5-2016

Statistical Modeling of the Temporal Dynamics in a Large Scale-Citation Network

Luis Javier Ek Jr.

University of Arkansas, Fayetteville

Follow this and additional works at: http://scholarworks.uark.edu/etd

Part of the Applied Statistics Commons

Recommended Citation

Ek, Luis Javier Jr., "Statistical Modeling of the Temporal Dynamics in a Large Scale-Citation Network" (2016). Theses and Dissertations. 1549.
http://scholarworks.uark.edu/etd/1549

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, ccmiddle@uark.edu.
Statistical Modeling of the Temporal Dynamics in a Large-Scale Citation Network

A thesis is submitted in partial fulfillment of the requirements for the degree of Masters of Science in Statistics and Analytics

by

Luis Javier Ek Jr
University of Belize
Bachelor of Science in Mathematics, 2013

May 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Qingyang Zhang
Thesis Director

Dr. Giovanni Petris
Committee Member

Dr. Mark Arnold
Committee Member

Dr. Avishek Chakraborty
Committee Member
Abstract

Citation Networks of papers are vast networks that grow over time. The manner or the form a citation network grows is not entirely a random process, but a preferential attachment relationship; highly cited papers are more likely to be cited by newly published papers. The result is a network whose degree distribution follows a power law. This growth of citation network of papers will be modeled with a negative binomial regression coupled with logistic growth and/or Cauchy distribution curve. Then a Barabási Albert model based on the negative binomial models, and a combination of the Dirichlet distribution and multinomial will be utilized to simulate a network that follows preferential attachments between newly added nodes and existing nodes.
Acknowledgements

I would like to thank everyone at University of Arkansas for being so helpful throughout the three years of my master studies. Thanks to all the faculty and staff for enabling my education goals and providing the opportunities for me to be successful. A special thanks to Dr. Qingyang Zhang for guiding me through the process of this thesis. Final statement of special gratitude goes to my family back in Belize, for being supportive and caring.
# Table of Contents

1 Chapter 1  
1.1 Graph theory ........................................... 2  
1.2 Citation networks of papers ................................ 4  
1.3 Our citation network ...................................... 5  

2 Chapter 2  
2.1 Logistic growth curve ...................................... 7  
2.2 Cauchy cumulative distribution ............................ 8  
2.3 Negative binomial regression ............................... 9  
2.3.1 Sampling from a NB .................................. 11  

3 Chapter 3  
3.1 Preferential attachment/Barabási Albert (BA) model ... 12  
3.2 Multinominal distribution .................................. 13  
3.3 Dirichlet distribution ....................................... 13  
3.3.1 Formula .............................................. 14  
3.3.2 Dirichlet distribution as a conjugate prior to multinominal distribution 14  

4 Chapter 4  
4.1 Modeling paper publishing ................................. 16  
4.1.1 Paper publishing on weekdays and weekends ......... 18  
4.1.2 Models for weekends and weekdays paper publishing ... 20  
4.2 Modeling paper citations/outdegree ...................... 23  

5 Chapter 5  
5.1 Model adequacy ........................................... 29  
5.1.1 Models for paper publishing ........................... 29  
5.1.2 Models for the mean number of citations of papers .... 30
6 Chapter 6

6.1 Simulation of future papers with the negative binomial model with logistic growth curve and cauchy distribution curve ........................................... 31

6.2 Simulation of future paper’s citations with the negative binomial models with logistic growth curve ................................................................. 33

6.3 Network Simulation for First two years based on proposed BA model .......................... 34

6.3.1 Paper simulation ................................................................. 34

6.3.2 Network Simulation .............................................................. 37

6.3.3 Simulation results and remarks ................................................ 40

7 Conclusion ................................................................................. 42

References ..................................................................................... 43
Introduction

A citation network of papers describes the number of citations that occur within a data set of papers. Thanks to the fact that papers are stored on the web server, people can easily create algorithms to construct such networks. The size of this type of networks is quite large due to a high number of publications in the past years. Furthermore, the size of such network keeps growing with the passing of time, since papers are still being published. The growth of such network is what intrigues us. From past literature Barabási (2016), and Xie et al. (2015), we know that these networks are not just the result of randomness, but that of “preferential attachment”. In this paper, we are going to model the mean number of papers and their respective number of citations published from a large citation network obtained from Stanford Website of large networks. This will be done with negative binomial regression coupled with either a logistic growth curve or a Cauchy distribution curve. Then the Dirichlet distribution and multinomial distribution will be used to simulate a preferential attachment model.

The break down of the chapters are as follows. Chapter 1 will discuss in depth what a network is and also provide some theory on networks. It will end by describing the network we are going to use. In Chapter 2 the models to model the number of published papers and their respective citations of the large citation network of papers will be stated and discussed. Chapter 3 will describe the Barabási-Albert model, along with two distribution that will help in simulating a network. Chapter 4 will provide the estimates of the models discussed in chapter 2. In chapter 5 we will assess the adequacy of the models, and in chapter 6 we are going to simulate our network based on our proposed Barabási-Albert model with Dirichlet and multinomial distribution.
Chapter 1

1.1 Graph theory

In this section of chapter one we will look at basic graph theory to familiarize the readers with it. We will start by stating the mathematical definition of the term graph. According to Kolaczyk and Csárdi (2014) a graph is defined by:

**Definition 1.1.** A graph $G = (V, E)$ is a mathematical structure consisting of a set $V$ of nodes (also commonly called vertexes) and a set $E$ of edges (also commonly called links), where elements of $E$ are unordered pairs $\{u, v\}$ of distinct nodes, $u, v \in V$.

