A Measurement-driven Analysis of Information Propagation in the Flickr Social Network

Meeyoung Cha
MPI-SWS
Campus E1 4
Saarbrücken, Germany
mcha@mpi-sws.org

Alan Mislove
MPI-SWS
Campus E1 4
Saarbrücken, Germany
amislove@mpi-sws.org

Krishna P. Gummadi
MPI-SWS
Campus E1 4
Saarbrücken, Germany
gummadi@mpi-sws.org

ABSTRACT

Online social networking sites like MySpace, Facebook, and Flickr have become a popular way to share and disseminate content. Their massive popularity has led to viral marketing techniques that attempt to spread content, products, and ideas on these sites. However, there is little data publicly available on viral propagation in the real world and few studies have characterized how information spreads over current online social networks.

In this paper, we collect and analyze large-scale traces of information dissemination in the Flickr social network. Our analysis, based on crawls of the favorite markings of 2.5 million users on 11 million photos, aims at answering three key questions: (*a*) how widely does information propagate in the social network? (*b*) how quickly does information propagate? and (*c*) what is the role of word-of-mouth exchanges between friends in the overall propagation of information in the network? Contrary to viral marketing "intuition," we find that (*a*) even popular photos do not spread widely throughout the network, (*b*) even popular photos spread slowly through the network, and (*c*) information exchanged between friends is likely to account for over 50% of all favorite-markings, but with a significant delay at each hop.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Human factors, Measurement

Keywords

Information dissemination, cascades, social networks, viral marketing, Flickr

1. INTRODUCTION

Online social networking has recently become a popular way to share and disseminate information. Users of websites like MySpace, Flickr, and Facebook create networks of friends. They share, find, and disseminate content at a massive scale. Every minute, ten hours of video are uploaded to YouTube [31]; Flickr contains over two billion photos [26]. As a result of their massive popularity, these sites have been exploited as a platform for the viral marketing of content, products, and political campaigns. For instance, major movie studios place trailers for their movies on MySpace; US presidential candidates ran online political campaigns on YouTube; and individuals and amateur artists promote their songs, artwork, and blogs through these sites, all hoping to reach millions of online users. Despite the excitement, neither the characteristics of information propagation in social networks nor the mechanisms by which information is exchanged are well understood.

One of the distinguishing features of online social networks is the potential for information dissemination along the social links, i.e., information propagation among friends in social networks, one hop at a time. It is widely believed that such user-to-user exchanges, also known as "word-of-mouth" exchanges, can spread content, ideas, or information *widely* and *quickly* throughout the network. In fact, a number of research efforts [7, 8, 11, 12, 24, 30] have proposed viral marketing campaigns to exploit the word-of-mouth effect. In 2007, $1.2 billion was spent on advertisement in online social networks worldwide, and this amount is expected to triple by 2011 [4]. However, to date, there is little data publicly available on viral propagation in the real world and only a few studies [1, 10, 15] have been conducted to characterize how information spreads over current online social networks.

In this paper, we collect and analyze large-scale traces of information dissemination in the Flickr social network. Flickr, founded in 2004 and acquired by Yahoo! in 2005, is an online social network for sharing photos. We crawled the social network to gather information from 2.5 million Flickr users and 33 million links between them. To capture the dynamics of information propagation, we crawled the social network for 104 consecutive days. In addition, we collect information about what photos these 2.5 million users marked as favorites and when they did so. We conduct an in depth study of these data sets to determine how pictures spread through the Flickr social network.[1]

We analyzed the data to answer three key questions:

1. How widely does information spread in the Flickr social network? Do popular pictures gather fans from different parts of the network or is their popularity limited to a certain region?

2. How quickly does information spread through the social network? How long after the upload of a photo do fans mark it as a favorite?

3. Does information in Flickr flow along its social network links? What fraction of a photo's fans discovered the photo through a friend? How long does this process take? By what other mechanisms do fans discover their favorite photos?

[1]The data traces used in this paper are shared for the wider community use at `http://socialnetworks.mpi-sws.org/`.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Our data analysis reveals several interesting facts about information propagation in Flickr. We find that most information does not spread widely throughout the network: while it is intuitive that many "personal interest" photos of an individual's family and friends will have highly localized appeal, we find that even the more popular photos have substantially limited popularity outside the immediate network neighborhood of the uploader. We also find that information spreads slowly in the network; even the most popular photos exhibit a slow, steady growth in popularity over a long period of time (1-2 years). Our findings suggest that over 50% of users find their favorite pictures (i.e., pictures they bookmark) from their friends in the social network. However, there is a significant delay (often several months) in the propagation of information across friend links.

Our findings are in conflict with our initial expectation that information would spread widely and quickly in a viral fashion across the social network. Our observations about high content locality (i.e., even top popular photos did not spread widely) may be related to the burnout process in the theory of information diffusion [25, 27, 30]. The slow pace of information propagation might reflect the challenges in finding relevant information from an overwhelming volume of information individuals get exposed to, even from their immediate friends. Our results, while preliminary, may have implications not only for the designers of viral marketing campaigns, but also for the mechanisms for finding information that social networking sites offer, and promote.

The rest of the paper is organized as follows. We describe our measurement methodology and introduce our data set in Section 2. We present characteristics of the Flickr social network that are relevant to information propagation in Section 3. We present analyses of how picture popularity is distributed over the network topology in Section 4 and how they evolve over time in Section 5, respectively. Section 6 investigates the role of social links in information propagation. In Section 7, we discuss possible explanations for our findings. We summarize related work in Section 8 and conclude in Section 9.

2. MEASUREMENT METHODOLOGY

In this section, we introduce the Flickr website and describe our data collection methodology.

2.1 Flickr

Flickr is a photo sharing site with social networking features, where users can create friend relationships with one another and share photos. Users can create networks of friends, join groups, send messages to other users, comment on photos, tag photos, and choose their favorite photos. To use most of these features, users must create a Flickr account and they must be logged-in to Flickr. Flickr provides users with privacy control over photos they upload, allowing photos to be classified as either private, visible only to their friends, or, the default, public.

Flickr allows users to create two types of links: links to favorite photos (called *favorites* in Flickr) and links to other users (called *contacts* in Flickr). We refer to users in the contacts list as *friends* in this paper. Like bookmarks, users may "favorite-mark" a photo to archive and share interesting photos with others. We refer to users who include a photo in their favorite photos list as *fans* of that photo. The list of a user's favorite photos and the list of a user's contacts are both available from a user's profile page. This paper focuses on these two features of Flickr.

