Characterizing Social Cascades in Flickr

Meeyoung Cha
MPI-SWS
Campus E1 4
Saarbrücken, Germany

Alan Mislove
MPI-SWS
Campus E1 4
Saarbrücken, Germany

Ben Adams
MPI für Informatik
Campus E1 4
Saarbrücken, Germany

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

ABSTRACT

Online social networking sites like MySpace and Flickr have become a popular way to share and disseminate content. Their massive popularity has led to the viral marketing of content, products, and political campaigns on the sites themselves. Despite the excitement, the precise mechanisms by which information is exchanged over these networks are not well understood.

In this paper, we investigate *social cascades*, or how information disseminates through social links in online social networks. Using real traces of 1,000 popular photos and a social network collected from Flickr, and a theoretical framework borrowed from epidemiology, we show that social cascades are an important factor in the dissemination of content. Our work provides an important first step in understanding how information disseminates in social networks.

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, epidemiology

1. INTRODUCTION

Online social networking has recently become a popular way to share and disseminate information. Users in websites like MySpace, Facebook, Flickr, and YouTube connect to each other with the purpose of finding and exchanging content. Their massive popularity has led to the viral marketing of content, products, and political campaigns on these sites. For instance, major movie studios place trailers for their movies on MySpace; US presidential candidates run 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, the precise mechanisms by which information is exchanged over these networks are not well understood.

One of the distinguishing features of online social networks is information dissemination along social links. Content in the form of ideas, products, and messages spreads across social connections like a virus: one person discovers new content and shares it with a few of their friends, who share it with a few of their friends, and so on. We call this spreading of a piece of content along links in a social network a *social cascade*.

Studies related to social cascades go as far back as the 1950s [13, 15]. Seminal work on persuasive communication, the branching process, and the diffusion of innovations spawned extensive literature in sociology, economy, social psychology, political science, marketing, and epidemiology [3, 6, 8, 12]. One of the key challenges in these studies has been obtaining large-scale data on the spread of content and the underlying social contact structure. Today, the wealth of online social networking data available on the Web provides a unique opportunity to understand the interplay between social networks and content dissemination.

To investigate the role of social links in the spread of information, we collected traces of content dissemination from Flickr [4], an online social network for sharing photos. Using this data, we explore the following questions:

- Does content in Flickr spread along links in the social network?
- What are the properties of content dissemination in Flickr (e.g., how long after being exposed to a piece of content do users tend to propagate it)?
- Can existing epidemiological models characterize the information dissemination observed in Flickr?

Our collected traces contain (a) information about the evolution of the social network and (b) the favorite photos for all Flickr users in our crawled network. For each of the favorite photos of a given user, we have the timestamp at which the user marked the photo as a favorite. Thus, if one of the user's friends had previously marked the photo as a favorite, we can infer that information, in the form of a photo, was disseminated from one user to the other.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WOSN'08, August 18, 2008, Seattle, Washington, USA.
Copyright 2008 ACM 978-1-60558-182-8/08/08 ...$5.00.

In our analysis, we study the high-level properties of how information is disseminated over the Flickr social network. We explore two key properties of the social cascades in Flickr: how long it takes for the content to spread to the first user and how long it takes to spread to subsequent users. Understanding these two properties offers intuition as to how fast content spreads in Flickr. Finally, we apply an epidemiological model to estimate the basic reproduction number R_0 of a photo, a threshold parameter that can predict whether a photo will flourish or flounder in the network.

The remainder of this paper is structured as follows. We describe various mechanisms of content dissemination in online social networks in Section 2. In Section 3, we introduce the dataset and demonstrate the significance of the social network to content discovery. We present an epidemiological model and examine the intrinsic characteristics of Flickr's social cascades in Section 4. Finally, we discuss issues for future work and conclude in Section 5.

