

An Empirical Validation of Growth Models for Complex Networks

Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi,
Peter Druschel, and Bobby Bhattacharjee

1 Introduction

Complex networks arise in a variety of different domains, including social networks [2], Internet topologies [11], Web connections [3], electrical power grids [30], and networks of brain neurons [6]. Despite their disparate origins, these networks share a surprising number of common structural features such as a highly skewed (power-law) degree distribution, small diameter, and significant local clustering. The link structure of these networks has received significant research attention, and the resulting understanding of their structure has led to popular search algorithms like PageRank [28] and HITS [15].

In this paper, we wish to understand the dynamic processes that lead to the observed structures. A number of different network growth models have been proposed—e.g., the preferential attachment [5], the random walk model [32], and the common neighbors model [26]—which all lead to graphs with similar structural properties. Unfortunately, none of these growth models have been validated using large-scale data from real networks. It is not known if the models predict how real networks actually grow; that is, how, when, and where links are added to a network.

A. Mislove (✉)
Northeastern University, Boston, MA, USA
e-mail: amislove@ccs.neu.edu

H.S. Koppula
Cornell University, Ithaca, NY, USA
e-mail: hema@cs.cornell.edu

K.P. Gummadi • P. Druschel
Max Planck Institute for Software Systems, Kaiserslautern–Saarbruecken, Germany
e-mail: gummadi@mpi-sws.org; druschel@mpi-sws.org

B. Bhattacharjee
University of Maryland, College Park, MD, USA
e-mail: bobby@cs.umd.edu

In his editorial on the future of power-law networks research, Mitzenmacher [25] argues that power-law research must move from observing and modeling power-law behavior to the challenging problem of model validation.

In this paper, we apply growth data from four different real-world networks towards understanding the processes that lead to the observed structures. We have repeatedly crawled large social networks (Flickr and YouTube) daily for over four months. Our crawls have generated datasets that contain 10 million and 14 million new links in Flickr and YouTube, respectively. We have also downloaded and analyzed the link formation history between pages in the English language Wikipedia. This dataset covers over six years of growth history, encompassing almost 40 million links between almost 2 million pages. Finally, we use successive snapshots of the autonomous system (AS) level Internet topology graph to observe inter-AS links being created. This dataset covers over three years of AS topology evolution, representing over 75,000 new links.

Using these four datasets, we investigate the link formation processes in each of these networks. Our primary goal is to understand the underlying processes that lead to the observed static structural properties. We examine a number of previously proposed models, and test how well the properties of links created by the models match our observed data. In particular, if these properties match, then we have some assurance that the model might describe the underlying growth process in the “real” network (some other models may also describe the same process). However, if not, then we can assert that the model does *not* describe the dominant growth mechanism in the network under study.

Our analysis shows that links are created by nodes in direct proportion to their degree (as predicted by the preferential attachment model) and that nodes in directed networks tend to quickly respond to incoming links by creating a link in the reverse direction. However, we show that the preferential attachment mechanism alone is insufficient in explaining how a node picks the recipients for its links. Rather, our experiments point to a strong proximity bias: Nodes tend to connect to nearby nodes in the network much more often than would be expected by using preferential attachment alone. Additionally, we find that among the mechanisms which use local rules to form new links, those which select new links based on the indegree of the destination tend to have higher accuracy.

We believe our work is an important first step towards empirical validation of the processes underlying network formation and growth. It is likely that the simple proximity bias models we have explored will not capture the fine-grained dynamics of real systems since these depend on domain-specific parameters. However, our work is directly useful in constructing networks that reflect both global and local characteristics of real-world networks. Better structural and growth models are also useful for network analysis and planning. For appropriate networks, such models can be used in the design of search algorithms (e.g., by pre-identifying nodes that are likely to be hubs), in data mining (e.g., by identifying potential nodes for placing data monitors), and in system evaluation (e.g., by allowing networks to be simulated at arbitrary sizes).

2 Background and Related Work

In this section, we provide background on work related to complex information networks.

2.1 *Static Structure of Complex Networks*

A long thread of research examines the structure of various complex networks. Researchers have shown that many real-world networks are *power-law networks*, including Internet topologies [11], the Web [5, 20], social networks [2], neural networks [6], and power grids [30]. In such networks, the probability that a node has degree k is proportional to $k^{-\gamma}$. In addition to power-law degree distributions, these networks have been observed to share a number of common structural properties, such as a small diameter and significant local clustering. For more detail on these networks, we refer the reader to the survey by Newman [27].

2.2 *Structural Growth Models*

Researchers sought to explain the intriguing similarity in the high-level structural properties across networks of very different scales and types by hypothesizing that the networks are the result of a few common growth processes at work. Many models of these processes have been proposed and analyzed to explain the generation of complex networks.

The well-known preferential attachment model [5], where nodes acquire links in proportion to their current degree, has been shown to result in power-law networks. Preferential attachment, as proposed by Bárábási, is a global process whereby nodes create new links based only on the degree of the destination. Many extensions to the preferential attachment model have been proposed, such as to add a tunable level of clustering [13].

Another class of models that produce power-law networks is based on local rules, such as the random walk model [31, 32], where nodes select new neighbors by taking random walks; the common neighbors model [26], where nodes select new neighbors by picking nodes with whom they share many friends in common; and the finite memory model [16], where nodes eventually become inactive and stop receiving any new links.

Additionally, Eiron and McCurley [10] note that many complex information networks have a natural hierarchical structure (such as the hierarchical nature of URLs for the pages in the Web). They propose a new model for constructing such networks which takes into account this hierarchical structure, and they show that

this approach more closely matches the observed networks' link structure. For a more detailed treatment of all of these models and others, we refer the reader to Mitzenmacher [24].

It is important to note here that these processes are, by and large, intuitive models that can explain the observed structural properties of the networks. But, they have not been validated using empirical data and they have not been shown to occur in practice. Mitzenmacher [25] poses this as one of the biggest challenges facing the future of power-law research. One of the contributions of this paper lies in evaluating how well these processes predict what actually occurs in different real-world networks at scale.

