Generative Model To Construct Blog and Post Networks In Blogosphere

by

Amit Karandikar

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
2007

ABSTRACT

Title of Thesis:

Generative Model To Construct Blog and Post Networks In Blogosphere

Author: Amit Karandikar, Master of Science, 2007

Thesis directed by: Dr. Anupam Joshi, Professor
Department of Computer Science and
Electrical Engineering

Web graphs have been very useful in the structural and statistical analysis of the web. Various models have been proposed to simulate web graphs that generate degree distributions similar to the web. Real world blog networks resemble many properties of web graphs. But the dynamic nature of the blogosphere and the link structure evolving due to blog readership and social interactions is not well expressed by the existing models.

In this research we propose a model for a blogger to construct blog graphs. We combine the existing preferential attachment and random attachment model to generate blog graphs which are type of scale-free networks. The blogger is modeled using read, write, idle states and finite read memory. The combination of these techniques helps in evolution of time stamped blog-blog and post-post network through citations within the blog-blog network. Other parameters like the growth function and the randomness in reading and writing posts help in the formation of graphs with different structural properties.

We empirically show that these simulated blog graph exhibits properties similar to the real world blog networks in their degree distributions, degree correlations and clustering coefficient. We believe that this model will help researchers to evaluate and analyze the properties of the blogosphere and facilitate the testing of new algorithms.

Keywords: blogs, generative models, power law, scale-free networks.

DEDICATION

Dedicated to Aai, Baba and Anand.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my graduate advisor Dr. Anupam Joshi. His suggestions, motivation and advice have proved very valuable in this work.

I take this opportunity to thank Dr. Finin, Dr. Yesha and Dr. Oates for graciously agreeing to be on my thesis committee. Dr. Finin and Dr. Yesha were instrumental in providing new ideas, pointers to related work and timely feedback on all aspects of my work.

I thank the folks at Nielson BuzzMetrics, ICWSM, TREC and WWE for making the large blog dataset available for research.

A special note of thanks to Akshay Java and Pranam Kolari for suggestions, the long discussions and willingly providing the needed lab resources on time. I would also like to thank my friends in eBiquity lab who were very enthusiastic about my work and have provided suggestions and constructive criticism.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

In this chapter, we present a quick introduction to the blogosphere. We will discuss the need for simulating the blogosphere and present the formal thesis definition.

1.1 Blogs and the blogosphere

Recently blogs have emerged as a medium for expression, discussion and sharing information on various topics on the web. The authors of the blogs are referred to as *bloggers*.

A blog is a user-generated website where the entries (often called blog posts) are made and displayed in a reverse chronological order. Blogs often provide reviews/discussions about an event or topic such as food, politics, or local news; some function as more personal online diaries. A typical blog combines text, images, and links to other blogs, web pages, and other media related to its topic.

The ability for readers to leave comments in an interactive format is an important feature that encourages discussions among bloggers. Blogs have become a new way to publish information at a global level, engage in discussions with a very large audience and eventually leads to the formation of online web communities. The pictures of the London underground bombing were first posted on some blogs in London and were later picked by the newspapers. Newspapers and news websites have begun to report events as discussed by various bloggers. Due to the growing influence of this specialized publishing infrastructure of blogs, this subset of the web sphere is popularly known as the *blogosphere*. Today, blogosphere encompasses primarily textual blogs, although some focus on photographs (photoblog), videos (vlog), or audio (podcasting), and are part of a wider network of *online social media* [1].

Blogs often discuss the latest trends and echo with reactions to different events in the world. The collective wisdom present on the blogosphere can be invaluable for market researchers and companies launching new products. Various researchers are trying the collect useful information from this large information source of blogs. Some of the recent areas of research are tracking the opinions and bias in blogs [2], modeling spread of influence in blogs [3], finding communities in blogs [4], sentiment analysis in blogs [5] and so on.

1.2 Motivation - why do we need blog graphs?

Structural analysis of the blogosphere requires lot of efforts in setting up the experiments. Generally, the tasks involved are:

1. Gathering the real world blog data by crawling the blogosphere and sampling the right blogs as a representative set for the study.
2. Preprocessing and data cleaning as described by Leskovec et al [6] (section 4.2).
3. Splogs or spam blog elimination as described by Kolari et al [7–9] is necessary as the splogs heavily affect the structural properties of the blog graphs like the degree distributions and average degree.

Other research involving community detection based on link structure, several temporal analyses of blog graphs need the creation of blog graphs from the real world data. The availability of unbiased samples of blogosphere is important for speeding the development and testing of methods and algorithms.

We believe that the blog graphs and post graphs will help the research community in blogosphere for creating large synthetic blog graphs. By analyzing these blog graphs, we can answer several questions like the community structure in blogosphere [4], spread of influence [3], opinion detection and formation [2], friendship networks [10, 11] and the formation of information flow networks [6].

In this work, we would thus like to define the process to create the blog-blog networks and the post-post networks with properties close to the blogosphere. This would help in the testing of algorithms developed for analyzing blogosphere at both; macroscopic and microscopic level. Currently we have done an analysis of the general properties of the blogosphere such as degree distributions, reciprocity, clustering coefficient, degree correlations, diameter, hop plot and so on. Advanced properties of social networks such as centrality measures, betweenness, centrality closeness, centrality eigenvector and so on are not considered for this study.

1.3 Thesis contribution

Our aim is to model the blogosphere by constructing blog-blog network and the time stamped post-post network; maintaining the known structural properties of the blogosphere. The proposed model captures the linking patterns arising in the blogosphere through *local interactions*. Local interactions refers to the interaction of the bloggers among the other blogs that are generally connected to them either by an inlink or an outlink.

The thesis contribution can be briefly stated as follows:

1. To propose a generative model for a blog-blog network using preferential attachment and uniform random attachment model [12] by closely modeling the interactions among bloggers.

2. To generate a post-post network as part of the generative process for blog graphs by creating links from one blog post to another.

The proposed model achieves the following goals:

1. Models the properties observed in real world blog graphs; mainly the degree distributions, degree correlations, clustering coefficients, average degree, reciprocity and the connected components.

2. Models the properties of the post network similar to the real world post network. The post network is sparse compared to the blog network and is characterized by the links per posts.

In addition we also hope to see how the parameters used for the model such as the number of readers and writers affect the properties of the blog graphs. The information diffusion model [6] and the model by Kumar et al in [13] do not study the different properties like degree distributions, average degree, degree correlations and reciprocity for the blog graphs. To the best of our knowledge, there exist no general models that can generate the blog-blog network and a post-post network that possess the properties observed in the real world blogs.

Chapter 2

BACKGROUND AND RELATED WORK

Graph analysis and models to synthetically generate the graphs have been popular for web graph analysis. Often these graphs models talk about special type of networks called the small world networks and scale free networks [14]. Further, we will present the recent work in analysis of the blogosphere and comparison of the statistical properties of blogosphere to the web. In this chapter, we will review various methods and techniques used in creating generative models for the web and also the recent research that suggest some approaches to model blog graphs.

2.1 Small world network

A *small world network* [15] is a class of random graphs where most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps. A small world network, where nodes represent people and edges connect people that know each other, captures the small world phenomenon of strangers being linked by a mutual acquaintance. The social network, the connectivity of the Internet, and gene networks all exhibit small-world network characteristics. The *small world phenomenon* [16] (also known as the *small world effect*) is the hypothesis that everyone in the world can be reached through a short chain of social acquaintances. This was first proposed by John Guare in his famous book *Six Degrees of Separation* [17].

2.2 Scale free networks

A *scale free network* [14, 15] is a special kind of complex network [12] because many "real world networks" fall into this category. The term "real world" refers to any of various observable phenomena which exhibit network theoretic characteristics (e.g., social network, computer network, neural network). As discussed by Newman [14] the term "scale-free" refers to any functional form $f(x)$ that remains unchanged to within a multiplicative factor under a rescaling of the independent variable x. In effect this means power-law forms, since these are the only solutions to f(ax) = bf(x), and hence "power-law" and "scale-free" are, for our purposes, synonymous.

These networks have certain non-trivial topological features that do not occur in simple networks. In scale-free networks, some nodes act as "highly connected hubs" (nodes with high degree), although most nodes are of low degree. *Scale-free networks' structure and dynamics are independent of the network size*. This is an important consideration in modeling such networks.

2.2.1 Properties of scale free networks

Few distinguishing properties of scale free networks as reported by Reka Albert et al [15] can be summarized as follows:

1. The degree distribution follows a power law relationship:

$$P(k) = k^{-\gamma} \tag{2.1}$$

   where k is degree of the node in the network and the exponent γ varies approximately from 2 to 3 for most real networks.

2. Scale-free networks are more robust against failure. This means that the network is

more likely to stay connected than a random network after the removal of randomly chosen nodes.

3. Scale-free networks are more vulnerable against non-random attacks. This means that the network quickly disintegrates when nodes are removed according to their degree.

4. Scale-free networks have short average path lengths [15]. Bollobas et al [18] proved that diameter of the scale free networks asymtotically can be expressed as:

$$D \approx logN/loglogN \tag{2.2}$$

where N is the total number of nodes in the network.

2.3 Generative models for the web

Mainly three types of generative models [19] for the web have been studied as follows.

2.3.1 Random graph models

Early work in web graph modeling was done by Erdos-Renyi (ER) model [20] for random graph generation. Such a graph is constructed by starting with an initial set of n nodes and randomly connecting a new node to one of the existing nodes at each time step. The Watts and Strogatz (WS) model [21] introduced a small world structure with short average path length and high clustering coefficient. Both these models *fail to exhibit the scale-free graph structure* (power law in degree distributions) [19] that has been observed in real world networks. Hence the two models are not suitable for modeling the blogosphere.

2.3.2 Preferential attachment graph models

The first proposal for a *preferential attachment* [22] mechanism to explain power law distributions was made by Herbert Simon in 1955. The basic idea behind preferential attachment is the "rich get richer" phenomenon. The book on complex graphs [12] describes a model to obtain the power law degree distributions in directed graphs using vertex step (adding a new node) and the edge step (adding new edges); and also provides a detailed mathematical analysis.

The Barabasi and Albert (BA) model [23] made the notion of preferential attachment popular by applying it to graphs. In this model, when a new node is added, it does not link to a randomly selected node, but to a preferentially chosen node that is already highly referenced (or linked). At each time step the probability Π that the new vertex would link to a vertex i with degree k_i is given as:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.3}$$

A more practical realization of the above is described by:

$$\Pi(k_i) = \frac{k_i + a}{\sum_j k_j + a} \tag{2.4}$$

For values $a > 0$, the above formulation avoids the problem of having zero probability for a vertex to be linked (as in case of vertex with no prior inlinks). The BA model accurately models the high connectivity nodes but *failed at modeling the high density of low connectivity nodes*. The BA model uses a *linear function* for modeling the growth of the graph while most real networks exhibit [**?**] [24]. Bollobs et al [25] have proposed a directed scale-free model for the web based on the BA model. This model is able to model few nodes with very high indegree but low outdegree and vice-versa. Vazquez [26] shows that preferential attachment is the natural outcome of growing network models based on "local rules". The term local rules refers to the evolution rules that involve a vertex and its neighbors.

2.3.3 Hybrid graph models

Pennock et al [27] studied the link distributions of communities in web graphs and found that the degree distribution in specific communities deviates from a power law. This model presumes that every vertex has at least some baseline probability of gaining an edge. Both endpoints of edges are chosen according to a mixture of probability α for preferential attachment $(1-\alpha)$ for uniform random attachment. In this approach, the probability that the new node connects to an existing node i is given by:

$$\Pi(i) = \alpha.\frac{k_i}{2mt} + (1-\alpha).\frac{1}{m_0+t} \tag{2.5}$$

where α = mixture model parameter, k_i = connectivity of node i, m_0 is the initial number of nodes and t = number of timestamps.

According to Pennock, simple preferential attachment model for web leads to pure "rich get richer" phenomenon - resulting in a nearly pure power law distribution over connectivities. Due to the addition of uniform random attachment component, the poorer sites (with some luck) can get rich too. This can be viewed as two common behaviors of web page authors: (a) creating links to pages that the author is aware of because they are popular, and (b) creating links to pages that the authors is aware of because they are personally interesting or relevant, largely independent of popularity. This is true for the blogosphere as well, hence this model is better suited as a baseline for our work.

A number of studies have used this model for generating synthetic web graphs and communities. Tawde et al [28] describe generating web graphs by embedding communities to get the desired properties of the web. This is one such model based on the Pennock model.

2.4 Model using random walk on graphs

A generative model can also be described using *random walks* [29] on a graph. In case of web graphs this is also called as the *random surfer model*. Given a graph G = (V, E), a random surfer traverses the graph starting from a random initial point and moving to its neighbors such that if u, $v \in V$ the probability that the walk moves to v is given by,

$$P_{uv} = \begin{cases} \frac{\beta}{d^+(u)} + \frac{1-\beta}{|V|}, & \text{if } (u,v) \in E; \\ \frac{1-\beta}{|V|}, & \text{otherwise} \end{cases} \tag{2.6}$$

where $d^+(u)$ is the out-degree of vertex u. Since G represents a web graph, in order to ensure that the graph is irreducible[1] and aperiodic[2], a random jump probability $\beta(< 1$, typically 0.8 to 0.9$)$ is added.

Thus, due to random teleports, the surfer follows the link to the neighboring node or moves to a random node in the graph. β also guarantees that the random surfer does not get stuck in a sink. For the limit $\beta \to 1$ the model follows a pure random walk, while with $\beta \to 0$ the probability of reaching any node is constant which is $\frac{1}{|V|}$.