2.2 Data collection methodology

We describe our Flickr data collection methodology. Since we were interested in studying the dynamics of information flow in a large-scale social network, we needed to collect (*a*) the evolving state of the social network and (*b*) evidence of information propagation from one user to another. Here we describe our methodology for collecting both types of information.

In order to collect the state of the social network, we crawled a significant subset of the Flickr online social network. We started with a randomly selected Flickr user and followed all of the friends links in the *forward* direction in a breadth first search fashion. In this way, we collected a "snowball" sample of the Flickr social network. The list of contacts for each user is publicly visible in Flickr, and we used the Flickr API to reduce the necessary bandwidth for crawling. Our snowball sample is part of a large weakly connected component of the entire Flickr network that is reachable from the seed user; we call our sampled data the *Flickr social network graph*.

To capture the dynamics of friends' relationships, we launched a complete crawl of the social network graph once per day. We visited all users in the previous day's social network graph and recorded any newly created or removed friend links or users.

Finally, to collect evidence of information flow over a social network, we collected information on the favorite photos of the Flickr users. Favorites photos are publicly visible from each user's profile page. We used the Flickr API to download the list of favorite photos for all known users, based on the final snapshot of our Flickr social network graph. Because Flickr provides information about the exact timestamp when a user marked a photo as a favorite, this allowed us to recreate favorite marking events over the dynamically evolving social network graphs. Moreover, for favorite markings that occurred during our repeated crawls of the social network, we know the state of the social network at the time the favorite marking took place. From this, we can examine the social network factors which influenced the favorite-marking user.

2.3 Data description

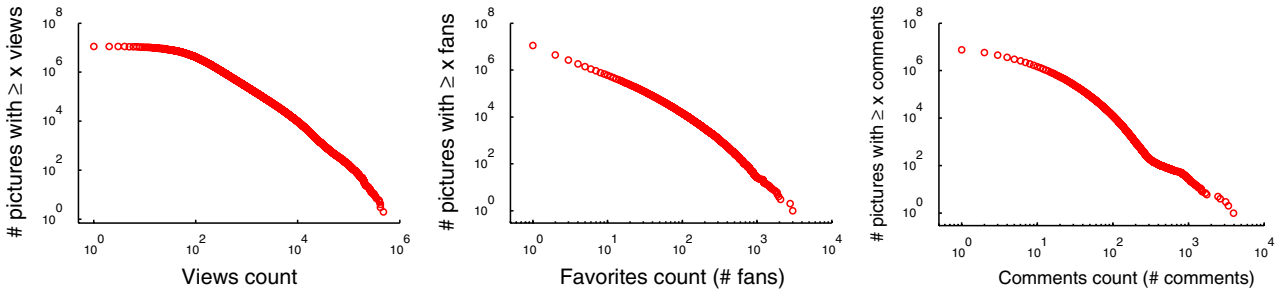
We crawled the Flickr social network graph once per day for the period of 104 consecutive days from November 2–December 3, 2006 and February 3–May 18, 2007. We observed 2.5 million Flickr users and 33 million links, an estimated 25% of the entire Flickr network. We refer the readers to our previous work [21, 22] for data analysis on the general properties and growth patterns of the network. In this paper we also collected information about 34 million favorite-markings over 11 million distinct photos.

Table 1: Summary of Flickr data set

Time period	104 days (starting Nov 2, 2006)
# Links	17,034,807 to 33,140,018
# Users	1,620,392 to 2,570,535
# Photos	11,195,144
# Favorite marks	34,734,221

2.4 Limitations

Although this data gives us a unique opportunity to examine information spread dynamics over social links, it has several limitations. First, our methodology does not take into account any deleted favorite marking. However, informal reports from Flickr users suggest that users rarely delete any of their favorite marked photos. Furthermore, this does not affect our analysis results because deleted favorite-markings can no longer spread via a social network.



(a) Views (b) Fans (c) Comments

Figure 1: Picture popularity distribution

Second, we do not know how a user was exposed to a photo prior to favorite-marking. There are many possibilities: visiting friends' pages, browsing featured photos, visiting external web pages, or finding via search engine. However, we do not know which mechanisms are responsible for which users' favorite-markings. Thus, we use a heuristic to infer whether a user found photos through social links.

Finally, we can only observe the network, but we cannot manipulate it. We are not able to make changes to the Flickr website or run tests in a controlled environment. For example, we cannot test whether Flickr would see an increase in traffic if we were to add a specific feature or create links between certain photos or users.

3. NETWORK TOPOLOGY AND PICTURE POPULARITY

In this section, we present an overview of the Flickr social network. We first describe the small-world pattern seen in the Flickr network topology, which is characterized by a strong local clustering and a small diameter. These structural properties are important because they indicate how widely information can propagate through the network. After presenting structural properties, we describe various popularity metrics that could be used for pictures in Flickr, and discuss why we chose to focus on the number of fans.

3.1 Social network topology

We begin by examining the degree distributions of the 2.5 million users. We construct the Flickr social graph such that each node represents a Flickr user and edges between nodes represent friend links. A user can unilaterally declare and point to any other user as a friend in Flickr, i.e., friends links are unidirectional. Thus, we represent the network as a directed graph. We refer to the number of friends a user declares and points to as the *outdegree* of the user and the number of other users that point to the user as a friend as the *indegree* of the user. In Flickr, most links are reciprocal; 68% of the links are bidirectional. The Pearson's correlation coefficient between a node's indegree and outdegree is 0.76, indicating that a user with many outdegree links is also likely to have many indegree links.

Figure 2 shows the in- and out-degree distributions of the 2.5 million users. A majority of users are connected to only a few other users. In fact, 55% of the nodes have just 1 outgoing link and 90% of them have an outdegree smaller than 10. The average outdegree was 14. However, a few nodes have tens of thousands of links; the node with the highest number of outgoing links has 26,342 friends. The indegree distribution is similar, but the maximum indegree is smaller than the maximum outdegree.

