Models for the growth of the Web

Chi Bong Ho  Yihao Ben Pu  Alexander Tsiatas
December 6, 2007

Introduction

There has been much work done in recent years about the structure of the Web and other large information networks. As the Internet, and social network gain importance and popularity, and other data are being represented in graph form (such as collaboration graphs or the Erdos number), researchers have noted basic properties that many such information networks exhibit, and they have proposed models for generating similar graphs with such properties. However, there are still many features of naturally-occurring and developing networks that are unexplored by graph generation models. In this paper, we propose several extensions on existing models that reflect additional properties of large information networks. Particular attention will be given to the Web and its properties.

Erdos-Renyi random graphs

A common starting point for the discussion of probabilistically generated random graphs is a model proposed by Erdos and Renyi in 1959 (Erdos & Renyi, 1959). The model considers a graph $G_{n,p}$ with $n$ nodes, and for each pair of nodes $(u, v)$, there is an edge connecting $u$ and $v$ in $G$ with probability $p$. Erdos and Renyi analyzed their model, showing that the average degree of a node is $np$, and the resulting degree distribution is a normal distribution. Depending on $p$, the graph may or not be connected, as the edge probability controls the density of edges in the graph. Specifically, there is a phase transition when $p$ is on the order of $\frac{\ln n}{n}$: if $p$ is smaller, then the graph will almost surely be disconnected, and if $p$ is larger, then the graph will almost surely be a single component.

For reasons we shall soon see, the Erdos-Renyi model may be simple and easy to analyze, but it is ill-suited for applying to modern information networks. The resulting properties of $G(n, p)$ are not exhibited in most real-world large networks, so we turn to other models of generating graphs that exhibit properties that match real data more closely.

Power laws and the preferential attachment model

One common feature in large naturally-grown networks is the power-law degree distribution. Scientists have been noticing this distribution for many years, including word frequency in European languages, populations of cities, and now social networks such as Facebook, collaboration graphs, and the Internet (Mitzenmacher, 2004). The power-law distribution is notable because it is not the result of any limit theorem in probability theory. In fact, many of the key laws of large numbers in probability theory, such as Chernoff's bound, suggest distributions with exponentially decaying tails on the...
probability distribution. The central limit theorem is a result that yields a normal distribution when independently, identically distributed random variables are summed. Normal distributions occur widely in nature (for example, a person’s height, or SAT scores, and $G(n, p)$), so it is striking that networks tend to result in a different distribution.

A power-law distribution for a random variable $X$ has a probability density of the form $f(x) = cx^\alpha$ for some $\alpha < 1$ and $x > 0$. It is often called a “heavy-tailed” distribution because $f(x)$ does not decay exponentially as $x$ increases (like the normal distribution does). Instead, $f(x)$ decays polynomially, and when samples are drawn from this distribution, there are higher frequencies of larger values of $X$ than one would expect from a normal distribution.

A power-law distribution can be easily recognized, because it forms a straight line when plotted on logarithmic axes. If $f(x) = cx^\alpha$, then $\log f(x) = \log cx^\alpha = \log c + \alpha \log x$. Thus, when $f(x)$ is plotted on logarithmic axes, it will appear as a straight line with slope $\alpha$.

One model that generates a graph with a power-law degree distribution is preferential attachment, as demonstrated by Albert and Barabási (Barabasi & Albert, 1999). The preferential attachment model is a graph growth model that starts with just one node and a self-loop. Then, the graph is grown by adding one node at a time. When a new node enters the graph, there will be some probability $\alpha$ that it will link preferentially to an existing node; that is, select an existing node to link to with probability proportional to its degree. With probability $1 - \alpha$, the new node will link to an existing node selected uniformly at random from all existing nodes.