2.3 *Empirical Validation of Growth Models*

Some recent work compared snapshots of the same network at different points in time to verify the growth processes. Newman [26] examined the properties of two scientific collaboration networks and found evidence of preferential attachment in both. Peltomäki and Alava [29] examined a scientific collaboration network and a movie-actor network and found evidence of sublinear preferential attachment. Jeong et al. examined citation and coauthorship networks and found that nodes received links in proportion to their degree. Nowell et al. [22] investigated coauthorship networks in physics to test how well different graph proximity metrics can predict future collaborations.

Our work shares similar goals and methodology as the above studies. However, the datasets we use are orders of magnitude larger than the ones used before. Moreover, our data allows us to analyze network growth at very small time scales. We analyze daily snapshots of Flickr and YouTube networks and weekly snapshots of the Internet topology. For Wikipedia, we have sufficient data to create the snapshot of the network at the precise time a new link is established. Since the growth models rely solely on the current network structure to predict new link formation, having frequent snapshots of the network is crucial to validating the models with high accuracy.

Other work has studied the high-level properties of graph evolution, looking for evolution trends at the global level. For instance, Leskovec et al. [21] examined the evolution of a number of real-world graphs, including collaboration networks and recommendation networks. They found that the graphs tend to densify over time and that the average path length shrinks over time (instead of growing in proportion to the number of nodes). This line of work is largely complementary to our work, as we focus on the local link formation phenomena which might lead to these global observations.

Analysis of our detailed growth data, reveals that the preferential attachment model by itself cannot explain new link formation. We believe that the datasets we gathered (and plan to release publicly) represent a significant first step forward in the creation and validation of generative models for complex networks.

2.4 *Explanatory Growth Models*

Some recent studies, particularly on online social networks, have proposed explanatory models of the network growth. Unlike structural growth models, which try to model growth solely as a function of the network structure, explanatory models seek to account for the underlying sociological factors that cause the links to be established. For example, an explanatory growth model for Flickr, a photo-sharing social network, would be based on an understanding of how users behave when sharing pictures.

Examples of work on explanatory growth models include Kumar [19], who divided users into ones who are active and passive and presented a model describing their behavior. Kumar et al. also observed the early evolution of the blogosphere [18] and found that it is rapidly increasing in both scale and connectedness. Jin et al. [14] presented a model of social networks based on known human interactions. Backstrom et al. [4] looked at two snapshots of group membership in LiveJournal and presented a model for the growth of user groups over time. Kossinets and Watts [17] used an inferred social network from an email trace to show that new links in the network are more likely to be established between nodes close in the network. Finally, Chang et al. [9] proposed a model for the growth in connectivity of the Internet topology.

Compared to structural growth models, explanatory models are more detailed, but they also tend to be specific to the network being investigated. For example, the reasons why ISPs connect to each other in the Internet topology are very different from the reasons why users in Flickr connect to each other. By being agnostic to these factors, structural growth models are inherently less accurate. But they are far more general and can be compared across different types of networks. In this paper, we focus only on structural growth models.

3 *Measurement Methodology*

We now describe the data presented in this paper and the methodology we used to collect it. Whenever appropriate, we describe limitations of the measurement methodology.

3.1 *Flickr and YouTube*

We begin by describing in detail the methodology for collecting data on online social networks. For the two social networks we consider, we were unable to obtain data directly from the respective site operators. So, we chose to crawl the user graphs using the public Web interface provided by the sites. Below, we first describe the challenges and limitations of obtaining data in this manner, and then we describe the datasets we collected.

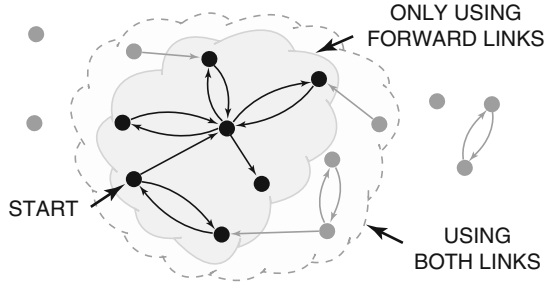


Fig. 1 Users reached using different links. Using only forward links crawls the inner cloud; using both forward and incoming links crawls the entire WCC (dashed cloud)

Crawling the Entire Graph

The primary challenge in crawling large graphs is covering the entire connected component. At each step, one can generally only obtain the set of links into or out of a specified node. In the case of online social networks, crawling the graph efficiently is important since the graphs are large and highly dynamic. Common algorithms for crawling graphs include breadth-first search (BFS) and depth-first search.

Crawling directed graphs, such as Flickr and YouTube, presents additional challenges over undirected graphs. In particular, most graphs can only be crawled by following links in the forward direction (i.e., one cannot directly determine the set of nodes which point *into* a specific node). Using only forward links does not necessarily crawl an entire weakly connected component (WCC); instead, it explores the connected component reachable from the set of seed users. This limitation is typical for studies that crawl online networks, including measurement studies of the Web [7].

Figure 1 shows an example of a directed graph crawl, where the users reached by using just forward links are shown in the inner cloud, and those discovered using both forward and incoming links are shown in the outer dashed cloud. Using both forward and incoming links allows us to crawl the entire WCC, while using only forward links results in a subset of the WCC. Both Flickr and YouTube can only be crawled using forward links.

Crawling Methodology

Using automated scripts on a cluster of 58 machines, we crawled the social network graphs of Flickr and YouTube once per day. We chose these sites because they represent different types of online social networking sites and because it is possible to crawl the entire network once per day. More details on our methodology and its limitations can be found in [23]. Here, we discuss the methodology and limitations that are relevant to the growth data.

We started each crawl by selecting a single known user as a seed. In each step, we retrieved the list of friends for a user we have not yet visited and added these users to the list of users to visit. We then continued until we exhausted the list, thereby performing a BFS of the social network graph, starting from the seed user. We performed these crawls once per day for each network. On each day, we revisited every user we had previously discovered, in addition to all nodes that were reachable from the seed node, and recorded any newly created or removed links and nodes.

