the network and m_+ positive links and m_- negative links are made between the new node and existing nodes in the network. There can only be one type of link between any two nodes. The *positivity*, p, of a signed network is the fraction of links in the network that are positive such that $p = \frac{m_+}{m_+ + m_-}$. In many signed social networks, $p = 0.8$, meaning 80% of the links in the network are positive [14,15]. However, to simplify our analytical study of growth, we set $p = 0.5$. Thus $m = m_+ = m_-$, such that m positive links and m negative links are added at each time step.

The *positive attachment kernel*, $\Pi_+(i)$, is the probability that a new positive link will attach to node i. The *negative attachment kernel*, $\Pi_-(i)$, is the probability that a new negative link will attach to node i. We assume that some variation of preferential attachment is occurring so we suggest that the positive and negative attachment kernels are related to the degree of the node. However, it is unclear how to weight the positive and negative degree of a node in an attachment kernel given the difference in semantic value of the two types of links. We propose several variations of signed preferential attachment and discuss their implications.

2.1.1 Simulation Methods

Here we present an overview of our simulation of growing signed networks. The user-inputted parameters for the program are the desired number of time steps, the value of m, and the attachment method. Each node has the following attributes: its positive and negative degree, its individual positive and negative attachment kernels (determined by the inputted attachment method of the network), its age, and a list of its neighbors. The network begins as a complete network of size $m_0 = 2m$, meaning that every possible link between the m_0 nodes exists. As links are added their signs alternate between positive and negative such that half of these links are positive and half are negative.

At each time step one new node is added to the network. Then m positive links and m negative links are added between the new node and existing nodes. The signs of these links alternate such that one positive link is added, then one negative link, then one positive link, and so on. When adding a link we first generate a random number between 0 and 1[1]. The positive and negative attachment kernels are, by definition, normalized so that the sum of the positive attachment kernels of all the nodes in the network at any given time is 1, and likewise for the negative attachment kernels. Consider each node as a bin on the number line from 0 and 1 where the width of each bin is equal to the node's positive or negative attachment kernel. Thus the random number corresponds to a particular bin which corresponds to a particular node, where the probability that the random number will land in that bin is equal to that node's positive or negative attachment kernel.

To add a positive link we iterate through the list of existing nodes, adding a node's positive attachment kernel to a running sum, then stop when the running sum reaches or exceeds our random number. The node at which we stop is selected to participate in the new positive link. Therefore a node's positive attachment kernel determines the probability that the node gains a positive link. To add a negative link, the same process is used, except using a node's negative attachment kernel. Before a new link is added we

[1] All random numbers were generated using the Ran.java class from Numerical Recipes [23].

verify that a link (positive or negative) does not already exist between those two nodes. If a link does exist between those nodes we return to the first step of the algorithm by generating a new random number and selecting a new target node.

This process of node and link addition terminates when the desired number of time steps is reached. The program offers options for outputting positive, negative, and joint degree distributions as well as other statistics about the resulting network.

2.2 Separate Preferential Attachment

The first method of signed attachment we propose is *separate preferential attachment.* The motivation for this method was to produce a simple base case for signed network growth that is almost analogous to unsigned network growth. The preferential attachment is "separate" in the sense that the positive attachment kernel depends only on the positive degree and the negative attachment kernel depends only on the negative degree. Thus an incoming positive link ignores a node's negative degree and prefers to connect to nodes with a high positive degree. Likewise, an incoming negative link ignores a node's positive degree and prefers to connect to nodes with a high negative degree. We define the positive and negative attachment kernels as follows:

$$\Pi_+(i) = \frac{k_{+i} + 1}{\sum\limits_j (k_{+j} + 1)} \tag{2.1}$$

$$\Pi_-(i) = \frac{k_{-i} + 1}{\sum\limits_j (k_{-j} + 1)}. \tag{2.2}$$

The +1 is added to the degree so that a node with a degree of zero still has a small chance of gaining a link. It is especially important to consider nodes with zero degree in a signed network because even if a node has a non-zero total degree, it may still have positive degree or negative degree of zero. In our simulation, every node has a positive and negative degree of at least m, so no node will ever have a degree of zero. However, we introduce the $+1$ in anticipation of fitting this model to a real-world data set which might contain nodes with degree zero.

Since the positive and negative attachment kernels act separately, essentially two unsigned networks are placed on top of each other. Thus we expect the positive and negative degree distribution for a network grown with separate preferential attachment to look the same as an unsigned network grown with preferential attachment. The joint probability distribution $P(k_+, k_-)$ is unknown and to be determined analytically in the next section.

2.2.1 Analysis of Separate Preferential Attachment

Because the attachment kernels for separate preferential attachment are so simple we are able to construct a theoretical prediction for the positive and negative degree distributions and verify that they are consistent with the degree distribution of an unsigned network. The joint probability distribution, $P(k_+, k_-, t)$, is the probability that a randomly selected node has positive degree k_+ and negative degree k_- at time step t. We use a rate equation approach to construct a master equation for the change in the joint probability distribution from time t to time $t+1$ [3]. To see what the network will look like in the long time limit, we solve this equation for the steady state[2].

To do this we consider two events. Event A is when a positive link is added to a given node at time $t+1$ and event B is when a negative link is added to a given node at time $t+1$. Let $P(A)$ be the probability that event A occurs, $P(\overline{A})$ be the probability that event A does not occur, $P(B)$ be the probability that event B occurs, and $P(\overline{B})$ be the probability that event B does not occur. There are three ways that a node can arrive at degree of (k_+, k_-) at time $t+1$. First, a node with degree $(k_+ - 1, k_-)$ can gain a positive link. Second, a node with degree $(k_+, k_- - 1)$ can gain a negative link. Lastly, a node that already has degree (k_+, k_-) can gain no new links. So we say that the probability

[2] The work in this section was conducted by Shadrack Antwi. For full details and results refer to [3].

that a node will have degree (k_+, k_-) at time $t+1$ is the sum of the probabilities of these three situations:

$$\begin{aligned} P(k_+, k_-, t+1) =& P(k_+ - 1, k_-, t) \cdot P(A)P(\overline{B}) \\ &+ P(k_+, k_- - 1, t) \cdot P(B)P(\overline{A}) \\ &+ P(k_+, k_-, t) \cdot P(\overline{A})P(\overline{B}). \end{aligned} \tag{2.3}$$

The probability that a link will be added is the product of the attachment kernel and m because each link that is added at a time step is a chance that a given node will be chosen. Using our definitions of the attachment kernel, we can write this in terms of k_+, k_-, and m:

$$\begin{aligned} P(k_+, k_-, t+1) =& P(k_+ - 1, k_-, t) \cdot m \frac{k_+}{\sum\limits_j (k_{+j} + 1)} \left[1 - \frac{k_- + 1}{\sum\limits_j (k_{-j} + 1)} \right]^m \\ &+ P(k_+, k_- - 1, t) \cdot m \frac{k_-}{\sum\limits_j (k_{-j} + 1)} \left[1 - \frac{k_+ + 1}{\sum\limits_j (k_{+j} + 1)} \right]^m \\ &+ P(k_+, k_-, t) \cdot \left[1 - \frac{k_- + 1}{\sum\limits_j (k_{-j} + 1)} \right]^m \left[1 - \frac{k_+ + 1}{\sum\limits_j (k_{+j} + 1)} \right]^m . \end{aligned} \tag{2.4}$$

From the master equation we can produce a recurrence relation for $P(k_+, k_-)$. We assume that the network reaches a steady state in the long time limit. To find this steady state we want to calculate the probability that a node with degree (k_+, k_-) will transition to a new degree combination.