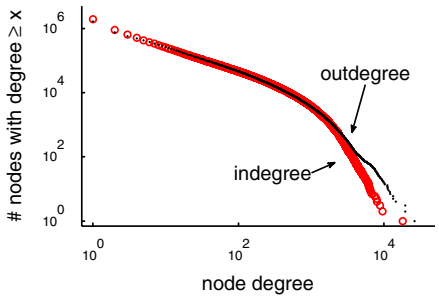


Figure 2: Node degree distribution

We examine two important structural properties of the social network graph: path lengths and clustering coefficient. The maximum path length between any two nodes in the network (i.e., diameter) is 27, while the average path length is 5.67. The clustering coefficient measures how tightly the neighbors of a node are interlinked. It can range from 0 to 1, where 0 means there is no connection between neighbors and 1 means a clique. For well-connected nodes in Flickr the clustering coefficient is typically between 0.05 to 0.1, and for poorly connected it is typically between 0.2 and 0.4 [22].

The small-world network properties have implications for information flow. For example, users with high indegree and outdegree can potentially receive and transmit information more widely. The observed small-world network structure indicates that most Flickr users are separated by only a few hops and that information can be retrieved over short network paths [29]. In summary, the Flickr social structure exhibits properties that promise wide-spread dissemination of popular information throughout the network.

3.2 Picture popularity

In Flickr, users can view pictures, leave comments on pictures they find interesting, or favorite mark pictures they like. Consequently, there are different metrics possible for ranking pictures; they can be ranked based on their popularity in terms of the number of views, fans, or comments. Figure 1 shows the popularity distributions for the 11 million pictures based on these three metrics. All three distributions show a heavy-tailed distribution. For example, the distribution of fan popularity shows that millions of pictures have fewer than 10 fans in the entire network, while an order of magnitude fewer pictures (252,126) have more than one thousand fans. This means that only a small fraction of pictures achieve high popularity and thus have the potential to spread widely through the social network.

Views of pictures tend to be two-orders of magnitude higher than fans or comments. To understand the relationship between the number of views, comments, and fans, we measure the strength of a linear relationship between these values (i.e., correlation coefficient). The correlation coefficient ranges from -1 to 1, where values close to 1 suggest a positive linear relationship, values close to -1 suggests negative linear relationship, and values close to 0 suggest there is no linear relationship between the data points. The number of views is not strongly correlated with the number of comments (0.13) or the number of fans (0.23). On the other hand, the number of comments pictures receive is highly correlated with the number of fans (0.60). This may be because leaving comments on pictures or marking them as favorites takes time. So users are likely to leave comments or favorite-mark only those pictures they find interesting, which would explain the high correlation between the number of fans and the number of comments.

More interestingly, the correlation between views and fans is weaker for popular pictures; the correlation coefficient decreases to 0.21 for pictures with over 100 fans and falls to 0.13 for pictures with over 1,000 fans. The low correlation between the number of view and fans could indicate that users find many of the pictures they view uninteresting. The low correlation may also arise because any web user can view a Flickr picture, but users need to register with Flickr and login to favorite-mark or comment on pictures.

In this paper, we focus on the fan popularity of pictures. Specifically, we study how widely and quickly pictures are favorite-marked throughout the network. We believe our analysis demonstrates how pictures that are of interest to users spread through the social network. However, our findings might not apply to the spread of picture views, because we do not have information about when which user viewed which photo in Flickr. (We only know the total view counts per photo.)

4. TOPOLOGICAL DISTRIBUTION OF PICTURE POPULARITY

We investigate how widely all the fans of a particular picture is distributed over its social network topology. The key question we want to answer is, how widely does information in the form of a favorite-marking propagate through the social network? We first analyze whether the popularity of Flickr pictures is global or confined to local regions in the network. We then study the distribution of fans as a function of their distance from the uploaders.

4.1 Local versus global picture popularity

To understand how well the local popularity of pictures in different regions of the network correlates with their global popularity, we determine the most popular pictures (we call this set a *hotlist*) in several local neighborhoods and compared them with the global hotlist of pictures. We assume that if pictures spread widely throughout the network then we will see a good match between the local and global hotlists of pictures. However, if the popularity is localized to a specific region and shows strong topological correlation, the global and local hotlists will vary substantially.

For the test, we randomly picked 250 users (or seed nodes) from the set of 2.5 million users who have favorite-marked at least one photo, and identified the top 100 pictures from the neighborhood of each seed node. We visited the 4-hop neighborhood around each of these seed nodes, based on the final snapshot of the network. Table 2 shows the neighborhoods of our 250 seed users. Increasingly, more nodes are reached as we increase the neighborhood boundary. Within 4-hops, our seed nodes reached on average 1,563,500 nodes, which is nearly 36% of our entire Flickr social network.

Table 2: Seed node neighborhood sizes

Distance	Min.	Med.	Avg.	Max.
1-hop	6	1,377	1,379	2,816
$\leq$ *2 hops*	2,785	199,330	174,100	290,671
$\leq$ *3 hops*	283,001	1,050,400	938,880	1,159,636
$\leq$ *4 hops*	880,051	1,625,482	1,563,500	1,667,054

For each of the neighborhoods, we identified the top 100 pictures based on the number of fans from that region and compared the list with the globally popular top 100 pictures. We then determine the number of photos that appear in both lists, the "overlap." Figure 3 shows the overlap between the two lists. Along the horizontal axis, we sort the neighborhoods based on the number of common photos. The one-hop neighborhood plot shows that, for 233 of the 250 local regions, there was no overlap between the local and global hotlists. The largest overlap was 19 pictures. The overlap between local and global hotlists increases as we consider wider neighborhood boundaries. Hotlists based on a 2-hop neighborhood boundary showed on average 8 globally popular pictures; while the 3- and 4-hop neighborhoods showed much larger overlaps of 39 and 70, respectively.

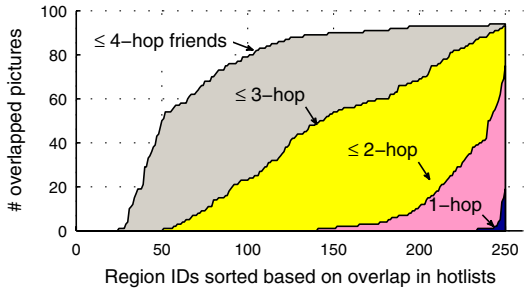


Figure 3: Resemblance in local and global hotlists

We make two key observations. First, the difference between global and local hotlists indicates that different pieces of information are popular among the different social network regions. Based on 1- or 2-hop neighborhoods, hotlists of pictures are localized. That is, these pictures were not widely available throughout the network. Second, focusing on the 4-hop neighborhood, we are able to see a high overlap in global and local hotlists. This is because this neighborhood covers a large number of nodes—36% of the entire graph. This also means that information is reachable within few hops (i.e., small world network).

