Growth of the Flickr Social Network

Alan Mislove
MPI-SWS
Campus E1 4
Saarbrücken, Germany

Hema Swetha Koppula
IIT Kharagpur
Kharagpur, India

Krishna P. Gummadi
MPI-SWS
Campus E1 4
Saarbrücken, Germany

Peter Druschel
MPI-SWS
Campus E1 4
Saarbrücken, Germany

Bobby Bhattacharjee
Computer Science Department
University of Maryland
College Park, MD

ABSTRACT

Online social networking sites like MySpace, Orkut, and Flickr are among the most popular sites on the Web and continue to experience dramatic growth in their user population. The popularity of these sites offers a unique opportunity to study the dynamics of social networks at scale. Having a proper understanding of how online social networks grow can provide insights into the network structure, allow predictions of future growth, and enable simulation of systems on networks of arbitrary size. However, to date, most empirical studies have focused on static network snapshots rather than growth dynamics.

In this paper, we collect and examine detailed growth data from the Flickr online social network, focusing on the ways in which new links are formed. Our study makes two contributions. First, we collect detailed data covering three months of growth, encompassing 950,143 new users and over 9.7 million new links, and we make this data available to the research community. Second, we use a first-principles approach to investigate the link formation process. In short, we find that links tend to be created by users who already have many links, that users tend to respond to incoming links by creating links back to the source, and that users link to other users who are already close in the network.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Measurement

WOSN'08, August 18, 2008, Seattle, Washington, USA.

Keywords

Social networks, measurement, growth

1. INTRODUCTION

Online social networking sites such as MySpace [15], Orkut [17], and Flickr [5] are now among the most visited sites on the Web. Users of these sites form social networks, which are powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides a unique opportunity to study the characteristics and dynamics of social networks at large scale.

To date, most measurement and analysis of online social networks has focused on the properties of static network snapshots. Despite the different goals and purposes of the various online social networking sites, the underlying social networks have been shown to exhibit a surprising number of common structural features, such as a highly skewed (power-law) degree distribution, a small diameter, and significant local clustering [1, 13]. This intriguing similarity suggests that the same underlying network growth processes may be at play in the different sites.

A proper understanding of these growth processes can provide insights into the observed network structure, allow predictions of future network growth, and enable simulation of systems on social networks of arbitrary size. However, most work on growth processes for large-scale networks has focused on theoretical models, instead of deriving the growth properties from empirical data. For example, two of the popular theoretical growth models are the Barabási-Albert model [2], where users connect to other users in proportion to the destination's popularity, and the random walk model [19, 20], where users connect to other users who are already close in the network.

In this paper, we use a first-principles approach, based on empirical data, to understand the growth processes that lead to the observed network structure. We collect large-scale network growth data from Flickr, a popular online social network. We crawled Flickr once per day for a period of three months, and we have observed 950,143 new users join and over 9.7 million links being formed. Our data covers a 58% growth in Flickr user population, and we make our dataset available to the research community.

Our analysis shows that new links are created and received by users in direct proportion to their current number of links, and that users tend to quickly respond to incoming links by

creating a link in the reverse direction. Additionally, our analysis reveals a strong proximity bias when users select other users to link to: users tend to connect to nearby users in the network much more often than would be expected when using previously proposed global processes.

We believe our work is an important first step towards understanding the processes that shape the structure of online social networks. Our work is directly useful in constructing synthetic networks that reflect both global and local characteristics of online social networks. Moreover, our collected data may lead to better structural and growth models, which are useful for network analysis and planning. Such models can be used in the design of search algorithms (e.g., by pre-identifying users that are likely to be hubs), in data mining (e.g., by identifying candidate users for placing data monitors), and in system evaluation (e.g. by allowing networks to be simulated over a wide range of sizes).

The remainder of this paper is organized as follows. Section 2 provides background and related work on the growth of online social networks. Section 3 describes our methodology for obtaining growth data from Flickr, and Section 4 details our analysis. Section 5 discusses implications of our study and we conclude in Section 6.