Figure 2.2 shows the web of possible degree transitions for a node with degree (k_+, k_-). The node can gain more positive links and move left along the web or gain more negative links and move right along the web. Consider a node transitioning from a degree of

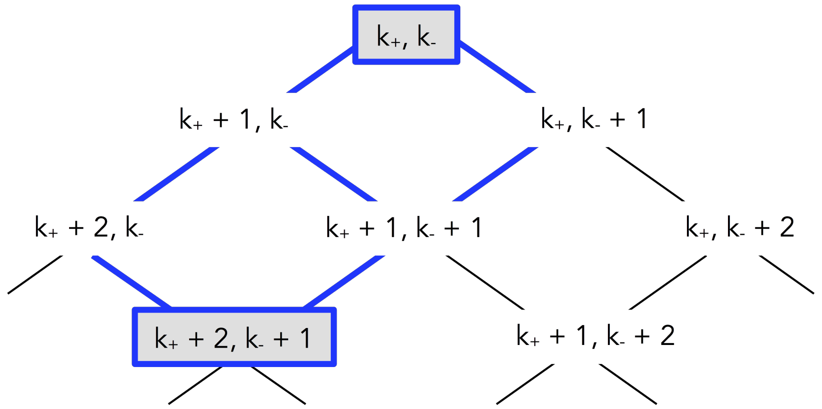


Figure 2.2: Binary tree of possible paths for a node from degree (k_+, k_-). Blue lines show possible paths from (k_+, k_-) to $(k_+ + 2, k_- + 1)$

(k_+, k_-) to $(k_+ + 2, k_- + 1)$. It could first add 2 positive links then 1 negative link, or 1 positive link then 1 negative link then 1 positive link, etc. Because the positive and negative attachment kernels are separate, each of these paths has the same probability of occurring, so to calculate the probability of this transition we count the number of possible paths. Using this observation we can solve the recurrence relation resulting in the following exact analytic expression for the joint probability distribution:

$$P(k_+, k_-) = \frac{\Gamma(k_+ + k_- - 2m + 1)}{\Gamma(k_+ - m + 1)\Gamma(k_- - m + 1)} \frac{\Gamma(k_+ + 1)\Gamma(k_- + 1)}{\Gamma(\frac{c}{a} + k_+ + k_- + 1)}\lambda, \tag{2.5}$$

where $a = 1 + \frac{1}{2m}$, $c = 2a^2 + 2a$, and $\lambda = \frac{2a\Gamma(\frac{c}{a}+2m)}{\Gamma(m+1)\Gamma(m+1)}$.

In the asymptotic limit of large k_+, k_- the joint probability distribution reaches the following steady state:

$$P(k_+, k_-) \sim \frac{(k_+ k_-)^m}{(k_+ + k_-)^{4+2m+1/m}}. \tag{2.6}$$

If we take the distribution of positive degree and negative degree separately we obtain the following asymptotic positive and negative degree distributions:

$$P(k_+) \sim {k_+}^{-3-1/m} \tag{2.7}$$

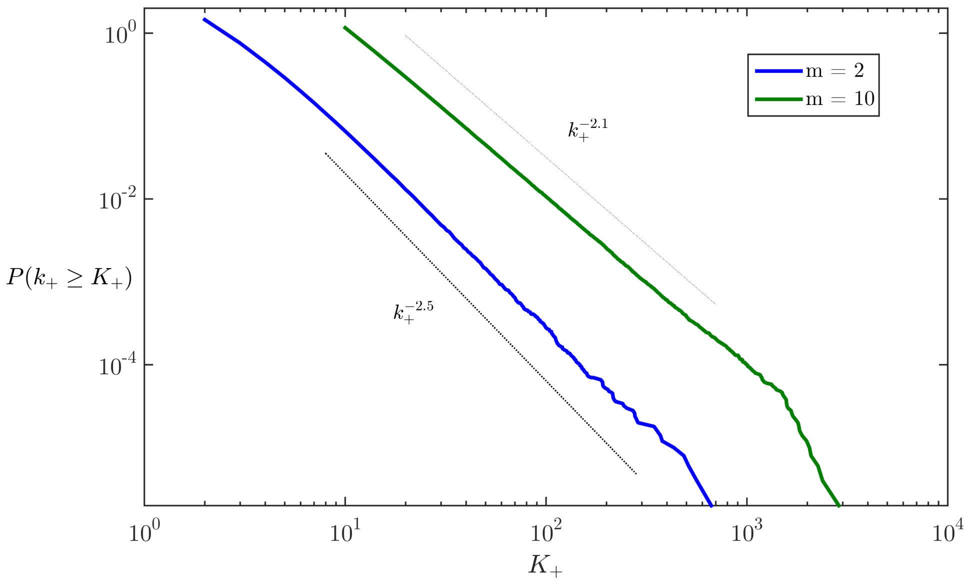


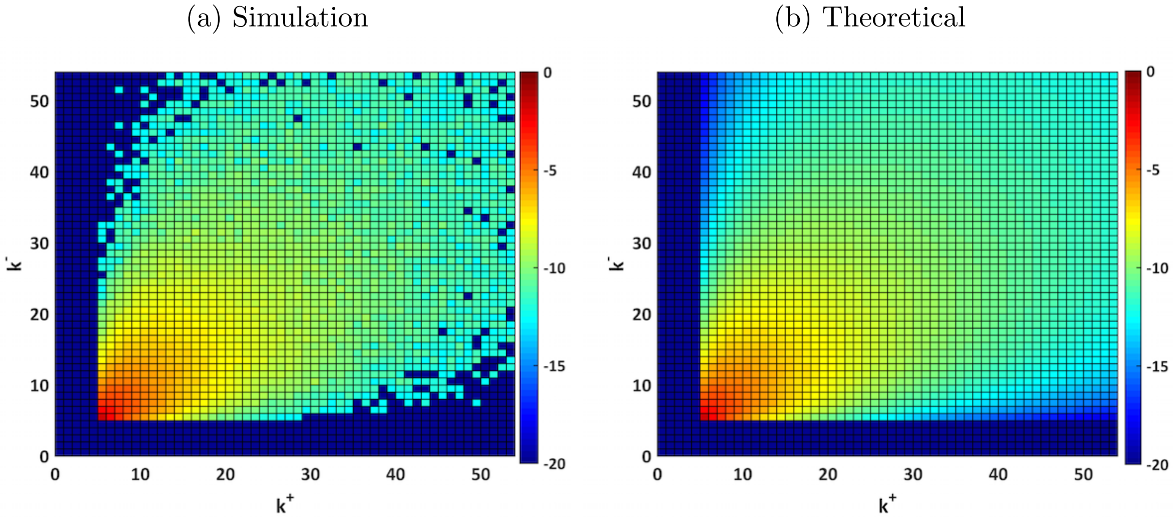
Figure 2.3: Reverse cumulative degree distribution for a network simulation grown with separate preferential attachment with $m = 2$ and $m = 10$ at $t = 500,000$. Dotted lines are theoretical prediction from Equation 2.7.

$$P(k_-) \sim {k_-}^{-3-1/m}. \tag{2.8}$$

These are identical to that of an unsigned network grown by the BA model with the variation of the attachment kernel being proportional to degree + 1 [3].