Since the sites do not provide the time of creation for any node or link, our growth data for the social networks has a granularity of one day for the links we observed being created. As a result, we cannot determine the exact time of link creation or the order in which links were created within a single day. Moreover, new nodes cannot be observed until they become connected to one of the nodes we have already crawled. Additionally, in the rest of the paper, we only examine links that we observed being created. In other words, we may discover a new node that has a few established links, but we do not examine these previously established links in our growth analysis, as we did not observe them being created.

Flickr

Flickr (www.flickr.com) is a popular photo-sharing site based on a social network. Flickr's social network is directed, as users can create links to other users without any approval from the destination. Flickr exports an API that we used to conduct the crawl.

We crawled the Flickr network daily between November 2, 2006, and December 3, 2006, and again daily between February 3, 2007, and May 18, 2007, representing a total of 104 days of growth. During that period of daily growth observations, we observed over 10.7 million new links being formed and discovered over 680,000 new users. This represents, relative to the initial network snapshot, over 42% growth in the number of users and over 63% growth in the number of links.

YouTube

YouTube (www.youtube.com) is a popular video-sharing site that includes a social network. Similar to Flickr, YouTube exports an API, and we used this feature to conduct our crawls.

We crawled the YouTube network daily between December 10, 2006, and January 15, 2007, and again daily between February 8, 2007, and July 23, 2007, representing 201 days of growth. Between the two date ranges of our crawls, YouTube changed its policy to require confirmation from the destination of a link (previously, this approval was not required). Thus, YouTube changed from a directed network to an undirected network between our two observation periods. To properly deal with this significant change in policy, we treat the two YouTube networks separately—we denote the first set of growth data covering the directed graph as *YouTube-D* and the second set representing the undirected graph as *YouTube-U*.

The YouTube-D dataset represents the growth of a directed network over a period of 36 days. During that period of daily growth observations, we observed over 540,000 new links being formed and discovered over 130,000 new users. This represents, relative to the initial network snapshot, over 13% growth in the number of users and over 12% growth in the number of links.

The YouTube-U dataset represents the growth of an undirected network over a period of 165 days. We observed the network grow by over 11.7 million links and over 1.8 million users. This represents, relative to the initial network snapshot, over 129% growth in the number of users and over 173% growth in the number of links.

Crawling Limitations

There are two limitations to our crawls of Flickr and YouTube. First, we were only able to crawl using forward links, which does not necessarily result in an entire WCC. Second, we only crawled the single, large WCC; there may be users who are part of small clusters not connected to this WCC at all. In this section, we evaluate the number and characteristics of users who were missed by our crawl.

We performed the following experiment using Flickr. We used the fact that the vast majority of Flickr user identifiers take the form of *[randomly selected 8 -digit number]@N00*. We generated 100,000 random user identifiers of this form (from a possible pool of 90 million) and found that 6,902 (6.90%) of these were assigned usernames. These 6,902 nodes form a random sample of Flickr users.

Among these 6,902 users, 1,859 users (26.9%) had been discovered during our crawl. Focusing on the 5,043 users *not* previously discovered by our crawl, we conducted a BFS starting at each user to determine whether or not they could reach our set of previously crawled users. We found that only 250 (5.0%) of the missed users could reach our crawled set and were definitively in the WCC. While we cannot conclusively say that the remaining 4,793 (95.0%) missed users are not attached to the WCC (there could be some other user who points to them and to the WCC), the fact that 89.7% of these have no forward links suggests that many are not connected at all.

Thus, we believe that our crawl of the large WCC, although not complete, covers a large fraction of the users who are part of the WCC. Further, our experience with the randomly generated Flickr user identifiers indicates that (at least for Flickr) the nodes not in the largest WCC tend to have very low degree—in fact, almost 90% of them have no outgoing links at all.

3.2 Wikipedia

Wikipedia (www.wikipedia.org) is a popular online encyclopedia that allows any user to add or edit content. Wikipedia makes its entire edit history available on a monthly basis, and we downloaded the edit history of the English language Wikipedia as of April 6, 2007.

To extract the graph of links between Wikipedia pages, we use the following method: For each link in the current snapshot, we determine the time when this link was first created. We then construct a graph using these derived links and the associated timestamps. This method allows us to remove the effects of page vandalism, where malicious users sometimes overwrite entire pages, thereby temporarily removing all of the links from vandalized pages.

Since Wikipedia allows pages to redirect to other pages, we configured our tool to follow the redirects and treat a link to a redirect page as if it was a link to the destination page. Thus, if page *A* originally linked to *B* at time *t*, but later, *B* was set to redirect to *C*, we treat this like a link from *A* to *C* established at time *t*. This allows us to handle multiple layers of redirect pages, as well as large-scale naming convention changes.

Since the data represents the complete history of a complex network, we exclude startup effects by limiting our analysis to the recent history. This is similar to previous studies [8, 26]. In particular, we only consider links created between January 1, 2005, and April 6, 2007, a period of 826 days. During this period, we observed over 1.1 million new pages and over 33 million new links, representing 169% growth in the number of pages and 500% growth in the number of links relative to the snapshot on January 1, 2005.

3.3 Internet Topology

The Internet can be viewed as a collection of *autonomous systems* (AS), where each AS represents a single administrative domain (typically, an ISP). The inter-domain routing protocol of the Internet, BGP, uses unique AS numbers to allow ASes to advertise their connections to their neighbors. The union of these advertisements forms an undirected graph representing the AS-level connectivity of the Internet.

We used the AS topology graphs collected by CAIDA [1] to study the evolution of the AS network. CAIDA creates weekly (monthly for the first two years) snapshots of the AS topology using a number of BGP monitoring machines. We downloaded the entire history of their measurements, which covers the period from January 5, 2004, until July 9, 2007. The AS topology evolution data therefore covers 1,282 days of growth. During this period, the number of ASes in the network grew from 9,978 to 25,526, a growth of 155%. Similarly, the number of AS links grew from 29,504 to 104,824, a growth of 255%.