Fortunato et al [30] provide a detailed analysis of distribution of PageRank (which is also the stationary probability of the random walk process described) at the two limits. Blum et al [31] describe the random surfer model for generating web graphs and theoretically relate it to the preferential attachment model. Mitzenmacher [19] and Bonato [32] provide a detailed review of generative models for the web and discuss the properties studied.

---

[1] G is strongly connected

[2] gcd of lengths of all cycle = 1

2.5 Generative models for the blogosphere

One of the earliest works to study the structure of the blogosphere and describe a generative model was by Kumar et al [13]. Their findings indicates a sudden growth of blogging during 2001 and a continuous expansion ever since. According to this study, the degree distribution of the Blogosphere is slowly converging to the power law exponent of -2.1, which is typical of many scale-free networks including the Web. Additionally, by comparing the size of the strongly connected component (SCC) and the community distribution of blog graphs since 1999, they find that there was a significant growth in both around 2001. To verify this phenomenon, the authors suggest the use of a *randomized Blogspace*. Such a blog graph is constructed by rewiring the destination of each link according to a uniform probability distribution. This random blog graph was found to exhibit the formation of a SCC but did not have a significant community structure.

More recently Leskovec et al [6] discussed algorithms for finding patterns of information cascades in blog networks. Cascades refer to the topological paths in the flow of information. A very interesting insight offered in this paper is that contrary to what one may expect, the probability a post of being referenced does not decay exponentially but follows a power-law with a slope -1.5. This means that a large number of references (if any) to the post take place in a very short time after the post is made available. Moreover, the most common pattern of information cascaded observed by the authors was *tree-like structures* and *star networks*, indicating short chains of conversation on the blogosphere. Finally, they describe a generative model that produces information cascades that are very similar to those observed on the blogosphere. Their model is similar to the epidemic and virus propagation SIS (susceptible, infected, susceptible) model. Creation of a post is similar to the infected state, and the *infected blog* can influence (or infect) the neighboring blogs to write a post. Once a blog returns to the susceptible state after being influenced, it can again be influenced by a neighbor. Thus, this model generates new *posts* and thereby produces information cascades. While the SIS model can be used to study information cascades, it does not help to model blog and post networks with reciprocity and degree correlations as observed in the blogosphere. Cascades [6] that represent flow of information are shown in figure 2.2.

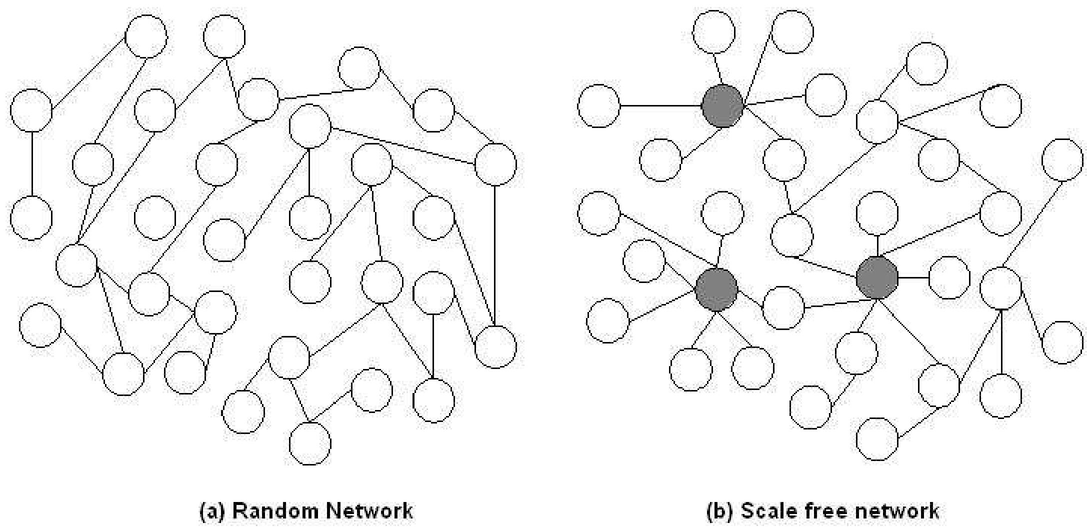


FIG. 2.1. Random network and scale free network
(fi gure courtesy: Wikipedia)

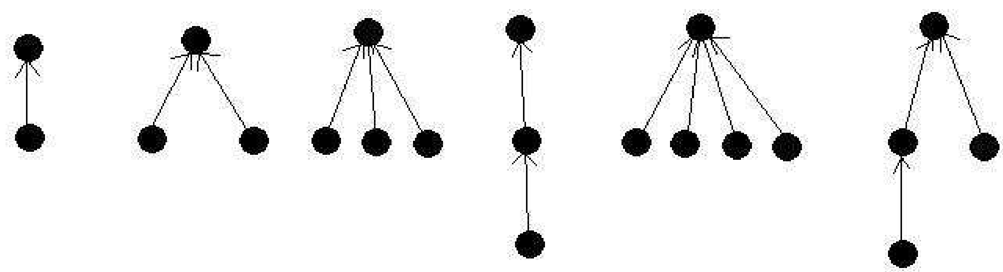


FIG. 2.2. Cascades (link structures) observed in blog graphs
figure courtesy [6]

Chapter 3

DEVELOPING A MODEL FOR BLOGOSPHERE: DESIGN CONSIDERATIONS

We believe that a model for blogosphere should reflect the natural tendency of bloggers to read and write posts, link to some of the posts that he/she read recently and liked. We will first look at some of the statistics from the survey of the blogger characteristics. We will later discuss how these characteristics are important in the modeling of blogosphere.

3.1 Defining a blog network and post network

We define a blog and post network as follows:

Blog network is defined as a network of blogs obtained by collapsing all directed post links between blog posts into directed edges between blogs as shown in figure (b). Blog networks give a macroscopic view of the blogosphere and help to infer a social network structure, under the assumption that blogs that are "friends" link each other more often.

Post network is formed by ignoring the posts' parent blogs and focus on the link structure among posts only. Each post also has a timestamp of the post associated with it. Post networks give a microscopic view of the blogosphere with details like which post linked to which other post and at what time.

The blog and post networks in blogosphere are shown in figure 3.1. We can see that a

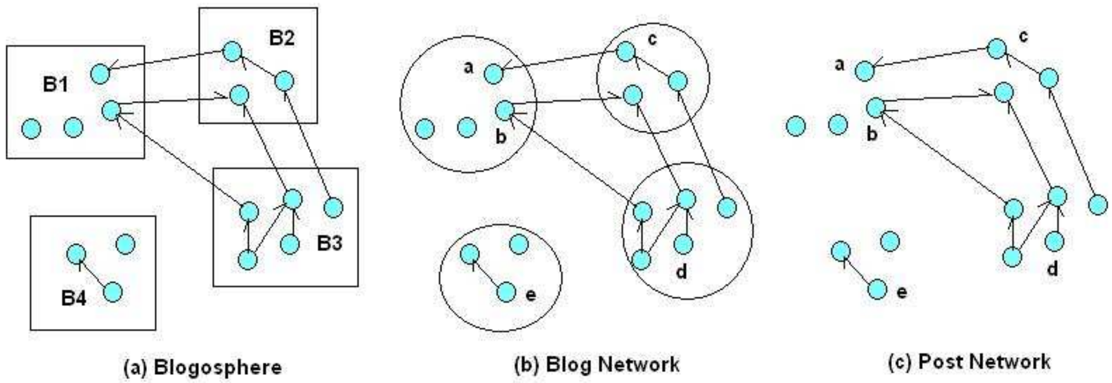


FIG. 3.1. Blogosphere, blog network and post network representation
fi gure courtesy [6]

blog node has higher degree than the post node in general.

3.2 Characterizing the blogger

Here are some finding of the general tendency of the bloggers as surveyed by Pew Internet and American Life Project [33]:

3.2.1 Blog writers are enthusiastic blog readers

> "90% of bloggers say they have read someone else's blog. Frequent updates to one's own blog seem to beget frequent reading of others' material. Bloggers who post new material at least once a day are the most likely group to check on other blogs on a daily basis."

3.2.2 Most bloggers post infrequently

> "While many of the most popular blogs on the internet post material frequently, even multiple times per day, the majority of bloggers do not post nearly so

often."

3.2.3 Blog readership

"Another way to ascertain readership is through blogroll or friend lists, which list links to other blogs. Two in five bloggers (41%) keep a blogroll on their blog, while 57% say they do not provide such a list. Bloggers who post new material daily are more likely to have a blogroll"

3.3 Characterizing the blogosphere resulting from blogger interactions

We now describe how the observations in section 3.2 affect the design of our model. We will discuss the observations and assumptions in details in the following subsections.

3.3.1 Creation of new blogs

As observed in different models proposed for modeling real world networks [12, 25, 29], the new blogger (blog node) may join the network as follows:

1. Not link to the existing network at all

2. Link to a friends blog (random node)

3. A well known blog according to his/her interests (popular blog or authoritative blog - a blog node with high indegree).

Hence our model uses this combination of schemes for connecting the new node. In the rest of our discussion, *blog* and *blogger* should be considered synonymous and the exact meaning is evident from the context.

3.3.2 Linking in blogosphere

As observed in the earlier section, generally bloggers read several posts and tend to link to some of the posts that they read recently. It is difficult to estimate how many posts bloggers read and how many of those do they link to. The only "observable behavior" for the posts read by a blogger in terms of graph analysis is the creation of a link to the read post (destination). We model this behavior by having the blog node to keep track of the recently read posts. In the write state, links are created to these posts from read memory. Here we introduce an idea of "read memory" for our blogger model. In practice, it is hard to justify the size of this read memory and we will empirically decide the size. Leskovec et al [6] show that any post gathers most of its inlinks within 24 hours of post time. This intuitively means that an interesting post is read and highly linked immediately after it is published. We approximate this behavior by linking to some recent (within a fixed window) of the visited blog when our blogger visits any blog node.

3.3.3 Blogger neighborhood

Most *active bloggers* tend to subscribe to the well known blogs of interest and read the subscriptions regularly thus forming a blog readership [e.g. Bloglines[1], Feedburner[2] or RSS feed readers]. Bloggers also follow the blogs in their blogroll (that evolves with time) often and further follow the links (inlinks or outlinks) from those blogs. Blogroll generally consists of "related blogs" or the friend blogs. Hence we see that the blogger interactions are largely concentrated in the neighborhood. We define the neighborhood of the blog node as the nodes connected by either inlinks or the outlinks from the given node.

[1] http://www.bloglines.com
[2] http://www.feedburner.com

3.3.4 Use of emerging tools in blogosphere

The blog readership is also affected by external factors like various new tools. Emergence of tools to identify and track the popular blogs like the blog search engines, feed aggregators and popular subscription information make the blogs more easily available to the wide audience. For example, Technorati[3] provides blog search engine, Digg[4] and del.icio.us[5] provide book marking of user content, Feeds that Matter[6] (FTM!) helps find the popular blog feeds, BlogPulse[7] provides automated trend discovery system, tag cloud[8] representations and news sources etc). Hence there is some probability that a blogger may read at a blog post at random. We believe that these reads are not totally random but biased towards the popularity of the existing but unknown blogs as the tools use some ranking algorithms similar to PageRank to model popularity. Thus with a random probability our model links to a any popular blog post.

3.3.5 Conversations through comments and trackbacks

Most conversations about a topic in the blogosphere happen through comments and trackback for the post. The exchange of links among bloggers due to comments and trackbacks leads to higher reciprocity (reciprocal links) in the blogosphere than the random networks. Bloggers tend to link to the blogs to which they have linked in the past either through comments, trackbacks or general readership. Hence we consider the neighbor of blog A as the blog that links to A or the blog that A links to. We expect these local interactions to provide for a higher clustering coefficient (as observed in blogosphere) than the random networks. In the blog neighborhood, the probability of linking a particular neigh-

[3] http://technorati.com
[4] http://digg.com
[5] http://del.icio.us
[6] http://morpheus.cs.umbc.edu/bloglines/
[7] http://www.blogpulse.com
[8] http://en.wikipedia.org/wiki/Tag_cloud

bor is proportional to indegree of the neighbor according to the preferential attachment mechanism.

3.3.6 Activity in the blogosphere

Not all bloggers are "active" (either reading or writing) at all times. Only a small portion of blogosphere is *active* and the rest can be termed as *idle*. Again, it is challenging to define the activeness of blogosphere as it depends on several factors like the number of readers/writers for that topic, the events in the external world (e.g. London bombings, tsunami and earthquakes) and the "buzz" related to that topic. One of the direct measures to model activity of the blogosphere is from the number of links that get created every time unit. We use a super linear growth function to model the activity as defined by Leskovec et al [24].

We consider outlinks from a blog as the measure of an *active blog writer*. If a blog has more outlinks linking to other blog posts, then it indicates that the blogger has read the post that was linked to. This is because an active writer will naturally look for more interesting sources to link to. The reverse may not be true that the blogger who reads a lot also writes more. Hence we assume that the blogger (blog node) with high outdegree as the one that writes more often. To prevent the model being completely biased towards the high outdegree blogger, we have some probability that any random blogger is given a chance to write a post. Our model allows the new post of a blog to connect to any previous post of the same blog, since this behavior is observed frequently where a blogger refers to an old post in the new discussion.