2.2.2 Simulation Results

To verify our exact theoretical prediction for the degree distribution we simulated a signed network grown with separate preferential attachment using the method described in Section 2.1.1. Figure 2.3 shows the reverse cumulative positive degree distribution for two simulations of a network with separate preferential attachment. The two simulations were run with different values of m to verify that the degree distribution depends on m. The dotted lines are the theoretical prediction from Equation 2.7. We see that this network is indeed scale-free. While the full joint probability distribution is in three dimensions,


(a) Simulation

(b) Theoretical

Figure 2.4: Joint probability distribution plots of log $P(k_+, k_-)$ vs. k_+, k_- with $m = 5$ at $t = 300,000$ for a network grown with separate preferential attachment from network simulation (a) and exact analytic expression (b). Figure from [3].

this plot looks only at the positive degree distribution so that we may observe it in two dimensions. A negative degree distribution would be identical.

Figure 2.4 shows the joint probability distribution in three dimensions. The x and y axes are positive degree and negative degree respectively and the color is the log of the probability of that degree combination. The two plots show agreement between the simulation and the exact theoretical prediction.

2.2.3 Cross Correlation

The relationship between a node's positive degree and its negative degree is an important result of a network grown with preferential attachment. Studying the resulting cross correlation of positive and negative degrees in a network gives us important insight into how the network was grown.

In a network grown with separate preferential attachment, older nodes have the greatest chance of receiving positive links and also have the greatest chance of receiving negative links. Thus we expect the positive and negative degree of the nodes in the network to be

k_{max}	Simulation	Exact Analytic Expression
50	0.7195	0.7206
55	0.7291	0.7281

Table 2.2: Cross correlation $C(k_+, k_-)$ from a network simulation and from the exact analytic expression. k_{max} is the maximum range of k_+ and k_- used to compute the cross correlation.

positively correlated.

Let σ_{k_+} be the standard deviation of k_+ and σ_{k_-} be the standard deviation of k_-. The mean positive and negative degrees are $\overline{k_+}$ and $\overline{k_-}$ respectively. Due to limitations of computing large factorials in Matlab we choose a k_{max} and restrict our computation to nodes with positive and negative degrees at or below our chosen k_{max}. We define the cross correlation between a node's positive degree k_+ and its negative degree k_- as follows:

$$C(k_+, k_-) = \frac{1}{\sigma_{k_+}\sigma_{k_-}} \left[\sum_{k_+=m}^{k_{max}} \sum_{k_-=m}^{k_{max}} (k_+ - \overline{k_+})(k_- - \overline{k_-}) P(k_+, k_-) \right]. \tag{2.9}$$

The values of σ_{k_+}, σ_{k_-}, $\overline{k_+}$, and $\overline{k_-}$ can be calculated using the exact expression for $P(k_+, k_-)$ allowing us to calculate the exact analytic cross correlation. Table 2.2 shows the resulting cross correlation computed using Equation 2.9 for a simulated network and verifies that they are consistent with our exact analytic expression for cross correlation.

2.3 Other Attachment Methods

Separate preferential attachment offered a base case for signed network growth and its simplicity allowed for thorough analytic investigation. However separate preferential attachment fails to fully incorporate the meaning of signed links into its method for growth. When a user makes a connection with another user, reviews a product, or votes on an issue they will most likely make their decision based on both types of available information: positive links and negative links. But how are these two attributes to be weighed in the decision? Do positive links help you just as much as negative links hurt you–or does one type of link carry more weight? Does the old saying hold true that "any press is good press"? The following sections propose several more complex methods for signed preferential attachment with the purpose of modeling these possible phenomena.

2.3.1 Ratio Preferential Attachment

Ratio preferential attachment incorporates the idea that positive links are intrinsically good and negative links are intrinsically bad. Thus the probability that a node will gain positive links is positively correlated to the number of positive links it has and negatively correlated to the number of negative links it has. Likewise the probability that a node will gain negative links is positively correlated to the number of negative links it has and negatively correlated to the number of positive links it has. We define the positive attachment kernel to be proportional to the ratio of positive degree to negative degree and define the negative attachment kernel to be proportional to the ratio of negative degree to positive degree as follows:

$$\Pi_+(i) = \frac{\frac{k_{+i}+1}{k_{-i}+1}}{\sum\limits_j \left(\frac{k_{+j}+1}{k_{-j}+1}\right)} \tag{2.10}$$

$$\Pi_-(i) = \frac{\frac{k_{-i}+1}{k_{+i}+1}}{\sum\limits_j \left(\frac{k_{-j}+1}{k_{+j}+1}\right)}. \tag{2.11}$$

By taking the ratio of positive to negative degree when deciding which nodes will receive new positive links, having a lot of positive links will increase your chances of gaining positive connections and having a lot of negative links will decrease your chances of gaining positive connections. Likewise, having a lot of negative links will increase your chances of gaining negative connections and having a lot of positive links will decrease your chances of gaining negative connections.

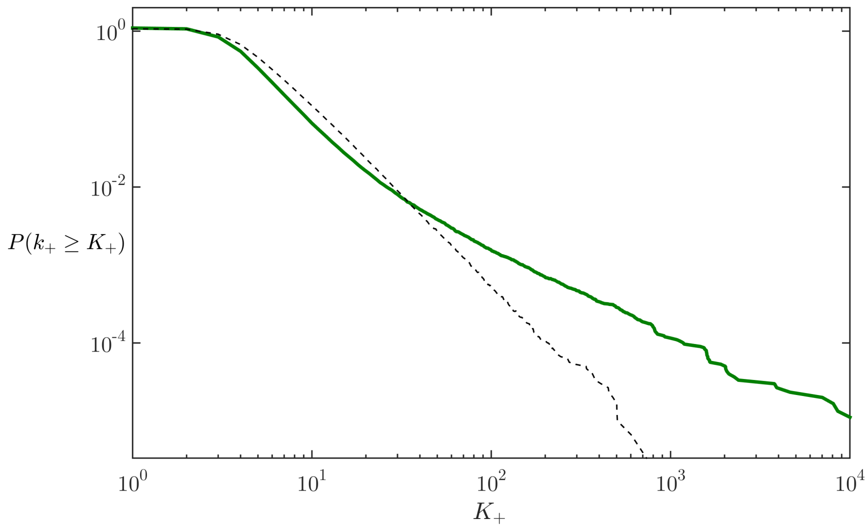


Figure 2.5: Reverse cumulative degree distribution for a network simulation grown with ratio preferential attachment with $m = 5$ at $t = 300,000$. Dotted curve is a degree distribution for an unsigned scale-free network.