2. DISSEMINATION MECHANISMS

The proliferation of Internet-based communications helps users conveniently connect with others online to share, organize, and find a massive amount of content. The video sharing website YouTube serves over 100 million videos a day [16] and the photo sharing site Flickr — the subject of our study — contains over two billion unique photos [14]. There are a number of mechanisms by which users locate content on these sites:

- **Featuring:** Certain content is placed strategically within the website to catch users' attention, such as on the front page. Other examples are the top 100 videos in YouTube's "most viewed" list and the list of photos selected by Flickr as "interesting."[1]

- **External links:** Users can find content through links from external websites, blogs, emails, and other mechanisms external to the site itself. According to a previous study [2], 47% of YouTube videos have links from external websites. This indicates that external links also serve as a common mechanism for users to find content.

- **Search results:** Users may search for specific content. Content metadata such as titles, tags, and descriptions are used by search engines embedded in the sites to find relevant content.

- **Links between content:** Sites often provide links to related content when users are browsing. For instance, YouTube pages contain thumbnails of related videos and video responses, and Flickr pages contain thumbnails of other photos by the same photographer.

- **Social network:** Users share content with their online social contacts. Different mechanisms exist on different sites, such as "channel subscriptions" on YouTube and "favorite photos" on Flickr. Regardless, these mechanisms allow a user's contacts to find content which the user has found interesting or useful.

[1] `http://www.flickr.com/explore/`

A large and rapidly growing number of content sharing websites support online social networking features, allowing users to exploit their social contacts to find and spread information. In this paper, we focus only on the effect of the social network on the dissemination of content in Flickr. Undoubtedly, other mechanisms are also at play in other networks, but studying their influence requires a richer dataset and is beyond the scope of this paper.

3. MEASUREMENT METHODOLOGY

In this section, we introduce our dataset and the relevant characteristics of Flickr, examine the growth in popularity of two sample photos, and describe the methods we employ for identifying social cascades.

3.1 Flickr datasets

We use existing Flickr social network data [9, 10] for analysis. The dataset contains daily snapshots of the large weakly connected component (WCC) of the Flickr social network, which covers approximately 25% of the entire Flickr user population (the remaining users were not connected to the large WCC). The dataset covers the evolution of the Flickr social network over approximately 100 days of growth. During this period, the Flickr social network grew from 1.6 million to over 2.5 million users. We refer the reader to these papers for a discussion of the dataset's limitations.

To get empirical data on how information propagates in Flickr, we focus on the "favorite photos" feature in Flickr, which allows users to maintain a list of their favorite photos on the site. Each user's favorite photos list is publicly visible, and photos from this list are shown to the users' contacts when they log into Flickr. We refer to users who include a photo in their favorite photos list as *fans* of that photo.