4.2 Distance from fans to picture uploaders

Motivated by high content locality, we look at the distance between fans and photo uploaders. We examine two quantities: (*a*) the fraction of fans that are located 1, 2, or 3 or more hops away from the uploaders and (*b*) the fraction of nodes that become fans of the pictures (as a function of the distance from the uploader). Calculating these distances for all 34 million favorite-markings on a 2.5 million node network is time-intensive. Therefore, we chose a subset of pictures from our data set for analysis. To pick representative pictures, we chose all pictures that were uploaded after November 2nd, 2007, and all of their fans as our target set. This set includes 3 million pictures and 10 million favorite-markings.

We first investigate the distance from fans of photos to their uploaders. Table 3 shows the percentage of fans that are k hops away from uploaders. We observe strong locality across all popularity level. For less popular photos, 91% of all fans are within 2 hops of the uploaders. But even for top popular photos, 81% of all fans are within 2 hops of uploaders. It is intuitive that globally less popular photos exhibit strong locality, since these are typically personal photos of family and friends, which are by definition interesting primarily to people pictured in the photo and those who know them. However, it is surprising to us that popular pictures with more than 500 fans also show a high level of content locality.

Table 3: Percentage of fans in k-hops distance from uploaders

# Fans	1-hop away	2-hops away	3+-hops away
1-5	60	31	9
6-100	55	32	13
101-300	43	42	15
301-500	37	46	17
501-	32	49	19

One potential concern with the above analysis is the vastly different k-hop neighborhood sizes around different uploaders, making it hard to compare the distribution for different photos. We therefore also calculated the fraction of the each k-hop neighborhood that became fans of the photo. Thus, we visited each and every user that is k-hops from the respective uploaders and counted how many of them have favorite marked the picture. Because visiting neighborhoods is also time-intensive, we limited our study to pictures with more than 100 fans. Table 4 shows the fraction of an uploader's k-hop friends that are fans. Across all popularity levels, 1-2% of neighbors eventually become fans of the uploaders' pictures. In 2- and 3-hop neighborhoods, a much smaller fraction of nodes become fans. This suggests that propagation of favorite-marked photos throughout the network is limited and that photos rarely spread beyond the immediate vicinity of the uploaders.

Table 4: Percentage of fans for uploaders out of k-hop friends

# Fans	1-hop away	2-hops away	3-hops away
101-300	1.77	0.08	0.001
301-500	1.39	0.12	0.004
501-	1.14	0.17	0.009

4.3 Summary

In this section, we presented data analysis of 34 million favorite markings of 11 million pictures in the Flickr social network. We examined the correlation between the locations of favorite marking and the social structure of users, based on the topology that existed on the last day of our crawl. We found that different sets of pictures are popular in different parts of the social network and that photo fans are closely located to the uploaders. These observations demonstrate that information does not propagate widely in the Flickr social network.

5. TEMPORAL EVOLUTION OF PICTURE POPULARITY

In this section, we study how the popularity of pictures evolves over time, i.e., how quickly fans favorite mark pictures after they are uploaded. We first study how four sample pictures accumulated fans over time to identify various patterns of growth in picture popularity. We then analyze the dominant patterns in the long-term evolution of popularity across all pictures.

5.1 Patterns of popularity growth

To understand patterns in the evolution of picture popularity, we focused on the pictures with the most number of fans. We examined the growth patterns in the 30 most popular pictures in Flickr. We also visited each photo's web page to gather additional information about potential external links to the photo, for instance, whether the photo had been featured on the front page of Flickr or whether it had received a photography award.

From these 30 pictures, we selected four representative photos and show how their popularity evolves over time in Figure 4. At a first glance, the four photos show different growth curves. However, their growth curves exhibit one or more of the following three distinct growth phases, which we call *active-growth*, *surge-increase*, and *sluggish*. Photo A experienced an active-growth in popularity over the course of 563 days, eventually acquiring 2,144 fans. Photo B's popularity evolved differently from Photo A's. Photo B stayed dormant and unknown to many Flickr users for nearly 272 days after it was uploaded; it gained only 10 fans by day 230. Then on day 272, the photo won a national photo contest and was linked from several external web pages. Almost immediately the photo witnessed a growth-spike and received 244 new fans on a single day! Over the next 50 days the photo continued to gain fans, but on day 325 the photo again became sluggish.

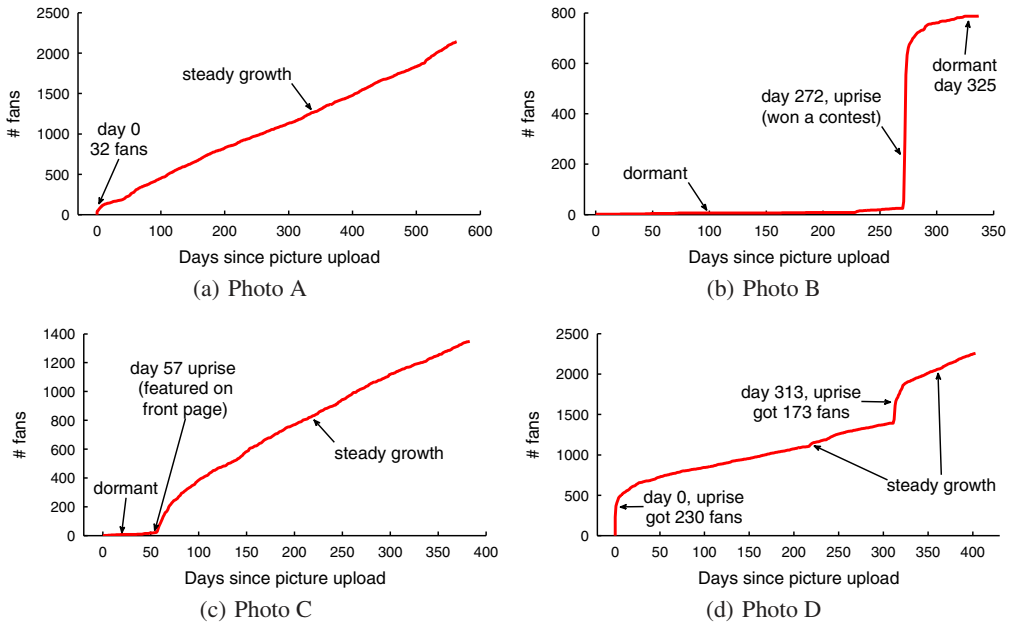
The popularity growth of Photo C shows sluggishness, surge-increase, and active-growth. After it was uploaded, the photo went through a sluggish period before day 40. As the photo slowly started gaining fans, it was featured on the front page of Flickr on day 57.[2] Once it was featured, the photo saw a spike in its popularity; it obtained over 200 fans by day 70. Then, it shifted into a long period of active, steady growth for nearly an year. Finally, Photo D's popularity shows two surge-increases, once at day 0 and another at day 313, and both increases are followed by extended periods of steady growth. On day 0, the photo was featured on Flickr's Explore page. The initial growth rate of photo popularity is influenced by how users link to the photo's uploader. The uploaders of Photo A and Photo D had 5,178 and 4,893 incoming links, respectively. This may explain why their pictures became instantly known to others as soon as the photo was uploaded. In contrast, the uploaders of Photo B and C had substantially fewer incoming links (289 and 966 respectively), and thus initially experienced much slower growth in popularity.