Figure 2.5 shows the reverse cumulative degree distribution for a simulated network with ratio preferential attachment. For comparison we also plot a degree distribution for an unsigned scale-free network. It is possible that an earlier stage of the network exhibited scale-free properties, however, in the long time limit this method does not result in a scale-free network. In the infinite limit the network results in one gigantic node that has a probability of gaining a new link almost equal to 1. This is a known phenomenon for preferential attachment methods where the degree is raised to a power greater than 1 [12].

2.3.2 Weighted Product Preferential Attachment

Weighted product preferential attachment is based on the principle that any press is good press. This means that your probability of gaining positive links is positively correlated with your negative degree. For the positive attachment kernel we wish to weight the

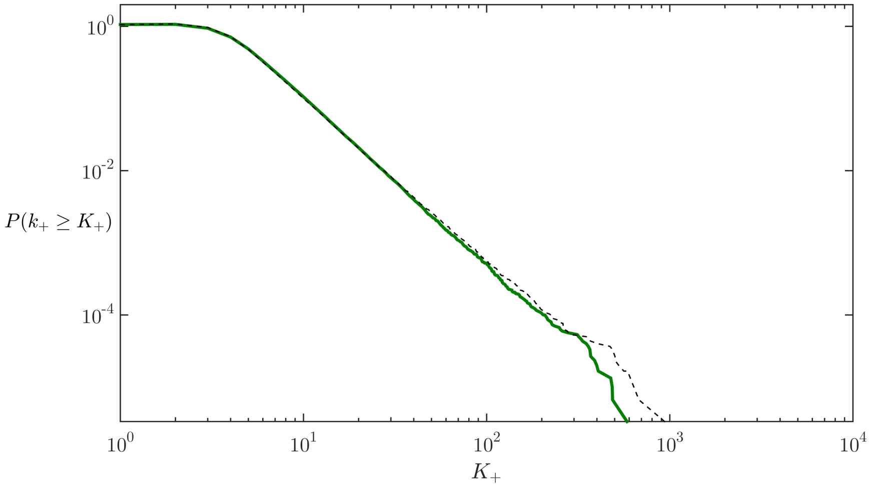


Figure 2.6: Reverse cumulative degree distribution for a network simulation grown with weighted product preferential attachment with $m = 5$ at $t = 300,000$. Dotted curve is a degree distribution for an unsigned scale-free network.

positive degree more heavily to account for the different meaning of positive and negative links so we scale the positive degree by an exponent $\beta > 0.5$ and scale the negative degree by $1 - \beta$. By choosing these exponents so that they sum to 1 we avoid the "gelation-like" effect that occurred in ratio preferential attachment when a degree is scaled by an exponent greater than 1. For the positive attachment kernel we choose to scale the positive degree by 0.7 and the negative degree by 0.3. The negative attachment kernel is symmetric such that the positive degree is scaled by 0.3 and the negative degree is scaled by 0.7. The resulting positive and negative attachment kernels are as follows:

$$\Pi_+(i) = \frac{(k_{+i}+1)^{0.7}(k_{-i}+1)^{0.3}}{\sum\limits_j (k_{+j}+1)^{0.7}(k_{-j}+1)^{0.3}} \tag{2.12}$$

$$\Pi_-(i) = \frac{(k_{+i}+1)^{0.3}(k_{-i}+1)^{0.7}}{\sum\limits_j (k_{+j}+1)^{0.3}(k_{-j}+1)^{0.7}}. \tag{2.13}$$

This model did result in a scale-free network in the long time limit (see Figure 2.6).

Thus it is the first method for signed preferential attachment that considers the meaning of signed links while still resulting in a scale-free network. This model also has the potential to be adapted by changing the weighting of the positive and negative degrees. We expect that a modification of this model with slight changes in the weighting of the degrees would still result in a scale-free network, offering variability when fitting to a real-world network.

Chapter 3

Data Fitting

3.1 Background and Assumptions

The previous chapter defined several models for signed network growth. In this chapter we discuss data fitting methods to select the best fitting model for a given real-world network. Determining the rules for growth in a real-world network will enable prediction of the future behavior of the network. This information is very important for understanding networks and has benefits for targeted advertising, user-retention efforts, and determining the quality of products. Here we discuss and test existing methods for measuring preferential attachment in temporal unsigned network data sets and suggest modifications for signed network data sets.

When choosing a fitting method for a given data set we must consider the number of time points available. Some data sets are very detailed and have a time stamp denoting exactly when each link was added to the network. Others might only have one or a few snapshots of the network structure at a given moment in time. Clearly, using more time points will result in a more accurate fitting; however, we also consider methods that only require a few time steps in order to fit a broader range of data sets.

We use the Epinions social network as a case study for measuring preferential attachment [16]. Epinions.com is a consumer review platform where users are paid to write reviews on a large collection of products. Users have the ability to personalize the reviews they see by “trusting” and “distrusting” other users. The Epinions social network consists of the trust connections (positive links) and distrust connections (negative links) between users. The data set gives the connections made by $131,828$ users on the website from January 2001 to August 2003. These connections are represented by $841,372$ directed links with a value of $+1$ or -1 and the date the link was added. Because the links are directed we only consider the in-degree of a node, the number of links pointing towards that node. For simplicity of computation, we only consider the network’s giant connected component, which consists of $648,623$ links. This data set has been used previously to infer unknown trust connections from existing trust connections [13,18,19]. However, these studies lack an understanding of the impact of negative links on preferential attachment.

This chapter provides background on existing methods for measuring preferential attachment in unsigned networks. We present our attempts to reproduce these methods and discuss the limitations that made some of these methods very sensitive to parameter choices. We then discuss adaptations of the methods to alleviate these shortcomings. Finally we propose methods for measuring preferential attachment in signed networks.

3.1.1 Nonlinear Preferential Attachment

The attachment kernels of growing networks were first measured in 2003 by Jeong et al. [11]. Their method fits each node’s change in degree between two time points (t_0 and t_1) to the attachment kernel of the network. The attachment kernel is assumed to be nonlinear, where the linear attachment kernel defined in the BA model (Equation 1.8) is modified by including a scaling exponent $\alpha > 0$:

$$\Pi(i) = \frac{k_i^\alpha}{\sum\limits_j k_j^\alpha}. \tag{3.1}$$

They calculate the proportion of new links added between t_0 and t_1 to nodes that originally had degree k at t_0. Jeong et al. assume that when a short interval between t_0 and t_1 is used, the functional form of $\Pi(i)$ does not depend on t_0 or t_1. Then they fit the change in degrees from t_0 to t_1 to Equation 3.1 and approximate the value of α for a given data set. They recalculate α with different t_0 and t_1 values to show that the fitting is not dependent on t_0 or t_1.