Chapter 4

GENERATIVE MODEL FOR CONSTRUCTING BLOG AND POST NETWORKS

The simple preferential attachment model as proposed by Barabassi [23] is a good starting point for obtaining the power law degree distributions in an *undirected network*. However, this model does not define the process for obtaining power law distributions for indegree and outdegree in a *directed network*. Also, basic preferential attachment does not capture the real blog graph characteristics like the reciprocity and clustering coefficient that results from local interactions. Our model clearly needs to be a directed model for preferential attachment as the link from a post to another is a directed link.

The model proposed by Pennock et al [27] is more suitable as a base model that includes preferential attachment along with some random attachment factor. Though this model helps to capture the random behavior, it does not capture the local interactions among nodes in the graph very well, and fails to explain preferential attachment model for directed graphs.

A directed model for preferential attachment is proposed by Linyuan Lu et al [12]. This model helps to obtain power law degree distributions in a directed graph and formulate the relation between exponent values and their model parameters. One disadvantage of this model is that the nodes with zero indegree never get inlinks and those with zero outdegree never form outlinks. This leads to mainly two giant components; one having most outgoing

Table 4.1. Symbols for the proposed model

Symbol Name	Explanation
rR	Probability of *random reads*
rW	Probability that the *writers* are selected at *random*
pD	Probability that the new nodes *do not connect* to the network
g	Growth function exponent
ts_i	i^{th} step of the graph evolution
RM_j	Read Memory of the blogger *j*, FIFO queue of finite length
M	Initial number of blog nodes
N	Total blog nodes to create
p(k, j)	k^{th} post of a blog node b_j
e(t)	Expected number of edges at step *t* (according to the growth function)

links and the other with most incoming links. Lu at al [12] have given an improvement for their original model called "alpha-attachment model" which reduces the drawbacks of the earlier approach. We have modified this model to reflect local interaction among the bloggers by *changing preferential attachment among all existing nodes to just neighbors of a node*.

4.1 Symbol and Notations

The main input parameters for the model are given in **bold** in table 4.1.

4.2 Proposed model

We will formally define our model using mainly four parameters as input to vary the properties of the generated graphs - **rR, rW, pD and g**.

Initialization of the network:

Start with M $(M << N)$ nodes and graph G(V, E) such that $V = M$, $E =$. The read memory of all blog nodes is empty at the start and fills as the bloggers read in the read phase.

We perform the following operations in every step:

1. **Add new blog node:** A new blog node b_i is added at each step ts = t_i

    - b_i creates its first post *p(1, i)*
    - With a probability pD, the new node remains disconnected from the existing network. Otherwise creates one link to an existing node b_j by randomly and independently choosing b_j in proportion to its **indegree**. Add a new directed post link *[p(1, i), p(L, j)]* to the graph G, along with step t_i. Note that $1 \leq j < i + M$, L is a recent post of blog node b_j.

2. **Read posts:**

    - When a blogger is in *read state*, blogger b_j (reader) reads the blog posts written by other bloggers by choosing step (a) or (b) below. The blogger may read one or more posts based on the number of links to be created as defined by the growth function.

        a. With probability rR, the blogger b_j visits a randomly and independently chosen existing blog b_k **in proportion to its indegree; irrespective of current read location**.

        b. Otherwise the blogger b_j visits a node by a **preferential walk from its current location** to the neighbor[1] by choosing neighbor b_k randomly and independently in proportion to its **indegree**.

    - **Read the recent post** *p(L, k)*, L is a recent post on b_k. *p(L, k)* is added to read memory RM_j which is a FIFO queue. In case of queue overflow, the oldest blog-post pair in RM_j is overwritten. The blogger in the read state continues to

[1]a node connected by an outlink or an inlink

read until a write operation (when blogger comes back to the homepage - node i for b_i), $1 \leq j, k < i + M$

3. **Write posts:** E edges are added every step according to the equation:

$$E = e(t) - e(t-1) \tag{4.1}$$

where expected edges at step *t*, e(t) = $n(t)^g$, $1 < g < 2$. We cap the value of E to 20 since we do not want it to grow to a large number as $N \to \infty$. This is because it is unrealistic that a post would have very large number of outlinks (except for the autogenerated spam blog posts). The value 20 is an approximate value observed from the power law distribution for post outlinks [6].

   Writer is selected using either (a) OR (b) in write state:

   a. With probability rW, the writer node b_j and the destination (read) node b_k are chosen with **uniform random probability**. Links are added from b_j to b_k by creating the new post at b_j; linking to a recent post of b_k.

   b. Otherwise the writer node b_j is chosen preferentially - **randomly and independently in proportion to its outdegree**. E links are added from new post in b_j to the existing blog network by using destination post from RM_j.

   Note that $1 \leq j < i + M$

4. **Idle Bloggers:** The blogger that do not perform either read or write are considered as idle in that step.

4.3 Preferential attachment in blog neighborhood

Let k_i be the indegree of the node b_i. The preference is calculated same as in the alpha attachment model [12] (Chapter 3):

$$P(k_i) = \frac{(k_i + A)}{\sum_j (k_j + A)} \tag{4.2}$$

where b_j is the *neighbor* of b_i (b_j has linked to b_i in the past or vice versa), $0 < A < 1$

This simple model based on "neighborhood" preferential attachment and random reads and writes generates the blog-blog network and post-post network with properties close to real blogosphere. We will empirically validate the model in the next chapter and discuss how the change in the parameters affects the observed properties of the graphs.

Note that in the model all nodes do not perform writes in each step. Thus the nodes that do not write can be considered to be idle nodes. The idle nodes are important in the modeling of blogosphere because if all the nodes perform the writes in every step then we do not help the "rich get richer" necessary for the power law distributions.

4.4 Memory and time efficient implementation

We have optimized the read and write states of our model. In general, we assume that all the nodes (or a majority) of the nodes in the network perform the read operation thus filling the read memory of each (RM_i). Since read memory is typically small, the subsequent read operations overwrite the earlier reads. The read memory is actually accessed to fetch the destination blog post only in the write phase of the writer node. Hence we optimize the model by performing the read operation only after we select the blog writer in the writer state.

The intuitive idea behind this optimization is as follows: the only observable behavior that the blogger has read a post is when he/she links to that post. As the read memory is typically small (discussed later) too many reads will overwrite the blog read memory. Thus for the blogger that have a very low probability of being selected for writing due to their low outdegree, the read states would waste the time and memory resources for the algorithm.

Using this optimization, the algorithm is as follows:

1. Add new blog node
2. Select writer nodes
3. Allow the selected writers to perform reads and then add links to the network.

The optimization helps the model to be 24 times faster than the earlier model. Also this approach helps the model to scale from few hundred of nodes to a million blog nodes. We have observed the same results in degree distributions and clustering coefficient with both the approaches. We have not made an attempt to prove it as this is not the focus of the research.

Chapter 5

EXPERIMENTAL ANALYSIS AND RESULTS

In this chapter we present the results of the experiments performed for developing the final model and to evaluate the model. We studied the properties of two large blog datasets available for researchers namely WWE[1] 2006 and ICWSM[2] 2007. We have listed the properties of these datasets.

The model is implemented using JAVA. Testing and experiments were performed using Perl scripts. The setup was run on a Linux machine with about 4 GB RAM.

5.1 Comparison of properties of generative models and blogosphere

In this section, we will give an overview of the properties of networks generated by various models and also list the known properties of the blog graphs along with suitable references. Table 5.1 shows that the properties of the simulation are quite close to the observations in blogosphere as compared to ER and BA models.

5.2 Characteristics and properties of real world blog graphs

The properties measured by different research work on the same dataset sometimes vary mainly due to the kind of preprocessing done on the dataset or the way in which blogs

[1]Workshop on the Weblogging Ecosystem: http://www.blogpulse.com
[2]International Conference on Weblogs and Social Media: http://www.icwsm.org/data.html

Table 5.1. Properties of models and blogosphere

Property	ER model	BA model	Simulation	Blogosphere
Type	undirected	undirected	directed	directed
Degree distribution	Poisson refer [14]	Power Law refer [15]	Power Law	Power Law refer [6, 34]
Slope [inlinks,outlinks]	N/A	[2.08,-]	[1.7-2.1,1.5-1.6]	[1.66-1.8,1.6-1.75]
Avg. degree	constant (for given p)	constant (adds m edges)	increases	increases
Component distribution	N/A (undirected)	N/A (undirected)	Power Law	Power Law [6]
Correlation coefficient	-	1 (fully preferential)	0.1	0.024 (WWE)
Avg clustering coeff.	0.00017	0.00018	0.0242	0.0235 (WWE)
Reciprocity	N/A (undirected)	N/A (undirected)	0.6	0.6 (WWE)

are sampled. The presence of splogs also greatly affects the degree distributions and the other properties of blog graphs. For instance, the slope of the indegree distribution for the WWE dataset before splog elimination is -1.6 and after splog removal is -2.0. Hence we measured the properties of these datasets again with the same assumptions and techniques that we used to measure our simulation results. This made our analysis and measurements consistent over all datasets.

5.2.1 Properties of Neilson Buzzmetric dataset

Leskovec et al [6] studied the properties of Neilson Buzzmetric dataset which is one of the largest available datasets in blogosphere. These properties are listed in table 5.2.

As seen from the table, the post network is very sparse compared to the blog network with just 205,000 links among 2.2 million posts. 98% of the posts are isolated, and the largest connected component accounts for 106,000 nodes, while the second largest has only 153 nodes.

5.2.2 Properties of TREC and WWE BlogPulse dataset

Table 5.3 lists the statistics from Shi et al [34] for the BlogPulse dataset. The TREC[3] and BlogPulse datasets contain some isolated vertices as well.

It is clear from table 5.3 that the BlogPulse dataset is much larger than the studied

[3]Text REtrieval Conference, Blog-Track 2006

Table 5.2. Properties of the Neilson Buzzmetric dataset

Blog network properties	
Number of Blogs	$45,000$
Correlation coefficient (scatter plot: inlinks vs outlinks)	0.16
Indegree power law exponent	-1.7
Outdegree power law exponent	$-$
Post network properties	
Number of posts	2.2 million
Total number of post links	205,000
Size of largest connected component	106,000
Indegree power law exponent	-2.1
Outdegree power law exponent	-2.9

TREC dataset. In spite of the large difference in sizes, the indegree distribution is seen to be fairly constant. This gives an hint about the "scale free" nature of the blog graphs. The paper does not provide the values for the outdegree exponents but confirms that it also observes the power law distribution.

5.3 Dataset properties and simulation results

We studied the properties for two large blog datasets namely ICWSM 2007 and WWE 2006 as shown in table 5.4. The table also contains the simulation results for our model. We eliminated the spam blogs from WWE dataset using the techniques pointed by Pranam et al [7–9]. The WWE dataset was largely biased toward LiveJournal[4], MySpace[5] and few other blogs. Hence we ignored all post links to and from these blogs.

The degree distributions for ICWSM and WWE in tables 5.4 and 5.5 are measured after splogs elimination.

The simulations in tables 5.4 and 5.5 use the following parameters: rR = 0.15, rW =

[4]http://www.livejournal.com

[5]http://www.myspace.com

Table 5.3. Properties of BlogPulse and TREC blog network

Blog network properties	BlogPulse (WWE06)	TREC
Number of Blogs	1.4 million	33,385
Blog-to-blog links	1.1 million	198,141
Average degree (extremely sparse)	4.924	– –
Indegree distribution	-2.18	-2.12
Outdegree distribution	– –	– –
Average shortest path length	9.27 (143,736 blogs)	7.12 (16,432 blogs)
Largest WCC[1] size	107,916	15,321
Largest SCC[2] size	13,393	2,327
Clustering coefficients (including splogs in both datasets)	0.0632	0.0617
Reciprocity	3.29%	4.98%
[1]Weakly connected component [2]Strongest connected component		

Table 5.4. Comparison of blog network properties of datasets and simulation

Blog network properties	ICWSM 2007	WWE 2006	Simulation
Total blogs	159,036	650,660	650,000
Total blog-blog links	435,675	1,893,187	1,451,069
Unique blog-blog links	245,840	648,566	1,158,803
Average degree	5.47	5.73	4.47
Indegree distribution	-2.07	-2.0	-1.71
Outdegree distribution	-1.51	-1.6	-1.76
Degree correlation coefficient	0.056	0.002	0.10
Diameter	14	12	6
Largest WCC size	96,806	263,515	617,044
Largest SCC size	4,787	4,614	72,303
Clustering coefficients	0.04429	0.0235	0.0242
Percent Reciprocity	3.03	0.6838	0.6902

Table 5.5. Comparison of post network properties of datasets and simulation

Post network properties	**ICWSM 2007**	**WWE 2006**	**Simulation**
Total posts	1,035,361	1,527,348	1,380,341
Total post-post links	1,354,610	1,863,979	1,451,069
Unique post-post links	458,950	1,195,072	1,442,525
Average outlinks per post	1.30	1.22	1.051
Average degree	2.62	2.44	2.10
Indegree distribution	-1.26	-2.6	-2.54
Outdegree distribution	-1.03	-2.04	-2.04
Degree correlation coefficient	-0.113	-0.035	-0.006
Diameter	20	24	12
Largest WCC size	134,883	262,919	1,068,755
Largest SCC size	14	13	3
Clustering coefficients	0.0026	0.00135	0.00011
Percent Reciprocity	0.029	0.021	0.01

0.35, pD = 0.10, g = 1.06. We will discuss the properties in tables 5.4 and 5.5 in sections 5.4 and 5.5.