3.4 Summary

Table 1 shows the high-level statistics of the data we gathered. The network sizes vary by over three orders of magnitude. Similarly, other metrics, such as the average number of links per node and the yearly growth rate also vary greatly between the

Table 1 High-level statistics of the network growth data

Network type	Flickr		Wikipedia		YouTube-D		YouTube-U		Internet	
	Directed		Directed		Directed		Undirected		Undirected	
Days of observed growth	104		825		36		165		1,281	
Resolution of link creation	Day		Second		Day		Day		Month/week	
Fraction of links symmetric	62%		17%		79%		–		–	
Initial number of nodes	1,620,392		695,353		1,003,975		1,402,949		9,978	
Final number of nodes	2,570,535		1,892,691		1,137,638		3,218,658		25,526	
Growth in number of nodes	42%		169%		13%		129%		155%	
Norm. growth rate (nodes/year)	242%		54%		145%		525%		31%	
Initial number of links	17,034,807		6,637,456		4,391,336		6,783,917		29,504	
Final number of links	33,140,018		39,953,145		4,945,382		18,524,095		104,824	
Growth in number of links	63%		500%		12%		173%		255%	
Norm. growth rate (Links/Year)	455%		120%		215%		822%		43%	

networks. Despite these differences, as our analysis later shows, the growth of these complex networks shows a number of commonalities.

4 Validation of Network Growth Models

Our goal in this section is to use our network growth data to validate existing models of network growth. In particular, we study how well the empirical data matches the predictions of growth models that have been proposed. Our findings can be summarized as follows. First, all of our data is consistent with the predictions of the popular preferential attachment model. Second, there are some properties in our datasets that cannot be explained by that model alone. Third, models that consider network proximity as a factor in link creation predict the empirical data better than preferential attachment. Fourth, no single proximity-based model best predicts link creation in all four of our datasets, but those which take into account the indegree of the destination tend to have higher accuracy. Fifth, reciprocation is a significant factor in the link creation of directed networks.

It is important to note that we can only study how well a particular model predicts the link creation that occurs in the empirical data. We fundamentally do not know why new links were established; we can only observe the source and destination of new links. Thus, we cannot ultimately prove or disprove any particular model; we can only examine the correlation between the observed data and what each model would predict. Nevertheless, knowing how well different models predict link creation in the data can improve our understanding of network evolution and can provide clues as to the actual underlying processes.

4.1 *Growth Dominates Network Evolution*

In all of the networks we examined, we found that link addition was significantly more frequent than link removal. In particular, we found that in Flickr, link additions exceeded link removals in our datasets at a rate of 2.43:1. Similar characteristics were observed in the other networks we studied: In YouTube-U, the ratio of link additions to removals was 3.71:1, and in the Internet, we found that the ratio was 2.06:1. Unfortunately, we did not record removed links for the YouTube-D dataset, and we are unable to estimate the fraction of removed links in Wikipedia due to the effects of page vandalism (i.e., vandalized pages often have their entire text, and therefore all of their outgoing links are replaced and then added back).

In summary, in the networks in which we were able to record link removals, we observed that link addition significantly exceeded link removal. Thus, in the rest of this paper, we focus only on how links are added to growing networks, and we leave examining link removal for future work.

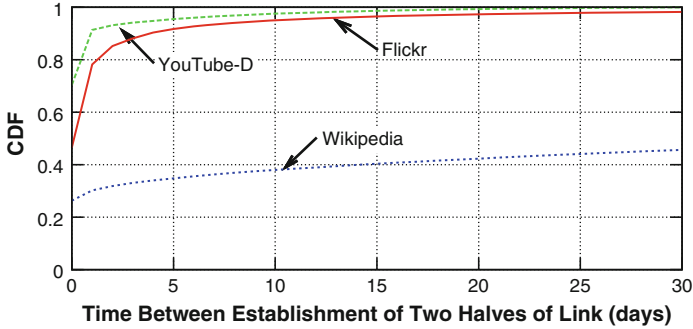


Fig. 2 CDF of time between establishment of the two directed links of a symmetric link. In both Flickr and Youtube, links are quickly reciprocated

All of the networks we observed showed a high growth rate: Normalizing for different observation periods across the networks reveals an average growth rate of between 31% and 525% per year in terms of nodes and a growth rate of between 43% and 822% per year in terms of links. These rapidly growing networks offer us a unique opportunity to observe new link creation.

4.2 Reciprocation

We begin by first examining *reciprocation*, a growth mechanism that exists only in directed graphs. Reciprocation occurs when the creation of a directed link between two nodes causes the reverse link to be established. Since undirected graphs are, by definition, symmetric, reciprocation does not make sense in the context of undirected graphs. Reciprocation has been proposed as an independent growth mechanism for large-scale directed graphs [12, 33].

Since we do not know why links were established, we rely on the timing between the creation of the two directed links of a symmetric link to guess whether the creation of the first causally affected the second. Figure 2 shows the distribution of the time between the establishment of the two links of a given symmetric link in the three directed graphs (Flickr, YouTube-D, and Wikipedia) that we studied.

From Fig. 2, it is clear that in the two social networks we observed, users often respond to incoming links by quickly establishing a reciprocal link back to the source node. In fact, over 83% of all symmetric links we observed in both Flickr and YouTube-D were established within 48 hours after the initial link creation. This suggests that users tend to quickly reciprocate links, if they reciprocate at all. Thus, it is highly likely that the establishment of the first link in these networks prompted the creation of the reciprocal link. The Wikipedia data, on the other hand, indicates a lower degree of reciprocation; only 30% of the symmetric links in Wikipedia had both halves of the link created within 48 hours of each other.

Our data suggests that reciprocation is an independent mechanism shaping the growth of directed networks. The degree of reciprocation is dependent on the network: the two social networks show significant reciprocation, while Wikipedia shows reciprocation, but to a less significant degree.

