Beyond Friendship Graphs:
A Study of User Interactions in Flickr

Masoud Valafar, Reza Rejaie
Department of Computer & Information Science
University of Oregon
Eugene, OR 97403
{masoud,reza}@cs.uoregon.edu

Walter Willinger
AT&T Research Labs
180 Park Ave. - Building 103
Florham Park, NJ, 07932
walter@research.att.com

ABSTRACT

Most of the existing literature on empirical studies of Online Social Networks (OSNs) have focused on characterizing and modeling the structure of their inferred friendship graphs. However, the friendship graph of an OSN does not demonstrate what fraction of its users actively interact with other users, how these users interact, and how these active users and their interactions evolve over time. In this paper, we characterize indirect fan-owner interactions through photos among users in a large photo-sharing OSN, namely Flickr. Our results show that a very small fraction of users in the main component of the friendship graph is responsible for the vast majority of fan-owner interactions; moreover, these interactions involve only a small fraction of photos in Flickr. We also characterize some of the temporal properties of fan arrival. For example, we show that there is no strong correlation between age and popularity of a photo and that most photos gain a majority of their fans during the first week after their posting. Overall, our findings provide new insights into the fan-owner interactions among Flickr users.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems

General Terms

Measurement

Keywords

Online Social Networks, User Interaction, Measurement

1. INTRODUCTION

A majority of published empirical studies of OSNs have focused almost exclusively on characterizing various properties of the inferred friendship graph of a target OSN (*e.g.*, [5, 1, 4]). While these studies provide valuable information about the structure of friendship relations among users of an OSN, they generally ignore the fact that not all users may be equally active and that the level of user activity in OSNs is likely to be highly dynamic (*i.e.*, different sets of users are active at different point of time). We are aware of only two studies on characterizing some aspects of user interactions in OSNs [3, 2]. However, the findings of both of these studies are somewhat limited by the nature of the available data. In fact, there exists anecdotal evidence that large fractions of users in different OSNs (even users with apparently many friends) do not interact with other users (*i.e.*, are not active) Given this observation, we argue that identifying and characterizing the "active" portion of an OSN's friendship graph and its evolution over time would clearly be more meaningful than continuing with the current (over-)emphasis on characterizing static friendship graphs as a whole. In particular, the following important questions about user interactions have not been addressed in the existing OSN literature: *(i)* What fraction of users in an OSN actively interact with other users in the system? *(ii)* Do the active users form a core in the interaction graph? *(iii)* What are the temporal properties of interactions among users? The general lack of attention to user interactions in prior studies of OSNs is mainly due to the difficulties associated with capturing user interactions through measurement. OSNs do not provide any means to obtain this information from their server easily and have no incentive to make this information publicly available.

In this paper, we tackle the above questions by characterizing the indirect interactions (*i.e.*, relationship) between fans and owners of photos in a popular photo sharing OSN, namely Flickr. Our main findings can be summarized as follows: First, the extent of fan-owner interaction is very limited in Flickr. More specifically, a very small fraction of users are fans of a very small fraction of photos which in turn are owned by a very small fraction of users. Furthermore, the vast majority of fan-owner interactions (>95%) are between a small fraction of users in the main component (*i.e.*, largest component) of the friendship graph. Second, active users appear to form a core in the interaction graph. There is a clear correlation between the level of activity of a user as a fan and as an owner. The top 10% of fans and owners (80K users) that are responsible for 80-90% of fan-owner interactions in the systems exhibit 50% overlap and 15% reciprocation (*i.e.*, bi-directionality of fan-owner relationship). Focusing on a smaller percentage of highly ranked users leads to a significantly smaller overlap but much higher level of reciprocation. Third, while older photos can reach higher popularity, there is no strong correlation between age and popularity for a majority of photos. Newer photos ap-

WOSN'09, August 17, 2009, Barcelona, Spain.
Copyright 2009 ACM 978-1-60558-445-4/09/08 ...$10.00.

pear to reach their target popularity much faster than older photos. However, closer examinations revealed that most photos receive a majority of their fans during the first week after posting. Therefore, older photos experience a lower average fan arrival rate simply due to a longer inactive period.

The rest of this paper is organized as follows. Section 2 discusses our measurement methodology and describes our datasets. We explore the extent of fan-owner interactions among users and connectivity among active users in Section 3. Section 4 examines temporal characteristics of fan-owner interactions among active users. Finally, Section 5 concludes the paper and briefly describes our future plans.