Another important definition from Kolaczyk and Csárdi (2014) that will be useful in our paper is the following:

**Definition 1.2.** Order: the number of nodes $N_v = |V|$ and Size: the number of edges $N_e = |E|$.

Since this paper deals with a directed network; hence, the above definition will need to change from unordered pairs to ordered pairs. This means that $(u, v) \neq (v, u)$. And for $(u, v)$ it means that $u$ is linked to $v$ from left to right. Now with the above, we can create a trivial (toy) graph for illustration purposes.

$$G = (V, E) \tag{1.1}$$

where $V = \{1, 2, 3, 4, 5, 6, 7\}$, and $E = \{(3, 1), (4, 1), (6, 1)\}$. Here we notice that the order and size of our graph is 7 and 3, respectively. The following diagram visualizes our graph 1.1.

This diagram, will help in understanding two more properties of a graph that will be useful in this paper. These two terms are in-degree and out-degree of a node. Kolaczyk and
Csárdi (2014) defines these two terms as:

**Definition 1.3.** The in-degree ($d_{v}^{in}$) and out-degree ($d_{v}^{out}$) of a node $v$ in a graph is defined as the number of edges that are pointing in towards and out from a node $v$, respectively.

Thus, from Figure 1.1 we can state some of its properties.

\[
N_v = |V| \quad \quad \quad \quad N_e = |E| \\
= 7 \quad \quad \quad \quad \quad \quad 3
\]

The order of our graph is 7 and its size is 3.
\[ d_{vi}^{in} = \begin{cases} 
3 & \text{if } i=1 \\
0 & \text{otherwise} 
\end{cases} \]

\[ d_{vi}^{out} = \begin{cases} 
1 & \text{if } i=3,4,6 \\
0 & \text{otherwise} 
\end{cases} \]

Notice that computing these properties takes a couple of seconds, because the order of the graph is quite small. If the order and size of our graph increases, we will take much time counting, thus, we resort to software. There are many other properties of graphs, yet these are sufficient for our paper. The following section will introduce citation networks.

1.2 Citation networks of papers

In this section we are going to extend our notion of graphs. Also we will be referring to graphs and networks as being the same thing. Now, from the previous examples, we limited ourselves to a network at a specific time. Such networks are considered static networks, since the network is not in flux; the number of nodes and edges are constant. However, as soon as we introduce the concept of time within our system, then our network becomes a dynamic network. This means that as time passes by, nodes are introduced and edges might be formed or not, (Rajaraman, 2006). A citation network of papers fits the above description. The Xie et al. (2015) definition for citation network is as follows.

**Definition 1.4.** Citation network of papers is a directed graph, which describes the inter-citations between the papers. The network regards papers as nodes and contains a directed edge from paper \( i \) to paper \( j \), if \( i \) cites \( j \).

Thus, further properties of citation network of papers (but not limited to) are provided by Kas (2011).
1. Citation networks are directed graphs. The edges of this graph are defined by citations between two papers.

2. The network is acyclic because a paper can only cite previous published papers.  

3. Nodes and edges added to the network are permanent and are not removed as time goes by.

4. Indegree is characterized as the number of citations a paper receives and outdegree stands for the number of citations a paper has.

Another important aspect we want to understand about citation networks, is the evolution of the network over time. These networks are not just random networks that create edges at random (Barabási, 2016, p.6). As Barabási (2016) points out the nodes in a citation network are biased, because “in real networks, new nodes tend to form an edge with the more connected nodes”. This is referred to as preferential attachment and will be explained in chapter 3 when we discuss how to generate the edges for our simulated network.

1.3 Our citation network

With this background we can now describe the citation network of papers we will be working with. As mentioned in the introduction, our citation network was obtained from Stanford Large Network Dataset Collection website. The name of the citation network is Arxiv HEP-PH (high energy physics phenomenology) which comes from the e-print arXiv (Leskovec and Krevl, 2014). It consist of citations between papers within the datasets from January 1993 to April 2003. The order and size of this network are given as follows:

1\textsuperscript{1}directed acyclic network are also referred to bayesian networks. These network link nodes (random variables) based on causation relationship. Hence a citation network can be regarded as a bayesian network.
\[ N_v = |V| \quad \text{and} \quad N_e = |E| \]
\[ = 34,546 \quad \text{and} \quad 421,578 \]

Clearly the size and order of this network indicates a large graph. In a manner of a decade 34,546 papers were published and within them 421,578 citations occurred. The models to model the growth of the papers published during this decade along with the growth of the number of citation for every published paper will be addressed in the following chapter.
2 Chapter 2

In this chapter we shall discuss how to model the numbers of papers being published and the mean number of citations of each published papers.