In summary, the popularity of individual pictures evolves differently. However, their growth curves share three key common patterns: sluggishness, surge-increase, and active-growth. The relative importance of each pattern differs across photos. For instance, Photo B gained most of its fans through its surge-increase, whereas Photo C gained most of its fans during an active-growth. Interestingly, most of the 30 popular pictures showed an active-growth pattern that is linear over an extended period of time. This steady linear growth in popularity cannot be easily explained by traditional information diffusion theories [25, 27], which predict an exponential growth in popularity followed by a saturation or maturity. We discuss potential explanations for this discrepancy in Section 7.

5.2 Long-term trends in popularity growth

How does photo popularity evolve over a long period of time? Which growth pattern is dominant in a time period of a year or

[2]Flickr uses an internal algorithm to choose interesting pictures to feature.

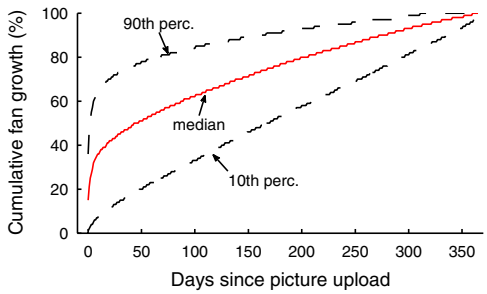


(a) Photo A (b) Photo B

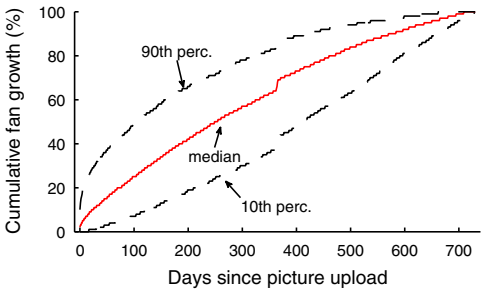
(c) Photo C (d) Photo D

Figure 4: Popularity growth pattern of sample photos

longer? To answer these questions, we examined the aggregate growth patterns of two subsets of photos: photos that are older than 1 year and photos that are older than 2 years. We considered only popular pictures that had at least 100 or more fans by the end of their first year and their second year. There were 5,346 and 897 such pictures. Similarly, we focused on only the fans acquired during the first or the second year after the picture was uploaded.



(a) Growth pattern over one year (5,346 pictures)



(b) Growth pattern over 2 years (897 pictures)

Figure 5: Popularity growth patterns over a long term period

We examine the long-term popularity trend of the two sets of photos. Figure 5(a) shows the growth in popularity of 5,346 pictures that are older than 1 year. The horizontal axis represents the age of the photo, or the time since the pictures' uploads. The vertical axis represents the fraction of fans a photo obtained by the given age, out of the total number of fans it obtained at the end of the first year. We show the 10th percentile, the median, and the 90th percentile growth rates for all 5,346 pictures for every day of the 1-year period. Similarly, Figure 5(b) shows the growth in popularity of pictures that are older than 2 years. This figure plots data for 897 pictures.

We make two key observations from the above figures about the long-term trends in popularity growth of pictures. First, in aggregate, many photos do show an active rise in popularity during the first few days after they are uploaded. Second, after the first few (10-20) days, most pictures, in aggregate, enter a period of steady linear growth. Surprisingly, the linear-growth is sustained over extended periods of time—the median growth rate does not show any sign of slowing down even after 1 or 2 years. As a result of this linear growth pattern, the fans these pictures obtained during the first few days accounts for an ever decreasing fraction of total fans. For a majority of pictures, over 40% of fans were acquired after the first 100 days. Conversely, our analysis suggests that Flickr users take a long period of time to find out about interesting pictures.

We also analyzed the trends in popularity growth for less-popular pictures, i.e., pictures with fewer than 100 fans (not shown in Figure 5). Unlike popular pictures that exhibit a steady-growth pattern in their fan population, a majority of the less-popular photos attract their limited fan population early on during their lifetime, and they become dormant after the first few months.

5.3 Summary

In this section, we studied how popularity grows over a long period of time. Existing models of information diffusion predict an exponential growth in popularity. In contrast, our data shows a steady linear growth pattern for the most popular pictures. Our data also reveals different paths to popularity, comprised of active-growth, surge-increase, and sluggishness.

Table 5: The extent of information flow through social links (i.e., social cascade) in Flickr

Popularity (# Fans)	Total pictures	Total fans	Social cascades				Cascades from uploaders			
			# Photos	Perc.	# Fans	Perc.	# Photos	Perc.	# Fans	Perc.
1-5	2,704,806	4,328,609	1,517,550	56%	2,197,522	51%	1,487,266	55%	2,111,551	49%
6-100	346,870	5,121,820	329,029	95%	2,834,704	55%	306,287	88%	2,307,155	45%
101-300	3,502	499,870	3,502	100%	273,596	55%	3,337	95%	171,085	34%
301-500	154	54,773	154	100%	27,849	51%	147	95%	15,251	28%
501-	29	20,113	29	100%	8,686	43%	28	97%	4,017	20%
Total	3,055,361	10,025,185	1,850,264	61%	5,342,357	53%	1,797,065	59%	4,609,059	46%

6. INFORMATION PROPAGATION VIA SOCIAL LINKS

In this section, we first describe the various mechanisms by which people can find content in Flickr. We then investigate the role played by one specific mechanism, namely, the exchange of information between friends, in propagating favorite-markings over the Flickr social network.