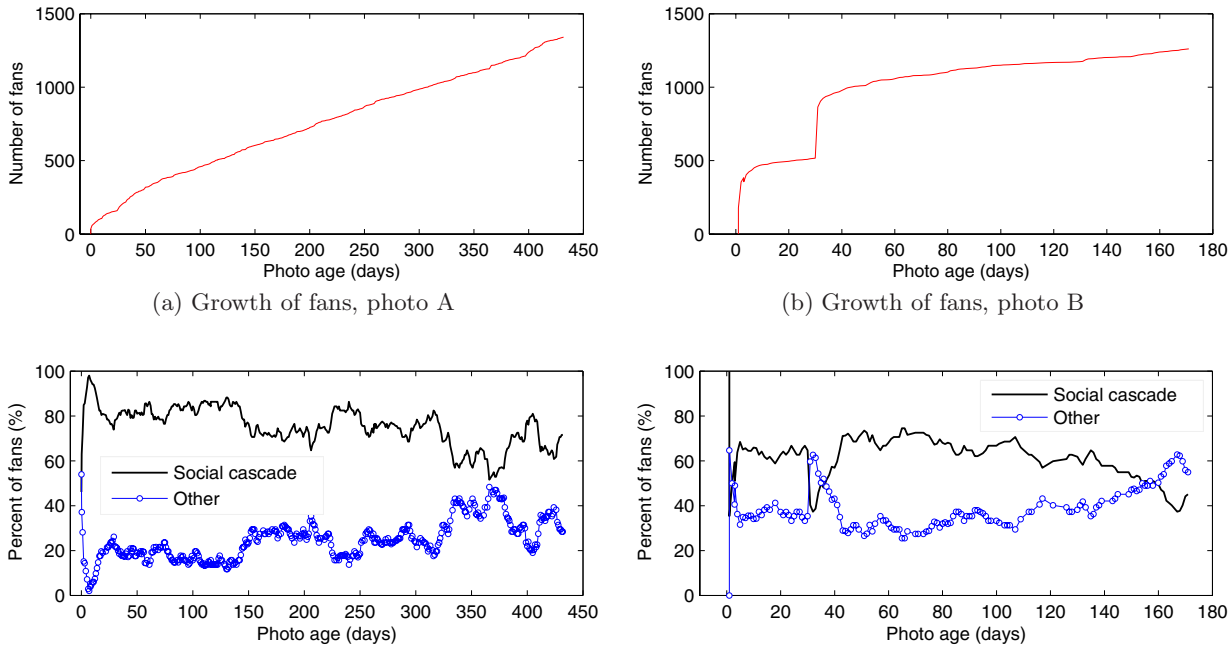
For each user in the above datasets, we used the Flickr API to download the list of that user's favorite photos, as well as the timestamp at which the user marked the photo as a favorite. In total, our photo dataset contains information covering 34,734,221 favorite markings over 11,267,320 distinct photos.

3.2 Social cascades

We only have data on the Flickr social network for 100 days [10] and do not know the state of the social network before or after this span of time. Therefore, we only focus on the photo favorite markings which occurred during these 100 days and do not consider any favorite markings which occurred outside of this time period.

Without a page view trace or asking the user directly, unfortunately, we are unable to say for sure how users in Flickr found photos. To estimate the influence of the social network in information dissemination, we use the combination of the social network state and the timestamps of favorite markings to take an educated guess about how a given user found a given photo. In particular, we say that user A found a photo P through the social network if and only if

- There exists some user B who also marked P as a favorite, and
- B included photo P on his favorite list before A included photo P on his favorite list, and
- B was a contact of A before A included photo P on his favorite list.



(a) Growth of fans, photo A

(b) Growth of fans, photo B

(c) Breakdown of new fans, photo A

(d) Breakdown of new fans, photo B

Figure 1: Evolution of the number of fans of photos A and B. The bottom plots show the fraction of new fans that are part of a social cascade. Both photos show strong evidence of social cascades.

In short, this means that B was A's contact before A found the photo, and B had already found the photo. If all of these conditions hold, then we consider the photo to have propagated across the $B \rightarrow A$ social link. Note that there may exist multiple such users from whom A could have found the photo – in this case, we consider all of the links as having been used. In other words, we assume A was exposed to the photo by all of these users.

3.3 Popularity Growth of Two Sample Photos

To ground our discussion of social cascades, we pick two popular photos (shown in Figure 2), and examine the growth in their number of fans over time.

(a) Photo A

(b) Photo B

Figure 2: Sample photos from Flickr

We show the number of fans over time for these photos in Figures 1(a) and 1(b), respectively. The horizontal axis represents time since upload of the photo, representing the photo's age on Flickr. Photo A shows steady linear growth, reaching 1400 fans over the course of 430 days. In contrast, photo B obtains approximately the same number of fans in a much shorter period of time, 180 days. Photo B shows two surges in popularity, one at day 1 when approximately 500 users become fans, and another at day 30. Growth is relatively slow in the intervening periods.

The difference in the pattern of fan growth between photos A and B may reflect different methods of information dissemination. Picture A's slow and steady fan growth may illustrate the social cascade pattern, in which users find their favorite photos from their contacts. Picture B's sudden surges in fan growth may illustrate the impact of featuring or external links, where photos are exposed to a large set of random users and increase their likelihood of being bookmarked.