When $\alpha = 1$, the network uses linear preferential attachment resulting in a power law degree distribution with $P(k) \sim k^{-3}$. When $\alpha > 1$, the network experiences a "gelation-like" effect where one gigantic node has a probability of gaining a new link almost equal to 1 [12]. These networks are not scale-free and their degree distributions do not reach a steady state. For $\alpha < 1$, the network uses sublinear preferential attachment which is shown to result in a stretched exponential degree distribution [13].

Jeong et al. approximated the scaling exponent for the Internet and citation network of published papers to be $\alpha = 1 \pm 0.1$, showing that these networks use linear preferential attachment. The co-authorship network of neuroscientists and the co-starring network of movie actors are shown to have sublinear preferential attachment with scaling exponents of $\alpha = 0.79 \pm 0.1$ and $\alpha = 0.81 \pm 0.1$ respectively [11]. These results show that the nonlinear preferential attachment model can achieve a better fit than the linear preferential attachment model for some real-world networks. Next we discuss variations of this method to better fit temporal network data sets to the nonlinear preferential attachment model.

3.2 Least Squares Fitting

The first method for preferential attachment measurement we investigate in detail uses least squares regression to fit the attachment kernel to a node's change in degree between two time points. Kunegis et al. expand on the method proposed by Jeong et al. and measure nonlinear preferential attachment using least squares fitting for over 40 online networks, showing that the scaling exponent depends on the type of network (e.g. social, rating, or interaction networks), leading to a better understanding of the social processes underlying preferential attachment [13]. In signed networks, their method ignores the sign of a link and treats all positive and negative connections between users as unsigned links.

Kunegis et al. propose that the two time points (t_0 and t_1) be chosen such that t_1 is the most recent state of the network and t_0 contains 75% of the links that are present at t_1. This differs from the choice of time steps by Jeong et al., which requires a short interval between t_0 and t_1. These choices for time points offer an advantage over some other data fitting methods because network data may be obtained by taking two snapshots of the network rather than identifying the exact time that each link was added.

The method used by Kunegis et al. performs least squares fitting on the logarithm of the change in degrees. We find that a parameter choice for logarithmic least squares fitting can drastically change the result so we suggest variations of least squares fitting that offer different weightings of the data.

3.2.1 Logarithmic Least Squares Fitting

We begin our study by attempting to reproduce the logarithmic least squares fitting method for measuring preferential attachment achieved by Kunegis et al. [13]. Kunegis et al. chose to perform the least squares fitting on the logarithmic degrees. They suggest that this avoids the over-weighting of noisy high degrees in the network.

We define k_i^0 to be the degree of node i at t_0, k_i^1 to be the degree of node i at t_1, and the change in degree of node i to be $\Delta(i) = k_i^1 - k_i^0$. Note that we use our own notation for clarity. We make the assumption that the following method of preferential attachment is used to grow the network:

$$\Pi(i) = \frac{(1+k_i^0)^\alpha}{\sum\limits_j (1+k_j^0)^\alpha}. \tag{3.2}$$

The +1 is added so that nodes with degree zero have a nonzero attachment kernel. We define the normalization term to be $c(t)$ and treat it as a constant $c(t_0)$. Then $\Pi(i) = (1 + k_i^0)^\alpha c(t_0)$. This attachment kernel should be proportional to the change in degree of each node:

$$\Delta(i) \propto (1 + k_i^0)^\alpha c(t_0). \tag{3.3}$$

Kunegis et al. suggest adding a regularization parameter λ to $\Delta(i)$. This ensures that the value is nonzero so we can we take its logarithm even if the node does not gain any links between t_0 and t_1. We search for an α such that

$$\ln[\Delta(i) + \lambda] \propto \ln[(1 + k_i^0)^\alpha c(t_0)]. \tag{3.4}$$

Thus we use the following expression over all nodes:

$$\ln[(1 + k_i^0)^\alpha c(t_0)] - \ln[\Delta(i) + \lambda]. \tag{3.5}$$

To simplify the expression we let $c(t_0) = e^\beta$. Then we arrive at the least squares minimization expression:

$$\min_{\alpha,\beta} \sum_j \left(\alpha \ln[1 + k_i^0] + \beta - \ln[\Delta(i) + \lambda]\right)^2. \tag{3.6}$$

We test this method for measuring preferential attachment by applying to temporal data from a simulated network where $\alpha = 1$ is known. We minimized the expression for the change in degrees of the network using the Matlab function fminsearch which implements the Nelder-Mead method. Since the expression is linear in its unknown variables, this minimization method is guaranteed to find a global minimum. Using $\lambda = 0.1$ as suggested by [13] we measured $\alpha = 1.2900$ for a simulated unsigned network grown using the attachment kernel from Equation 3.2. Because the value of the parameter λ was chosen arbitrarily, we investigated its effect on the resulting measurement of α. We ran the method for values of λ between 0.01 and 1.00 and list the results in Table 3.1. The range

λ	α
0.01	1.7692
0.05	1.4370
0.10	1.2900
0.20	1.1378
0.30	1.0453
0.40	0.9779
1.00	0.7526

Table 3.1: Resulting α values using logarithmic least squares fitting for various λ values.

of α measurements for these small shifts in the value of λ presents a deficiency in this method. The arbitrary choice of λ leads to incorrectly classifying the same network as sublinear, linear, or superlinear. This shortcoming leads us to consider data fitting methods that do not take the logarithm of the degree and thus do not require the parameter λ.

3.2.2 Basic Least Squares Fitting

The logarithmic least squares fitting method introduced an unnecessary arbitrary parameter that made the fitting unreliable. We return to a basic least squares fitting method and suggest variations of how to group and weight the degrees.

For a simple least squares fitting method, we search for an α such that

$$\Delta(i) \propto (1 + k_i^0)^\alpha c(t_0). \tag{3.7}$$

Note that we have removed the λ parameter because $\Delta(i)$ is permitted to be zero if we are not taking its logarithm. Then we minimize the following over all nodes:

$$(1 + k_i^0)^\alpha c(t_0) - \Delta(i). \tag{3.8}$$

To simplify the expression we let $c(t_0) = \beta$. Then we arrive at the least squares mini-