4.3 Preferential Attachment

Preferential attachment [5], colloquially referred to as the “rich get richer” phenomenon, is a growth model in which new links in a network are attached *preferentially* to nodes that already have a large number of links. Under preferential attachment, the probability that a new link attaches to a given node is proportional to the node’s current degree.

To examine whether preferential attachment predicts the observed growth data, we calculated how the number of new links per day varies with the node degree. If preferential attachment is taking place, we would expect to see a linear correlation between the degree of a node and the number of new links it creates or receives. However, it is important to note that a linear correlation is a necessary but not sufficient condition for the validity of the preferential attachment mechanism, as other mechanisms could also result in such a linear correlation. For example, the “connecting nearest neighbors” model [32] has been shown to also exhibit such a linear correlation.

Figure 3 plots this distribution in log–log scale for each of the five networks we studied. For the three directed graphs, we separately plot the number of new links created and received, with respect to the node’s current outdegree and indegree.

Undirected Networks

For the two undirected networks, YouTube-U and the AS-level Internet, we show how the degree of a node correlates with the number of new links per day. We find a strong linear correlation between the current degree and the number of newly created links in both of the networks.

Directed Networks

For the three directed networks, we separate the preferential attachment model into two aspects: *preferential creation* and *preferential reception*. Preferential creation describes the mechanism by which nodes *create* new links in proportion to their outdegree, and preferential reception describes the mechanism where nodes *receive* new links in proportion to their indegree. This distinction is consistent with previously proposed models of preferential attachment on directed graphs [8].

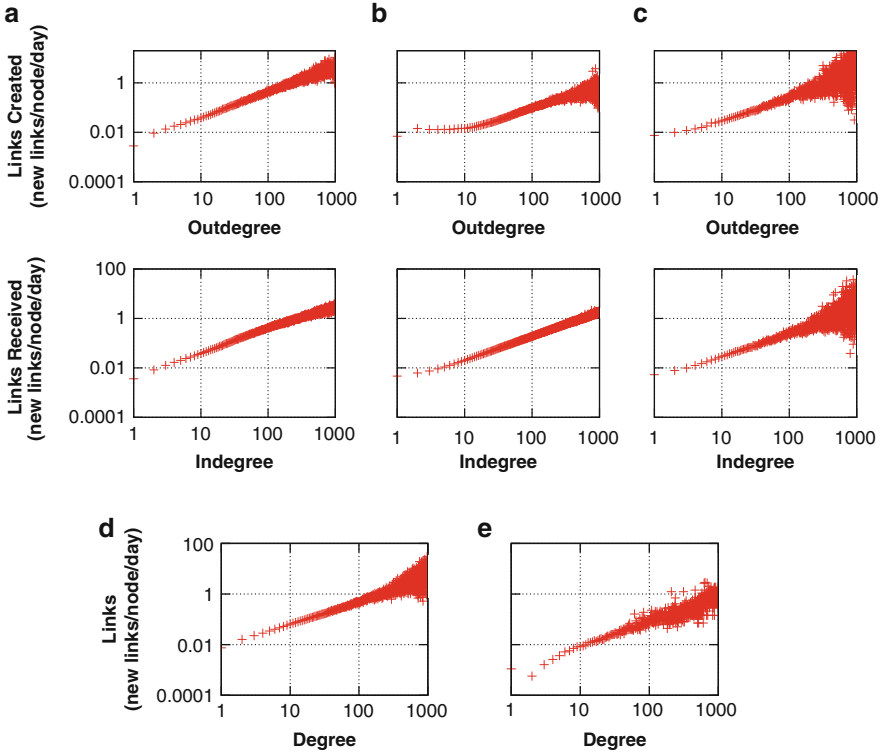


Fig. 3 Log-log plot of degree versus number of new links per day. For directed graphs (graphs a–c), separate plots are shown for outdegree (top) and indegree (bottom). All networks show strong evidence of preferential attachment. (a) Flickr, (b) Wikipedia, (c) YouTube-D, (d) YouTube-U, (e) Internet

It is important to understand why we separate preferential attachment into preferential creation and preferential reception for directed networks. Preferential attachment was originally defined for undirected graphs [5] and therefore does not distinguish between node indegree and outdegree. However, in the directed networks we study, link creation is very different from link reception. Nodes are in complete control over their outgoing links, since they decide who they link to, but they are not in control of their indegree, since it depends upon who they receive links from.

For the three directed networks, Flickr, Wikipedia, and YouTube-D, we separately examine how the current outdegree and indegree of a node is related to the number of newly created and received links per day. Figure 3 shows that the outdegree of nodes is linearly correlated with the number of new links created per node per day. This is a necessary, but not sufficient condition for the validity of the preferential creation mechanism. Figure 3 also shows, for the three directed networks, that the increase in node indegree is linearly correlated with the current

indegree of the node. Similarly, this is a necessary condition for the validity of the preferential reception mechanism.

Discussion

Our data shows that a necessary condition for preferential attachment, a linear correlation between the degree of a node and the number of new links, is present in all five networks. However, this alone is not sufficient to claim that preferential attachment *is* the mechanism that is causing the growth, as other, different mechanisms could also result in this correlation. In the next section, we examine more closely the growth data to look for further evidence of preferential attachment.

4.4 Proximity Bias in Link Creation

In this section, we consider the distance in the network among the nodes connected by a new link and consider if preferential attachment can explain our observations. In particular, we examine the shortest path distance between the source and destination of newly created links, before a new link is created between them. If preferential attachment is the underlying mechanism, then the observed distance distribution between nodes should match that predicted by the model.

Over 50% of the links in all five networks are between nodes that have, a priori, some network path between them.¹ For these links, Fig. 4 shows the cumulative distribution of shortest-path hop distances between source and destination nodes for newly created links. It reveals a striking trend: Over 80% of such new links in Flickr connect nodes that were only two hops apart, meaning that the destination node was a friend of a friend of the source node. Similarly, this fraction is over 42% in YouTube-D, over 50% in Wikipedia, over 45% in YouTube-U, and over 57% in the Internet topology.