2. MEASUREMENT METHODOLOGY

Flickr is a popular OSN for photo sharing. Individual users can post their own photos, view photos posted and owned by other users, become fan of posted photos (i.e., tag them as their "favorite" photos), and comment on posted photos. In essence, Flickr users can indirectly interact with one another through posted photos, as opposed to directly interacting by exchanging messages.

2.1 Representing Fan-Owner Interactions

We use a detailed representation of fan-owner interactions (or relationships) through their photos in Flickr as shown in Figure 1. Fans are grouped on the left, owners are grouped on the right, and photos are grouped in the middle column. Note that a user may appear both as a fan and as an owner. Each fan has one or more *favorite* photos. An edge from fan C to photo p indicates that p is one of C's favorite photos and thus represents an indirect interaction between fan C and the owner of photo p. An edge from photo p to owner A simply indicates that p is owned by user A. Fans, photos and owners are then separately ranked in descending order, based on their level of "activities" (or amount of interactions) which we define as follows:

- *Activity of users as fans* is determined by the number of favorite photos per fan (*i.e.*, outgoing degrees of fans in Figure 1);
- *Activity of photos* is determined by the number of fans (or "popularity") per favorite photo (*i.e.*, incoming degrees of photos in Figure 1); and
- *Activity of users as owners* is determined by the number of "favored" photos (that is, photos with one or more fans) posted by each owner (*i.e.*, incoming degrees of owners in Figure 1).

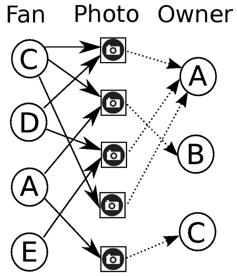


Figure 1: Indirect fan-owner interactions

This representation of indirect fan-owner interactions in Flickr clearly separates the roles of a user as a fan and as an owner, and illustrates the key role individual photos play in this context. Note that we do not consider a user as "active" if they only browse through user photos without declaring any photo as favorite or posting some favored photos. The reason for this is twofold. First, we are unable to capture appropriate measurements for studying such browsing activities, and second, our focus is on user interactions (or relationship) that enhances the overall value of an OSN. Characterizing other types of user interactions remains as a future work.

2.2 Data Collection

Flickr provides a well documented API[1]. We leverage this API to query the Flickr server and obtain (publicly available) information about fan-owner interactions among users using the following two strategies:

Crawling Owned Photos: To identify the list of *fans* for individual photos posted by user u, we first have to query the server to obtain the IDs of all photos owned by u. Then we need to issue a separate query to the server for each photo owned by u to obtain the user IDs of all the fans of the photo and associated timing information (*i.e.*, when the fan declared the photo as her "favorite"). This approach discovers fan-owner interactions from the owner side and provides timing information. However, it is inefficient and slow – it requires a separate query for individual photos, even though a majority of the discovered photos do not have any fans.

Crawling Favorites Photo List: For a given user u, we can query the server to obtain the IDs of favorite photos (along with the ID of their associated owners). This process discovers fan-owner interactions from the fan side without providing any timing information. However, this approach is very efficient because the number of required queries is proportional to the number of users (which is much smaller than the number of photos), and each query discovers some new fan-owner interactions.

2.3 Datasets

Similar to many other OSNs, Flickr limits the rate with which a user can query the server. This limit for Flickr is 10 queries/second. This limit on the rate of queries, coupled with the inefficiency of the first approach (*i.e.*, crawling owned photos), makes the second approach (*i.e.*, crawling favorite photo lists) a very appealing alternative for data collection. We have collected a dataset with each of the above two measurement approaches for capturing fan-owner interactions as follows:

Dataset I (Interactions of Random Users): Selecting random users in Flickr is feasible since user IDs have a well known format that consists of a six-to-eleven digit prefix, followed by "@N0" and a one-digit suffix (*e.g.*, 1234567890@N02). Using this feature, we identified about 122K random Flickr user IDs and collected their user-specific attributes, including their posted photos, associated fans and their arrival times, and favorite photos and associated owners. This collection represents photos that are posted by a random set of users and thus provides a representative sample of fan-owner interactions in Flickr through these photos [2].