2. BACKGROUND AND RELATED WORK

In this section, we provide background on work related to the growth of online social networks.

2.1 Growth models

Researchers sought to explain the intriguing similarity in the high-level structural properties across networks of very different scales and types (including social networks) by hypothesizing that the networks are shaped by a set of common growth processes. For example, the power-law degree distribution often found in these networks can be produced through *preferential attachment*, where new links tend to attach to already-popular nodes.

One class of growth models uses global processes to determine the source and destination of new links. The well-known Barabási-Albert (BA) model [2] has been shown to result in networks with power-law degree[1] distributions. In the BA model, new links are attached to nodes using a probability distribution weighted by node degree. Many extensions to the BA model have been proposed, e.g. to add a tunable level of clustering [7].

Another class of models that produce power-law networks are based on local rules, such as the random walk model [19, 20], where nodes select new neighbors by taking random walks, and the common neighbors model [16], where nodes select new neighbors by picking nodes with whom they share many friends in common. Both of these models exhibit preferential attachment (since high-degree nodes end up being selected more often), but with higher levels of local clustering than the BA-model [20].

For a more detailed treatment of these and other models, we refer the reader to Mitzenmacher [14]. In this paper, we use detailed data from a large-scale online social network to look for evidence that supports either of these two classes of models.

[1]A node's *degree* is the number of links the node has to other nodes. For directed networks, we distinguish between *indegree* (the number of incoming links) and *outdegree* (the number of outgoing links).

2.2 Empirical data

Some recent work compared snapshots of the same network at different points in time to examine the growth processes. Newman [16] examined the properties of two scientific collaboration networks and found evidence of preferential attachment in both. Peltomäki and Alava [18] examined a scientific collaboration network and a movie-actor network and found evidence of sub-linear preferential attachment. Jeong et al. [8] examined citation and co-authorship networks, and found that nodes received links in proportion to their degree. Kumar [10] divided users from two online social networks into groups who are active and passive, and presented a model describing their behavior. Finally, Kossinets and Watts [9] 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.

Other work has used empirical social network data to predict user behavior. For example, Lerman and Jones [11] used a small data sample from Flickr and found that the social network is used to locate new content in the site. Nowell et al. [12] investigated co-authorship networks in physics to test how well different graph proximity metrics can predict future collaborations.

Our work shares similar goals and methodology with many of the above studies. However, the dataset we use is orders of magnitude larger than the ones used before. Moreover, our data allows us to analyze network growth at very small time-scales, as we have daily snapshots of the Flickr network.

3. MEASUREMENT METHODOLOGY

We begin by describing in detail the methodology for collecting data on the Flickr online social network. We were unable to obtain data directly from the Flickr site operators. So we chose to crawl the user graphs using the public web interface. Below, we first describe the challenges and limitations of obtaining data in this manner, and then we describe the dataset we collected.

3.1 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 adjacent to a given user. 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 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 user which point *to* a specific user). 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 [3].

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

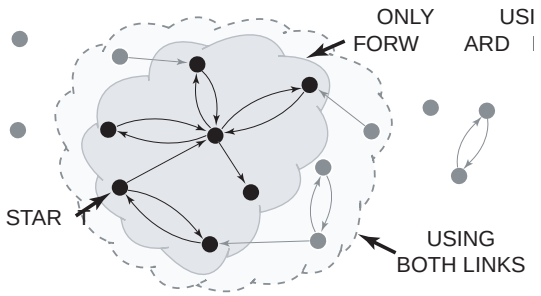

Figure 1: Users reached using different links. Using only forward links crawls the inner cloud; using both forward and incoming links crawls the entire WCC (dashed cloud).

in a subset of the WCC. In Section 3.4, we will show how this limitation affects our data.

3.2 Crawling methodology

Using automated scripts on a cluster of 58 machines, we crawled the social network graphs of Flickr once per day. Flickr exports an API that we used to conduct the crawl. More details on our methodology and its limitations can be found in [13]. Here, we discuss the methodology and limitations that are relevant to the growth data.