One might wonder whether in small diameter networks like the ones we observe, this high level of proximity in link establishment is simply a result of preferential attachment. This is plausible, since the high-degree nodes that preferential attachment prefers tend to be close to many nodes. To test this hypothesis, for each newly created link, we computed the expected distance from the source to the destination, if the destination is chosen using the preferential attachment mechanism (or preferential reception, for directed graphs). Figure 4 also plots this distribution for each network.

In all five networks that we study, the observed distances between the source and destination of links show a significant bias towards nearby nodes, relative to what preferential attachment or preferential reception would predict. In fact, in Flickr, Wikipedia, and YouTube-D, we found that the number of new links connecting

¹For directed networks, we only count directed paths.

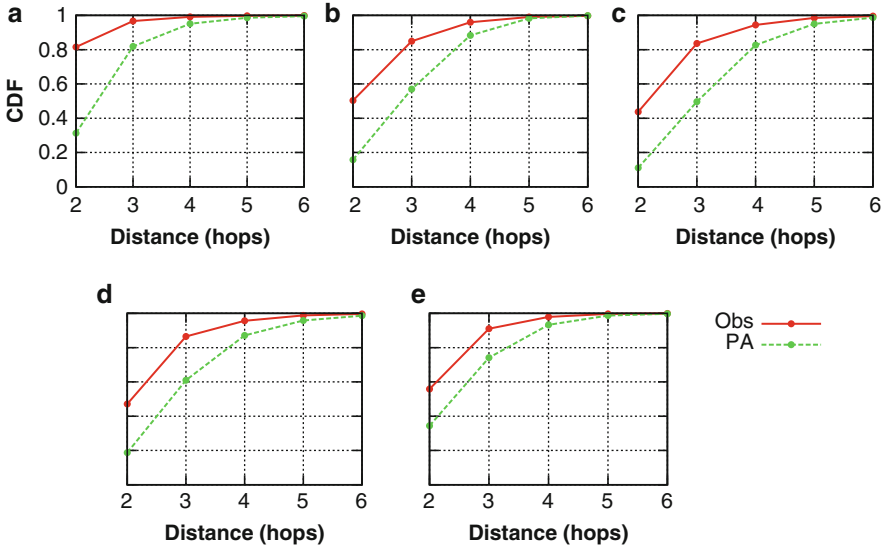


Fig. 4 CDF of distance between source and destination of observed links (*Obs*). Also shown is the expected CDF from preferential attachment (*PA*). The numbers in parenthesis are the fraction of all new links connecting nodes that had, a priori, some path between them. All networks show a proximity bias that is not predicted by preferential attachment. (a) Flickr (83%), (b) Wikipedia (58%), (c) YouTube-D (68%), (d) YouTube-U (82%), (e) Internet (54%)

2-hop neighbors in the empirical data exceeded that predicted by preferential reception by a factor of three.

This result shows that the new links created in the networks cannot be explained by a preferential attachment mechanism alone. Nodes are far more likely to link to nearby nodes than preferential attachment would suggest. This result is consistent with the previous observations on static networks which showed that the clustering coefficient was significantly higher than would be predicted by preferential attachment. In the next section, we focus on how nodes choose which nearby node to link to.

4.5 Mechanisms Causing Proximity Bias

Next, we examine network growth models that are known to have a stronger bias towards proximity than preferential attachment. To make the analysis tractable, we focus on new links that occur between nodes that are two hops apart. Such links account for over 45% of the links in all networks. We consider preferential creation, combined with five different proposed mechanisms for selecting the destination of a newly established link:

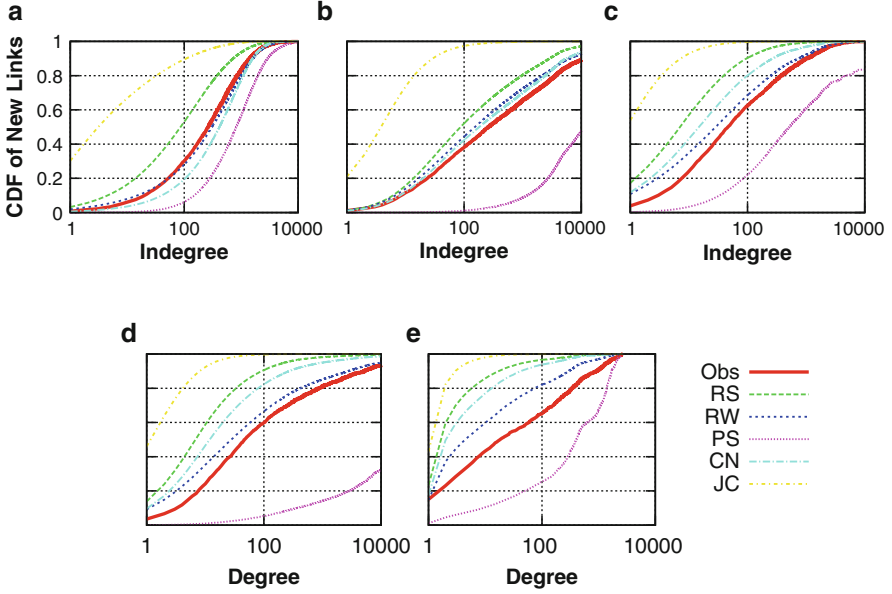


Fig. 5 CDF of nodes receiving new links by indegree. Plots are shown for observed data (Obs) and simulated mechanisms: random selection (RS), random 2-hop walk (RW), preferential selection (PS), common neighbors (CN), and Jaccard's coefficient (JC). The observed data does not match any one mechanism, suggesting that different mechanisms are at play in different networks. (a) Flickr (b) Wikipedia (c) YouTube-D (d) YouTube-U (e) Internet

- *Random selection (RS)*, where a node chooses the destination randomly from its set of two-hop neighbors. This mechanism serves as a baseline for evaluating the other mechanisms.
- *Random two-hop walk (RW)*, where a node performs a random two-hop walk to find the destination [32].
- *Preferential selection (PS)*, where a node chooses from its set of two-hop neighbors preferentially according to the nodes' indegrees. This is similar to preferential attachment, except that a node only considers its two-hop neighbors [22].
- *Common neighbors (CN)*, where a source makes a weighted random choice among its set of two-hop neighbors. The likelihood that a given candidate is chosen is proportional to the number of neighbors the source shares with the candidate [26].
- *Jaccard's coefficient (JC)*, where a source makes a weighted random choice among its set of two-hop neighbors. Here, the likelihood that a given candidate is chosen is proportional to the number of neighbors the source shares with the candidate divided by the candidate's indegree [22].