This model will result in a power-law degree distribution. One way of analyzing this model is by representing the degree of a node as a continuous variable, as done by Barabási et al. (Barabasi, Albert, & Jeong, Mean-field theory for scale-free random networks, 1999). In this case, it is possible to write a differential equation for the change in degree of a node $i$. Let $d_i$ represent the degree of node $i$ (represented as a continuous quantity), and we have

$$\frac{dd_i}{dt} = \frac{1 - \alpha}{t} + \frac{\alpha d_i}{t}$$

This equation arises based on the probabilities: with probability $1 - \alpha$, an existing node will receive one more link at time $t$. Using the continuous approximation, this is now modeled as each node receiving $\frac{1}{t}$ degree. With probability $\alpha$, each node receives degree proportional to its existing degree. The solution to the differential equation is of the form

$$d_i(t) = \frac{At^\alpha - (1 - \alpha)}{\alpha}$$

for some constant $A$. If we note that a node originally has no degree when it enters the graph at time $s$, then the initial condition is $d_i(s) = \frac{As^\alpha - (1 - \alpha)}{\alpha} = 0$, which gives $A = (1 - \alpha)s^{-\alpha}$. Substituting this in for $A$, the formula becomes
Then, if we consider how many nodes have degree \( d_i(t) > d \), algebraic manipulations yield:

\[
\frac{i}{t} < \left( \frac{\alpha}{1-\alpha} d \right)^{-\frac{1}{\alpha}}
\]

This is the fraction of nodes that have degree \( > d \). Translating this into a probability density, we can simply take a derivative to derive the fraction of nodes with degree exactly \( d \):

\[
\frac{1}{1-\alpha} \left( \frac{\alpha}{1-\alpha} d + 1 \right)^{\frac{1}{\alpha} - 1}
\]

This formula is a power-law distribution in \( d \). Note that this derivation is based on approximating the degree as a continuous random variable, but it is possible, but more complicated, to rigorously analyze the model using the true discrete random variables (Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, & Upfal, 2000).

The preferential attachment model is interesting because it is simple, but it still has a sound explanation and motivation when thinking about the growth of a large network. In terms of the Web, it makes sense that new pages are created often, and pages are more likely to link to other well-known and established pages that already have many links. It also makes sense that this is not the only growth mechanism; pages can link to other interesting pages that do not necessarily have high degree, motivating the \( 1 - \alpha \) probability of uniform attachment. The model is also simple to reason about (at least with differential equations), making extensions of the model feasible. Furthermore, it is rather simple to implement. In all, preferential attachment is a good model for generating graphs with power-law degree distributions. However, other features of large graphs are not present in graphs that are generated by preferential attachment.

The striking difference between a graph grown with preferential attachment and real-world information networks is the overall topology. Since one edge is created for each node, the resulting graph will have \( n \) edges and \( n \) nodes, and, barring the original self-loop, it will be a tree. This is very much unlike the Web and other large networks, which are much denser and contain many cycles. The generated graph will also be a single connected component, while large information networks tend to have several small components in addition to one giant component. Also, nodes in the generated graphs are permanent; whereas Web sites often disappear as time goes on.

Another undesirable property is that the high-degree nodes in the graph are generally always the oldest nodes, since they have more opportunities to be linked to. Thus, the usual “rich-get-richer” scheme becomes an “old-get-richer” scheme. Of course, this does not necessarily happen in the real world. Google is a very high degree node that entered the Web in 1997, years after many other nodes were already there. This is not just limited to the Web: research papers that are cited often are not necessarily the oldest ones, and certainly this is not the case for actors appearing in the same film, since actors have a limited lifespan for collaboration.
Other properties of natural networks are not present in graphs grown by preferential attachment. One such feature is clustering. In many networks, there is a high clustering coefficient: if there are three nodes $u, v, w$ such that $(u, v)$ and $(v, w)$ are edges, then it is likely that $(u, w)$ is also an edge. This cannot happen in the preferential attachment model.

The preferential attachment model is mostly associated with generating undirected graphs, but it can also be used to generate directed graphs: when adding a node to the graph, give it a directed edge to an existing node with the same probabilities as the undirected case. While this is useful, since the Web is very much a directed graph, the resulting generated graph will be a directed acyclic graph. Furthermore, while the in-degree distribution will be a power-law, the out-degree distribution will be uniform, since each node only receives one out-link upon its arrival to the graph, and no further out-links are allocated. Thus, the in-degree distribution will be uniform, while the Web exhibits power-laws in both the in- and out-degree distributions.

Lastly, one feature of the Web graph that has been well-documented is the “bow-tie” structure of the Web (Broder, Kumar, Maghoul, Raghavan, Stata, & Wiener, 2000). A Web crawl revealed that the Web graph can be partitioned into several sections: a large strongly-connected component, then a set of nodes OUT that can be reached from the SCC, but cannot reach the SCC, a set of nodes IN that can reach the SCC but cannot be reached from the SCC, and other sections (smaller components, tendrils, tubes). This structure is inherent in the Web, but no model explores generating a graph with this feature.