We started the first 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. On each subsequent day, we revisited every user we had previously discovered, in addition to all users that were reachable from the these users, and recorded any newly created or removed links or users.

Since Flickr does not provide the time of creation for any user account or link, our growth data has a granularity of one day. 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 users cannot be observed until they become connected to one of the users we have already crawled. In the rest of the paper, when we refer to "newly created links", we are referring to links that we observed being created. In other words, we may discover a new user that has a few established links, but we do not treat these previously established links as "newly created", as we did not observe them being created (i.e., we do not know when these previously established links were created).

3.3 Dataset

We crawled the Flickr network daily between November 2nd, 2006 and December 3rd, 2006, and again daily between February 3rd, 2007 and May 18th, 2007, representing a total of 104 days of growth. During that period of daily growth observations, we observed over 9.7 million new links being formed and discovered over 950,000 new users. This represents, relative to the initial network snapshot, over 58% growth in the number of users and over 63% growth in the number of links. Table 1 shows the high-level statistics of the data we gathered.

All of the data considered in this paper is available to the research community. The data has been anonymized in order to ensure the privacy of the social network users. A detailed description of the data format and downloading instructions are available at
`http://socialnetworks.mpi-sws.org`

Days of Observed Growth	104
Fraction of Links Symmetric	62%
Initial Number of Users	1,620,392
Final Number of Users	2,570,535
Growth in Number of Users	58%
Normalized Growth Rate in Users per Year	242%
Initial Number of Links	17,034,807
Final Number of Links	33,140,018
Number of Observed Created Links	9,792,634
Growth in Number of Links	63%
Normalized Growth Rate in Links per Year	455%

Table 1: High-level statistics of the Flickr growth data.

3.4 Limitations

There are two limitations to our crawl of Flickr. 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. In this section, we evaluate the number and characteristics of users who were missed by our crawls.

We performed the following experiment. 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 existing usernames. These 6,902 users form a random sample of Flickr users.

Among these 6,902 users, 1,859 users (26.9%) had been discovered during our crawls. Focusing on the 5,043 users *not* previously discovered by our crawls, 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 crawls 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 the users not in the largest WCC tend to have very low degree — in fact, almost 90% of them have no outgoing links at all.

4. GROWTH CHARACTERISTICS

In this section we use the collected data to explore the processes that underlie the growth of the Flickr social network. From our data, we extract three processes that appear to shape the growth in Flickr.

We found that link additions in Flickr exceeded link removals in our datasets at a rate of 2.43:1. Thus, in the rest of this paper, we focus only on how links are added to Flickr and leave examining link removal to future work.

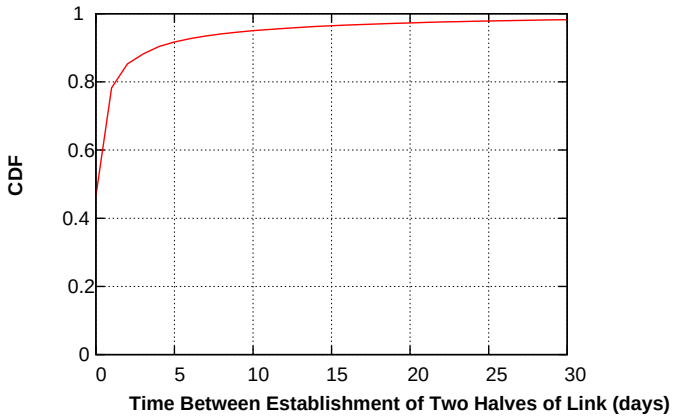

Figure 2: CDF of time between establishment of the two directed links of a symmetric link. In Flickr, links are quickly reciprocated.