5.4 Definitions, computation techniques and analysis

In this section, we will precisely define the properties used for the analysis in tables 5.4 and 5.5. These are some of the standard properties of the graphs as defined in various research papers like [6, 14, 15, 18, 24]. Some properties have been evaluated using tools provided by Jure Leskovec [35].

5.4.1 Average degree of the graph

To compute the average degree in our graphs, we consider the undirected version of the directed graph. The average degree of the undirected graph G (V, E) is defined as the ratio of total edges to the total nodes in the graph.

$$deg(G) = E/V \tag{5.1}$$

As seen in tables 5.4 and 5.5, the average degree of our simulated blog graph is 4.47 and for post graph is 2.10, which is close to the observed degree in the real datasets. It is easy to see from the average degrees that the post network is much spare compared to the blog network.

5.4.2 Degree distributions

The degree distribution, $P(k)$, gives the probability that a selected node has exactly k links. Degree distribution is basically a function describing the total number of vertices in a graph with a given degree. Formally, the degree distribution is

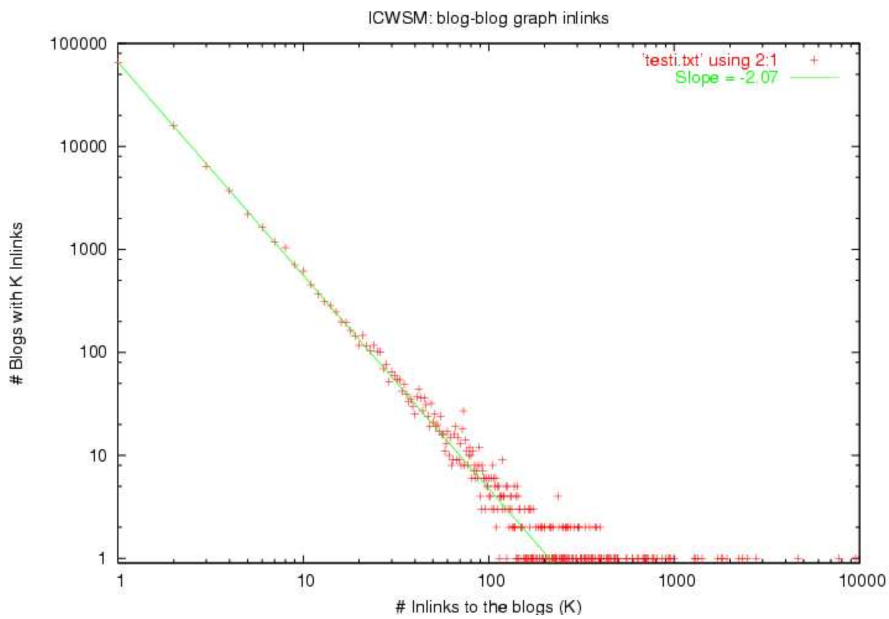
$$P(k) = \sum_{v \epsilon V | deg(v) = k} 1 \tag{5.2}$$

where v is a vertex of the graph G(V, E), and $deg(v)$ is the degree of vertex v. The degree distributions for directed graphs are given by the indegree and outdegree distributions. The *indegree* is the number of incoming links to a given node (and vice versa for *outdegree*). The degree distribution for the scale free networks follows a power law with exponents. Figure 5.1(a) and 5.1(b) shows the inlinks distribution for blog and post networks of ICWSM dataset. Similarly figure 5.2 shows the inlinks distribution for the simulated network.

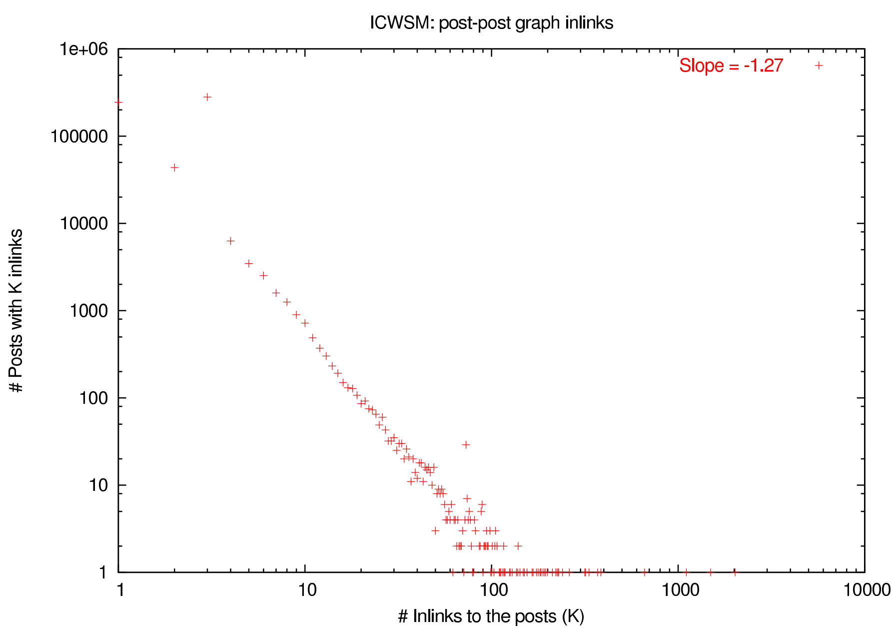
5.4.3 Degree correlations and scatter plot

Scatter plots are generally used to show how much one variable is affected by another. The relationship between two variables is called their correlation. In directed graphs, degree correlations can be used to see the relation between the indegree and outdegree of the graph.

As observed by Leskovec et al [6], the attention (number of in-links) a blog gets is

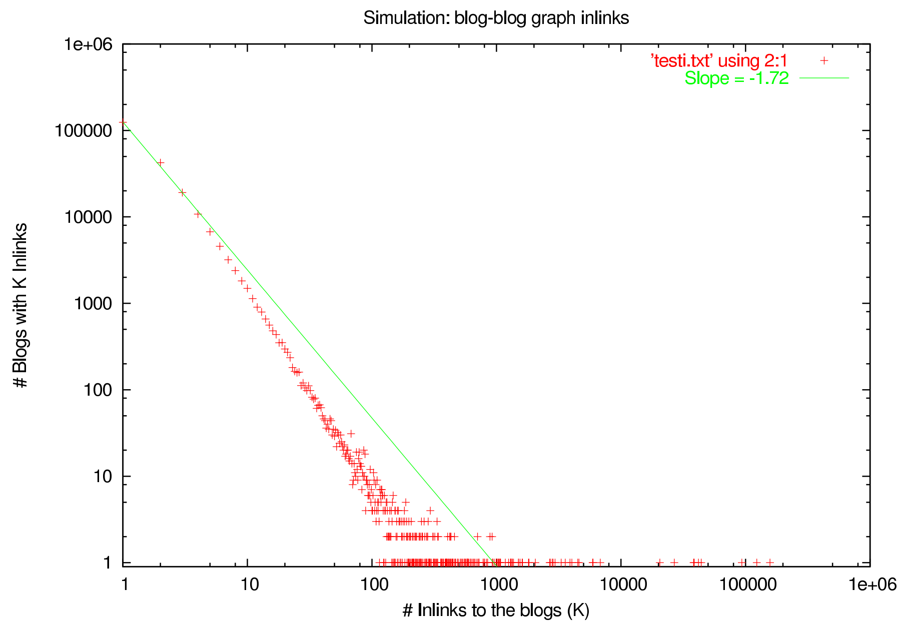


(a) Blog network

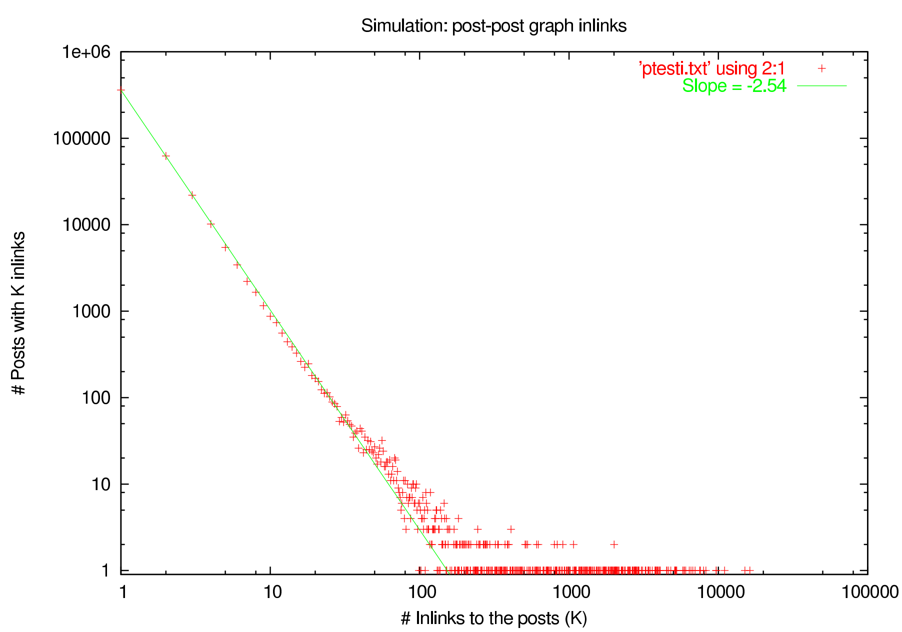


(b) Post network

FIG. 5.1. ICWSM dataset: Inlinks distribution



(a) Blog network



(b) Post network

FIG. 5.2. Simulation: Inlinks distribution

not correlated with its activity (number of outlinks). We also observed in the ICWSM and WWE datasets that there exists very less correlation between the indegree and the outdegree for blogs as well as posts. Randomness in reading and writing the blog posts is necessary to model the low correlation coefficient.

The algorithm for computation of correlation coefficient is used from [36].

From figure 5.3(a) we see that the outdegree of the blogs is not highly correlated with its indegree. The correlation coefficient for our simulated blog graph is 0.10. In our simulation scatter plot shown in figure 5.3(b) correlation coefficient is small which indicates that the indegree and outdegree are not correlated.

5.4.4 Diameter of the network

Diameter of a network is the number of links in the shortest path between the furthest pair of nodes. Small world networks typically have a very small diameter compared to the size of the network. As seen in tables 5.4 and 5.5, the diameter of the blog network is always smaller than the corresponding post network. The diameter of our simulation is 6 and 12 for blog and post network respectively.

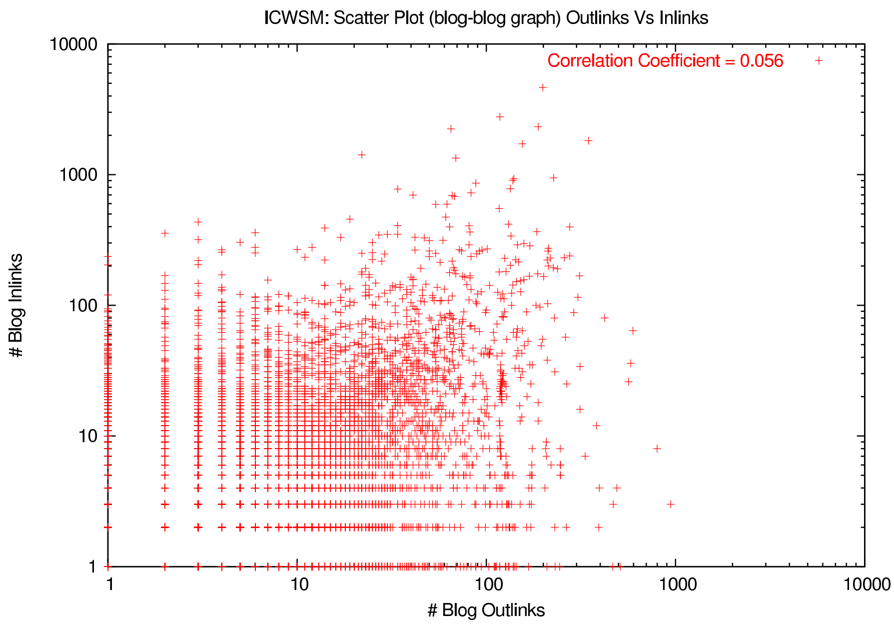
The diameter of the web graph [37] is given by the formula:

$$< l >= 0.35 + 2.06 log(N) \tag{5.3}$$

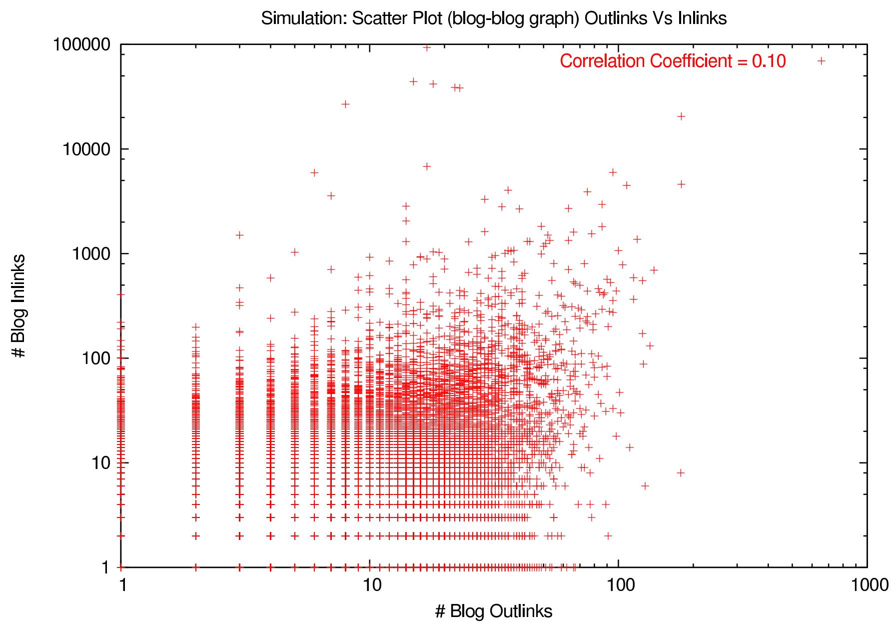
The diameter of our blog graphs using above equation is approx. 12 which is comparable to the observed diameter.