Variations on the preferential attachment model

Others have come up with variations on the preferential attachment model to better mirror features of real-world graphs. One such model, due to Bollobás et al., generates a directed graph with power-law in- and out-degree distributions (Bollobas, Borgs, Chayes, & Riordan, 2003). They proposed a model that grows a directed graph with five parameters $\alpha, \beta, \gamma, \delta_{in}$, and $\delta_{out}$. The model starts with an initial directed graph $G_0$ with $t_0$ edges. Then, at each time step from $t_0$ onward until some time horizon $T$, one edge is added to the graph, and in some cases one vertex is added as well. Thus, at time $t$, the graph $G$ will have exactly $t$ edges. The edges are added in the following manner:

- With probability $\alpha$, a new vertex $v$ is created, and one edge is added leading from $v$ to an existing vertex $w$, where $w$ is chosen among all existing nodes with probability proportional to the sum of its in-degree and $\delta_{in}$.
- With probability $\beta$, no new vertex is added, but two vertices $v$ and $w$ are chosen, and an edge is added leading from $v$ to $w$. In this case, $v$ is chosen among all existing nodes with probability proportional to the sum of its out-degree and $\delta_{out}$, and $w$ is chosen with probability proportional to the sum of its in-degree and $\delta_{in}$.
- With probability $\gamma$, a new vertex $w$ is created, and one edge is added leading from an existing vertex $v$ to the new vertex $w$, where $v$ is chosen among all existing nodes with probability proportional to the sum of its out-degree and $\delta_{out}$. 
Of course, for this to make sense, the sum $\alpha + \beta + \gamma$ must be 1.

Bollobás et al. go on to prove two theorems about the properties of graphs arising from this model of growth. The first shows that the resulting distributions of both in- and out-degrees both are well-approximated with power laws (modulo some edge cases with the parameters). Their next result shows that for all nodes with a fixed in-degree, their out-degree distribution is well-approximated with a power law, and for all nodes with a fixed out-degree, their in-degree distribution is well-approximated with a power law. Finally, the paper proposes another model, without analysis, incorporating two fitness distributions in the model: a distribution for fitness for in-degree as well as another distribution for out-degree. Thus, all original and subsequently added nodes receive two fitness values from these distributions, and when edges are added, the nodes selected are chosen with a probability dependent on the fitness values.

This model does seem to have some well-grounded motivation. Out-links are added to nodes in the same way as preferential attachment, and in this model, in-links are also added using a variation of preferential attachment. This does make sense, because there are plenty of Web sites out there that work as “link repositories” or information sources that point to many other Web pages, and they tend to link to even more if they seem to be of interest. Also of note is the fact that in the real Web, pages can make links at any time; they are not limited to one link upon site creation, with no further links allowed. Thus, this model by Bollobás et al. better models these traits of the Web. Also, now that existing nodes are allowed to link to each other, the graph will no longer necessarily be a tree (plus a self-loop) or a DAG. This is much more accurate than the original preferential attachment model. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components. However, this model does not take into account the fact that networks often have multiple components.

Four years later, Borgs et al. proposed and analyzed a preferential attachment model based on fitness (Borgs, Chayes, Daskalakis, & Roch, 2007). In their model, they start with a fitness distribution, and they start with one node (and a self-loop), with a fitness value drawn from the distribution. Then at every subsequent time step, one vertex is added (with a corresponding fitness value sampled from the fitness distribution), and it links to an existing node $v$ with probability proportional to the product of $v$’s fitness and $v$’s degree. Borgs et al. go on to prove that the fraction of edges with endpoints on nodes within a range of fitness values is independent of the size of the graph, and that the degree distribution on nodes with a specified fitness value is indeed a power law.

Borgs et al. found that their graph generation can be characterized into three phases:

- At first, the behavior of the graph generation mimics preferential attachment closely.
- After a while, the high-degree nodes are no longer the older nodes; the fitness values affect the distribution considerably.
- Eventually, there will be a constant fraction of links that are added to newly created nodes with higher fitness values.
Especially after the graph reaches the third phase, the model captures more of the processes that appear to be influencing the Web graph. The most important is that the “old-get-richer” scheme of the traditional preferential attachment model is now apparently gone, and the model allows for the case that a newer Web page attracts links due to its innovation, even though it may be new compared to many other existing sites. The model also preserves the necessary power-law degree distributions. In all, this model is an improvement, but it still does not explore clustering, the “bow-tie” model, or the possibility of multiple components in the graph.