2.1 Logistic growth curve

The logistic growth curve is a simple model for population (not necessarily people) growth that reaches an horizontal asymptote. The logistic growth curve is given as follows.

\[
u(t) = \frac{\alpha_1}{1 + \exp(\alpha_2 + \alpha_3 t)} \tag{2.1}\]

Where:

1. \(u(t)\) represents the number of objects (in our case papers or nodes) published at time \(t\).
2. \(\alpha_1\) is the asymptote which the population (the numbers of papers being published) tends to as \(t\) goes to infinity.
3. \(\alpha_2\) stands for the number of papers at time \(t = 0\) relative to the asymptotic size, \(\alpha_1\).
4. \(\alpha_3\) controls the growth rate of our populations (in this case the number of papers).

The \(\alpha\)'s are going to be estimated by using least square method.

The logistic growth curve will be used to model the trend for the number of papers published along with their citations during 1992-2000. This trend will be then used as a covariate for a negative binomial regression. This will be shown in chapter 4.
2.2 Cauchy cumulative distribution

The Cauchy cumulative distribution looks like the logistic growth curve, so it will serve the same purpose as logistic growth curve. But first it needs to be modified. The standard Cauchy distribution function has no scalar parameter, since its between \([-1, 1]\), now if we want to increase this range, we need to introduce a scalar component \(c\). Thus the modified Cauchy cumulative distribution to be used as is given as follows:

\[
c(t) = c \left( \frac{1}{2} + \frac{1}{\pi} \arctan \left( \frac{t - a}{b} \right) \right)
\]  

\( (2.2) \)

Where:

1. \(a\) is the shape parameter
2. \(b\) is the scale parameter
3. \(c\) is a scalar parameter.

Both \(a, b\) and \(c\) are going to be estimated by using least square method.

The Cauchy cumulative distribution function (CDF) will be used to model the trend for the number of papers published during 1992-2000. Then this modified “Cauchy” trend will be used as a covariate for a negative binomial regression. The results produced will be compared to that of the logistic growth curve and the superior one will be chosen. This will be shown in chapter 4.
2.3 Negative binomial regression

In modeling counts, the Poisson distribution is the one mostly used (Hilbe, 2011, p.140). This model implicitly assumes that the sample average and sample variance are close, which can be easily violated in many real problems. Hilbe (2011) says “Overdispersion is, in fact, the norm” (p. 140). Due to this, the most commonly used model for counts of this nature is the negative binomial regression (NB). There are many forms on how to derive this model (p.5), however we will stick to the probabilistic nature of the model.

First we start by stating the definition of the negative binomial distribution.

**Definition 2.1.** Negative binomial distribution counts the number of success in a sequence of i.i.d trials, with success rate being p, before a specific number of failures occurs,r.

If $Y \sim NB(p, r)$ then the PMF of $Y$ is defined as follows

$$f(y|p, r) = \binom{y+r-1}{y} (1-p)^r p^y$$

(2.3)

Where:

$$y = 0, 1, 2, 3, 4, ....$$

The expected and variance of $Y$ is given by the following

$$E[Y] = \mu = \frac{pr}{1-p}$$

(2.4)

$$Var[Y] = \mu + D\mu^2 = \mu + \frac{1}{r}\mu^2$$

(2.5)
Here D is referred to as the dispersion parameter. Note that as D approaches zero, Y starts to follow a Poisson distribution; hence the Negative Binomial is a generalization of Poisson. Furthermore, an NB can be represented by a Poisson model with the mean $\lambda$ following gamma distribution.

Now if we have observations $Y$ (count numbers) which if we assume to follow a negative binomial distribution, then we can model it with a categorical regression model known as the negative binomial regression (NB). The general NB regression model known in literature is as follows:

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$  \hspace{1cm} (2.6)

where Agresti (2007, p.66) states that:

1. $\log(\mu)$ is known as a link function. This specifies a relationship between the mean of our random component Y to its predictors x.

2. $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ is an affine linear combination of k predictors as in regular linear regression.

With the above regression model the number of paper published and mean number of citations of each published papers will be modelled. This will be accomplished by using the a trend, either 2.1 or 2.2 as its covariate as shown below:

$$\log(\mu) = \alpha + \beta u(t)$$  \hspace{1cm} (2.7)

$$\log(\mu) = \alpha + \beta c(t)$$  \hspace{1cm} (2.8)

where the coefficients of equation 2.7 and 2.8 are going to be estimated using generalized linear regression software.
2.3.1 Sampling from a NB

\[ \log(\mu) = \alpha + \beta \times TREN D \]  \hspace{1cm} (2.9)

\[ \mu = \exp(\alpha + \beta \times TREN D) \] \hspace{1cm} (2.10)

Once equation 2.7 or 2.8 has been estimated, it will be used to estimate the mean count value for a response random variable \( Y \sim NB(r, p) \), as shown in equation 2.10. As equation 2.4 indicates, the values of \( r \) and \( p \) can be expressed in terms of \( D \) and \( \mu \), respectively (shown below). Once these values are present, random numbers can be simulated for every possible value of \( \mu \) and \( D \) (\( D \) remains constant) at time \( t \).

\[ r = \frac{1}{D} \] \hspace{1cm} (2.11)

\[ p = \frac{\mu D}{1 + \mu D} \] \hspace{1cm} (2.12)
Chapter 2 provided the means to model the number of papers along with its citations. Yet we need three additional concepts in order to meet our goals.