We examined newly established links in all networks that connect nodes that were previously two hops apart. We then calculated the expected indegree distribution of

Table 2 *Prediction accuracy of two-hop link creation mechanisms relative to the baseline random selection mechanism. While no one mechanism appears to be the most accurate across all networks, random walk and preferential selection tend to have higher accuracy*

	RS	RW	PS	CN	JC
Flickr	0.17%	2.0	1.1	1.2	1.2
Wikipedia	0.15%	2.9	2.9	1.3	0.7
YouTube-D	0.35%	1.6	1.5	1.1	1.0
YouTube-U	0.59%	1.7	1.3	1.1	1.4
Internet	0.53%	1.9	4.1	1.1	0.5

nodes that would have been selected using each of the five mechanisms above. We then compared the results to the distribution in the empirical data. Figure 5 plots these distributions for each network.

From Fig. 5, we can see that no one mechanism closely matches the empirical data in all networks. In fact, in two of the networks (Flickr and Wikipedia), the random walk mechanism most closely matches the observed data. However, in the other three networks, the results are less conclusive. To better quantify how well the various mechanisms predict the selected destination of new links, we calculated the accuracy of each mechanism, in the same manner as previous studies [22]. Thus, for each newly created link, we calculated the fraction of time each mechanism correctly predicted the selected destination. The results are shown in Table 2, relative to the random selection model.

The accuracy results in Table 2 show that no one model dominates in terms of accuracy across different networks. However, closely examining the results reveals that the two mechanisms which take into account the indegree of the destination (RW and PS) do tend to have higher accuracy. This suggests either that different mechanisms may be at play in different networks or that the actual mechanism driving link creation is not among the ones we evaluated or that the actual mechanism is a complex combination of some of the mechanisms we tested. This result is not surprising, though, as each of the networks represents a different system, and it is unlikely that one single mechanism would describe the link creation behavior in all of them.

4.6 Summary

In this section, we closely examined network growth data from five different networks and compared the empirical data to the predictions of previously proposed growth mechanisms. We found evidence of reciprocation as a mechanism in directed networks. We also found that nodes tend to create and receive links in proportion to the outdegree and indegree. However, we found that the preferential attachment

mechanism did not accurately predict the proximity bias among nodes connected by new links in any of the empirical datasets. All networks showed a stronger bias towards proximity between new sources and destinations than would have been predicted by preferential attachment. Upon closer examination of these links, we found that no single proximity model we examined appears to accurately predict this proximity across all networks, suggesting that further research into growth mechanisms is necessary. In the next section, we discuss some of these future directions and describe the implications of our findings.

5 Discussion

In this paper, we have used empirical growth data from multiple large-scale complex networks to test if previously proposed growth models actually are at play in these networks. We have chosen to focus on the preferential attachment model because it is simple and has been suggested as the underlying growth mechanism in different contexts. Clearly, preferential attachment leads to global degree distributions of the type observed in many diverse networks, and absent other data, it is an attractive choice for researchers to explain static snapshots of crawled networks.

5.1 *Is Proximity Fundamental?*

We believe that some notion of proximity is inherent in the link creation processes underlying large networks. As a network grows larger, it is increasingly unlikely that nodes are influenced by knowledge of the global degree ranking when choosing their neighbors. In many networks, it may not even be possible to discover the global degree ranking of nodes, knowledge of which is required for pure preferential attachment. Other mechanisms that rely on global properties are equally unlikely because of technical and policy issues with computing global metrics.

In the networks we have examined, the bias towards proximity can be explained by considering the node discovery mechanisms available to users and the factors that constrain them. In the social networks (YouTube and Flickr), the primary mechanism available to users for exploring the network is to walk their neighborhood. This might explain our observation in Flickr and YouTube that there is a much stronger bias in link creation towards nearby nodes than would be predicted by preferential attachment alone, yet there still is a bias towards high-degree nodes (see Table 2). On Wikipedia, semantically closer pages are likely to be proximal in the network, leading to a proximity bias in link creation.

The Internet AS graph is fundamentally different because each AS consists of many different routers and there is a significant cost associated with creating new links. A model for AS link formation is given in [9], and our observations are consistent with the reasoning therein. The AS graph is naturally “tiered” with many

small stub ASes interconnected by a few large backbone providers (who also tend to have high connectivity/degree). AS link creation is often constrained by financial, technical, and geographical factors: For most stub ASes, links to far away ASs tend to be costly (especially if the geographic distance is large) and are unlikely to be profitable since the upstream provider already provides transit to reach these ASes. Such links only make sense in specific cases where business relationships mandate a specific inter-AS peering. Thus, stub ASes tend to connect to their nearby backbone AS providers, and the resulting AS graph shows proximity bias coupled with strong preferential selection.

5.2 Proximity Mechanisms

While our growth data cannot *assert* which mechanism is at play when links are formed, it can be used to *disprove* existing hypotheses. Perhaps unsurprisingly, we find that the simplest mechanisms (such as global preferential attachment) are not sufficient to explain our observations. In particular, we have shown that to explain the empirical growth data, we must include some notion of proximity in the growth models. While proximity has been previously suggested as a factor in link creation in large networks, we believe we are the first to provide empirical data from multiple large-scale networks to support this conjecture.