[1] http://www.flickr.com/services/api/

[2] We noticed that the obtained information for a very small

Table 1: Dataset I - Interactions of random users

	# users	# fans	# owners	# photos	# favored photos	# favorite photos
Singletons	101,210	2,638	1,230	835,970	3,734	24,078
MC_f users	21,127	4,053	5,075	2,646,139	142,391	532,333

Using these 122K randomly selected users as seeds, we also crawled the friendship graph by progressively obtaining the friend lists of known users. This allowed us to identify the main component of the friendship graph (denoted by MC_f) and determine which subset of the randomly selected users are part of MC_f. This analysis revealed that while the MC_f consists of about 4,200K users, only around 21K of our randomly selected users are located within the MC_f (with the rest being mostly singletons[3]). Since only 21K of our randomly selected users (*i.e.*, 1 out of 6) are located within MC_f, the total population of users in Flickr is approximately 6 times the size of the main component or about 25 million users.

Dataset II (Interactions of MC_f Users): To capture a more complete snapshot of fan-owner interactions among users in MC_f, we crawled the friendship graph (*i.e.*, using the friend lists of individual users) to identify its main component (MC_f). We collected the list of favorite photos (and their owners) for all the users in MC_f as well as any new user that we discover as an owner of a favorite photo. Since we discover edges of the interaction graph that are associated with reachable fans in MC_f, we miss those interactions that are associated with singleton fans or unreachable fans within the main component. However, we argue that the percentage of these missing interactions can be expected to be very small. For one, only a very small fraction of fans (2.6%) are singletons, and second, a crawl of the friendship graph tends to reach a significant portion of MC_f due to the large number (some 21K) of randomly selected seeds within MC_f. Table 1 presents the number of randomly selected users in Dataset I that are singleton or MC_f users in separate rows. It also shows the number of users that are fans and owners. Furthermore, Table 1 reports the total number of photos posted by each type of users, and a subset of these photos that are favored or favorite. Table 2 shows the total number of MC_f users in Dataset II, number of users that are fans or owners, and the number of favorite photos associated with these users.

Table 2: Dataset II - Interactions of MC_f users

# users	# fans	# owners	# favorite photos
4,140,007	821,851	1,044,055	31,495,869

3. CHARACTERIZING INTERACTIONS

3.1 Extent of Fan-Owner Interactions

To examine the extent of fan-owner activity, we first focus on Dataset I and then validate our findings using Dataset II. We are interested in determining the portion of "active" photos as well as in identifying and locating the fractions of associated users that are "active" in their roles as a fan or as an owner.

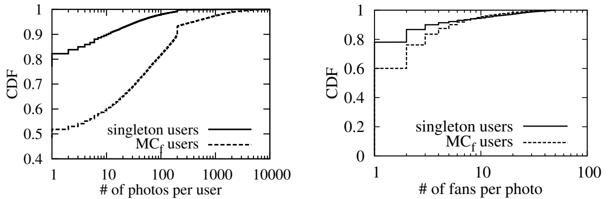
Active Photos: The 120K randomly selected users collectively posted 3,482K photos; of those, around 836K were posted by singleton users and 2,646K by MC_f users, *i.e.*, MC_f users contribute three times more photos than singleton users. Figure 2(a) depicts the distribution of the number of all photos (with or without fans) posted by MC_f users and singleton users in Dataset I. This figure shows that around 48% of MC_f (18% of singleton) users post more than one photo.Furthermore, the number of posted photos by individual MC_f users varies across a wider range (2 to 10K photo/user) compared to singleton users (2 to 1K photo/user). The sudden change at 200 photo/user for MC_f users is due to a Flickr-imposed 200-limit for the number of posted photos by regular users. Users with more than 200 photos are considered "professional" users and are expected to pay a fee for using Flickr.

To examine interactions, we are only interested in posted photos that are "active," *i.e.*, have at least one fan. From Table 1, we see that these active photos make up a very small fraction of the total number of posted photos, namely 3K (0.4%) of photos owned by singletons and 142K (5.3%) of photos owned by MC_f users. *This demonstrates that the vast majority of active photos is owned by MC_f users.*

Active Owners: We consider a user in her role as owner to be "active" if she has at least one photo with a fan. Table 1 demonstrates that out of 101,210 singleton and 21,127 MC_f users in the random datasets, only 1,230 (1.2%) and 5,075 (23%) are active owners, respectively. Moreover, Table 1 reveals that those 1,230 singleton active owners have 3,734 fans while the 5,075 MC_f active owners have a total of 142,391 fans. *This shows that more than 97% of fan-owner interactions are associated with active MC_f owners.*