3.1 Preferential attachment/Barabási Albert (BA) model

Barabási (2016) states that a BA model creates a network with respect to time by adding a new nodes at time $t$ and then linking this node to existing nodes based on the nodes’ degree (in our case, indegree/or number of citations). Hence, the nodes with higher in-degree are more likely to be linked to this new node. This is exactly what happens in citation network of papers; highly cited papers are more likely to be cited by new published papers. A formal definition from Barabási (2016, p.8) is as follows:

**Definition 3.1.** BA model: the network starts with $m_0$ nodes, the links between which are chosen arbitrarily as long as each node has at least one link. The network develops following two steps:

1. Growth: At each timestep a new node(s) is added with $m$ ($\leq m_0$) links that connect the new node to $m$ nodes already in the network.

2. Preferential Attachment: The probability $p_i$ that a link of the new node connects to node $i$ depends on the degree (or in-degree $k_i$)

   \[ p_i = \frac{k_i}{\sum_j k_j} \]  

From Srinivasan (2013, p.2), a BA model will have the following property:
Property. The degree (in or out) distribution resulting from a BA model will follow a power law of the form \( d(k) \sim k^{-\beta} \), where \( d(k) \) is the number of nodes of degree \( k \). If the graph is undirected or directed, then \( \beta \) will have a value of 3 or 2, respectively.

This idea of preferential attachment will be used to create a dynamic network. However, we will not use Barabási preferential attachment probability function 3.1. Instead, we are going to use the Dirichlet distribution to sample probabilities and then generate edges with the multinomial distribution. These two distributions are discussed as follows.

3.2 Multinominal distribution

A multinomial distribution is an extension to the binomial distribution where each trial has more than 2 possible outcomes. This distribution occurs whenever we have \( n \) independent objects and we want to place them into \( k \) categories. If \( Y = (y_1, y_2, \cdots, y_k) \), a vector of size \( k \), follows a multinomial distribution, then its PDF is defined as:

\[
f(y_1 = n_1, y_2 = n_2, \cdots, y_k = n_k) = \frac{n!}{n_1!n_2! \cdots n_k!} \prod_{i=1}^{k} \pi_i^{n_i}
\]  

(3.2)

Note, that the sum of all \( y_i \)'s is \( n \), and the sum of all \( \pi_i \)'s is 1.

This distribution will be helpful in placing links between a new published paper and all previous published papers. However we need to estimate the values of \( \pi_i \). To do this, we will need the following distribution.

3.3 Dirichlet distribution

The Dirichlet distribution is usually described “as a probability distribution of PMFs” (Hortensius, 2012, p.1). Hortensius (2012) visualization of this distribution is to:

“consider a bag of 6 sided dice - each of them being unfair in different ways - each die will have its own PMF (it has a different probability distribution for the sides) which can be slightly different from the other dice. By drawing a die out from the bag you draw a PMF”.
Hence a Dirichlet distribution models the randomness of PMFs.

### 3.3.1 Formula

The Dirichlet distribution has only one parameter $\alpha$ which is a vector of positive real numbers. If $Y \sim \text{Dir}(\alpha)$, then its PDF is defined as follows:

$$f(x_1, x_2, \cdots, x_K|\alpha_1, \alpha_2, \cdots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i-1}$$  \hspace{1cm} (3.3)

Note that:

1. $\sum_{i=1}^{K} x_i = 1$, for every $x_i \in \mathbb{R}^+$

2. Area under $f$ gives the probability of vector $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$ In other words, the chance of obtaining a PMF

3. The dimension of the PDF is a hyperplane of $k-1$ dimensions in $K$-dimensional space because of 1 above.

### 3.3.2 Dirichlet distribution as a conjugate prior to multinomial distribution

From above, we know that a multinomial distribution describes the number of ways of placing $n$ objects into $k$ categories. Now if we denote the probabilities of each category as $p_i$, we sample $p_i$ from a $\text{Dir}_k(\alpha)$, since Dirichlet distribution is known to be conjugate prior of the multinomial distribution. The following shows this.
Proof. Let $X \sim \text{Multinomial}_k(n,p), p \sim \text{Dir}_k(\alpha)$ What is the density of $p$ given $X$?

By definition from bayesian theory

$$f(\theta|x) \propto f(x|\theta)f(\theta) \quad (3.4)$$

Hence,

$$f(p|x) = C \left( \prod_{i=1}^{k} p_i^{x_i} \right) \left( \prod_{i=1}^{k} p_i^{\alpha_i-1} \right) \quad (3.5)$$

$$= C \prod_{i=1}^{k} p_i^{\alpha_i+x_i-1} \quad (3.6)$$

$$= \text{Dir}_k(\alpha + \alpha) \quad (3.7)$$

Hence, the multinomial distribution with the Dirichlet distribution acting as its prior for all $\pi_i$'s, will be used to generate the links between a newly created paper and all previous papers. This will be shown in detail in chapter 6.
4 Chapter 4

In this chapter, we will use the theory from all previous chapters and model our citation network of papers. The end result of our modeling will be a simulated network that follows a BA model as the original model. The process will be broken into two parts. The first part will deal on how to model the number of papers being produced in a particular day from January 1992 to April 2002. The second Part will be to model the mean number of citations for papers published per day during this decade. Then in chapter 5 we will use part 1 and 2, along with the distributions in chapter 3 to generate a dynamic BA network.