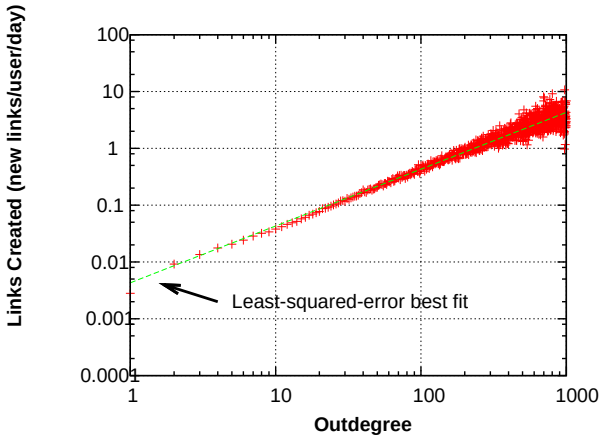

Figure 3: Log-log plot of outdegree versus number of new links created per day. Flickr shows strong evidence of preferential creation.

4.1 Reciprocation

We begin by examining *reciprocation*, a growth mechanism that occurs when the creation of a link from one user to another causes the reverse link to be established. Since Flickr is a directed network, the presence of a link from one user to another does not necessarily imply the presence of the reverse link. Reciprocation has been proposed as an independent growth mechanism for large-scale directed graphs [6, 22].

Since we do not know why links were established, we rely on the timing between the creation of the two directed halves of a symmetric link to guess whether the creation of the first causally affected the second. Figure 2 shows the cumulative distribution of the time between the establishment of the two halves of the symmetric links in Flickr. The data in Figure 2 covers the 62% of the links we observed being created that were symmetric (the other 38% of observed links did not have the reverse link created during our data collection period).

From Figure 2, it is clear that users often respond to incoming link creation by quickly establishing a reciprocal link. In fact, over 83% of all reciprocal link creations we observed occured 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 prompted the creation of the reciprocal link.

Since Flickr informs users by email of new incoming links, we hypothesize that many users tend to reciprocate links as a matter of courtesy. We investigated this behavior further by contacting the ten users with the highest number of incoming links in Flickr; of the six users who responded, all reported that they tended to reciprocate most incoming links quickly regardless of who the other user was.

4.2 Preferential attachment

Preferential attachment [2], colloquially referred to as the "rich get richer" phenomenon, is a growth model in which new links in a network are attached *preferentially* to users that already have a large number of links. For example, Barabási and Albert proposed a specific growth mechanism, called the BA model [2], which follows preferential attachment. Under the BA model, users are selected for new links in linear proportion to their degree (e.g., a user with seven links is seven times as likely to obtain a new link as a user with just a single link).

We separate the preferential attachment model into two aspects: *preferential creation* and *preferential reception*. Preferential creation describes the mechanism by which users *create* new links in proportion to their outdegree, and preferential reception describes the mechanism where users *receive* new links in proportion to their indegree. This distinction is consistent with previously proposed models of preferential attachment on directed graphs [4].

It is important to understand why we separate preferential attachment into preferential creation and preferential reception. Preferential attachment was originally defined for undirected graphs [2], and therefore does not distinguish between user indegree and outdegree. However, in Flickr, as well as other directed networks, link creation is very different from link reception. Users 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.

To examine whether preferential attachment (via preferential creation and preferential reception) is occurring in the observed growth data, we calculated how the number of new links per day varies with the user degree. If preferential attachment is taking place, we would expect to see a positive correlation between the degree of a user and the number of new links she creates or receives.

In Figures 3 and 4, we separately examine how the current outdegree and indegree of a user is related to the number of newly created and received links per day. Figure 3 shows that the outdegree of users in Flickr is linearly correlated with the number of new links created per user per day. The least-squared-error linear fit for link creation has a slope of 0.0044, implying that users create, on average, one new link per day for every 227 links they already possess. Similarly, Figure 4 shows that the increase in user indegree is linearly correlated with the current indegree of the user. The least-squared-error linear fit for link reception has a slope of 0.0027, implying that users receive, on average, one new link per day for every 370 links they already possess.