Active Fans: We consider a user in her role as a fan to be "active" if she has at least one favorite photo that is owned by another user. Table 1 indicates that only 2,638 (2.6%) of singleton users and 4,053 (18.4%) of MC_f users in our dataset are active fans. Moreover, those 2,638 active singleton fans have only a total of 24,078 favorite photos while the 4,053 active MC_f fans have 532,333 favorite photos. *This means that more than 95% fan-owner interactions are associated with active MC_f fans.*

In summary, the above findings about fan-owner interac-

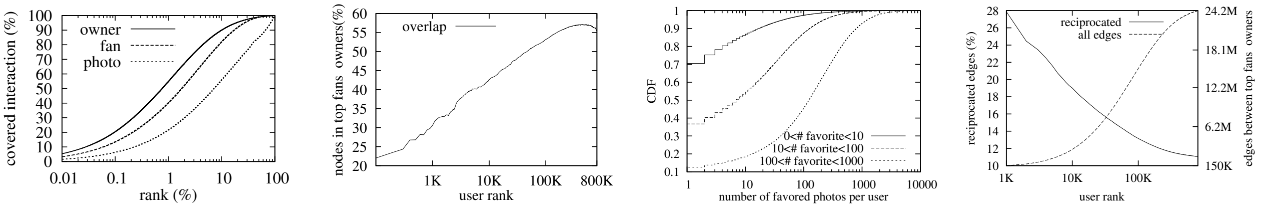


(a) Dist. of posted photos per user (b) Dist. of fans per photo

Figure 2: Characteristics of fan-owner interactions for randomly selected users (Dataset I)

fraction of collected photos (< 0.01%) was inconsistent. For example, some photos had a very old posting time, or a posting time that occurred after the arrival of some fans. We removed these photos from Dataset I.

[3] A negligible fraction of random users are part of small partitions and thus they are ignored.



(a) Contribution of top fans, owners, and photos in interactions

(b) Common users between user top x fans and top x owners

(c) Dist. of number favored photos for users with certain number of favorite photos

(d) Reciprocation of fan-owner interactions among top x users

Figure 3: Characteristics of fan-owner interactions between users in the main component (Dataset II)

tions collectively conclude that the vast majority (more than 95%) of interactions occur among users in the main component of the friendship graph (*i.e.*, MC_f). To validate our results, we generated all the above distributions and numbers across the MC_f users in Dataset II. This exercise confirmed that the above results based on the 21K randomly selected MC_f users accurately represents the behavior of 99% of MC_f users, *i.e.*, our random samples in Dataset I are representative for MC_f users.

3.2 Extent of Fan-Owner Interactions in MC_f

Given that almost all the fan-owner interactions occur between MC_f users, we are interested in a more detailed understanding of these users in their roles as fans and owners. Toward this end, we rely on Dataset II for our analysis because it provides a more comprehensive view of interactions among MC_f users.

Locality of Interactions: To quantify the nature of fan-owner interactions, Figure 3(a) depicts the number of fan-owner interactions that are associated with the top active fans, owners and photos. It shows that the top 10% of owners and fans are responsible for 90% and 80% of all interactions, respectively. However, the top 10% of photos are associated with only 55% of all the interactions. In essence, the distribution of contributions of fans and owners to interactions is significantly more skewed than the popularity of photos, *i.e.*, the range of values for the contribution of fans and owners is two orders of magnitude larger than the range of popularity for photos.

Overlap Between Active Fans & Owners: One interesting question is *"whether there is any correlation between the activity of a user as a fan and as an owner?"* To answer this question, Figure 3(b) shows the percentage of users that are common between the top x fans and top x owners. The figure demonstrates that around 30% of the top 1K fans are among the top 1K owners. The percentage of overlap monotonically increases until it reaches a maximum of around 57% for the top 200K fans and then slightly drops.