We look for evidence of social cascades in the growth of popularity in these two photos. For each new fan, we determine whether one of that fan's contacts was already a fan (in accordance with our definition in Section 3.2). If such a previous fan exists, we place the new fan in the "social cascade" group. Otherwise we place the new fan in the "other" group.

Figures 1(c) and 1(d) show the fraction of new fans that participate in social cascades over time. We make several observations. First, the "social cascade" group accounts for over half of new fans for both photos. This suggests that the social network plays a significant role in content dissemination. Second, we observe that the dominance of the "social cascade" group over the "other" group switches during the two popularity surges exhibited by photo B. This suggests that during these surges in popularity, other mechanisms such as linking from external sites or featuring are driving the rapid increase in fans.

Motivated by these preliminary findings from the case studies, we delve further into the dynamic patterns of information dissemination through social links in Flickr in the next section.

4. SOCIAL CASCADES IN FLICKR

We now examine the social cascades of photos in Flickr. To examine the social cascade patterns, we only consider the users' bookmarking events where users find content through social links (i.e., the "social cascade" group from Figures 1(c) and (d)). We disregard any user favorite markings which do not meet the definition in Section 3.2. Then, we focus on the set of popular photos which had at least 300 fans. This comprises approximately 1,000 photos in our trace, totaling over 15,000 unique fans and 35,000 favorite markings. We choose to focus only on popular photos due to the amount of data these photos contain. We leave examining social cascades of unpopular photos as future work.

In this section, we (a) describe a baseline model of information dissemination through social links, (b) examine the characteristics of the social cascade in the Flickr trace, and (c) estimate the basic reproduction number of photos to quantify how widely those photos can spread within Flickr.

4.1 Social cascade model

We now present a model of social cascades that is similar to those employed to study the spread of infectious diseases through human populations [12], viral marketing [3], and diffusion of innovation [13,15]. A social cascade begins when the first user includes the photo in his list of favorites. This event does not result from any social links, but is necessary to initiate the cascade.[2] After the initialization event occurs, the cascade continues along social links.

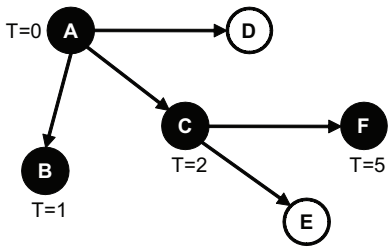


Figure 3: Illustrating example of a social cascade

Figure 3 illustrates a social cascade in a small example network of 6 users and 5 directional links. A social cascade proceeds in the direction of the arrows. For example, a directional link from user A to B indicates that B added A as his contact. Fans of the photo are indicated by black nodes. At time T = 0, user A marks a photo as a favorite. Over time, users having A as a contact, namely B, C, and D, are exposed to the photo. The first step of the cascade happens when B marks the photo as a favorite at time $T = 1$. The next step of the cascade happens when C marks the photo as a favorite at $T = 2$. User D remains unaffected by the photo and is indicated by a white node. This may be because D is not interested in the photo or has not yet been exposed to it. At $T = 5$, user F discovers the new photo because he is a contact of C, marks it as a favorite, and continues the social cascade. Thus, at the end of the cascade, users A, B, C, and F are infected, while users D and E are removed.

In our model, we assume that there is a set J of photos and a set U of users. Each user $u \in U$ is in one of three possible states for each photo $j \in J$: Initially, each user is susceptible to every photo (S_j^u). Once a user is exposed to a photo, they transition into one of two possible end states: If they include the photo in their list of favorites, they become infected (I_j^u); otherwise, they become removed (R_j^u) from the social cascade and do not propagate the photo.