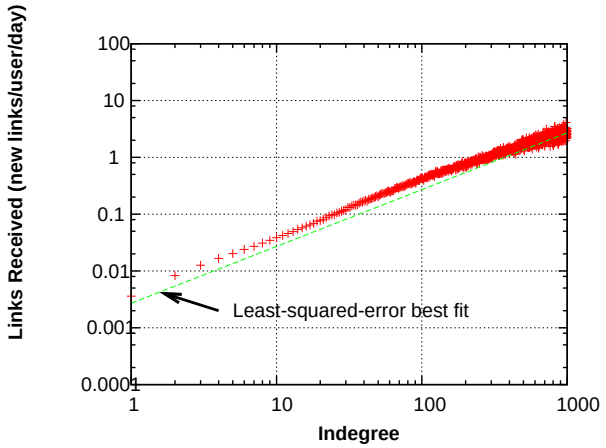

Figure 4: Log-log plot of indegree versus number of new links received per day. Flickr also shows strong evidence of preferential reception.

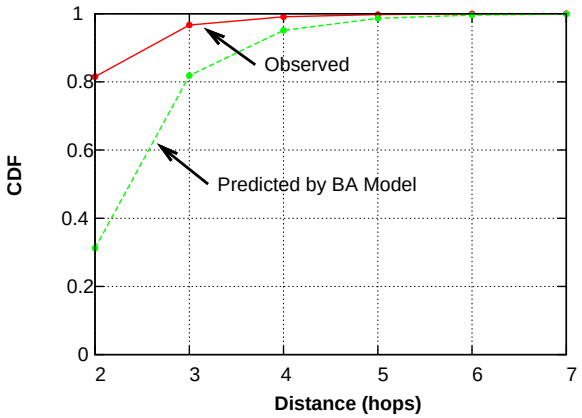

Figure 5: CDF of distance between source and destination of observed links. Also shown is the expected CDF from the BA model. A proximity bias exists that is not predicted by the BA model.

Our data shows that both preferential creation and preferential reception is occurring in Flickr, as there is a positive correlation between the number of links users have and their probability of creating or receiving new links. However, this alone is insufficient to claim that any specific mechanism (such as the BA model) is the mechanism that is causing the growth, as a number of different mechanisms could also result in this correlation. In the next section, we more closely examine the growth data to look for further evidence of specific growth mechanisms.

4.3 Proximity bias in link creation

In this section, we take a closer look at our growth data to look for evidence of specific global or local mechanisms that lead to preferential attachment. We look for evidence of models based on local rules by focusing on the distance between newly-linked users. Specifically, we examine the shortest path distance between the source and destination of newly created links, before a new link is created between them. If, for example, the BA model is the underlying mechanism, then the observed distance distribution between users should match that predicted by the model. Otherwise, if we see a stronger bias towards close users, it may suggest that users follow local, rather than global, rules for selecting the destinations for new links.

Over 50% of the observed new links in Flickr are between users that have, a priori, some network path between them (the remainder of the observed new links are between users which are, a priori, disconnected). For these new links among already connected users, Figure 5 shows the cumulative distribution of shortest-path hop distances between source and destination users. It reveals a striking trend: over 80% of such new links connect users that were only two hops apart, meaning that the destination user was a friend-of-a-friend of the source user before the new link was created.

One might wonder if, in small diameter networks like Flickr, this high level of proximity in link establishment is simply a result of preferential attachment. This is plausible, since the high-degree users that preferential attachment prefers tend to be close to many users. 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 BA model. Figure 5 also plots this distribution for each network.

The observed distances between the source and destination of links shows a significant bias towards nearby users, relative to what the BA model would predict. In fact, we found that the number of new links connecting two-hop neighbors in the empirical data exceeded that predicted by the BA model by a factor of three.

This result shows that while new link formation in Flickr follows preferential attachment, the link creation process cannot be explained by the BA model alone. Users are far more likely to link to nearby users than that model would suggest. This result is consistent with previous observations on static networks, which showed that the clustering coefficient was significantly higher than predicted by the BA model.