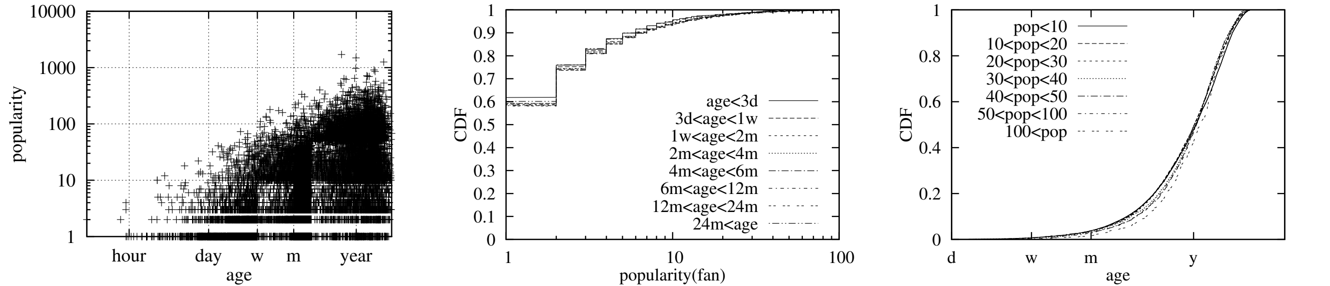
Figure 3(c) depicts the distribution of the number of favored photos among three groups of MC_f users with different number of favorite photos: weakly active fans (with less than 10 favorite photos), moderately active fans (with 10 to 100 favorite photos), and very active fans (with 100 to 1000 favorite photos). In essence, this figure shows the correlation between the level of activity of a user as a fan and as an owner. It clearly demonstrates that the most active fans are most likely to be among the active owners.

Reciprocity of Interactions: Finally, we examine the level of reciprocation in fan-owner relationships among active MC_f users. Interactions between two users are reciprocal if they have bidirectional fan-owner relationship. Figure 3(d) depicts the number of fan-owner interactions between the top x owners and top x fans as well as the percentage of these interactions that are reciprocal. This figure demonstrates that the total number of interactions exponentially increases with x while the percentage of reciprocated interactions rapidly drops. The highest level of reciprocation is 28% which occurs among the top 1000 fans and owners.

4. TEMPORAL PROPERTIES

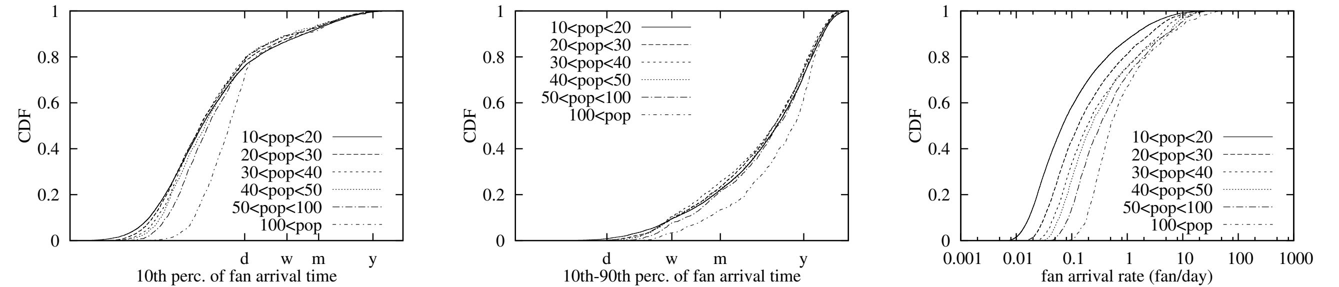
Intuitively, fan-owner interactions are highly dynamic since fans arrive over a period of time with certain patterns. The purpose of this section is to examine the temporal properties of fan-owner interactions. Toward this end, we are interested in characterizing how the popularity of individual photos changes over time (*i.e.*, the pattern of fan arrival for individual photos). We can only conduct these temporal analysis for photos collected from the sampled users in Dataset I since fan arrival time is only available for these photos. Intuitively, when a photo is posted, its popularity monotonically increases following a certain "pattern" until it has attracted a majority of its fan. After this period, more fans may arrive at a lower rate which results in a slow increase in popularity. As a result, older photos have more opportunities to attract fans, and thus, are more likely to have higher popularity than more recently posted photos. We leverage the following four properties of individual photos to capture their pattern of fan arrival: *(i)* popularity of a photo (*i.e.*, total number of fans), *(ii)* time of arrival of the 10th-percentile of fans after posting, *(iii)* time between the arrival of the 10th-percentile and 90th-percentile of fans, *(v)* age of a photo. The time of arrivals of the 10th-percentile of fans reflect how fast initial fans arrive after the posting of the photo. The time between the 10th- and 90th-percentile of fan arrivals captures how quickly a photo gains most of its popularity without being too sensitive to the arrival times of the first or last few fans.