6.1 Dissemination mechanisms

In Flickr, people can find pictures through various mechanisms. We list some of the important ones below.

- **Featuring.** Flickr officially provides two key features: one is the front page and the other is the Explore page[3] which is the list of photos selected by Flickr as "interesting."
- **Search results.** Users may search for specific content within Flickr. Content meta data such as titles, tags, and descriptions are used by the Flickr photo search engine to identify relevant content.
- **Links between content.** Flickr provides links between picture pages that allow users to easily navigate the website. Examples are "sets," which are groups of similar pictures by the same uploader, and "pools", which include photos uploaded by different users, but that have common themes.
- **External links.** Users can reach Flickr photos from external websites, blogs, emails, and other mechanisms external to the Flickr website.
- **Social network.** Flickr users find new pictures that are uploaded or shared by their friends.

In the rest of this section, we focus on the dissemination of content via social network links in Flickr. Undoubtedly, other mechanisms are also at play, but studying their influence requires a richer data set and is beyond the scope of this paper.

6.2 Identifying social cascades

Information can travel widely through a social network one-hop at a time via word-of-mouth exchanges between friends in the network. We refer to such information dissemination as a *social cascade*. Without a page view trace or asking the user directly, we cannot say for sure how users in Flickr found photos. So we used a heuristic to infer the cases when Flickr users are likely to have discovered pictures using their friends links. Our heuristic uses the state of the social network at the time of the favorite-marking to make an educated guess about how a new fan might have found the photo. In particular, we say that user A found a photo P through the social network if and only if there exists a user B who is a friend of A such that:

- B also marked P as a favorite,
- B included photo P on his favorite list before A included photo P on his favorite list, and
- B was a friend of A before A made photo P his favorite.

The above conditions state that B must be A's friend before A found the photo, and that B must have already favorite-marked the photo before A found it. If all of these conditions hold, then we consider the photo to have propagated from B to A via a social link. Our definition may identify multiple friends from whom A could have found the photo. For this work, we assume A received information about the photo from all of these users. Finally, Flickr users cannot mark the pictures they uploaded as favorites. For the purpose of our analysis, we consider uploaders as fans of their pictures by default.

To apply the above test, we need to know the social network graph at the time when photos are marked as favorites. We use the data from our daily snapshots of the Flickr network to recreate the state of the network on each of the 104 days it was crawled. Thus, we are able to infer the role of social links in transmitting favorite-marking information only for the favorite markings made during the 104 days. We further limited our analysis to the set of photos that were uploaded during the period of our crawl as this guarantees that its entire popularity history is known to us. This leaves us with 10,025,185 favorite markings over 3,055,361 pictures.

6.3 The role of social cascades

We examined the fraction of favorite markings that spread through social links in Flickr. Table 5 summarizes the role of social cascades in the spread of favorite markings. We grouped pictures based on their popularity to determine whether we observe social cascades for both popular and unpopular pictures. Table 5 shows the number of pictures in each popularity level and their number of favorite markings (i.e., fans). Out of 3 million pictures, nearly 2.7 million of them (88%) obtained no more than 5 favorite markings, while 3,685 pictures gained more than 100 fans during the daily crawl period. This skewed popularity of pictures also matches the heavy-tailed popularity distribution over a long-term period (Section 3.2).

Columns grouped as *social cascade* in Table 5 represent the set of pictures and fans that were identified as being part of a social cascade. Overall, out of the more than 10 million total favorite markings, 5 million or 53% of all favorite markings were propagated through social links, suggesting that a crucial role is likely played by the social network in information propagation. The remaining fans might have found the pictures through various other mechanisms. Social cascade plays a significant role in propagating information for both popular and unpopular pictures. The fraction

[3] http://www.flickr.com/explore/

of social cascade favorite markings remains consistently high, varying from 43% to 55% across the different popularity ranges. Thus, social links play an important role in transmitting information, independent of its popularity.

While aggregate statistics show that a consistently high fraction of social cascades in the spread of favorite markings, individual pictures vary from this pattern. Figure 6 shows that varying fractions of favorite markings are part of social cascades for different pictures. The horizontal axis represents the photo popularity, i.e., the number of fans a photo received, and the vertical axis represents the percentage of favorite markings that spread through social links for the group of pictures with similar popularity. The horizontal axis is in a log-scale. The fraction of social cascade-based favorite-marking is shown for every 10th percentile values. The three solid lines indicate the minimum, the median, and the maximum values. The median plot shows similar values as observed in Table 5. Across all popularity ranges, individual pictures benefit from information propagation through social links to varying degrees.

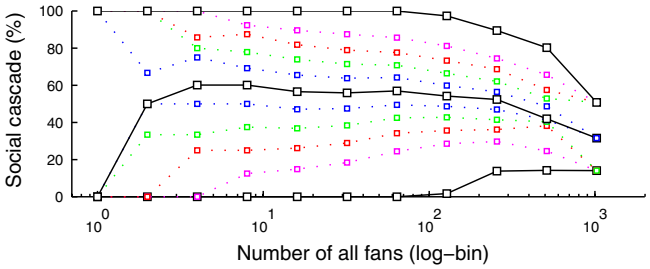


Figure 6: Probability of information flow through social links across picture popularity: Plot shows every 10th percentile probabilities including minimum, median, and maximum.

Social cascades are based on word-of-mouth and could potentially reach far and wide throughout the network. To understand how widely social cascades spread beyond the first hop from the uploader, we show the fraction of social cascade-based favorite markings that are just 1-hop away from the uploaders in the rightmost columns in Table 5, denoted *cascades from uploaders*. For pictures with 1 to 5 fans, 2.11 million out of 2.19 million social cascade-based favorite markings were from fans 1-hop away from the uploaders. For popular pictures with more than 500 fans, only 4,017 out of 8,686 favorite markings were from fans 1-hop away from the uploaders. These observations suggest that uploaders play a crucial role in the social cascades of less popular pictures, while social cascades of popular pictures spread information beyond the immediate vicinity of the uploader.

6.4 Peer pressure in photo favorite marking

We check whether a user's tendency to favorite-mark a picture is influenced by the number of friends who have previously favorite-marked the same picture. If one has many friends who declare that they "like" a given picture, is the user more likely to mark the picture as a favorite in the future? To answer this question, we focused on the set of 3,685 pictures with more than 100 fans. We examined the number of times users were exposed to those pictures through their 1-hop friends and counted how many of them later became fans of those pictures. Figure 7 shows the result, where the probability of becoming a fan is shown as a function of the number of friends who have already favorite marked the picture. The plot shows the average value.