4.4 Summary

In this section, we closely examined network growth data from Flickr and compared the empirical data to the predictions of a previously proposed growth mechanism. We found evidence of reciprocation as a mechanism causing link creation. We also found that users tend to create and receive links in proportion to their outdegree and indegree, respectively. However, we found that the BA model mechanism did not accurately predict the proximity bias among users connected by new links. We observed a stronger bias towards proximity between new sources and destinations than would have been predicted by the BA model alone. In the next section, we discuss some future directions and describe the implications of our findings.

5. DISCUSSION

In this paper, we have used empirical growth data from a large-scale complex network to test if previously proposed growth models actually are at play. We have chosen to focus on preferential attachment because it is simple and has been suggested as the underlying growth mechanism for a variety of real-world networks. 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. However, we observed that the BA model (a global mechanism which follows preferential attachment) does not account for the proximity bias we observe in Flickr's link creation.

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 users are influenced by knowledge of the global metrics when choosing their neighbors. In many networks, it may not even be possible to obtain a global view of the network, due to technical and policy issues with computing global metrics.

In Flickr, the bias towards proximity can be partially explained by considering the discovery mechanisms available to users and the factors that constrain them. The primary mechanism available to users for exploring the network is to walk their neighborhood. In particular, there are very few global metrics for users, such as a "most popular users" list. This might explain our observation that there is a much stronger bias in link creation towards nearby users than would be predicted by the BA model alone, yet there still is a bias towards high-degree users.

Other social networks may have different policies and features available to users. For example, the YouTube [21] social network provides a list of the most viewed and most subscribed-to users. The presence of these global-view lists may change the ways in which users create links, as they provide a birds-eye view of the social network. We leave investigation of how different site mechanisms affect the link creation process to future work.

6. SUMMARY AND FUTURE WORK

In this paper, we studied the link formation processes that drive the growth of online social networks. We collected and analyzed detailed growth data from Flickr, and compared our empirical observations to the predictions of previously proposed models. Our analysis shows that the link formation processes follow the well-known preferential attachment model, but that global mechanisms (such as the BA model) alone are insufficient to explain the observed proximity between link sources and destinations.

We believe that this work opens up new 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 make available to the research community could be used to guide the development of new models based on empirical data.

7. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th World Wide Web Conference (WWW'07)*, Banff, Canada, 2007.

[2] A.-L. Bárabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web: Experiments and Models. In *Proceedings of the 9th International World Wide Web Conference (WWW'00)*, Amsterdam, May 2000.

[4] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, , and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physics Review E*, 74, 2006.

[5] Flickr. `http://www.flickr.com`.

[6] D. Garlaschelli and M. Loffredo. Patterns of link reciprocity in directed networks. *Physics Review Letters*, 93, 2004.

[7] P. Holme and B. J. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65, 2002.

[8] H. Jeong, Z. Neda, and A.-L. Barabasi. Measuring preferential attachment for evolving networks. *Europhysics Letters*, 61, 2003.

[9] G. Kossinets and D. J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311:88–90, 2006.

[10] R. Kumar, J. Novak, and 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, Aug 2006.

[11] K. Lerman and L. A. Jones. Social Browsing on Flickr. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*, Boulder, CO, Mar 2007.

[12] D. Liben-Nowell and 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, Nov 2003.

[13] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and 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.

[14] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[15] MySpace. `http://www.myspace.com`.

[16] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physics Review E*, 64, 2001.

[17] Orkut. `http://www.orkut.com`.

[18] M. Peltomäki and M. Alava. Correlations in bipartite collaboration networks. *Journal of Statistical Mechanics*, P01010, 2006.

[19] J. Saramaki and K. Kaski. Scale-free networks generated by random walkers. *Physica A*, 341:80, 2004.

[20] A. Vásquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physics Review E*, 67, 2003.

[21] YouTube. `http://www.youtube.com`.

[22] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physics Review E*, 74, 2006.