The analysis in the previous section revealed some insights into how proximity affects the growth of complex networks. While our results are not conclusive, it appears that growth models that take into account the indegree of the destination (e.g., preferential selection and random walk) match the data more closely than other models. Moreover, preferential selection outperforms random walk only for the Internet AS graph.

6 Summary and Future Work

In this paper, we examined the link formation processes that govern the growth of large-scale information networks. We collected growth data from four different networks, including social networks, Wikipedia, and the Internet. We carefully analyzed this growth data, comparing the empirical observations to what would be predicted by the numerous proposed models. Our analysis shows that the link formation processes follow the well-known preferential attachment model but that this model alone is insufficient to explain the observed proximity between link sources and destinations. We then examined how well other models with local rules matched the observed data, and our findings suggest that different mechanisms are at play in different networks.

We believe that this work opens up many avenues for future research. In particular, the data we collected can be used to test other previously proposed growth

models to see how well they match the observations. Similarly, the data we obtain could be used to guide the development of new models based on empirical data. While it is unlikely that a single model can capture all of the complexity of a large-scale real-world system, closely analyzing the data may reveal new insights into the link formation processes.

References

- [1] <http://as-rank.caida.org/data/>
- [2] L.A. Adamic, O. Buyukkokten, E. Adar, A social network caught in the Web. *First Monday* **8**(6) (2003)
- [3] R. Albert, H. Jeong, A.-L. Bárábási, The diameter of the World Wide Web. *Nature* **401**, 130 (1999)
- [4] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, evolution, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, 2006
- [5] A.-L. Bárábási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999)
- [6] V. Braitenberg, A. Schz, *Anatomy of a Cortex: Statistics and Geometry* (Springer, Berlin, 1991)
- [7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web: experiments and models, in *Proceedings of the 9th International World Wide Web Conference (WWW'00)*, Amsterdam, 2000
- [8] A. Capocci, V.D.P. Servedio, F. Colaiori, L.S. Buriol, D. Donato, S. Leonardi, G. Caldarelli, Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia. *Phys. Rev. E, American Physical Society, College Park, MD* **74**, 036116-1–0361166 (2006)
- [9] H. Chang, S. Jamin, W. Willinger, To peer or not to peer: modeling the evolution of the internet's AS-level topology, in *Proceedings of the 25th Conference on Computer Communications (INFOCOM'06)*, Barcelona, Spain, 2006
- [10] N. Eiron, K.S. McCurley, Link structure of hierarchical information networks, in *Proceedings of the Third Workshop on Algorithms and Models for the Web-Graph (WAW'04)*, Rome, Italy, 2004
- [11] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'99)*, Cambridge, MA, 1999
- [12] D. Garlaschelli, M. Loffredo, Patterns of link reciprocity in directed networks. *Phys. Rev. Lett., American Physical Society, College Park, MD* **93**, 268701-1–268701-4 (2004)
- [13] P. Holme, B.J. Kim, Growing scale-free networks with tunable clustering. *Phys. Rev. E, American Physical Society, College Park, MD* **65**, 026107-1–026107-4 (2002)
- [14] E.M. Jin, M. Grivan, M. Newman, The structure of growing social networks. *Phys. Rev. E, American Physical Society, College Park, MD* **64**, 046132-1–046132-8 (2001)
- [15] J. Kleinberg, Authoritative sources in a hyperlinked environment, in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA'98)*, San Francisco, CA, 1998
- [16] K. Klemm, V.M. Eguiluz, Highly clustered scale-free networks. *Phys. Rev. E, American Physical Society, College Park, MD* **65**, 036123-1–036123-5 (2002)
- [17] G. Kossinets, D.J. Watts, Empirical analysis of an evolving social network. *Science* **311**, 88–90 (2006)

- [18] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, On the bursty evolution of blogspace, in *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*, Budapest, Hungary, 2003
- [19] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, 2006
- [20] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling the web for emerging cyber-communities. *Comput. Network* **31**, 1481–1493 (1999)
- [21] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data, Association for Computing Machinery, New York, NY* **1**, 1–41 (2007)
- [22] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, in *Proceedings of the 2003 ACM International Conference on Information and Knowledge Management (CIKM'03)*, New Orleans, LA, 2003
- [23] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in *Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07)*, San Diego, CA, 2007
- [24] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**(2), 226–251 (2004)
- [25] M. Mitzenmacher, Editorial: the future of power law research. *Internet Math.* **2**(4), 525–534 (2006)
- [26] M.E.J. Newman, Clustering and preferential attachment in growing networks. *Phys. Rev. E, American Physical Society, College Park, MD* **64**, 025102-1–025102-4 (2001)
- [27] M.E.J. Newman, The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003)
- [28] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web. Technical Report, Stanford University, 1998
- [29] M. Peltomäki, M. Alava, Correlations in bipartite collaboration networks. *J. Stat. Mech., IOP Publishing, Bristol, UK* P01010, 1–23 (2006)
- [30] A.G. Phadke, J.S. Thorp, *Computer Relaying for Power Systems* (Wiley, New York, NY, 1988)
- [31] J. Saramaki, K. Kaski, Scale-free networks generated by random walkers. *Phys. A* **341**, 80 (2004)
- [32] A. Vázquez, Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E, American Physical Society, College Park, MD* **67**, 056104-1–056104-15 (2003)
- [33] V. Zlatić, M. Božićević, H. Štefančić, M. Domazet, Wikipedias: collaborative web-based encyclopedias as complex networks. *Phys. Rev. E, American Physical Society, College Park, MD* **74**, 016115-1–016115-9 (2006)

Dynamics On and Of Complex Networks, Volume 2

Applications to Time-Varying Dynamical Systems

Mukherjee, A.; Choudhury, M.; Peruani, F.; Ganguly, N.;

Mitra, B. (Eds.)

2013, XIII, 343 p., Hardcover

ISBN: 978-1-4614-6728-1

A product of Birkhäuser Basel