4.1 Modeling paper publishing

In this section we will start by analyzing some basic characteristics of our network. We know from chapter 1 that the size and order of our network is 34,546 and 421,578, respectively. Next we look at the distribution of the papers being published throughout this decade rather than the network itself, which will look very cluttered.

![Distribution of Published papers from 1993-2003](image)

(a) Histogram

![Scatterplot](image)

(b) Scatterplot

Figure 4.1: Distribution of papers published per day in 1992-2002, and scatterplot of papers published per day from 1992-2002
Figure 4.1a shows us that the number of papers being published per day is rightly skewed. Throughout this decade, the number of papers published most often per day are between 1 and 5 papers, then its followed by 10-15 papers and so forth. The max number of papers published is 55 papers at day 3675 not shown in the histogram but shown in the scatter plot in Figure 4.1b. Figure 4.1b shows the daily number of published papers. In this plot we can see that there are two trends separated by a gap at the interval \(5 < y < 12\). The trend that is above \(y = 12\) is more cluttered and grows faster than the trend below \(y = 12\). Further more if we construct correlogram we obtain the following:

![Correlogram](image)

Figure 4.2: Correlogram for No. of papers published per day from 1992-2002

Clearly from the figure 4.2 there is a strong periodicity of 7, meaning there is high correlation between the number of papers published during this time. Due to this, we will break down our days into weekends and weekdays and model them separately.
4.1.1 Paper publishing on weekdays and weekends

Figure 4.3: No. of papers published on the weekdays of 1992-2002

Clearly, figure 4.3 shows a visible trend. The number of papers published on the weekdays grows gradually as time goes by. This trend will be modeled with both the logistic growth curve and the Cauchy distribution curve. Upon getting these trend models, we are going to use them as a covariate for a NB regression to model the number of papers published during the weekends. Then we are going to compare both NB regressions is and choose the one that fits best.
Here we notice that the number of papers published during weekends of 1992-2002 are lower than that of the weekdays. The max being around 13 papers. The trend here grows slower and is not steep. The same approach as the one used for modeling the trend for the number of papers during the weekdays will be used here.
4.1.2 Models for weekends and weekdays paper publishing

Here we provide the models that were obtained for modeling paper publishing during weekdays and weekends. The blue line and red line stands for the trend and NB model respectively. Both models fit the data appropriately, since the p-values of the coefficients in both regression models are significant ($\alpha = 0.05$), as shown in the section “Model Adequacy”

Weekdays

Figure 4.5: Models for the No. of papers published on weekends of 1992-2002

Figure 4.5 shows two models for the number of papers published per day from 1992-2002. The one to the left shows a negative binomial model with logistic growth curve and the right one shows a negative binomial model with a modified Cauchy distribution curve. The red lines in both plots is what interest us. Both of these two lines have the same pattern, yet as time goes to 3500 the NB model obtained with the Cauchy distribution curve seems to have a bit larger rate of increase than the NB model with logistic growth curve.
Nevertheless, based on the current data, there is not a significant difference on the two NB models; thus, both can be used to estimate the mean number of papers published during the weekends. The following are the models stated explicitly

Logistic growth trend model

\[
\begin{align*}
\alpha_1 & + \beta u_1(t) \\
\mu_1 & = 1.30541 + 0.09011u_1(t) \\
\end{align*}
\]

Negative binomial model with logistic growth trend

\[
\begin{align*}
\log(\mu_1) & = \alpha_1 + \beta_1 u_1(t) \\
\mu_1 & = 1.30541 + 0.09011u_1(t) \\
\end{align*}
\]

Cauchy distribution trend model

\[
\begin{align*}
\alpha_1 \left( \frac{1}{2} + \frac{1}{\pi} \arctan \left( \frac{t - \alpha_2}{\alpha_3} \right) \right) \\
\mu_1 & = 20.52 \left( \frac{1}{2} + \frac{1}{\pi} \arctan \left( \frac{t - 840.24}{982.40} \right) \right) \\
\end{align*}
\]

Negative binomial model with Cauchy distribution trend

\[
\begin{align*}
\log(\mu_1) & = \alpha_1 + \beta_1 c_1(t) \\
\mu_1 & = 1.28174 + 0.09171c_1(t) \\
\end{align*}
\]
Figure 4.6: Models for the No. of papers published on weekends of 1992-2002

Figure 4.6 shows two NB models just like figure 4.5. Here the two NB models (shown by the red lines) are very much alike. There is not a big difference between them. Hence either models can be used. The models are stated explicitly below.