Unlike the spread of most infectious diseases, the structure of Flickr suggests an interesting epidemiological phenomenon: A user will retain the same favorite indefinitely – there is no recovery – unless he removes the photo from his favorite list. While users can have multiple favorites, as the number of photos marked as favorites grows, the bookmarked photos are paged such that older ones have less chance of exposure to other users. For such pictures, the user effectively becomes no longer infectious and can be placed in the removed (R_j^u) class. In the epidemiological literature this replacement of infections is known as superinfection [11]. If our model holds in Flickr, we expect to see a certain correlation between the times users stay infected (I_j^u) and the node degree: While highly connected nodes are more likely to widely disseminate a picture, they are also likely to replace that picture with a different favorite very rapidly. So their transmission rate is high, but the duration of infection is short. Conversely, weakly connected nodes may have low transmission potential but a long duration of infection. The nodes that transmit most efficiently over long time periods may thus be of intermediate connectivity.

Another unique aspect of Flickr is the "social" behavior of users. In our example, 3 time steps were required before user F was infected, while users B and C needed only 1 and 2 time steps, respectively. These could be nothing more than randomly distributed waiting times. However, there is an extensive literature from sociology, social psychology, and mathematical physics concerning the binary decisions of users that exhibit "positive externalities" or "network effects" [6,7]. These studies describe how the probability that an individual will choose x over y increases with the relative number of others choosing x. Similarly the "threshold rule" posits that a switchover happens when sufficiently many others have adopted x in order for the perceived benefit of adopting a new innovation to outweigh the perceived cost. We will shortly explore this issue in Flickr.

Our modeling framework highlights a number of questions about the time and frequency parameters of social cascades. In particular, how fast is the social cascade and what fraction of the neighbors participate in the social cascade?

4.2 Cascade characteristics

We refer to the actors of our social cascades as *infectors* and *infectees*. We first examine two time-related parameters of the social cascade: the time to the first step of the cascade (the infector's perspective) and the duration of exposure to a photo before infection (the infectee's perspective). Both of these factors are of interest because they reflect how susceptible the network is to new content – or how fast content can spread through the network.[3]

First we focus on the infectors. For each photo, we identified the time lag between the point when the initializing

[2]Multiple simultaneous cascade initializations, such as those associated with featuring or external links, are also possible. We focus only on cascades with a single initialization event.

[3]Other factors that can affect the susceptibility are connectivity of the initializing node, attractiveness of the picture, currently dominant favorites, and the length of time for which the picture has been circulating. For this work, we focus on the network structure.

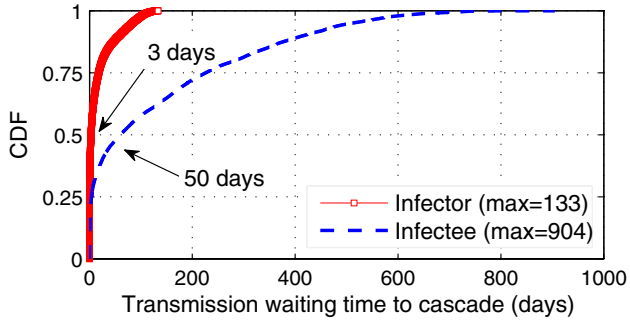


Figure 4: Latent time in social cascade events

user marks the photo as a favorite and the time at which the next user in the cascade marks the photo as a favorite. Since infection may persist indefinitely, the time to the first step of the cascade serves as an indicator of how fast content can spread through the Flickr network. The solid line in Figure 4 shows the cumulative distribution of time to the first step of the cascade in Flickr. We observe that 50% of first cascade steps happen in less than three days. This indicates new content can spread quickly from one user to another in the Flickr network. However, 20% of the first cascade steps happen take longer than a month. This could be because the picture was introduced in an isolated part of the network or because it initially faced strong competition from other pictures.

Now we focus on the infectees. For each social cascade infection, we computed the duration of exposure to a photo before the infection happens. To do so, we identify the earliest time when any contacts of the infectee marked the photo as a favorite; by subtracting this time from the time at which the infectee marked the photo as a favorite, we can determine the amount of time the infectee was "exposed" to the photo before marking it as a favorite himself. The dotted line in Figure 4 shows the cumulative distribution of the exposure time for infectees. We observe that 50% of cascade events occur within 50 days, an order of magnitude larger than the time before the first step of the cascade. Some cascades happen only after several years of exposure to the content, which is possibly due to infrequent user activity. It is also possible that some users have many contacts but only regularly check updates from a few of them.