New graph generation models

We propose several new models based on preferential attachment that will better capture a few key properties of the Web and other large networks. One important property of the Web is the fact that it is dynamic: Web pages come and go often, and they are also able to change their links if their owners decide it is appropriate. All the graph generating models discussed this far are limited in that they generate permanent nodes and edges. We will outline a model that takes this into account by allowing for a mutable link structure among fixed nodes, creating a directed graph that exhibits power-law in- and out-degree distributions.

Another property that we explore in further detail is clustering. The Internet tends to organize itself into different categories. One can argue that pages about a specific topic are more likely to link to other pages about the same topic. For example, Web pages with information on a topic as specific as lepidopterology (the study of butterflies) are likely to link to other butterfly-related web pages, and they are very unlikely to link to some other obscure topic such as Soviet textiles of the Cold War era. Other topics are more popular in general: a lot of websites tend to link to search engines, news articles, and e-commerce websites such as eBay. Also, consider a category of spam websites. This set is destined to be large, but websites from outside the spam category are unlikely to link to a spam website (at least, not intentionally).

We will propose a model that will reflect the following key ideas concerning clustering and categorization on the Web:

- The Web naturally gives rise to many dense clusters, often corresponding to a set of Web sites about a specific topic or category.
- Web sites within a category tend to link to other Web sites within categories.
- Categories have an inherent weight: the popularity of the topic determines its size.
- Inter-category links exist, but the probability of their existence is very dependent on the categories that contain the endpoints.
Our model will generate a grown graph that exhibits power-law in- and out-degree distributions as well as natural clustering mimicking the Web. This clustering is also prevalent in citation networks (papers about a specific topic tend to cite each other), the actor collaboration graph (actors can tend to star in a specific genre), and food chain networks (consider different ecosystems with dense interactions within the system, but sparse interactions between systems). Our model will therefore be more indicative of large, modern information networks than existing preferential attachment models.

In the following section, we first present our new Web graph models. Next, we present data from a round of simulation. Finally, we conclude with analysis of our model and suggestions for future work.

**Link mutation model**

As a first extension of previously-studied graph generation models, we first explore a simplified version of the preferential model with fitness proposed by Borgs et al. This model first creates a graph with a fixed number of nodes. Each node is assigned an out-degree and a fitness factor, generated by two probability distributions $\mu$ and $\nu$, respectively. Afterwards, the graph undergoes a series of mutations: at each time step, each edge is rewired to a different node, chosen with probability proportional to the product of its fitness and its in-degree.

This model deviates from the one proposed by Borgs et al. in a number of ways. First, our graph has a fixed size: no nodes are added with the passage of time. This contrasts with Borgs’ model as well as the traditional preferential attachment model, where nodes enter the graph sequentially. Second, our graph allows nodes to have multiple edges added as in-links at the same time step. In previous models, only one edge would be added at once. Now, since at each time step, all edges are mutated, a single node can attract many in-links at the same time. Additionally, a node can also lose in-links, another phenomenon that is not modeled by the previously mentioned models. The core concept from Borgs’s model that is retained is the scheme of using a fitness distribution and using it to determine the eventual in-link structure. We keep attraction factor of each node proportional to its fitness times its current in-degree, the same as the model proposed by Borgs et al.

**Attachment-based model with clustering**

Our second model incorporates the notion of clustering into the model previously presented by Bollobás et al. in 2003 by assigning categories to nodes. Initially, our model starts with a graph $G$ with no edges and \( i \) nodes, each belonging to a different category. Thus, there are \( i \) distinct categories. Each category $C$ is assigned a weight $w_c$, generated uniformly at
random on \((0, 1)\). The category weight will influence both the number of nodes and edges in category \(C\), as explained below. At each time step \(t\), \(G\) is modified as follows:

- With probability \(\alpha\), a new node \(u\) is added to \(G\). \(u\) is randomly assigned to a category \(C_u\), which is randomly chosen with probability proportional to the category's weight \(w_{C_u}\). Then, with probability \(\beta\), \(u\) creates a link to an existing node \(v\), which is randomly chosen from \(C_u\) with probability proportional to its in-degree + 1.
- With probability \(1 - \alpha\), a new link between two existing nodes is created. First, two categories, \(C_x\) and \(C_y\), are independently chosen with probabilities proportional to \(w_{C_x}\) and \(w_{C_y}\), respectively. Afterwards, node \(x\) is chosen from \(C_x\) with probability proportional to its out-degree + 1, while node \(y\) is chosen from \(C_y\) with probability proportional to its in-degree + 1, and a link \(x \rightarrow y\) is added to \(G\).