5.4.5 Connectivity: Size of strongly and weakly connected components

For a directed graph, there are two types of connected components: the weakly connected component (WCC) and the strongly connected component (SCC). A strongly (weakly) connected component is the maximal subgraph of a directed graph such that for



(a) ICWSM blog network



(b) Simulated blog network

FIG. 5.3. Scatter plots: Outdegree vs. Indegree

every pair of vertices in the subgraph, there is a directed (undirected) path from v_x to v_y. Thus the weakly connected component is a larger subgraph than the strongly connected component.

In our algorithm, the parameter pD (probability of new nodes that are disconnected from the existing network) can be used to vary the size of the WCC and SCC. The distribution strongly connected components in blog and post networks for WWE and simulated graphs is shown in figures 5.4(a) and 5.4(b). In both the figures we can see that there are large number of small components and small number of very large components. The results for our simulated graphs match closely with the observed curves from WWE dataset.

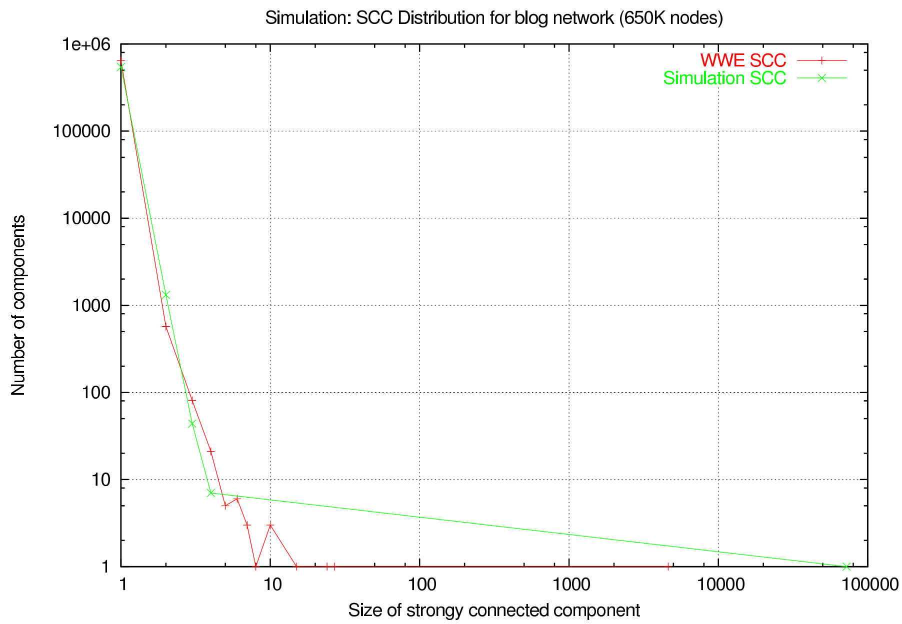
The WCC for the blog and post network do not match as closely as the SCC. Though the curves are of similar shapes the component sizes vary by large amounts. This is not the limitation of our approach and the available blog datasets have also sown variations in component sizes. To model the components as required, we use the parameter pD in the simulation that can help us model the components as required. Also there is no standard value for the component sizes that one may observe in the blogosphere. These components may vary according to the due to various aspects such as popularity of the topic, number of subscribers for the blog and so on. Figure 5.5 shows the WCC distribution for our model and the WWE dataset.

5.4.6 Average clustering coefficient

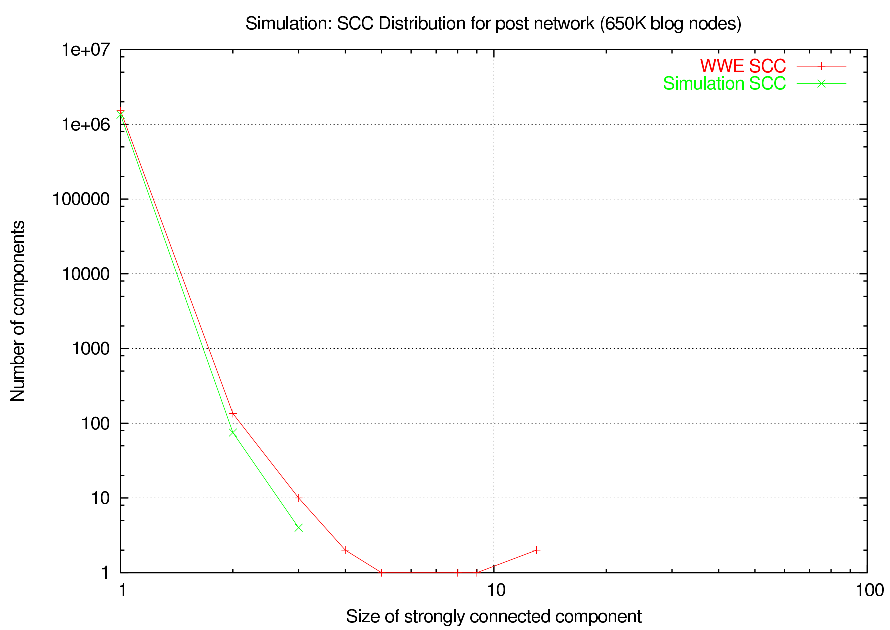
Suppose neighborhood N for a vertex v_i is defined as its immediately connected neighbors as follows:

$$N_i = v_j : e_{ij} \epsilon E \tag{5.4}$$

The degree k_i of a vertex is the number of vertices, $|N_i|$, in its neighborhood N_i. Then the clustering coefficient C_i for a vertex v_i is the ratio of number of links between the vertices within its neighborhood to the number of links that could possibly exist between

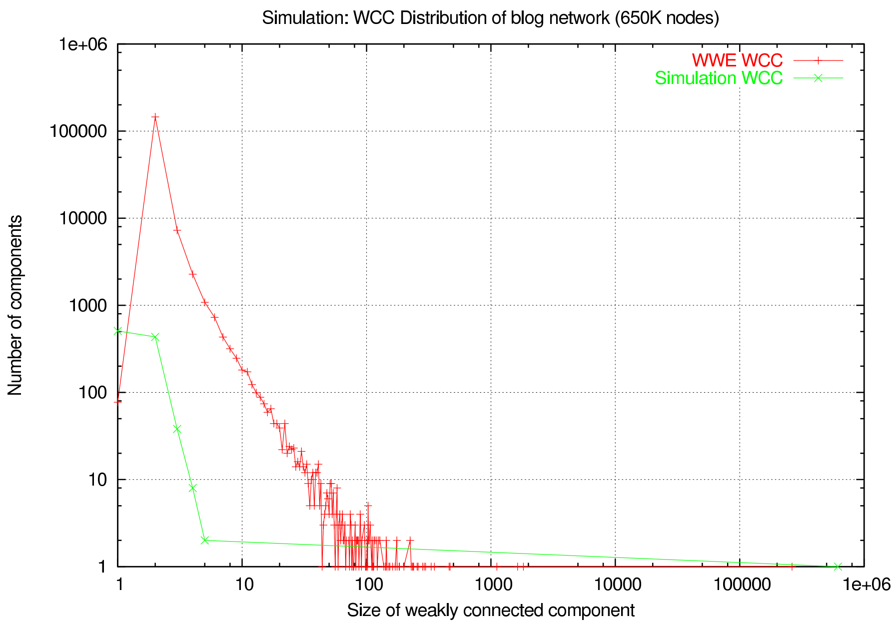


(a) SCC blog network

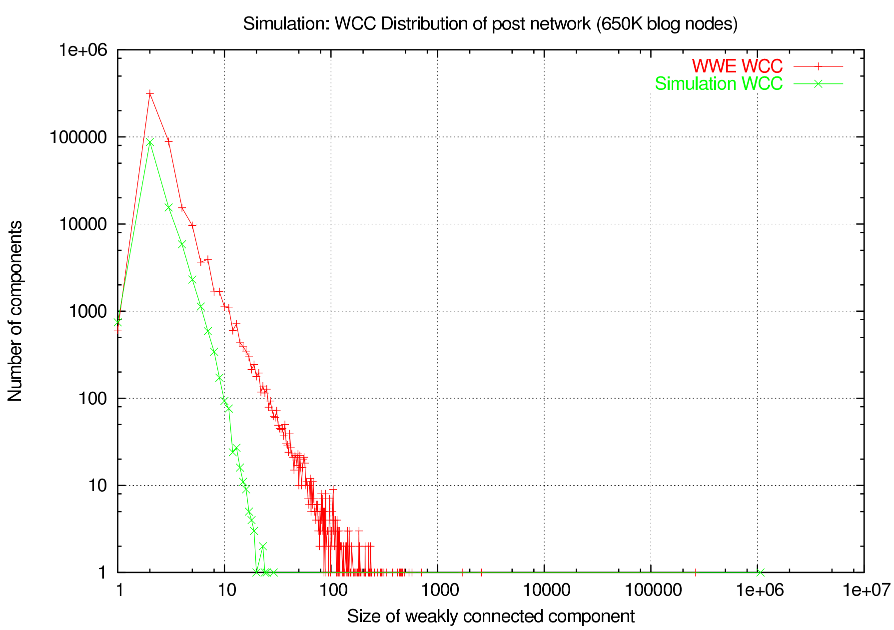


(b) SCC post network

FIG. 5.4. Distribution of strongly connected component (SCC) in WWE dataset and Simulation



(a) WCC blog network



(b) WCC post network

FIG. 5.5. Distribution of weakly connected component (WCC) in WWE dataset and Simulation

them. For a directed graph, e_{ij} is distinct from e_{ji}, and therefore for each neighborhood N_i there are $k_i(k_i - 1)$ links that could exist among the vertices within the neighborhood. Thus, the clustering coefficient of any vertex v_i for a directed graph is given as

$$C_i = \frac{|e_{jk}|}{k_i(k_i - 1)} : v_j, v_k \epsilon N_i, e_{jk} \epsilon E \tag{5.5}$$

The average clustering coefficient for the whole system is the average of the clustering coefficient for each vertex:

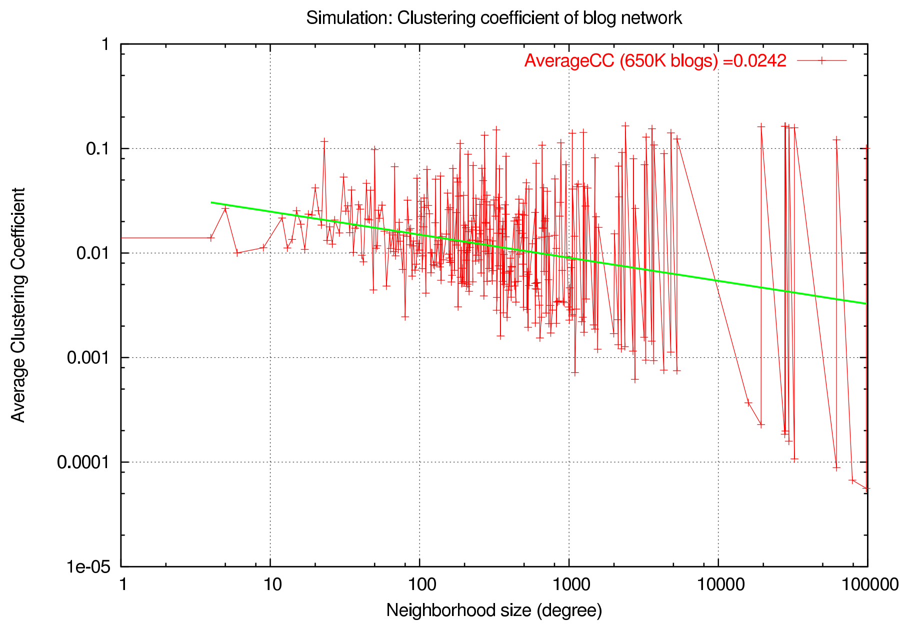
$$C_{avg} = \frac{1}{N} \sum_{i=1..N} C_i \tag{5.6}$$

The figure 5.6(a) shows the distribution of the clustering coefficients according to the neighborhood size for simulated graphs. These values and distributions are similar to those observed in the real datasets. For comparison of the distribution with the WWE dataset, please refer to figure 5.6(b).