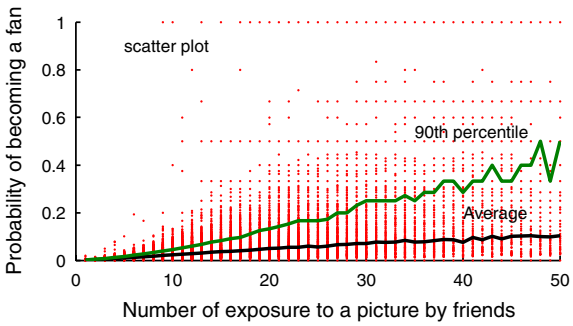


Figure 7: The probability of becoming a fan as a function of the number of friends who are already fans of the same picture

The probability of a user becoming a fan of a photo increases with the number of her friends who are already fans of the photo. That is, the behavior of favorite marking is influenced by peers in the social network. Backstrom et al. [3] also found a similar relationship in their study of LiveJournal and DBLP data, where the probability of joining a community increased as the number of friends who are already in the community increased.

6.5 Time taken for social cascade hops

Finally, we look at how long it takes for information to propagate along each hop of the social cascade. For each favorite marking that spread over a social link, we examine the time difference between when the favorite marking was made and the earliest time when one of the user's friends favorite-marked the same picture. When a friend's favorite marking precedes the time of link formation with the user, we take the time when the link was formed as the earliest possible time the user was exposed to the picture. Table 6 presents the time taken for information to propagate along a social link in units of days. We only consider favorite-marking of pictures with more than 100 fans, users are less likely to be exposed to pictures with few fans. This subset of data accounts for 190,353 favorite markings for 3,685 pictures.

Table 6 shows that some users discovered their friends' favorite photos and marked them as their favorites the same day that their friends favorite-marked them. While not presented in the table, our data shows that 35% of fans found their favorite marked pictures within a week after friends' favorite marking of the same picture. However, 50% of fans took over 60 days to favorite mark their friends' favorite-marked pictures. The average delay in information propagation across a social link is significantly higher exposure at 140 days. One user favorite marked what his friends had already favorite marked after 904 days. These observations indicate that favorite marking information takes a long time to spread across each link in the social network.

Table 6: Exposure time in days prior to favorite marking

# Pictures	# Fans	Min.	Med.	Avg.	Max.
3,685	190,353	0	60	140	904

6.6 Summary

We have examined to what extent information propagates through social links in Flickr. Our analysis suggests that (*a*) social network plays a notable role in Flickr, likely accounting for more than 50%

of the favorite marking and (*b*) individuals take a long time, typically 3 to 5 months, to favorite mark pictures that their friends had previously listed as favorites.

7. DISCUSSION

In this paper, we have made two key observations: first, most fans of a given picture are within a few hops of the picture uploader and second, pictures spread slowly throughout the social network. These observations contradict our expectations and need to be investigated thoroughly. In this section, we discuss two possible explanations of the high content locality and one potential explanation of delay in the social cascade.

In Section 4, we showed that most favorite-markings for a photo come from the 2-hop neighborhood region of its uploader, which is a small part of the entire Flickr social network. While it is not surprising that personal photos would exhibit high content locality, we have also seen that even popular photos with more than 500 fans have substantially limited popularity beyond a 2-hop neighborhood. One possible explanation for such high content locality is suggested by models of viral marketing. Watts and Peretti [30] describe a word-of-mouth marketing model as follows. It starts with "seeds" of individuals who spread information by infecting their friends, in a similar fashion to the spread of an infectious disease. The expected number of new infectious generated by each infected person is called the *reproduction rate* or R. If $R > 1$, each person is infecting more than one additional person and the number of infected people will grow exponentially, i.e., viral marketing is a success. When $R < 1$, initial seeds will quickly burn themselves out after several steps of information spreading. In this case, the final number of infected people will be approximately

$$\frac{N}{1-R} \tag{1}$$

where N is the number of initial seeds. In Flickr, the uploader is often the only seed who actively advertises a photo (i.e., $R > 1$), and the characteristics of R may change drastically beyond the immediate neighborhood of the uploader (i.e., $R \ll 1$). This may explain high content locality near the uploaders.

Another possible reason for high content locality might be related to homophily in social networks [2,6]. Homophily is colloquially described by the aphorism "birds of a feather flock together." It refers to the principle that contacts between similar people (i.e., "birds of a feather") occur at a higher rate than among dissimilar people [20]. In Flickr, homophily can be explained as follows: people who like each other's pictures tend to become friends and people who are friends tend to like each other's pictures, thereby ensuring that popularity of pictures is localized, even for top popular pictures.

To see if this is true, we examined what fraction of links, from a user A to a user B, were established *after* A had favorite marked some of B's pictures. From a random selection of 150,000 new links, we found that 27,546 or 18% of the links were formed after favorite marking the others' pictures. In 83% of the cases, A was previously only 2-hops away from B. Still, 17% of the remaining links indicate that users actively reach out to content creators who were beyond the friend-of-a-friend range. By strategically forming links, users can later visit the web pages of their new friends and follow up on their content.

In Section 5, we showed that even top popular photos took a long time to propagate from one friend to another. This delay may be related to the rate at which users are exposed to the new pictures bookmarked by their friends. In Flickr, users get a small number of updates about their friends' newly uploaded pictures when they log in. So the rate of information propagation may be limited by the frequency of user logins. In summary, our findings about high content locality and slow spread of information propagation through social links could potentially be explained by the burnout process in the theory of information diffusion [25, 27, 30] and the frquency with which users are exposed to the information.

8. RELATED WORK

We briefly review related work on theories in information diffusion, viral marketing, and the data analysis of information spreading on online social networks. Studies related to social cascades go as far back as the 1950s [25, 27]. Seminal work on persuasive communication, the branching process, and the diffusion of innovations spawned an extensive literature in sociology, economics, social psychology, political science, marketing, and epidemiology [19, 23, 28]. More recently, research on information diffusion has been conducted in light of viral marketing [7, 11, 24] and social networks [1, 10, 12]. Especially, Leskovec et al. [15] studied the cascade characteristics of purchases in a recommendation referral network of a large retailer website. They found that the distribution of the size of cascades followed a power-law distribution.