Since new nodes are created only with probability \(1 - \alpha\), we can analyze the expected number of nodes in \(G\) at time \(t\). The graph starts with \(i\) nodes, and at each time step \(k \in \{1, 2, 3, ..., t\}\), a new node is created with probability \(\alpha\). Let \(X\) be the random variable representing the number of nodes at time \(t\), and let \(X_k\) be the indicator variable that is 1 if a new node is created at time \(k\). This occurs with probability \(\alpha\), so \(E[X_k] = \alpha\) for all \(k\). Then, we have

\[
E[X] = E\left[i + \sum_{k=1}^{t} X_k\right]
\]

\[
= i + \sum_{k=1}^{t} E[X_k]
\]

\[
= i + \sum_{k=1}^{t} \alpha
\]

\[
= i + \alpha t
\]

We also analyze the expected number of edges. Let \(Y\) be the random variable representing the number of edges at time \(t\), and let \(Y_k\) be the indicator variable that is 1 if a new node is created at time \(k\). This occurs with probability \(\alpha \beta + (1 - \alpha) = 1 - \alpha (1 - \beta)\), so \(E[Y_k] = 1 - \alpha (1 - \beta)\) for all \(k\). Since \(G\) starts with no edges, we have

\[
E[Y] = E\left[\sum_{k=1}^{t} Y_k\right]
\]

\[
= \sum_{k=1}^{t} E[Y_k]
\]
It is worth noting the differences between our model and the model proposed by Bollobás et al. First, at each time step, the Bollobás et al. model also may add a new node with an in-link from an existing node with a certain (non-zero) probability. Our model excludes this feature because, barring link-spamming, a web site is generally not created solely with the purpose of being linked to by another, existing site. As we will claim in the next section, this excluded process is not crucial to forming the out-degree power law distribution in the web graph. The second difference is the assignment of categories to nodes. We conjecture that while pages on the Web are diverse in nature, most can be categorized by the functions they serve, e.g. corporate web sites, personal blogs, online stores and news sites. Without actually analyzing categories and their linking behavior on the web, we simply model it as a random process, with pages within the same category behaving similarly, i.e. using the same weights to choose nodes to link to. The third difference is that our model incorporates weights for each category, which similar to the idea presented in Borgs et al.

Simulation and analysis of data

Link mutation model

We simulated our link-mutation model with the following parameters. We chose the in-degree distribution $\mu$ and the fitness distribution $\nu$ to modified normal distributions. We let $\mu$ have mean $-5$ and standard deviation 20, and $\nu$ has mean $-4$ and standard deviation 10. Since these normal distributions can generate negative values, we take substitute any negative samples with zero. We let our graph $G$ undergo 20 full link mutations.

We ran this simulation on a graph of 10,000 nodes and got the following data:

Largest strongly connected component size: 8578
In-degree distribution plot (for this trial and an additional trial with 100,000 nodes and the same parameters:

Degree Distribution

Degree Distribution, |N| = 100,000

A nearly straight line appears in the log-log graph of the in-degree distribution for each trial, suggesting that a power-law distribution best models the true degree distribution. This is somewhat surprising considering how different our simplified model is compared to the preferential model.

It is worth noting that our model has its flaws. First, the number of mutations must not be too high, or else the graph will have a few very popular nodes with very high in-degrees, while the majority of the nodes will have zero in-links. This is because as time goes on, in-links are rewired, taking them away from some nodes. If a node does not have any in-links, it will never receive any more, because links are wired to nodes with probability proportional to the product of the fitness and the number of in-links. If the number of in-links is zero, then that node will receive a mutated edge with probability 0. Second, the use of normal distributions for out-degrees and fitness is quite arbitrary, and we do not have a compelling reason to use them over other distributions. An additional flaw is that our model has a static node set; i.e. no new nodes can enter the graph as in the traditional preferential attachment model as well as the previously-discussed models.