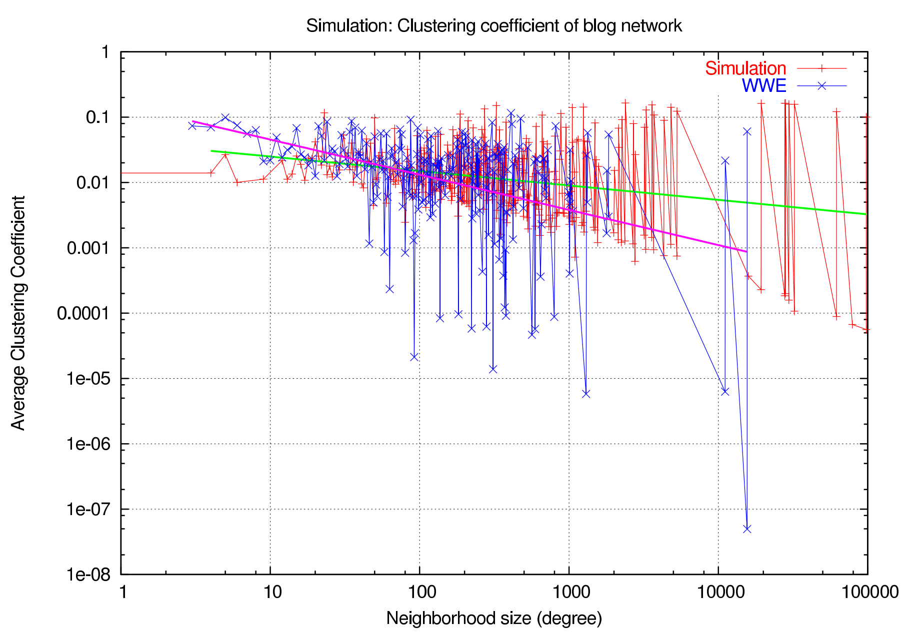
5.4.7 Edge reciprocity of the graph

A vertex pair (v_a, v_b) is said to be *reciprocal* [1, 38] if there are edges between them in both directions. The reciprocity of a directed graph is the proportion of all possible (v_a, v_b) pairs which are reciprocal, provided there is at least one edge between v_a and v_b. The reciprocity values (how often, when v_a links to v_b, v_b links to v_a) is a measure of *cohesion*, reflecting *mutual awareness* at a minimum, and potentially online interaction and dialogue.

In our model, the preferential random walk in read state induces reciprocity in the model because if v_a preferentially links to a neighbor node v_b, then there exists a high probability that node v_b may link to v_a preferentially at some later time (of course preferentially).



(a) Simulated blog graph



(b) Simulated blog and WWE blog graphs

FIG. 5.6. Distribution of clustering coefficients according to the size of neighborhood

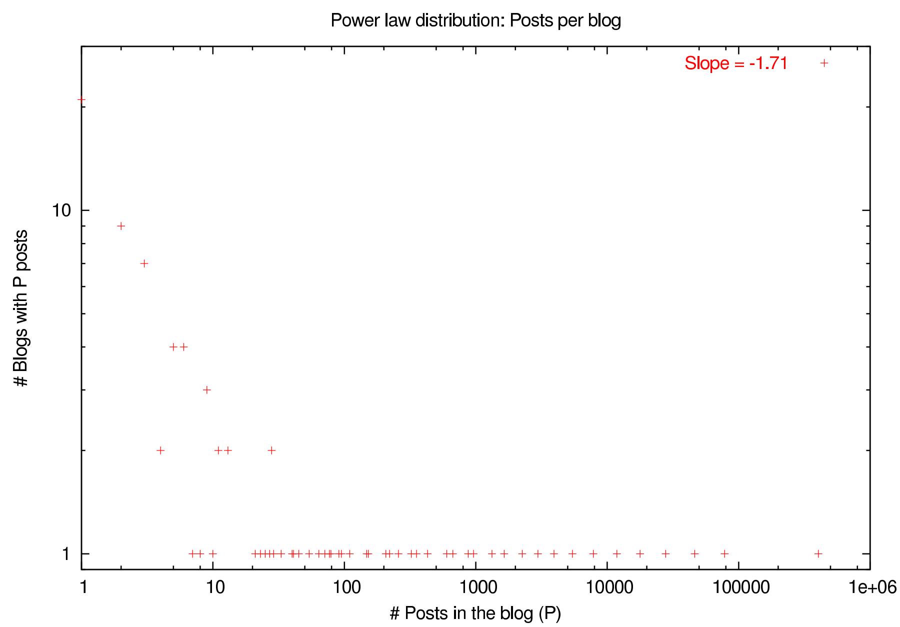


FIG. 5.7. Simulation: Power law distribution in post per blog

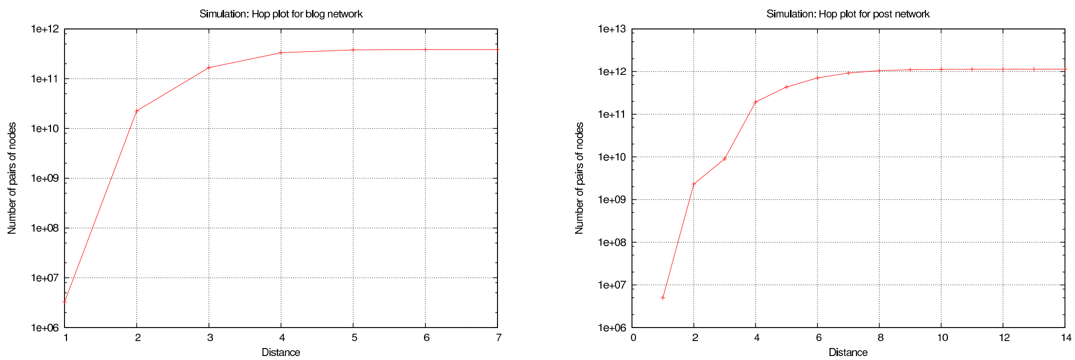
5.4.8 Posts per blog distribution

The post per blog is seen to follow a power law distribution [6].

The plot for the posts per blog for our simulation in figure 5.4.8 also shows a power law distribution with slope -1.71. The maximum number of posts per blog is not as high as the real datasets (highest number of posts seen from the plot is about 20). Again, this can be modeled using a more complex approach where we keep track of the current number of posts that are present at a blog node and considering this parameter in the selection of blog writers to induce post counts as seen in the blogosphere.

5.4.9 Hop plot

The hop plot in figures 5.8(a) and 5.8(b) shows the number of nodes that can be covered by every increase in the hop count. The hop plot becomes constant after certain number of hops which means that the entire network can be reached with that number of hops.



(a) Blog network (b) Post network

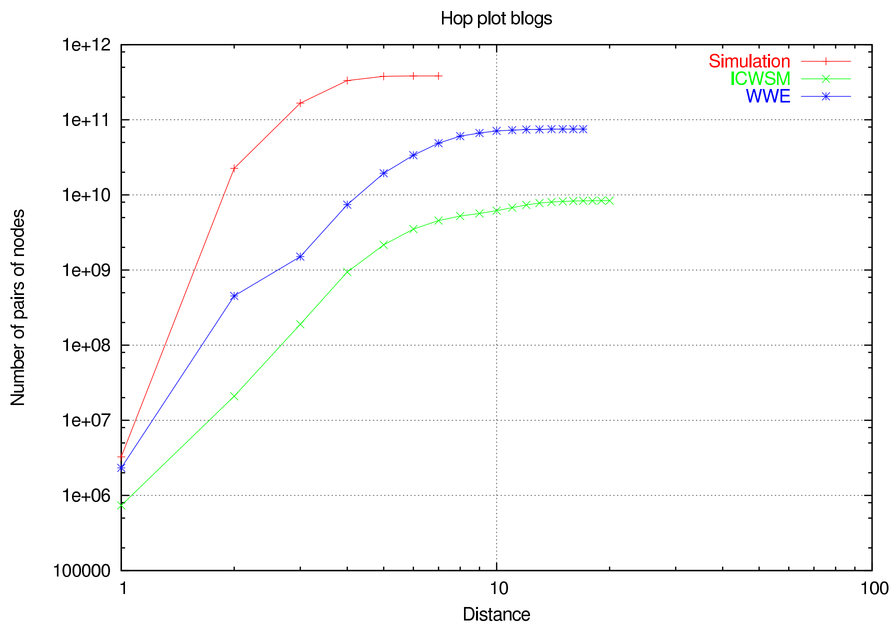
FIG. 5.8. Simulation: Hop plot

The hop plot for the post network at a larger hop count than the blog graph. This also indicates that the post network is sparse compared to the blog network along with the average degree of the post network.

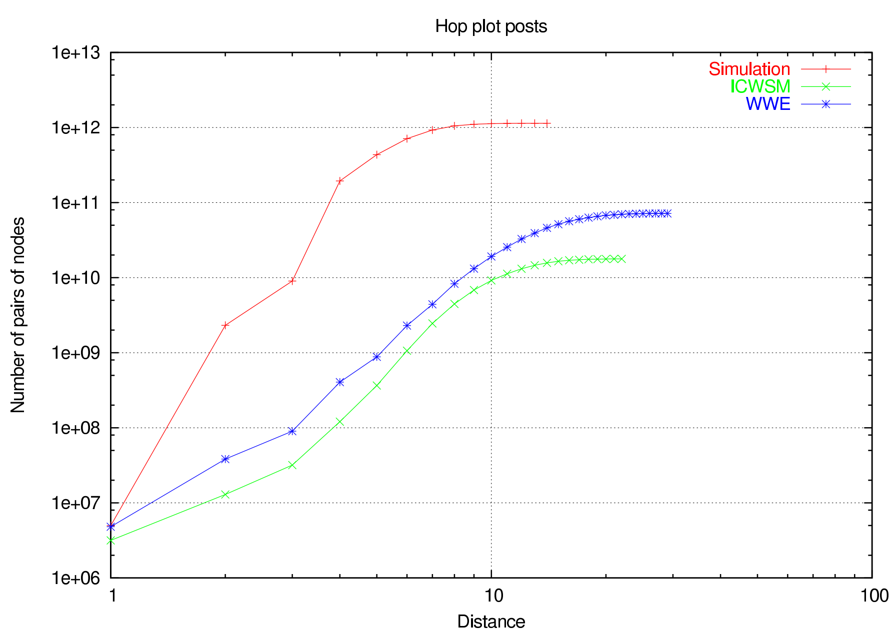
The comparison of the hop plots of the simulation with the ICWSM and WWE datasets please refer figure 5.9.

5.4.10 Information cascade structures

Various star-like and tree-like cascade structures are present in the simulated blog and post network as a result of the preferential reading in the blog neighborhood. These are found with abundance among the connected components as observed by Leskovec et al [6]. These structures are similar to those shown in figure 2.2. We have not done a detailed analysis of these structures as it is beyond the scope of this research.



(a) Blog network



(b) Post network

FIG. 5.9. Hop plot comparisons

Table 5.6. Selecting rR (Number of blogs = 100K)

rR	**1**	**0.5**	**0.15**	**0.0**
Blog inlinks distribution	1.77	1.79	1.76	1.80
Reciprocity of blog network	0.14	0.37	0.69	0.69

5.5 Selection of the parameters for the model

We will analyze the effect of various parameters in the evolution of graphs and try to see how the parameters affect the properties of the simulated network. We will see the effect of blog growth function on the properties of the graph.

5.5.1 Probability of random reads

In our model, we assume random reads to be the effect of various blog tracking and search tools that help a blogger discover new content in the blogosphere. We believe that these randomly read blogs or posts are selected preferentially based on the overall popularity (indegree) of the destination node. Again, we have assumed the number of inlinks as an approximation for popularity of the blog. From table 5.6 we see that as the random reads increases the reciprocity of the network decreases. In other words, to obtain higher reciprocity as seen the the blogosphere, the probability of random reads should be low so that maximum reads are preferential in nature. This makes sense because when blogger preferentially read and link to the “friend” blogs they in turn get links for the linked blogs in the same way.

Here, we chose to measure reciprocity for this experiment as intuitively reciprocity is directly affected by the way we perform the reads. Table 5.6 shows that the inlink distribution is fairly constant which hints at the scale free nature of our simulation.

Table 5.7. Selecting rW (Number of blogs = 100K)

rW	**1**	**0.50**	**0.35**	**0.0**
Blog outlinks distribution	-1.39	-1.70	-1.76	-2.36
Correlation coefficient of blog network	0.27	0.11	0.10	0.04

5.5.2 Probability of selecting random writers

Empirically from table 5.7 we find that the rW = 35% gives correlation coefficient similar to WWE reference dataset. The intuition behind introducing this parameter is to allow any random blogs to create outlinks and obtain inlinks so that the scatter plot and the correlation coefficient can be modeled as seen in the blogosphere.

We chose to measure the outlink distribution and correlation coefficient for this experiment as these are directly affected by the writes. We see that as rW decreases the slope of outlink distribution increases to reach the value of the Web (about -2.5) when $rW \rightarrow 0$. This makes sense because as rW decreases the probability of preferential writes increases.

5.5.3 Growth function for blogosphere

The number of links added every time stamp is an important parameter than decides the total edges in the simulated blog graphs. Both ICWSM and WWE datasets have similar values for links per post. Leskovec et al [24] studied various real graphs including the blog graphs and found that the average degree increases with the number of nodes in the graph. Also the number of edges seems to follow the super linear growth function w.r.t number of nodes. This super linear growth function (densification power law, or growth power law) can be given as:

$$E(t) = N(t)^g \tag{5.7}$$

where $E(t)$ = expected number of edges (links) at time t

$N(t)$ = total nodes at time t

Table 5.8. Effect of growth exponent on the degree distributions (100K blog nodes)

Growth function exponent (g)	**1.02**	**1.04**	**1.06**	**1.08**	**1.10**
Blog inlinks slope	-2.18	-1.95	**-1.73**	-1.61	1.46
Blog outlinks slope	-2.51	-2.07	**-1.79**	-1.69	1.53
Post inlinks slope	-2.51	-2.49	**-2.52**	-2.57	2.63
Post outlinks slope	-2.86	-2.42	**-2.18**	-1.75	1.59
Reciprocity	0.14	0.36	**0.69**	0.84	0.96
Clustering Coefficient	0.0124	0.0202	0.0376	0.0494	0.0574

g = the exponent for growth, $1 < g < 2$

The exponent g can be used to vary the activity of the blogosphere of our model. Empirically we found that $g = 1.04$ to 1.06 is a good value to model the blog and post graphs. This value helps to get the average degree as observed in the ICWSM and WWE blog and post graphs. Also this value of g gives similar average outlinks per post as seen in the reference datasets.