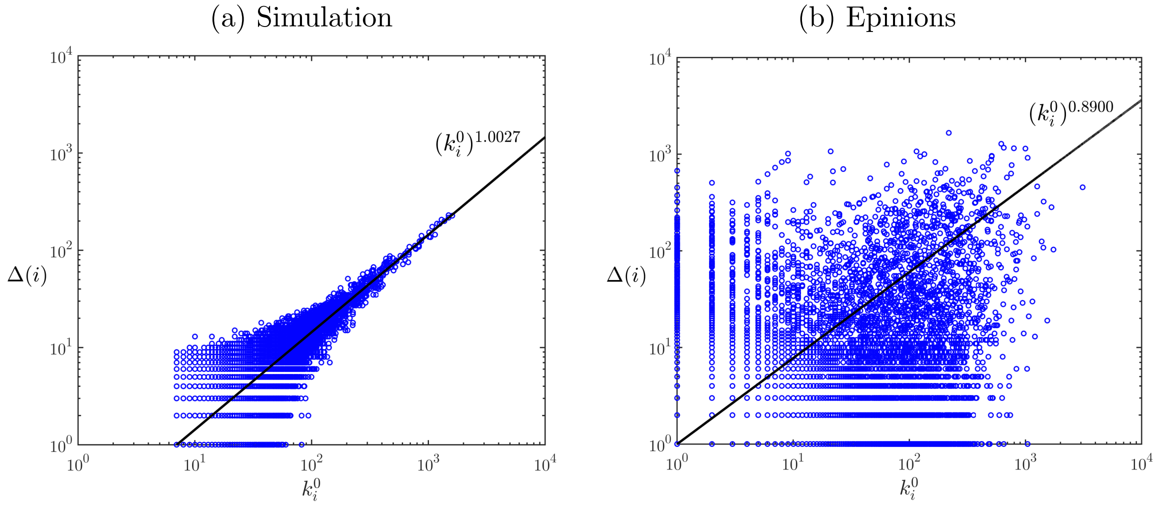


Figure 3.1: Change in degree from t_0 to t_1 for (a) a simulation grown with linear preferential attachment at $t = 300,000$ and (b) the Epinions social network. Black curve shows the fitting obtained by the basic least squares fitting method in Equation 3.9.

mization expression:

$$\min_{\alpha,\beta} \sum_j \left(\beta(1+k_i^0)^\alpha - \Delta(i)\right)^2. \tag{3.9}$$

Note that this minimization expression is nonlinear in its unknown variables and thus it is not guaranteed that our minimization method will find its global minimum. Our tests with simulated data and the Epinions data set are indeed able to find a minimum and visual inspection validates the fitting (see Figure 3.1).

We test this method on a simulated network grown using the preferential attachment method described in Equation 3.2 with $\alpha = 1$. The basic least squares method produced an estimation of $\alpha = 1.0027$ that is more accurate than the estimation produced by the logarithmic least squares method (see Table 3.2). We also test the basic least squares method on the Epinions data set and produce an estimation of $\alpha = 0.8900$ which differs significantly from the value of α suggested by Kunegis et al. [13].

Fitting Method	**Simulation** ($\alpha = 1.0$)	**Epinions**
Logarithmic Least Squares ($\lambda = 0.1$)	$\alpha = 1.2900$	$\alpha = 0.5525$
Basic Least Squares	$\alpha = 1.0027$	$\alpha = 0.8900$
Weighted Least Squares	$\alpha = 1.0330$	$\alpha = 0.9514$
Binned Least Squares	$\alpha = 1.0150$	$\alpha = 0.7384$

Table 3.2: Resulting α values using various fitting methods. The simulation is grown for $300,000$ time steps with a known value of $\alpha = 1$. The expected value for the Epinions data set is unknown.

3.2.3 Weighting of Data Points

Kunegis et al. suggest the importance of how the data points in a fitting are being weighted [13]. They suggest that the high degrees in the network are noisy and thus may not accurately describe the attachment kernel. On the other hand, there are significantly more low-degree rather than high-degree nodes in the network so taking each node as its own data point weights the fitting toward the attachment kernels exhibited by non-noisy low-degree nodes. It is possible that this skew due to the number of occurrences of nodes correctly weights the data.

Kunegis et al. proposes that the data must be weighted further to downplay the noisy high degrees. Let σ_k be the error in Δ for a node with degree k. Kunegis et al. suggest that σ_k increases with k. Define $y = \ln \Delta$. Kunegis et al. weight all y values equally, independent of k. Since $\Delta = e^y$, $\sigma_k = e^y \sigma_y$ by propagation of error. Because σ_y is assumed to be constant, $\sigma_k \propto e^y$ which increases with k. Thus high degree nodes are treated as having larger error bars. Then the least squares fitting of the logarithmic data can be performed without adjusting for error.

We wish to avoid taking the logarithm of degrees as it introduced the arbitrary parameter λ, so we present an alternative method to weight the high degree nodes less heavily. We assume that the change in degree Δ of a node with degree n is drawn from a binomial distribution with mean Δ_n. Each new link has probability p to connect to this node, increasing its change in degree. The variance of a binomial distribution is $\sigma^2 = \mu(1 - p)$.

For a large network each p is very small so the variance for a single node is $\sigma^2 \approx \mu$. Thus $\sigma^2 \approx \Delta_n$. This observation matches our intuition because high degrees have a large Δ_n and thus have a greater error. We correct for this error by weighting the data points by their variance in the least squares minimization:

$$\min_{\alpha,\beta} \sum_j \frac{\left(\beta(1+k_i^0)^\alpha - \Delta(i)\right)^2}{\Delta(i)}. \tag{3.10}$$

The results of this method to fit a simulated network and the Epinions data set are shown in Table 3.2. This method accurately estimates the preferential attachment of a simulated network where α is known. The estimation of α for the Epinions data set differs from previous fitting methods, further showing the impact of weighting on a data fitting.

Our next proposed variation involves a different assumption about the errors of the degrees. Here we wish to counteract the skew due to the large number of occurrences of low-degree nodes so that low degrees and high degrees are weighted equally. We do this by averaging the occurrences of a degree into one data point. We consider all the nodes at time t_0 and group them by their degrees into bins $n = 0, 1..., k_{max}$. Our data points for degrees at t_0 are k_n^0 for $n = 0, 1..., k_{max}$. For each bin we take the nodes in the bin and average the degree of those nodes at t_1 resulting in k_n^1 for $n = 0, 1..., k_{max}$. Thus our change in degree function is $\Delta(n) = k_n^1 - k_n^0$. We consider the following over all n:

$$(1+k_n^0)^\alpha c(t_0) - \Delta(n). \tag{3.11}$$

Then we arrive at the least squares minimization expression:

$$\min_{\alpha,\beta} \sum_{n=0}^{k_{max}} \left(\beta(1+k_n^0)^\alpha - \Delta(n)\right)^2. \tag{3.12}$$

The results of this method to fit a simulated network and the Epinions data set are shown in Table 3.2. This method also accurately estimates the preferential attachment of a simulated network where α is known. Again the estimation of α for the Epinions data set differs from previous fitting methods.

These methods present varying ways to weight the degrees in a data fitting method. No one method was shown to be superior to another. All resulted in different estimations for a data set with unknown preferential attachment, though we find the logarithmic least squares method to differ the most from the other methods. We conclude that the weighting of data points has significant impact on data fitting and should be further studied to obtain more accurate measurements of preferential attachment.

3.3 PAFit Maximum Likelihood Estimation

The previous section discussed the least squares fitting method which assumes the functional form of the preferential attachment method and uses two time points to fit the change in degree. This method is designed for data sets with limited time points or when power law preferential attachment is known to be present. This section discusses a more versatile method for estimating attachment kernels proposed by Pham et al. called PAFit maximum likelihood estimation [22].

The PAFit maximum likelihood estimation method iteratively fits temporal network data to the vector $\vec{A} = [A_0, A_1, ..., A_{k_{max}}]$, where A_k gives the relative probability that a node with degree k will gain a link such that $A_{k_i} \propto \Pi(i)$ for all i. We set $A_0 = 1$. Then, for example, a network with linear preferential attachment would have a vector $\vec{A} = [1, 2, 3, ...]$. The functional form of the attachment kernel need not be assumed because the method produces relative attachment probabilities to which a functional form for attachment can later be fitted.