However, our model does demonstrate a way to incorporate the changing structure of the Web into a graph generation model. It also avoids the undesirable “old-get-richer” scheme that is apparent in preferential attachment as well as the directed model proposed by Bollobas et al. Despite its simplicity, the ideas from our link-mutation model can be incorporated into other, richer models, adding the notion of the changing link structure of the Web into other models for generating Web-like graphs.

Additionally, the aspect of the preferential model where nodes enter the graph one at a time seems to be unnecessary in achieving an in-degree power law distribution. Our model is
able to replicate that distribution using a fixed-sized graph while using a “rich-get-richer” scheme with the fitness functions. Furthermore, the initial layout of the graph seems to be insignificant. Our graph starts out looking like the random graph model $G_{n,p}$, but evolves into a power-law graph given enough mutations. Lastly, a prolonged activity window of nodes does not seem to affect power law. In the preferential attachment model, nodes become inactive once they enter the graph. We can view each mutation of the graph as prolonging the activity window of the nodes, when they are allowed to modify their out-links to point to more attractive nodes.

Attachment-based model with clustering

We run simulations on our proposed model with $\alpha = 0.1$, $\beta = 0.2$, 10 categories, and 40,000 time steps. We draw the category weights from the uniform distribution on $(0,1)$. We then perform calculations on our model and compare their values with those of the Web. Specifically, we calculate the in-degree and out-degree distributions, the maximum strongly connected component size, the PageRank distribution, and the in- and out-degree distributions for the categories.

The in-degree and out-degree distributions of the graph are plotted above on logarithmic axes. Despite deviations from the Bollobas model, our model still produces power-law distributions for the in- and out-degrees. Furthermore, the PageRank distribution nearly resembles a power law, another property that makes it similar to the Web (Becchetti, Castillo, Donato, Leonardi, & Baeza-Yates, 2006). Below are other statistics from our simulation.
Our model successfully creates a bow-tie structure, with the majority of the bow-tie lying in the strongly connected component (SCC). Category degrees, defined as the number of links formed by two nodes from different categories, are also shown above. About two-thirds of all links generated are formed by two nodes from different categories, indicating rich interaction between categories. The remaining third of the links are formed within a category, suggesting that there is also clustering behavior within categories.

One of the biggest flaws of our model is the fixed number of categories. Our model does a good job forming a cluster for each category. However the number of clusters on the web increases with time, a property not reflected in our model. Another flaw is the disproportionate distribution of nodes in the bow tie; there are too few nodes in the OUT component, and too many in the SCC. The analysis of the web by Broder et al. in 2000 found that the SCC contained approximately 27.7% of the nodes, and IN and OUT were approximately the same size, containing 21.3% and 21.2% of the nodes, respectively. Finally, the degree distributions have a
exhibit power-law distributions with an exponent of approximately 1. This deviates from the observed power-law degrees of the Web: 2.1 for in-degree distribution and 2.7 for out-degree (Pennock, Flake, Lawrence, & Glover, 2002).

Still, this model considers many aspects of the Web that have not been explored much in previously-studied graph generation models. First off, this model takes into account natural clustering and categorization in the Web, a feature that is absent in many previous models. Furthermore, the graph topology in general looks more like the Web: multiple components can and do arise from this model, and a bow-tie structure emerges.

Conclusions

Our models successfully incorporate additional features of the Web, such as link mutation, the bow-tie structure, and clustering, which are unaccounted for in many previous models. Even with our additional complexity in the models, they still produce the power-law degree distributions which arise naturally in information networks of all types. In future work, more extensions of these models can be considered: for example, a new model could increase the number of categories as the graph grows, as increasing the number of clusters over time steps is more representative of Web growth. A model could also incorporate different degree distributions within clusters. Extension and improvements to our model can also be facilitated with mathematical analysis. Our models successfully modeled a few additional features of the Web, but there are always more possibilities to consider as we look for better models and understanding of the Web and processes that govern its structure.

References


Broder, A., Kumar, R., Maghoul, F., Raghavan, P. R., Stata, R. T., & Wiener, J. (2000). Graph structure in the web. 9th International WWW Conference.