Logistic growth curve trend model

\[
u_2(t) = \frac{\alpha_1}{1 + \exp(\alpha_2 \alpha_3 t)}
= \frac{3.787383}{1 + \exp(1.846790 - 0.002017 t)}
\] (4.9)

Negative binomial with logistic growth trend

\[log(\mu_2) = \alpha_2 + \beta_2 u_2(t)\] (4.11)
\[= -0.3901 + 0.4691 c_2(t)\] (4.12)
Cauchy distribution trend model

\[ c_2(t) = \alpha_1 \left( \frac{1}{2} + \frac{1}{\pi} \arctan \left( \frac{t - \alpha_2}{\alpha_3} \right) \right) \]  

(4.13)

\[ = 4.109 \left( \frac{1}{2} + \frac{1}{\pi} \arctan \left( \frac{t - 603.778}{1015.587} \right) \right) \]  

(4.14)

Negative binomial model with Cauchy distribution trend

\[ \log(\mu_1) = \alpha_1 + \beta_1 u_1(t) \]  

(4.15)

\[ = -0.3946 + 0.4705c_2(t) \]  

(4.16)

4.2 Modeling paper citations/outdegree

Now that we have means to estimate the mean number of papers published at a given time, the next objective is to estimate the mean number of citations made by each published paper. We start this section by looking at the nature of citations made by each paper published during 1992-2002. In doing so we will see the distribution of citations for the entire decade and growth of citations over time.
Figure 4.7: No. of Citations of papers published from 1992-2002

Figure 4.7, indicates that the number of citations (Outdegree) of papers published during 1992-2002 is skewed. We can also observe that papers are citing around 0-20 papers and 20-40 papers more often. However, the x axis goes to 400, which means that there are papers citing around 400 papers.
Figure 4.8: Distribution of citations/outdegree of papers published from 1992-2002.

Figure 4.8, shows, the growth of the outdegree of papers in our network from 1992-2002. We can clearly see that on the onset of our network, the mean outdegree of papers is not more than 5. Fast forward 5 years, we have an max outdegree bigger than 100. After 5 more years, we see that the max outdegree of papers has reach around 300. The following diagram shows the same data as figure 4.8, just that it is all combined in one plot and the x axis is reduced for a more focused view.
Figure 4.9: Distribution of citations/outdegree of papers published from 1992-2002.

So far we have seen the distribution of outdegree for our network. We noticed that it is a very skewed distribution. Next we are going to look at how the mean outdegree of the network grows throughout time.
Figure 4.10: No. of papers published on weekends of 1992-2002

Figure 4.10 shows us that the mean outdegree of the network grows gradually with increasing variance. To model the data, the same method as for modeling paper publishing was used. A trend was estimated with the logistic growth curve, where the independent variable $u_3(t)$ is the trend for mean outdegree and $t$ is time. Next this trend was used as a covariate for another NB model. The respective estimates are as follows.
Trend

\[ u_3(t) = \frac{\alpha_1}{1 + \exp(\alpha_2 + \alpha_3 t)} \quad (4.17) \]

\[ = \frac{17.160894}{1 + \exp(2.587327 - 0.001634 t)} \quad (4.18) \]

Our NB model

\[ \log(\mu_3) = \alpha_3 + \beta_3 u_3(t) \quad (4.19) \]

\[ = 0.6971 + 0.1386 u_3(t) \quad (4.20) \]

Note that only a logistic growth curve was used here since the scatterplot shows that the mean number of citations in our network grows very slowly.
5 Chapter 5

5.1 Model adequacy

In this chapter we will assess the adequacy of the models presented in Chapter 4. The following two pages shows the summary of each model. We are going to set \( \alpha = 0.05 \). The estimates for the parameters of each calculated trend and each calculated NB model have significant p-values; thus making the models adequate.

5.1.1 Models for paper publishing

The estimates for the models used to model the number of papers published during the weekdays are given as follows:

**Negative binomial model with logistic growth curve**

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_1 )</td>
<td>1.884e+01</td>
<td>2.432e-01</td>
<td>77.47</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>( \alpha_2 )</td>
<td>1.302e+00</td>
<td>6.033e-02</td>
<td>21.58</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>( \alpha_3 )</td>
<td>-1.505e-03</td>
<td>7.307e-05</td>
<td>-20.59</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 1: Trend Estimates for \( u_1(t) \)

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>1.305408</td>
<td>0.026088</td>
<td>50.05</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>trend</td>
<td>0.090106</td>
<td>0.001704</td>
<td>52.89</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 2: NB estimates for \( \log(\mu_1) \)

**Negative binomial model with cauchy distribution curve**

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_1 )</td>
<td>20.5240</td>
<td>0.2904</td>
<td>70.66</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>( \alpha_2 )</td>
<td>840.2387</td>
<td>49.4572</td>
<td>16.99</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>( \alpha_3 )</td>
<td>982.3999</td>
<td>33.5735</td>
<td>29.26</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 3: Trend Estimates for \( c_1(t) \)

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>1.281736</td>
<td>0.026712</td>
<td>47.98</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>trend</td>
<td>0.091710</td>
<td>0.001748</td>
<td>52.47</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 4: NB estimates for \( \log(\mu_1) \)
The estimates for the models used to model the number of papers published during the weekends are given as follows:

**Negative binomial model with logistic growth curve**

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>3.7873832</td>
<td>0.1141404</td>
<td>33.182</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>1.8467899</td>
<td>0.2390178</td>
<td>7.727</td>
<td>&lt;2.58e-14</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-0.0020166</td>
<td>0.0002655</td>
<td>-7.595</td>
<td>&lt;6.83e-14</td>
</tr>
</tbody>
</table>