The derivation of the PAFit maximum likelihood estimation defines variables that describe the links and nodes that are added at each time point then uses these to calculate the likelihood of observing a given network in the next time step. Let G_t be the state of the network at the end of time t after nodes and links have been added. We define $n(t)$

to be the number of nodes added during time t and $m(t)$ to be the number of links added during time t. A parameter vector $\vec{\theta}(t)$ determines the joint distribution of $n(t)$ and $m(t)$ and is assumed not to depend on $\vec{A}$. We let $\vec{\theta}_*$ denote the parameter vector specifying geometry for the network at $t = 0$.

The PAFit method first calculates the probability of observing the network G_t given the attachment kernel vector $\vec{A}$ and the previous state of the network at time $t-1$, G_{t-1}. This depends on two probabilities: first, the likelihood of $n(t)$ and $m(t)$ given G_{t-1} and $\vec{\theta}(t)$, and secondly, the likelihood of the current graph G_t given the previous graph G_{t-1}, $m(t)$, $n(t)$, and $\vec{A}$. Therefore,

$$P(G_t|G_{t-1}, \vec{A}) = P(m(t), n(t)|G_{t-1}, \theta(t)) \cdot P(G_t|G_{t-1}, m(t), n(t), \vec{A}). \tag{3.13}$$

Now that we can calculate the probability of one time point in the data we can compile these individual probabilities to find the likelihood of observing the entire data set given some $\vec{A}$. Then we maximize this likelihood to solve for $\vec{A}$. The likelihood of observing the entire data set is

$$P(G_1, G_2, ..., G_T|\vec{A}) = P(G_0|\vec{\theta}_*) \prod_{t=1}^{T} P(G_t|G_{t-1}, \vec{A}). \tag{3.14}$$

Note that the logarithm of this product will have the same maximum so we can instead maximize the following function:

$$l(\vec{A}) = \log P(G_0|\vec{\theta}_*) + \sum_{t=1}^{T} \log P(G_t|G_{t-1}, \vec{A}). \tag{3.15}$$

Using Equation 3.13 and the initial conditions for G_0 we obtain:

$$\begin{aligned} l(\vec{A}) = &\sum_{t=1}^{T} \log P(G_t|G_{t-1}, m(t), n(t), \vec{A}) \\ &+ \sum_{t=1}^{T} \log P(m(t), n(t)|G_{t-1}, \vec{\theta}(t)) \\ &+ \log P(G_0|m(0), n(0), \vec{\theta_*}) \\ &+ \log P(m(0), n(0)|\vec{\theta}(0)). \end{aligned} \tag{3.16}$$

Since we are maximizing the function by finding $\vec{A}$ we can ignore the last three terms because they do not depend on $\vec{A}$.

We define $n_k(t)$ to be the number of existing nodes with degree k at time t and $m_k(t)$ to be the number of new links that connect to nodes with degree k. Then the probability that a newly added edge at time t connects to a node with degree k is $p_k(t) = \frac{n_k(t)A_k}{\sum_{j=0}^{k_{max}} n_j(t)A_j}$. We represent $P(G_t|G_{t-1}, m(t), n(t), \vec{A})$ by a multinomial distribution to count the different ways links could be added:

$$P(G_t|G_{t-1}, m(t), n(t), \vec{A}) = \frac{2m(t)!}{\prod_{k=0}^{k_{max}} m_k(t)!} \prod_{k=0}^{k_{max}} p_k(t)^{m_k(t)}. \tag{3.17}$$

Substituting this equality into Equation 3.16 and dropping terms independent of A

Expected Value	**Estimated Value**
$\alpha = 1.0$	$\alpha = 0.920$
$\alpha = 0.7$	$\alpha = 0.661$

Table 3.3: Resulting α values using PAFit maximum likelihood estimation on two simulated networks grown for $300,000$ time steps with a known value for α.

we reduce the function as follows

$$\begin{aligned} l(\vec{A}) &= \sum_{t=1}^{T} \log \prod_{k=0}^{k_{max}} p_k(t)^{m_k(t)} && (3.18) \\ &= \sum_{t=1}^{T} \sum_{k=0}^{k_{max}} m_k(t) \log \left(\frac{n_k(t) A_k}{\sum_{j=0}^{k_{max}} n_j(t) A_j} \right) && (3.19) \\ &= \sum_{t=1}^{T} \sum_{k=0}^{k_{max}} m_k(t) \left[\log(n_k(t) A_k) - \log \left(\sum_{j=0}^{k_{max}} n_j(t) A_j \right) \right] && (3.20) \\ &= \sum_{t=1}^{T} \sum_{k=0}^{k_{max}} m_k(t) \log(n_k(t) A_k) - \sum_{t=1}^{T} m(t) \log \left(\sum_{j=0}^{k_{max}} n_j(t) A_j \right). && (3.21) \end{aligned}$$

To solve for $\vec{A}$ we begin with some arbitrary guess for $\vec{A}$ then update each A_k using the following recurrence relation for every time point:

$$A_k = \frac{\sum_{t=1}^{T} m_k(t)}{\sum_{t=1}^{T} \frac{m(t) n_k(t)}{\sum_{j=0}^{K} n_j(t) A_j}}. \tag{3.22}$$

Pham et al. prove that this recurrence relation converges.

We used code provided by Pham et al. to apply this method to data from two simulated networks grown with preferential attachment, one with a scaling exponent of $\alpha = 1$ and the other with $\alpha = 0.7$ [24]. We fit the resulting attachment kernel vector to the nonlinear preferential attachment kernel described in Equation 3.2. Table 3.3 displays the results from these data fittings and verify that the PAFit maximum likelihood estimation method can correctly measure nonlinear preferential attachment.

3.3.1 Signed Maximum Likelihood Estimation

The flexibility of the PAFit maximum likelihood estimation estimation makes it an ideal option for fitting signed networks. Here we propose an adaption of the PAFit maximum likelihood estimation method that fits both the positive and negative attachment kernels of a network.

Since signed attachment kernels are dependent on both positive and negative degrees we aim to fit the two dimensional parameter matrices $\mathbf{A}_+$ and $\mathbf{A}_-$. Each entry $A^+_{k_+,k_-}$ in $\mathbf{A}_+$ represents the relative probability that a node with positive degree k_+ and negative degree k_- will gain a positive link. Likewise each entry $A^-_{k_+,k_-}$ in $\mathbf{A}_-$ represents the relative probability that a node with positive degree k_+ and negative degree k_- will gain a negative link.