We chose to analyze the properties like degree distributions, reciprocity and clustering coefficient for analyzing the growth exponent. This is because the network growth directly affects these parameters. Empirically, we found that $g = 1.06$ gives the values of degree distribution and reciprocity close to the real datasets. This value of g is similar to that observed by Leskovec et al [24] to be 1.11 for IMDB movie database network, 1.08 for affiliation network and 1.12 for email network. Also from table 5.8 we observe that the reciprocity is directly proportional to the growth rate.

5.5.4 Probability of disconnected new nodes and the size of connected components

In modeling blog graphs, if we assume that a new node always links to a node in the existing network then all the nodes at any point become part of the WCC. Hence to sufficiently model the size of weakly connected components there should be a small probability that the new nodes do not connect the the network. This helps in modeling the weakly

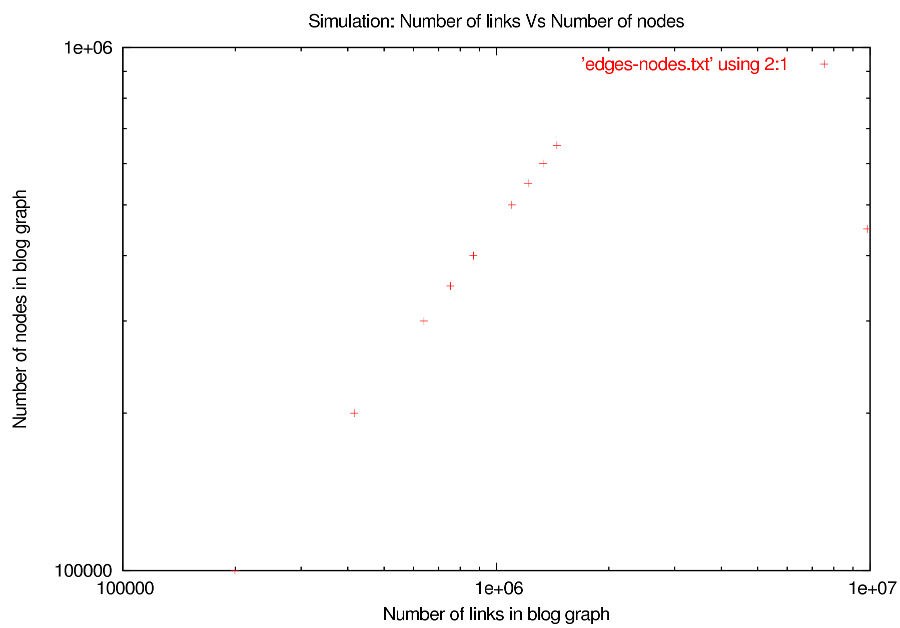


FIG. 5.10. Simulation: Number of edges Vs Number of nodes, g = 1.06 (super linear function)

connected components as well as the distribution of the SCC. These new nodes that are not initially connected do have a chance to get inlinks or even create outlinks due the random jump factor in read phase or the random writer selection in the write phase. Empirically, $pD = 0.10$ gives good distribution of the connected components and also the overall properties of the blog and post network.

The effect of pD is observed on 100K blogs and the other parameters as follows: g = 1.06, rR = 0.15, rW = 0.35. Please refer table 5.9 and figure 5.11 for better understanding on the formation of components. When pD = 0.0, the whole network becomes single huge WCC. As pD increases, the size of the largest WCC decreases and the size of second largest WCC increases. Similar observation has been made by Leskovec et al [6] for the Neilson Buzzmetric dataset. The change in pD does not affect the degree distributions to a great

Table 5.9. Effect of pD on the connected components (100K blog nodes)

pD	Blog inlinks	Blog outlinks	1^{st} SCC	2^{nd} SCC	1^{st} WCC	2^{nd} WCC
0	1.83	1.82	5,657	3	100,000	0
10	1.78	1.81	6,518	4	94,347	4
20	1.72	1.80	8,519	5	89,044	6
50	1.60	1.70	12,737	5	74,185	9

extent. This is a positive result as it means that the model can be used to vary the size of connected components as required without affect the distributions. Note that IL = Inlinks slope, OL = Outlinks slope, 1^{st} SCC = Largest SCC, 2^{nd} SCC = Second largest SCC and so on.

5.5.5 Hop plot in the presence of network components (selecting pD)

The connectivity of the components in the network or the nodes in the network in general affect the hop plot or reachibility of the network. Figure 5.12 shows the hop plot for different values of pD. As the value of pD decreases, the number of nodes reachable per hop increases.

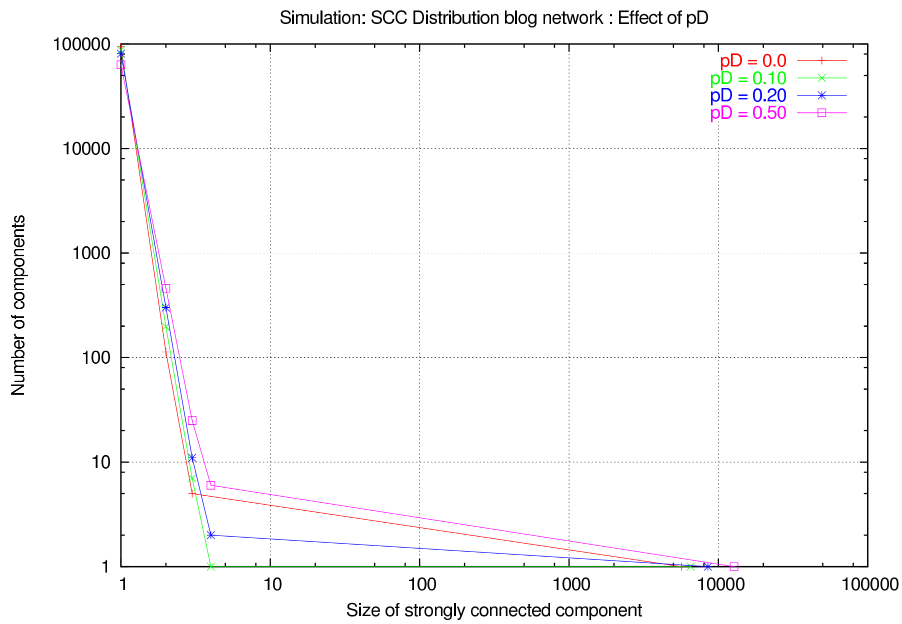
5.5.6 Size of blogger read memory

The blogger read memory should be sufficient to "remember" m post links as computed by the expected growth function. Hence size of read memory must be enough to hold the number of links as computed by the equation:

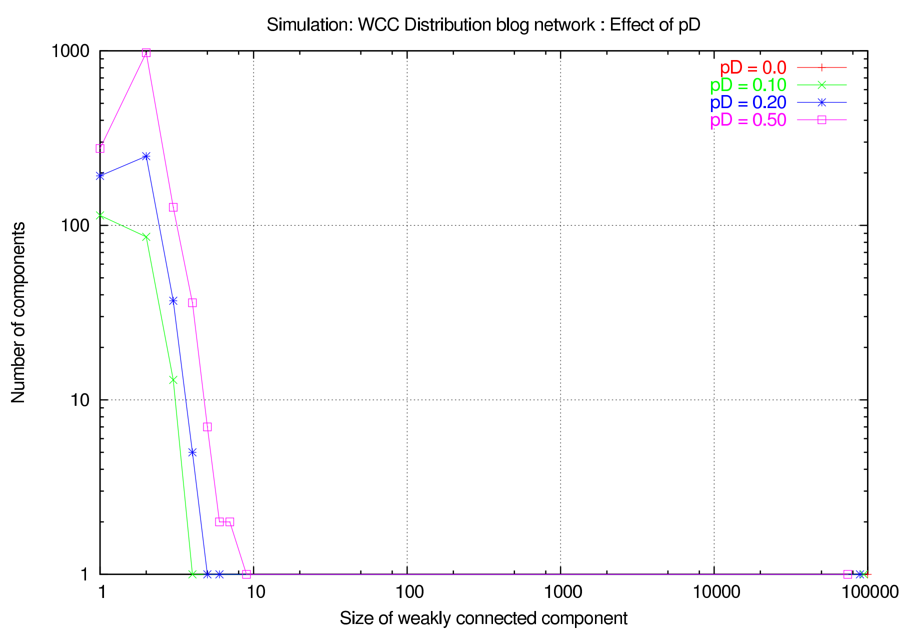
Memory size = Maximum number of new links that will be created in the final step:

$$MemSize = \sum_{n=1..N} floor(n^g) - \sum_{n=1..N-1} floor(n^g) \tag{5.8}$$

We have an upper bound on the MemSize as we have for the number of new links to be added to the graph per step. This will prevent from the memory size to grow to an



(a) SCC Distribution



(b) WCC Distribution

FIG. 5.11. SCC and WCC plots for pD = 0.0, 0.10, 0.20, 0.50

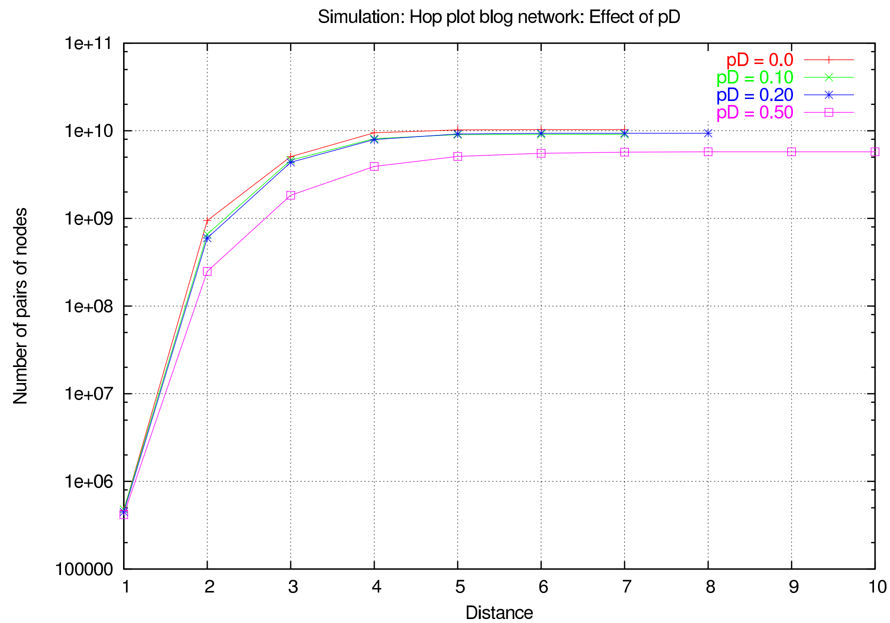


FIG. 5.12. Simulation: Hop plot for different values of pD

extremely large value as $N \to \infty$.

5.6 Effect of graph evolution on the properties

5.6.1 Average degree with graph evolution

Shi et al [34] illustrates that as the time duration increases, the average degree also increases with the number of nodes in the network. For a time period of $10, 20, 30$ and 40 days of the TREC crawl, the average degree $< k >$ increased as follows: $1.5, 2.0, 2.2$, and 2.65, over a maximum of about $11K$ blogs and $28K$ links. This shows that as the time duration increases, the average degree also increases. Our model also possesses this phenomenon due to the super linear growth function as seen from table 5.10.

Leskovec et al. [24] described the densification law prevalent in many real networks.

Table 5.10. Densification of blog graphs

Number of blogs	Average Degree (blog network)
150K	4.08
250K	4.21
350K	4.30
450K	4.36
550K	4.42
650K	4.47

Densification power law states that the networks are becoming denser over time, with the average degree increasing (and hence with the number of edges growing super-linearly in the number of nodes). This law naturally holds for our network as the super linear growth function is used to model the growth of links.