Table 5: Trend Estimates for $u_2(t)$

**Negative binomial model with cauchy distribution curve**

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>4.1089</td>
<td>0.1438</td>
<td>28.577</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>603.7782</td>
<td>102.7079</td>
<td>5.879</td>
<td>5.56e-09</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>1015.5869</td>
<td>74.5296</td>
<td>13.627</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 7: Trend Estimates for $c_2(t)$

**5.1.2 Models for the mean number of citations of papers**

The estimates for the models used to model the number of papers published during the weekends are given as follows

<table>
<thead>
<tr>
<th>Value</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>T</th>
<th>PVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>1.716e+01</td>
<td>3.679e-01</td>
<td>46.64</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>2.587e+00</td>
<td>1.674e-01</td>
<td>24.09</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-1.634e-03</td>
<td>8.746e-05</td>
<td>-18.68</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Table 9: Trend Estimates for $u_2(t)$
6 Chapter 6

In this chapter we are going to use the models that were established in chapter 4 and 5 for simulation.

6.1 Simulation of future papers with the negative binomial model with logistic growth curve and Cauchy distribution curve

The above diagrams show the simulated number of papers published during weekends and weekdays from 2002 to 2010. These values were obtained by sampling from their corresponding NB models (4.3, 4.11 respectively) as shown in section 3.2.1. Both diagrams, show that the number of papers simulated after 2002, grow extremely slow; they seem to reach an asymptote. The NB model flattens as soon as it reaches 2002. This tells us that our model is not following the trend of the original data, and this not good. Let’s compare
these results to the results obtain with a NB model with a Cauchy distribution trend.

Figure 6.2: Simulation of no. of papers published ten years after 2002 using NB model with Cauchy distribution curve

The above plot, clearly shows that a negative binomial model with Cauchy distribution curve (both 4.7 and 4.15) are superior than a negative binomial models with logistic growth curve. The NB simulation with Cauchy distribution, shown as a red line, does not immediately flattens after 2002 as its predecessor. Instead it keeps growing steadily. Hence, we will utilize the negative binomial model with Cauchy distribution to model the number of papers published in our network.
6.2 Simulation of future paper’s citations with the negative binomial models with logistic growth curve

The negative binomial model with logistic growth curve does a good job in modeling the mean number of citations for each paper in our network. This NB model, shown by the red line, does not flatten after 2002, instead it keeps growing slowly, which is a good indicator that our model is appropriate.
6.3 Network Simulation for First two years based on proposed BA model

Under this section, a network which follows preferential attachment will be simulated. The simulated network will have a life time of two years. Two parts will be required to build such network. In the first part, NB models 4.7, 4.15, and 4.19 will be used as suggested by section 3.2.1. With NB models 4.7 and 4.15, and their corresponding Dispersion parameters we can simulate number of papers “published” by some system. Now for each single paper we will use NB model 4.19 and its dispersion parameter to simulate its number of citations/outdegree. In part two the simulated data from part 1 will be used to simulate a network by employing a BA model built on the Dirichlet and multinomial distribution.

6.3.1 Paper simulation

![Simulation for no. paper publishing during the first 2 years](image)

(a) Weekends  
(b) Weekends

Figure 6.4: Simulation for no. paper publishing during the first 2 years

Figure 6.4 (a) and (b) indicate that the simulation of papers published from 1992-1994 is very close to the real data set. The simulation represented by the green lines follows closely to the real data set, represented by the black lines. These green lines were obtained by
using the NB models 4.7, 4.15. The following explains this. Models 4.7 and 4.15 were used to obtain the mean number of papers $\mu_t$ published at time $t$. This is represented by the red lines in figure 6.4. Now each 4.7 and 4.15 NB model has its own dispersion parameter (refer to tables in chapter 5). Given this, random numbers were generated by sampling from a negative binomial distribution with parameters $(\mu_t, D)$ as explained in chapter 3 under the section “Sampling from a negative binomial distribution”.

![Figure 6.5: Simulation for mean no. of citations of papers published for first 2 years](image)

Figure 6.5 shows that simulated mean number of outdegree for papers simulated during 1992-1994. These results were obtained by using 4.19 in the same manner as the previous plots. Note that the green lines starts different from the real data set (black line), but then comes close as time goes by. The difference here is that our NB model 4.19 starts creating outdegrees for the papers simulated during the first couple of days, whereas in the real data set the occurrence of citation/outdegree is till around 100 days.

When we combine the simulated data from the figures 6.4 and 6.5 the following table is
obtained. This table shows only 15 papers out of the 3650 papers simulated in 2 years (730 days)

<table>
<thead>
<tr>
<th>Row</th>
<th>Time</th>
<th>Paper.ID</th>
<th>OD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 11: Part simulated data

Note that at the beginning of our network $t = 1$, the outdegree of the papers are at 0. Then as $t$ increase the number of papers start having outdegree $\geq 0$. 

36
6.3.2 Network Simulation

Now with simulated results from the above table, a network will be simulated using a BA model, based on the Dirichlet distribution and the multinomial distribution. The process of the proposed BA model is outlined in the following steps.

1. At a given $t = i - 1$ there exist $N_{i-1}$ papers in our network. Denote $D^{in}$ as a vector for the indegrees of all $N_{i-1}$ papers. Note that at time $t=0$, this vector is a zero vector.