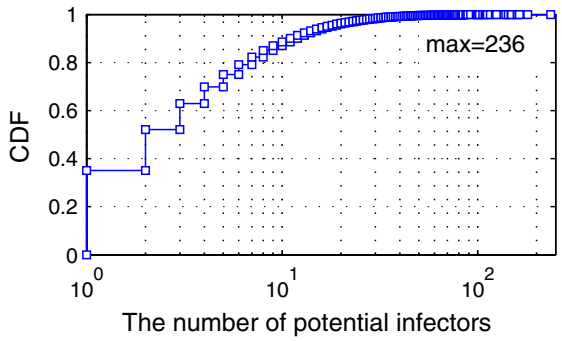


Figure 5: Network effects in a social cascade

Next, we investigate how the number of exposures affects the adoption rate. Specifically, we examine how many times users are exposed to a photo before they adopt it as a favorite. Figure 5 shows the cumulative distribution (in log scale) of the number of infected contacts at the time a user marks a photo as a favorite. We observe that 35% of social cascade events are influenced by a single infector; 20% of the events by two infectors; and the remaining 45% involve three or more potential infectors. For 10% of the events, the infectee had more than 10 contacts who had already marked the same photo as a favorite. Based on these findings, we plan to identify how the number of exposures to candidate favorite photos affects the rate of adoption.

4.3 The rate of cascade

In epidemiological models, the basic reproduction number R_0 is defined as the expected number of secondary infections resulting from a single infected individual in an entirely susceptible population. If $R_0 > 1$ then, on average, the one infected individual will infect more than one other individuals, and the epidemic will grow. Conversely, if $R_0 < 1$ then a small number of initial seeds will invariably fizzle out before many additional people are infected. Finally, $R_0 = 1$ is a special "critical" case where the outbreak changes its character from collapse to growth. Epidemiologists call this point the *epidemic threshold* and in public health the goal is to reduce R_0 below 1 in order to stop epidemics. HIV has an R_0 between 2 and 5; and measles has an R_0 between 12 and 18. Although the concept of R_0 is tied to populations that are entirely susceptible, it continues to hold as an approximation as long as the number of susceptible individuals is much larger than the number of infected or immune individuals.

The theory of epidemiological models shows that the basic reproductive number on a network is given by:

$$R_0 = \rho_0 \langle k^2 \rangle / \langle k \rangle^2 \qquad (1)$$

where $\rho_0 = \beta\gamma\langle k \rangle$ [8]. Here, β is the transmission rate, γ is the duration of infection, k is the node degree, and $\langle \cdot \rangle$ represents the mean value. However, in a model for photo circulation in Flickr, we assume that the natural duration of infection is equal to the lifetime of the user, which is very large in comparison to the timescale of the cascade. In this case, we can assume that a picture will definitely be shared between two connected nodes. If we then define σ_0 to be the probability that a person will adopt the picture when it is shared, we get $\rho_0 = \sigma_0 \langle k \rangle$.

An empirical estimate of the transmission probability of a picture σ_0 can be calculated by identifying an infected node and then counting the proportion of its connected nodes (i.e., social contacts in the reverse direction) that subsequently become infected. Knowing the transmission probability, we can then estimate the reproductive number $\hat{R}_0$ directly from Equation (1). The Flickr social contact network used in this study had a mean node degree, $\langle k \rangle = 14.7$, and high heterogeneity in the node degree distribution, $\langle k^2 \rangle / \langle k \rangle^2 = 48.0$. Similarly, an empirical value for the basic reproduction number R_0 can be assessed by counting the number of nodes directly infected by the initializing node. These will be underestimates because in the real network there is only a finite time for transmission before the picture is replaced by one of the many others in circulation.