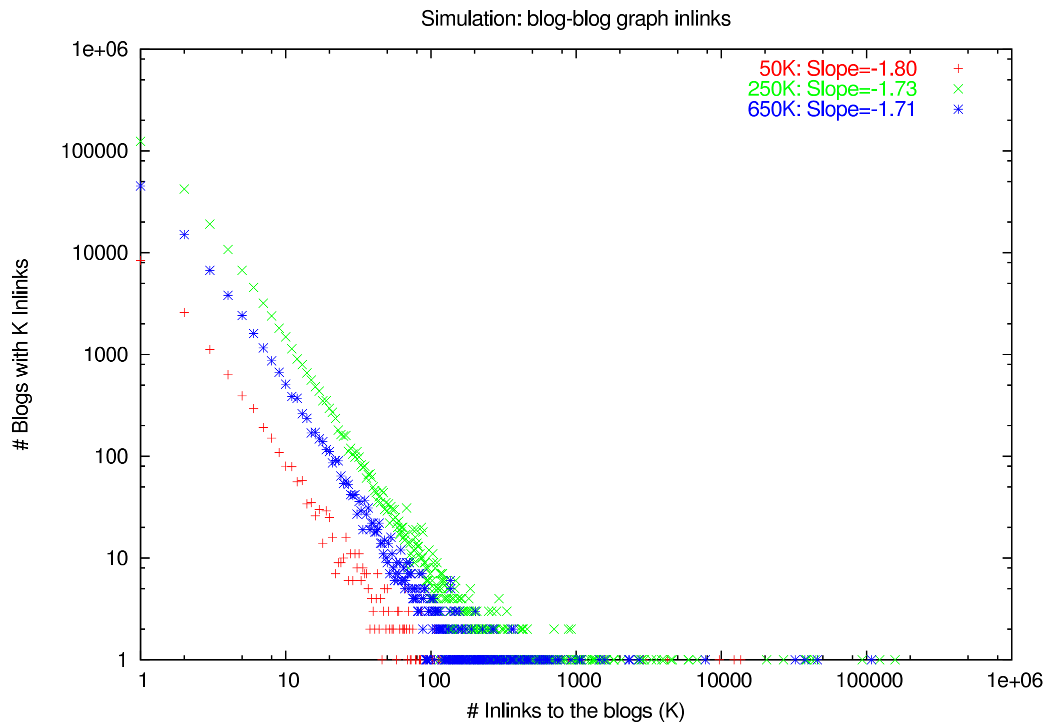
5.6.2 Increasing the number of blogs

Blog network and corresponding post network

Tables 5.11 and 5.12 shows the effect of increase in number of blogs on the parameters of the generated network. This is an important observation since some of the properties of the scale free network like the degree distributions should remain constant for any size of the network.

We will now compare the measurements in tables 5.11 and 5.12 to the observations by Shi et al [34] for various large blog datasets.

Effect of time of evolution on degree distributions: The distributions (both indegree and outdegree) are seen to remain fairly constant over time in real world data. Our simulated graphs match these characteristics. From the figures 5.13(a) and 5.13(b) we observe that the shape of the curve is very similar over time.



(a) Blog network

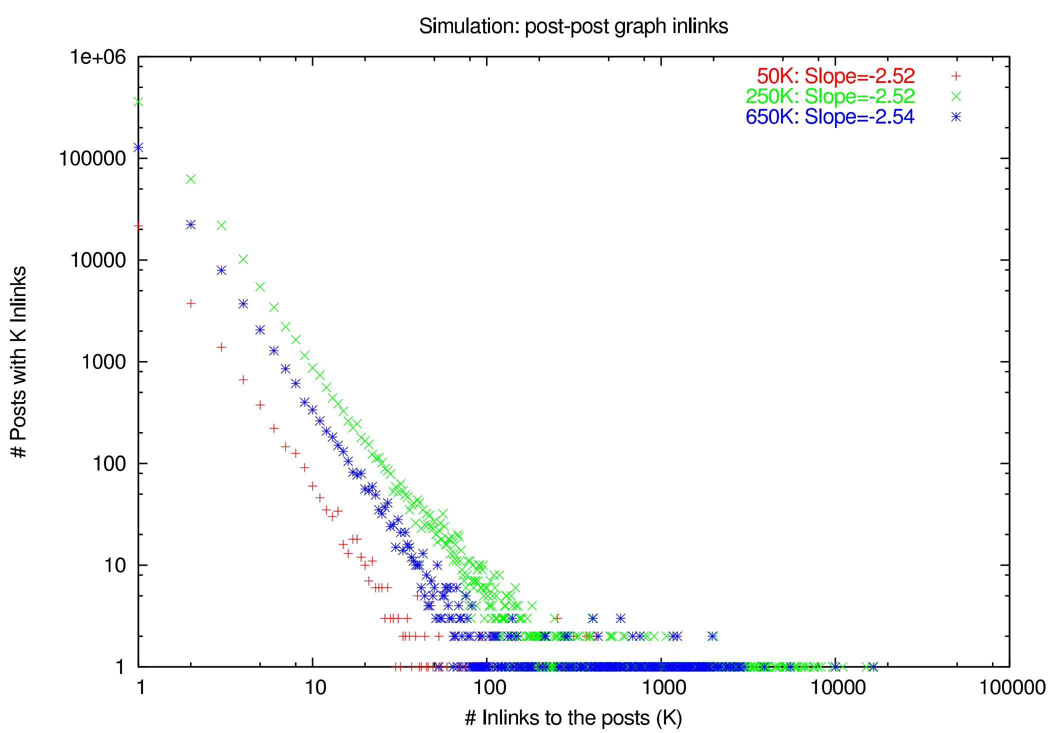


Table 5.11. Effect of increase in the number of blogs

Total blogs	**50K**	**250K**	**650K**
Total links	95,699	527,007	1,451,069
Indegree distribution	-1.8	-1.73	-1.71
Outdegree distribution	-1.83	-1.79	-1.76
Clustering coefficient	0.0390	0.0281	0.0242
Reciprocity	0.5913	0.6462	0.6902
Correlation coefficient	0.069	0.12	0.10
Average degree	3.82	4.21	4.47
Diameter	6	6	6

Table 5.12. Effect of increase in the number of posts

Total posts	100,353	519,347	1,380,341
Indegree distribution	-2.52	-2.52	-2.54
Outdegree distribution	-2.27	-2.18	-2.04
Clustering coefficient	0.00027	0.00019	0.00011
Reciprocity	0.006	0.014	0.01
Correlation coefficient	0.10	0.09	-0.006
Average degree	1.91	2.03	2.10
Diameter	11	12	12

5.6.3 Reciprocity with graph evolution

The reciprocity is seen to increase over time in the blogosphere [6]. This intuitively means with longer time the blogs get higher opportunity to reciprocate. Our simulation also shows similar characteristics for both blog and post network. For the blogs in table 5.11 we see reciprocity increasing from 0.5913, 0.6462 and 0.6902 ($50K$, $250K$ and $650K$ sizes respectively).

5.6.4 Clustering coefficient with graph evolution

The clustering coefficient is found to increase with time by Shi et al [34]. Our model *does not* clearly show this increase but the clustering coefficient remains fairly constant over time of evolution. At the same time one must note that increase in clustering coefficient will become clearly visible for a simulation over a million nodes (since inherently use a super linear growth function) or increasing the growth exponent g as shown in table 5.8. This characteristic can be simulated when we are specifically interested in having such blog graphs by increasing the growth exponent and decreasing the probability of new nodes that remain disconnected. But one should note that this would affect the other structural properties to some extent.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

We proposed and implemented a generative model to create blog graphs and post graphs resulting from the interactions among the bloggers using preferential attachment and uniform random attachment. We analyzed the important structural and statistical properties of the simulated blog and post networks and compared them with the largest available real world blog datasets. The model is based on the local interactions among bloggers in the blogosphere and closely resembles in the properties like degree distributions, degree correlations, reciprocity, average degree and clustering coefficient of the reference datasets. These properties have been verified for blog and post networks. We also show the presence of connected components which lead to the formation of link based community structures in blogosphere. This simulated blogosphere will definitely be useful to the research community by saving effort on data gathering and preprocessing for some of the experiments. Also another advantage of the model is that it will help researchers generate blog graphs with desired properties to test their algorithms by varying the input parameters for extrapolation. We believe that the work will be specifically useful for research in spread on influence due to the presence of reciprocal links, star and tree-like structures as shown in figure 2.2. The scatter plot for the indegree and outdegree, the distributions of the clustering coefficient and connected components also closely resemble the real blogosphere dataset.

The blog growth function in equation 5.7 helps to model the fast or slow growth as required for that section of blogosphere.

The analysis of blogosphere for the structural and analytical properties is a relatively new domain and there is huge scope to explore various properties. The absence of standard published results and detailed analysis of the blogosphere makes the modeling of blogosphere a challenging problem. Also we hope that more datasets for blogosphere will be made available in the future to facilitate research in this area.

6.2 Future work

For future work, we would like to analyze the temporal properties of the post links and closely model the popularity of the posts in the network. This model has the same drawback as the BA model where the earlier created blog gathers more inlinks. Also the early blogs are some of the largest blogs in terms of number of posts in the simulated network. This may be acceptable since the older blogs do have large post count as compared to most newly created blogs. The current model does not sufficiently model the "buzz" where some new post suddenly emerges as the most linked post for some period of time. This will require more careful modeling of the post network with attention to the post outlinks distribution, post scatter plot, methods to model outdegree of posts and posts per blog.

In short, modeling the temporal "buzz" would be an interesting addition to the model. More detailed analysis of the structures observed in these simulated graphs needs to be done which will be useful is analysis of spread of influence. There is always a trade off between the complexity of the model and ease of analysis.

REFERENCES

[1] J. P. Scott, *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

[2] A. Java, P. Kolari, T. Finin, J. Mayfield, A. Joshi, and J. Martineau, "The UMBC/JHU blogvox system," in *Proceedings of the Fifteenth Text Retrieval Conference*, November 2006.

[3] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," tech. rep., University of Maryland, Baltimore County, March 2006.

[4] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng, "Discovery of Blog Communities based on Mutual Awareness," in *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*, Computer Science and Electrical Engineering, University of Maryland, Baltimore County, May 2006.

[5] A. Kale, A. Karandikar, P. Kolari, A. Java, A. Joshi, and T. Finin, "Modeling trust and influence in the blogosphere using link polarity," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, Computer Science and Electrical Engineering, University of Maryland, Baltimore County, March 2007.

[6] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *SIAM International Conference on Data Mining (SDM 2007)*, 2007.

[7] P. Kolari, T. Finin, and A. Joshi, "Svms for the blogosphere: Blog identification and splog detection," in *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[8] P. Kolari, A. Java, and T. Finin, “Characterizing the Splogosphere,” in *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*, Computer Science and Electrical Engineering, University of Maryland, Baltimore County, May 2006.

[9] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, “Detecting spam blogs: A machine learning approach,” 2006. AAAI 2006 - AI on the Web.

[10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, “Group formation in large social networks: membership, growth, and evolution,” in *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 44–54, ACM Press, 2006.

[11] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, “Geographic routing in social networks,” *Proceedings of the National Academy of Sciences.*, vol. 102, no. 33, pp. 11623–1162, 2005.

[12] F. Chung and L. Lu, *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. Boston, MA, USA: American Mathematical Society, 2006.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, “On the bursty evolution of blogspace.,” in *Proceedings of the 12th international conference on World Wide Web*, pp. 568–576, 2003.

[14] M. Newman, “The structure and function of complex networks,” 2003.

[15] R. Albert, *Statistical mechanics of complex networks*. PhD thesis, 2001.

[16] J. Kleinberg, “The Small-World Phenomenon: An Algorithmic Perspective,” in *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[17] J. Guare, *Six Degrees of Separation*. 1995.

[18] G. Szabo, M. Alava, and J. Kertesz, “Shortest paths and load scaling in scale-free trees,” *Physical Review E*, vol. 66, p. 026101, 2002.

[19] M. Mitzenmacher, “A brief history of generative models for power law and lognormal distributions.”

[20] P. Erdos and A. Renyi, “On the evolution of random graphs,” *Mat. Kutato Int. Kozl 5*, pp. 17–60, 1960.

[21] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks.,” *Nature*, vol. 393, pp. 440–442, June 1998.

[22] H. A. Simon, “On a class of skew distribution functions.,” *Biometrika*, vol. 42, pp. 425–440., December 1955.

[23] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, p. 509, 1999.

[24] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Transactions on knowledge discovery from data*, vol. 1, p. 1, 2007.

[25] B. Bollobas, C. Borgs, J. Chayes, and O. Riordan, “Directed scale-free graphs,” in *SODA ’03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, (Philadelphia, PA, USA), pp. 132–139, Society for Industrial and Applied Mathematics, 2003.

[26] A. Vazquez, “Growing networks with local rules: preferential attachment, clustering hierarchy and degree correlations,” *Physical Review E*, vol. 67, p. 056104, 2003.

[27] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles, “Winners

don't take all: Characterizing the competition for links on the web," *Proceedings of the National Academy of Sciences*, vol. 99, no. 8, pp. 5207–5211, 2002.

[28] V. Tawde, T. Oates, and E. J. Glover, "Generating web graphs with embedded communities.," in *Workshop on Algorithms and Models for the Web-Graph - WAW*, pp. 80–91, 2004.

[29] U. Brandes and T. Erlebach, eds., *Network Analysis: Methodological Foundations [outcome of a Dagstuhl seminar, 13-16 April 2004]*, vol. 3418 of *Lecture Notes in Computer Science*, Springer, 2005.

[30] S. Fortunato and A. Flammini, "Random walks on directed networks: the case of pagerank," Apr 2006.

[31] A. Blum, T. H. C. Mugizi, and M. R. Rwebangira, "A random-surfer web-graph model," in *Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO06)*, 2006.

[32] A. Bonato, "A survey of models of the web graph," in *A survey of models of the web graph*, Combinatorial and Algorithm Aspects of Networking, 2004.

[33] A. Lenhart and S. Fox, "Bloggers: A portrait of the internet's new storytellers."

[34] X. Shi, B. Tseng, and L. Adamic, "Looking at the blogosphere topology through different lenses," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[35] J. Leskovec, "Garden graph analysis tools." http://www.cs.cmu.edu/ jure/pub/garden/.

[36] "Pearson correlation coefficient." http://www.stat.wmich.edu/s216/book/node122.html.

[37] R. Albert, H. Jeong, and A.-L. Barabsi, "The diameter of the world wide web," *Nature*, vol. 401, pp. 130–131, September 1999.

[38] S. Wasserman, K. Faust, and D. Iacobucci, *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.