2. Use $D^{in}$ as $\alpha$ of a Dirichlet distribution to sample a PMF $P$, which will be a vector of size $N_{i-1}$. However, $Dir(\alpha)$ where $\alpha$ is a zero vector is undefined, recall 3.3. Hence the starting values for vector $\alpha$ will all be 10. This value was chosen, because all the values of $\alpha$ are clustered in the center of the $k-1$ hyperplane, Frigyik B., Kapila A., and Gupta M (2010, p.4). Now $\alpha$ will change over time as the indegree of the nodes will grow, hence the $k-1$ hyperplane will move around but still remain concentrated (Frigyik et al., 2010, p.4)).

3. After step two has been completed, vector $P$ will represent the probability for each node at time $t = i - 1$ to be cited. In other words, the chance of being selected by a future paper, i.e creating a link (edge) between a paper simulated at time $t = i$ and time $t = i - 1$. To randomly generate a link(s) the following is performed:

   (i) Refer to the number of outdegree (number of citations) of each paper at time $t = i$ as $N_{od}$.

   (ii) If $N_{od}$ is 0, then jump to the following paper. If $N_{od}$ is 1, then generate a link with a multinomial distribution with parameter $P$. If $N_{od}$ is $\geq 2$, then $N_{od}$ links must be created with the multinomial distribution with parameter $P^*$. $P^*$ starts as $P$ but then it is updated whenever a $m^{th}$ paper is chosen from the $N_{i-1}$. The manner on how $P^*$ is updated is to use a Dirichlet $Dir(\alpha^*)$ where $\alpha^*$ is the vector of indegree of the $N_{i-1}$ papers after $m^{th}$ paper is removed.
4. After step three has been completed, update $D^{in} = \alpha$ as time goes by. Keep all nodes with indegree 0 as 10. Repeat the above procedure

Now with the above process, we obtain the following network.
Figure 6.6: Growth of simulated network in 16 days

(a) Day 2

(b) Day 4

(c) Day 6

(d) Day 8

(e) Day 10

(f) Day 12

(g) Day 14

(h) Day 16
6.3.3 Simulation results and remarks

The network simulated based on our proposed BA model for two years is shown in figure 6.6. Note that as time goes by, highly cited papers ("hot papers") represented by larger nodes start to appear. Hence, it appears that as time goes by the newly published papers citing published papers based on preferential attachment. Next, we will compare the distribution of the indegree of the simulated data against the real data.

![Histogram of Indegree](image1.png) ![log(indegree) vs log(freq)](image2.png)

(a) Histogram of Indegree (b) log(indegree) vs log(freq)

Figure 6.7: Simulated Network indegree distribution from proposed BA model

The figure 6.7 shows in-degree distribution for the simulated network model. Clearly we see from the histogram that the distribution appears to follow a power law function. This is corroborated on figure 6.7 (b) where the log of the frequencies and log of the in-degree were plotted, and a simple linear regression was run to determine the line of best fit. If indeed the simulated network follows preferential attachment then the scatterplot of \( \log(d^{in}) \) versus \( \log(freq) \), should follow a linear trend with slope approximately to 2. The following calculations provide a mathematical approach to the why this is true.
**Calculation.** Assuming the histogram above follow a power law, then from Chapter 3 under BA model

\[
d^\text{in}(k) = \alpha k^{-\beta}
\]  

(6.1)

where \(d^\text{in}(k)\) is in the frequency for the indegree of size \(k\), \(\alpha\) and \(\beta\) are constants that characterize the shape and scale of the power function.

Now if we take log on both sides, we express the product of our function as sums.

\[
\log(d^\text{in}(k)) = \log(\alpha) - \beta \log(k)
\]  

(6.2)

\[
= mk + c
\]  

(6.3)

This simple linear function can be estimated from simple linear regression. The equation of the line is shown in figure 6.7 (b), with slope \(\hat{\beta} \approx 2\). If we compare figure 6.7 results with the real network (shown below), we can notice that the shape of the distribution is similar.

The slope are slightly different, yet still close to the value of 2. This could have been to some unknown factors in the real network that were not considered. Nevertheless, our proposed BA model does a good job in adhering to the preferential attachment notion.

(a) Histogram of Indegree    (b) log(indegree) vs log(freq)

Figure 6.8: Real Network indegree Distribution following a BA model
7 Conclusion

The goal of this paper was simulate a network with a modified BA model. This was performed by first modeling the publications of papers and their citations in a large citation network of papers from Stanford Website of large networks with a NB regression with a Cauchy distribution curve and a NB model with logistic growth curve, respectively. Then the NB models along with a combination of Dirichlet and Multinomial distribution were used to simulate a network. The resulted network had a indegree distribution that indicated the presence of preferential attachment between the nodes. Hence the modeling features proposed in this paper are very suitable in simulating network that grows based on preferential attachment.

This paper can be extended by making further analysis on networks simulated by our BA model. Some of them include 1. Analyzing the number of triangles or motifs in our model. That is to see the what percentage of the papers follow the following rule: If paper i cites paper j, and paper k cites paper j, how likely is it for paper i to cite paper k 2. Describing the structure of the network such as but not limited to centrality, and, shortest path.

Another thing that can be performed, is to use unequal probability sampling instead of the multinomial distribution in creating links and then compare and contrast both models.