A number of studies focus on the interplay between social structure and information dissemination in real networks [2, 5, 9, 13, 16–18]. Amongst them, diffusion in blogsphere has been studied based on the keywords [10] as well as links embedded in blog posts [1]. Kossinets et al. studied a university email network to identify the information "backbone," where information has the potential to flow the quickest [13]. Anagnostopoulos et al. examined the spreading of picture tags in Flickr and developed a statistical test to distinguish social influence (causality) from correlation [2]. They found that the choice of tags used by Flickr users is not likely to be due to social influence. Gómez et al. studied the social network that instantly arise on the discussion threads in Slashdot website [9]. They studied the dissemination tree in respect to identifying how controversial a post was.

The most similar to our work is by Lerman and Jones [14]. They studied Flickr and also found that the social network played a significant role in photo propagation. They confirmed this by examining the correlation between the number of fans for 1,500 photos and the indegree of uploaders of those pictures. In this work, we have examined the influence of not only the uploaders, but also neighboring fans. We have also examined the detailed spatial and temporal growth patterns of photo popularity.

9. CONCLUDING REMARKS

This paper presented a data analysis of how picture popularity is distributed across the Flickr social network, and characterized the role played by social links in information propagation. We showed empirical evidence that (*a*) social links are the dominant method of information propagation, accounting for more than 50% of the spread of favorite-marked pictures; (*b*) information spreading is limited to individuals who are within close proximity of the uploaders; and (*c*) spreading takes a long time at each hop. As a result, we found that content popularity is often localized in the network and popularity of pictures steadily increases over many years.

While the popularity pattern observed is natural for many personal photos, we have also observed similar trends for popular photos with hundreds of fans. Our findings differ from from the common expectations about the quick and wide spread of word-of-mouth effect, and they need to be investigated thoroughly.

We would like to extend our work in many directions. First, we would like to understand the mechanisms of user behaviors that

leads to a substantial delay and high content locality in information propagation. Second, we are interested in developing tools and features in Flickr that can enable the full viral spread that the theory suggests is possible. For instance, content may propagate more quickly and widely in a push-based system, compared to the pull-based system used in Flickr. Third, we are interested in exploring opportunities for personalized recommendations in Flickr. We have seen that users are interested in local content (within 2-hop neighborhood), but it took a long time to for many users to reach that content. We would like to test the efficacy of recommending photos that are popular within one's local neighborhood, rather than from the entire user population (as Flickr currently provides in the Explore list).

10. ACKNOWLEDGMENTS

We thank Augustin Chaintreau, Anja Feldmann, Duncan Watts, Divesh Srivastava, Rasmus Pagh, Mikkel Thorup, Juan Antonio Navarro Pérez, Bryan Ford, Nuno Santos, Bimal Viswanath, Rose Hoberman, and anonymous reviewers, for their valuable comments.

11. REFERENCES

[1] E. Adar and L. A. Adamic. Tracking Information Epidemics in Blogspace. In *ACM Intl. Conf. on Web Intelligence*, 2005.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and Correlation in Social Networks. In *ACM SIGKDD*, 2008.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *ACM SIGKDD*, 2006.

[4] Bebra Aho Williamson. EMarketer Social Network Marketing: Ad Spending and Usage. 2007.

[5] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Online Social Networks: Sheer Volume vs Social Interaction. In *ACM IMC*, 2008.

[6] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback Effects between Similarity and Social Influence in Online Communities. In *ACM SIGKDD*, 2008.

[7] P. Dodds and D. Watts. A Generalized Model of Social and Biological Contagion. *J. of Theoretical Biology*, 2005.

[8] P. Domingos and M. Richardson. Mining the Network Value of Customers. In *ACM SIGKDD*, 2001.

[9] V. Gómez, A. Kaltenbrunner, and V. López. Statistical Analysis of the Social Network and Discussion Threads in Slashdot. In *Proc. of WWW*, 2008.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace . In *Proc. of WWW*, 2004.

[11] J. Hartline, V. S. Mirrokni, and M. Sundararajan. Optimal Marketing Strategies over Social Networks. In *Proc. of WWW*, 2008.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence Through a Social Network. In *ACM SIGKDD*, 2003.

[13] G. Kossinets, J. Kleinberg, and D. Watts. The Structure of Information Pathways in a Social Communication Network. In *ACM SIGKDD*, 2003.

[14] K. Lerman and L. Jones. Social Browsing on Flickr. In *Proc. of Int. Conf. on Weblogs and Social Media*, 2007.

[15] J. Leskovec, L. A. Adamic, and B. A. Huberman. The Dynamics of Viral Marketing. *ACM Trans. on the Web (TWEB)*, 2007.

[16] J. Leskovec and E. Horvitz. Planetary-Scale Views on a Large Instant-Messaging Network. In *Proc. of WWW*, 2008.

[17] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical Properties of Community Structure in Large Social and Information Networks. In *Proc. of WWW*, 2008.

[18] D. Liben-Nowell and J. Kleinberg. Tracing Information Flow on a Global Scale using Internet Chain-Letter Data. *Proc. Natl. Acad. Sci. USA*, 2008.

[19] R. M. May and A. L. Lloyd. Infection Dynamics on Scale-Free Networks. *Physical Review E*, 2001.

[20] M. Mcpherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

[21] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *ACM SIGCOMM Workshop on Online Social Networks*, 2008.

[22] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *ACM IMC*, 2007.

[23] R. Pastor-Satorras and A. Vespignani. *Epidemics and Immunization in Scale-Free Networks*. Wiley, Berlin, 2005.

[24] M. Richardson and P. Domingos. Mining Knowledge Sharing Sites for Viral Marketing. In *ACM SIGKDD*, 2002.

[25] E. M. Rogers. *Diffusion of Innovations*. Free Press, New York, 5th edition, 2003.

[26] TechCrunch, 2 Billion Photos on Flickr.

[27] T. W. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, Cresskill, N.J., 1995.

[28] D. J. Watts. A Simple Model of Global Cascades on Random Networks. *Proc. Natl. Acad. Sci. USA*, 2002.

[29] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and Search in Social Networks. *Science*, 294:1302–1305, 2002.

[30] D. J. Watts and J. Peretti. Viral Marketing for the Real World. *Harvard Business Review*, 2007.

[31] YouTube Fact Sheet. `http://www.youtube.com/t/fact_sheet`.