Popularity vs Age: One key question is *"Does the age of a photo affects its popularity?"* Figure 4(a) shows a scatter plot of the popularity and age of individual photos using a log-log scale. As expected, the range of possible popularity values widens with the age of photos. To examine the correlation between age and popularity of photos more closely, we divide all photos in Dataset I into several groups based on their age (e.g., less than 3 days, between 3 days and 1 week, etc.) and plot the distribution of popularity among photos in each group in Figure 4(b). Similarly, we divide all photos into several groups based on their popularity (e.g., less than 10 fans, between 10 and 20 fans, etc.) and plot the distribution of photos in each group in Figure 4(c). Surprisingly, the



(a) Correlation of popularity with age (b) Dist. of popularity for photos with different age (c) Dist. of age for photos with different popularity

Figure 4: Relation between popularity and age of photos in Flickr



(a) 10th- perc. of fan arrival time across photos with various popularity (b) 10th- to 90th-perc. of fan arrival time for photos with various popularity (c) Avg. fan arrival rate across photos with various popularity

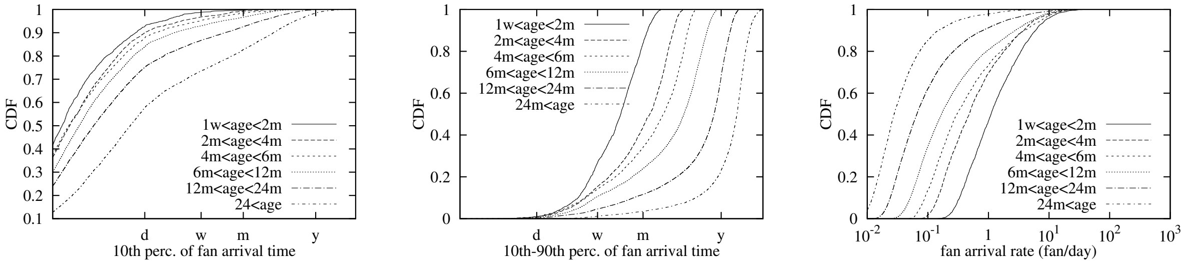
Figure 5: Effect of photo popularity on fan arrival pattern

resulting distributions are all very similar; the popularity of new photos that are just a few days old follows roughly the same distribution as the popularity of photos that are quite a bit older. These results reveal that age and popularity of a photo do not have a strong correlation, *i.e.*, knowing the popularity of a photo does not provide any strong indication about its age and vice versa. This observation appears to be in conflict with the effect of age on popularity of a photo as shown in Figure 4(a). A possible explanation of this unexpected finding is that the majority of favorite photos (>90%) have a popularity that is smaller than or equal to ten fans and are likely to reach that level within a few days. Below, we investigate fan arrival patterns in more detail to clarify these conflicting findings.

Fan Arrival Pattern: We now explore the effect of popularity and age of a photo on fan arrival pattern. Given that some of our metrics such as the 10th-percentile or 10th- to 90th-percentile of fan arrivals are not meaningfully defined for unpopular photos, we consider these metrics only for those photos that have at least 10 fans. Figures 2(b) shows that roughly 10% of photos have more than 10 fans, and Figure 3(a) reveals that these top 10% of photos are responsible for 60% of interactions in Dataset I. To examine the effect of popularity on fan arrival patterns, Figures 5(a) and 5(b) depict the distributions of the 10th-percentile and 10th- to 90th-percentile of fan arrival time across different groups of photos with a specific range of popularity, respectively. Both of these distributions follow roughly a similar shape across photos with different popularity. The only exception is for the very small fraction of photos (< 1%) with the highest popularity (>100). These photos typically require more time to attract the initial 10% of fans and to reach the 90th-percentile of their popularity. Since the times between arrival of the 10th- to 90th-percentile of fans follow the same distribution for most photos with different popularity, the average rate of fan arrivals for most photos is proportional to their popularity values as shown in Figure 6(c).