We define $m_+(t)$ to be the number of positive links added at time t and $m_-(t)$ to be the number of negative links added at time t. The parameter vector $\vec{\theta}(t)$ determines the joint distribution of $n(t)$, $m_+(t)$, and $m_-(t)$ and does not depend on $\vec{A}$. We let $\vec{\theta}_*$ denote the parameter vector for the network at $t = 0$. Then the likelihood of observing the network G_t given the previous network G_{t-1}, $\mathbf{A}_+$, and $\mathbf{A}_-$ is as follows:

$$\begin{aligned} P(G_t|G_{t-1},\mathbf{A}_+,\mathbf{A}_-) =& P(m_+(t),m_-(t),n(t)|G_{t-1},\theta(t)) \\ & \cdot P(G_t|G_{t-1},m_+(t),m_-(t),n(t),\mathbf{A}_+,\mathbf{A}_-). \end{aligned} \tag{3.23}$$

In the same manner that we achieved Equation 3.16 we combine the probabilities for each time step to calculate the probability of observing the entire data set arriving at the following maximizing function:

$$\begin{aligned} l(\mathbf{A}_+, \mathbf{A}_-) = \sum_{t=1}^{T} \log P(G_t | G_{t-1}, m_+(t), m_-(t), n(t), \mathbf{A}_+, \mathbf{A}_-) \\ + \sum_{t=1}^{T} \log P(m_+(t), m_-(t), n(t) | G_{t-1}, \vec{\theta}(t)) \\ + \log P(G_0 | m_+(0), m_-(0), n(0), \vec{\theta_*}) \\ + \log P(m_+(0), m_-(0), n(0) | \vec{\theta}(0)). \end{aligned} \tag{3.24}$$

The last three terms may be dropped because they do not depend on $\mathbf{A}_+$ or $\mathbf{A}_-$. We propose to solve for $\mathbf{A}_+$ and $\mathbf{A}_-$ by maximizing this function.

Our intuition from the models of signed preferential attachment presented in Chapter 2 allow us to make predictions about the positive and negative attachment matrices for signed networks. First we observe that the existence of separate matrices for positive and negative attachment enable fitting to a network that uses completely different rules for positive link attachment and negative link attachment. The models we propose in Chapter 2 have symmetry between their positive and negative attachment kernels. However, a real-world network may, for example, use separate preferential attachment for its positive links and ratio preferential attachment for its negative links.

If we suspect a signed preferential attachment method may be occurring in a network we can identify the functional form of the attachment kernel from its positive and negative attachment matrices. For example we can deduce the expected attachment matrices for a network with separate preferential attachment. The positive attachment matrix $\mathbf{A}_+$ for such a network would have all identical rows because the positive attachment kernel does not depend on negative degree. Likewise the negative attachment matrix $\mathbf{A}_-$ would have all identical columns because the positive attachment kernel does not depend on positive degree.

The PAFit maximum likelihood estimation method has the disadvantage of requiring a robust data set with many time points. It also requires many occurrences of each k_+, k_- combination, though this requirement can be relaxed by binning the degree combinations. Despite these disadvantages, the flexibility of the PAFit maximum likelihood estimation method to not have to assume the functional form of the attachment kernel offers a great benefit, making it the most viable option for signed network data fitting that we have studied.

Chapter 4

Conclusions

The use of network science in a diverse range of fields reveals the importance of studying and understanding network structure. By modeling the growth of networks over time we can offer accurate predictions for the future behavior of a network and extract meaning about the individual members in a network. Current research about network growth continues to improve our understanding of the rules for network evolution. This project contributes to the recent body of research relating to signed networks. We aim to study the effect of assigning a positive or negative value to links in the network.

The first stage of this project was to develop potential rules for signed network growth. We successfully proposed and analyzed the basic mechanism of separate preferential attachment and reached a solid understanding of its effect on the degree distribution and cross correlation of a network. We presented several other mechanisms for signed preferential attachment, each motivated by an intuition about social behavior. These preferential attachment mechanisms offer a foundation for future research on signed network growth and provide functional forms to which real-world data can be fit.

The next stage of the project was to investigate methods for fitting data from real-world signed networks. These fitting methods attempt to determine the form of preferential attachment used in a network so that we may classify networks and make predictions about their future behavior. We first reproduced the method of logarithmic least squares regression proposed by Kunegis et al. We show that this method is unstable due to a regularization parameter and suggest that the stable method of basic least squares regression is preferred. Lastly we investigate the PAFit maximum likelihood estimation method proposed by Pham et al. We conclude that the flexibility of this method makes it the ideal choice for signed network data fitting.

4.1 Future Work

This project offers a foundation for the study of growing networks and leads to several ideas for further research. Our proposals for signed network attachment serve as examples for the breadth of possible signed preferential attachment mechanisms that could be present in real-world networks. Further research could include developing more possibilities for signed preferential attachment and analyzing their viability.

Another important continuation of this project is to complete the development of a signed network fitting method. We present suggestions for signed network data fitting but have yet to fully implement a fitting method. Finally, future research on signed networks would greatly benefit from data collection for more real-world signed networks. For this project we found only one publicly available signed network data set of an adequate size. The collection of many real-world signed networks would open up great opportunity for the study and classification of signed network growth.

Bibliography

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[2] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 09 1999.

[3] Shadrack A. Antwi. *Dynamic Social Networks with Beneficial and Detrimental Interactions*. ProQuest Dissertations Publishing, 2015.

[4] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 10 1999.

[6] Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 02 2004.

[7] Phillip Bonacich and Paulette Lloyd. Calculating status with negative relations. *Social Networks*, 26(4):331 – 338, 2004.

[8] Timo Hiller. Alliance formation and coercion in networks. 2011.

[9] Timo Hiller. Friends and enemies: A model of signed network formation. *Available at SSRN 2371249*, 2013.

[10] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 10 2000.

[11] Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

[12] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85:4629–4632, Nov 2000.

[13] Jérôme Kunegis, Marcel Blattner, and Christine Moser. Preferential attachment in online networks: Measurement and explanations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 205–214, New York, NY, USA, 2013. ACM.

[14] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: Mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 741–750, New York, NY, USA, 2009. ACM.

[15] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370. ACM, 2010.

[16] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`, June 2014.

[17] Fredrik Liljeros, Christofer R. Edling, H. Eugene Stanley, Y. Aberg, and Luis A. N. Amaral. Social networks (communication arising): Sexual contacts and epidemic thresholds. *Nature*, 423(6940):606–606, 06 2003.

[18] Haifeng Liu, Ee-Peng Lim, Hady W. Lauw, Minh-Tam Le, Aixin Sun, Jaideep Srivastava, and Young Ae Kim. Predicting trusts among users of online communities:

An epinions case study. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08, pages 310–319, New York, NY, USA, 2008. ACM.

[19] Paolo Massa and Paolo Avesani. Controversial users demand local trust metrics: An experimental study on epinions. com community. In *Proceedings of the National Conference on artificial Intelligence*, volume 20, page 121. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[20] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[21] A. Rényi P. Erdős. On random graphs. i. *Publicationes Mathematicae*, 1959.

[22] Thong Pham, Paul Sheridan, and Hidetoshi Shimodaira. Pafit: A statistical method for measuring preferential attachment in temporal complex networks. *PLOS ONE*, 10(9), 9 2015.

[23] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge University Press, New York, NY, USA, 3 edition, 2007.

[24] Thong Pham and Paul Sheridan and Hidetoshi Shimodaira. *PAFit: Nonparametric Estimation of Preferential Attachment and Node Fitness in Temporal Complex Networks*, 2015. R package version 0.7.5.