Figure 6 compares the basic reproduction number R_0 obtained directly from the trace and the estimated value $\hat{R}_0$

calculated from Equation (1) using values for the transmission probability σ_0 directly obtained from trace. We observe reassuringly high correlation (Pearson correlation coefficient of 0.9765) between the two values across all photos. This is a promising result with regard to predicting the popularity of photos. Given the transmission probability of a picture derived from a short time series of user activity, we can then predict the expected spread of the photo for any network structure for which the node degree distribution k is known. This means that we can predict the reproduction number and the resulting spread of these 1,000 photos when they are adopted into other online social networks such as Facebook, Livejournal, and MySpace.

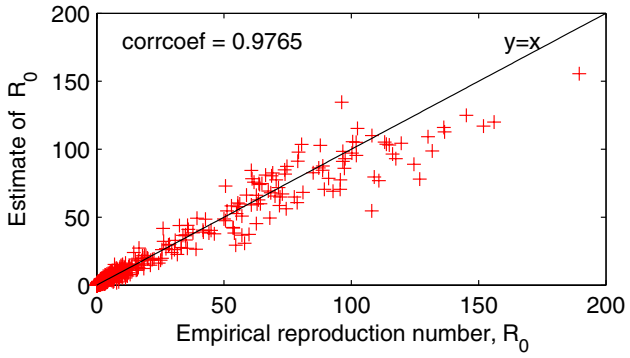


Figure 6: Estimating the reproduction number

4.4 Summary

We examined how information spreads through social links in Flickr. We found that social cascades are occurring in the spread of photos though the Flickr social network. We borrowed the theoretical framework of epidemiological models and calculated the basic reproduction number, an important threshold statistic determining whether a photo will flourish or flounder in the network. We found the basic reproduction number of popular photos to be between 1 and 190. This is much higher than very infectious diseases like measles, indicating that social networks are efficient transmission mediums and online content can be very infectious.

In our analysis of the level of heterogeneity and network effects, we found that photos have the potential to circulate rapidly within Flickr and individual users vary considerably in the frequency and the duration of exposure to a new picture before they adopt it. Future research with the methods developed here should take into account that other online social networks may have a quite different epidemic environment and dynamics due to the web design and community features.

5. CONCLUDING REMARKS

We examined content dissemination though social networks, which we call social cascades, using real traces collected from Flickr. We studied the influence of social contacts in the bookmarking of favorite photos, examined the dynamics of social cascade patterns, and applied an existing epidemiological framework to estimate the potential spreading capability of 1,000 popular photos. We found that social links are an active mechanism for disseminating information in online social networks.

Our work is one of the first studies to leverage real traces of a large online social networking website for studying the effects of social cascades. There are several directions that we wish to pursue as future work. We are interested in identifying and quantifying the impact of influential users [6] and network topology [5], as well as other factors that affect the spread of information. We would also like to compare the social cascade patterns with those of the offline human travel models in [1], to understand the similarities between online and offline dissemination. While this paper was limited to investigating social cascades due to the availability of data, future studies should examine other information dissemination mechanisms such as featuring, related content recommendations, and links from external sites.

6. REFERENCES

[1] D. Brockmann, L. Hufnagel, and T. Geisel. The Scaling Laws of Human Travel. *Nature*, 2006.

[2] M. Cha, H. Kwak, P. Rodriguez, Y-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. of ACM IMC*, 2007.

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

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

[5] A. J. Ganesh, L. Massoulié, and D. F. Towsley. The Effect of Network Topology on the Spread of Epidemics. In *Proc. of IEEE INFOCOM*, 2005.

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

[7] S. Liebowitz and S. Margolis. *Network Effects and Externalities*. The New Palgrave's Dictionary of Economics and the Law, MacMillan, New York, 1998.

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

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

[10] A. Mislove, H. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *Proc. of WOSN*, 2008.

[11] M. A. Nowak and R. M. May. Superinfection and the Evolution of Parasite Virulence. *Biological Sciences*, 1994.

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

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

[14] TechCrunch. *2 Billion Photos on Flickr*. `http://www.techcrunch.com/2007/11/13/2-billion-photos-on-flickr/`.

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

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