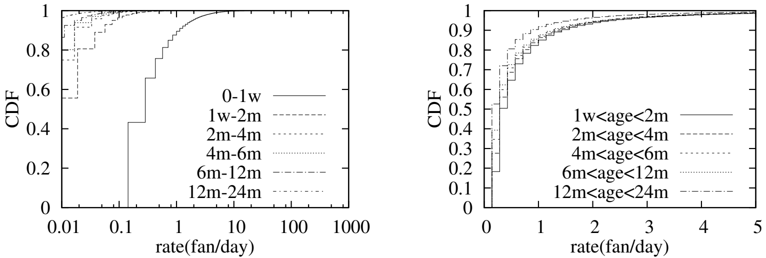
To explore the effect of a photo's age on the fan arrival pattern, Figures 6(a) and 6(b) show the distributions of 10th-percentile and 10th- to 90th-percentile of fan arrival time across photos with different ranges of age (same as Figure 4(b) but without the first two groups). These figures reveal two interesting points. First, older photos require a longer period of time to attract their initial 10 percent of fans. Second, the time to attract the majority of fans is directly proportional to the photos' age, *i.e.*, the older a photo is, the longer it takes to attract most of its fans regardless of its popularity. Since the distributions of popularity for photos with different age are very similar (as we showed in Figure 4(c)), the average rate of fan arrivals is significantly affected by the 10th- to 90th-percentile of fan arrival time which results in visibly larger average fan arrival rates for newer photos as shown in Figure 6(c). *In summary, the patterns of fan arrivals seem to be largely independent of popularity and age of photos. This in turn implies that older photos experience a lower average rate of fan arrival than more recently posted ones.*

Underlying Causes: An interesting question is "*Why do older photos experience a lower average fan arrival rate?*" To provide some insight into this issue, Figure 7(a) depicts the distribution of average fan arrival rate during different periods of a photos' life time. Each line in this figure contains all the photos whose ages are larger than the upper



(a) 10th-perc. of fan arrival time across photos with different ages (b) 10th- to 90th-perc. of fan arrival time for photos with different ages (c) Rate of fan arrival across photos with different ages

Figure 6: Effect of photo age on fan arrival pattern



(a) Dist. of fan arrival during different period of photo lifetime (b) Dist. fan arrival rate during the first week of photo life for photos with different age

Figure 7: Fan arrival rate at different intervals of photo life

end of the corresponding interval. This figure clearly demonstrates that the fan arrival rate during the first week of a photo's life time is at least an order of magnitude larger than during other intervals. To explore any effect of photo age on the average fan arrival rate during the first week after posting, Figure 7(b) shows the distribution of fan arrival rate during the first week across photos with different ages. This figure confirms that the age of a photo does not have an impact on its fan arrival rate. *These results collectively demonstrate that photos usually attract most of their fans during the first week after posting, and thus the lower average fan arrival rate for older photos is merely due to their longer inactive period in the system.*

5. CONCLUSIONS AND FUTURE WORK

In this paper, we study fan-owner interactions among Flickr users. Our results reveal that only a very small fraction of Flickr users are active as fans or owners, and most of these active users are located in the main component of the underlying friendship graph. Furthermore, active users in Flickr appear to form a core in the interaction graph that is responsible for the vast majority of fan-owner interactions. Analyzing the temporal properties of fan-owner interactions in Flickr reveals that there is no strong correlation between age and popularity of a photo and that a majority of fans arrive during the first week after a photo is posted. If our study of fan-owner interactions in Flickr is any indication of user interactions in other OSNs, the current emphasis by researchers on characterizing inferred friendship graphs of OSNs provides little insight into the nature of user interactions and their evolution over time.

The fact that only a small fraction of users are "active" in an OSN such as Flickr is promising because it suggests an efficient approach to capturing and characterizing the dynamic nature of a target OSN that targets the users in the OSN's core component rather than the entire user population. As part of our future work, we plan to leverage this idea to capture and characterize snapshots of the core component of an OSN's interaction graph and then conduct a multi-resolution analysis of these snapshots in time and space to examine the dynamic nature of real-world OSNs in great details. We also plan to look into other OSNs to characterize other types of interactions that can be captured through measurement (*e.g.*, text messaging, tagging of content).

6. REFERENCES

[1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *WWW*, 2007.

[2] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing Social Cascades in Flickr. In *WOSN*, 2008.

[3] H. Chun, H. Kwak, Y. Eom, Y. Ahn, S. Moon, and H. Jeong. Comparison of Online Social Relations in terms of Volume vs. Interaction: A case Study of Cyworld. In *IMC*, 2008.

[4] R. Kumar, J. Novak, and A. Tomkins. Structure and the Evolution of Online Social Networks. In *ACM SIGKDD*, 2006.

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