Characterizing User Interactions in Flickr Social Network

Masoud Valafar
DRP Report

***Abstract*—Online Social Networking (OSN) services have become among the most popular services on Internet and their growth has led to creation of lots of different applications and protocols. Most of these applications and protocols rely on findings of previous studies which were concentrated on analyzing and modeling of the structure of inferred friendship graph of social networks. However, serious questions have been raised about the significance of links in friendship graphs of different social networks. In this work, we present a measurement study on interactions occured in Flickr OSN. We show that a small portion of users consists a core and are responsible for most of the interactions on Flickr. We analyze the correlations between interactions and friendships, and observe that most of the interactions do not happen between friends. Furthermore, we investigate temporal properties of interactions and observe some insensitivity in results concerning the distribution of popularity (as a function of age) and age (as a function of popularity) of photos in Flickr. We see that this pattern emerges due to fast reaction of users to posted contents. Our results propose that links in friendship graph say little, if anything at all, about the level of activity of users in Flickr. Our findings also reveal patterns of interactions that can be used further in designing of new applications and protocols.**

I. Introduction

In the past few years, Internet has experienced a significant growth on Online Social Networking services. For example MySpace and Facebook together have more than 500 million users. Services provided by OSNs, loosely fall into two main categories, (*i*) social interaction and (*ii*) content sharing. Social interaction services allows users to provide profiles that contains some basic information about themselves, e.g. name, and age, and make friendship links with other users. Moreover, social networking services enable users to interact with each other, e.g via direct message passing or commenting on other users profile. OSNs such as Facebook and Orkut are very well-known examples in this categories. The content sharing services allow users to share their content with other users in the network. For example, MySpace is the favorite for individuals and bands to share their musical pieces, and YouTube enable users to publish their videos. Typically, an OSN provides a combination of both social interaction and content sharing services. Furthermore, OSNs allow users to create friendship links to other users. We refer to the graph which is created by representing users of a network as its vertices and the friendship links between the users as its edges, as *friendship graph*.

The ever-growing popularity of OSNs[1] has motivated characterization studies on OSNs. Such studies tend to shed light on the extent of the OSNs' impacts on the Internet and help improving OSNs performance by revealing their performance bottlenecks. However, such characterizations need accurate snapshots of friendship graph and users associated data and content. OSN administrators are unwilling to share this information due to security and privacy concerns. Therefore, the only viable data collection processes for emprical characterization of OSNs are crawling and sampling. Crawling is the process of progressively discovering about the users on a network and capturing their information. Sampling is the process of selecting a random (unbiased) set of users from a network whose properties represent the emtire population of an OSN. However, there are several challenges that should be addressed in order to use crawling and sampling [36].

Characterization studies on OSNs are mostly focused on the friendship graph and its evolution [27] [30] [3] [26]. There are only a few works investigating the properties of contents [19] [12] and even a smaller number on the interactions [14]. Furthermore, most of these studies do not carefully examine the accuracy of their data collection scheme.

In this study, we mainly focused on the interactions in Flickr OSN. Our target OSN is Flickr which is the largest image repository on the Internet (at the time of study). We collected unbiased information through sampling and crawling. We generated random user IDs and gathered and unbiased dataset. Using the sampled users as seed, we crawled the Flickr freindship graph and captured its main component. We conduct analysis to study (*i*) the degree of interaction across users, (*ii*) the correlation of interactions with friendship links, and (*iii*) interaction patterns. Our findings are as follows:

- First, we show that most of the active users are in WCC[2] of friendship graph. Furthermore, we demonstrate that different users in WCC show various degrees of interaction; such that highly active users form a dense component (core) that comprises a large portion of the interactions, and lowly active nodes are loosely connected to this core.

[1] Many of the OSN websites are among the top most visited websites according to [1]

[2] a maximal subgraph of a directed graph such that for every pair of vertices u, v in the subgraph, there is an undirected path from u to v and a directed path from v to u

- Second, we compare interaction graph[3] with friendship graph. We show that most of the interactions happen between users who are not friends. This fact undermines the implication of friendship graph in other applications and protocols such as [18].
- Lastly, we analyze the patterns of interaction on photos and show that most of the interactions happen in the first few days after upload of the photos. Interestingly, our results demonstrate that age and popularity of photos don't have a strong correlation.

The rest of this paper is organized as follows. In the next section we present an overview of OSNs. On Sections II and III, we introduce Flickr and explain our data collection and datasets. Section IV discusses the user activity in Flickr and Section V focuses on user activity in Flickrs main friendship component. In Section VI, we explain the correlation of interactions and friendship. We explore patterns on interaction in Flickr in Section VII and in the last section, we compile related works.

II. Characterizing Online Social Networks

In this section, we focus on different aspects of investigating OSNs. First, we discuss the main properties of OSNs. Then, we focus on OSN measurements and describe feasible methods of data collection, namely sampling and crawling. Finally, we investigate the implications of OSN characterizations.

A. OSN Overview

Users are the first class objects on OSNs. Upon joining the networks, users may provide some personal information, e.g. real name and location. These personal information are kept in users' profiles. OSNs also enable users to upload their own contents. Different OSNs may become famous for various services they provide for a specific type of content, e.g. YouTube is well-known for video sharing and Flickr is famous for photo sharing. On different OSN, access to users' profiles and contents is defined based on the network policies and owners preferences.

All the OSNs, as the term suggests, provide means for their users to connect and interact with each other. On the very basic form, users can become friends with each other. Friendship links may suggest an existing relationship between users or can be an indication of interest in another user's contents. Based on the OSNs servives and features, there can be various methods for users to interact with each other. The interaction can be direct, e.g. through a direct message exchange, or indirect, such as writing a comment on a photo in Flickr and Facebook.

Connections and interactions between users on OSNs can be demonstrated with an annotated graph. Each vertex on the graph represents a user and each edge represents the interaction or connection between them. For example, friendship connections can be demonstrated by a graph, simply by assigning a vertex to each user, and creating an edge between the vertices of any two users who are friends. The graph that is inferred accordingly is known as *friendship graph*. We can also represent various types of interactions with graphs. The graph that represents the users with vertices and a specific type of interaction with (weighted) edges is known as *interaction graph*.

Friendship and interaction graphs can be either directed or indirected. If the interactions or connections between users are mutual, the graph is undirected. For example, the friendship connections in Facebook and Orkut are mutual. But if the connections or interactions essentially happen in one direction, such as sending a message to another user, then the graph is directed. Moreover, the interaction graphs can be weighted, if a weight can be assigned to each interaction link as an indication of interaction degree, such as the number of times user A has sent messages to user B. Representing interactions and connections in a network by graphs is advantageous as the problems can be translated to graph analysis.

B. Measuring OSNs

In order to get a complete view of a specific type of interaction on OSNs, a complete snapshot of the interaction graph is needed. However, capturing such snapshots is a non-trivial task. OSN administrators are unwilling to reveal their data for security and privacy concerns. Furthermore, OSNs limits the access to their data by imposing limits on the number of queries an individual can send[4]. Such limitations significantly affect the speed of data collection process.

Data collection processes have one of the following two methods: (*i*) sampling, (*ii*) crawling. The former works by collecting random, thus unbiased, set of users from a network. The numerical ID space of some OSNs, allows generation of random user IDs. Flickr is an example of such OSNs. On contrary, user ID spaces for some OSNs are not numerical, e.g. YouTube. For this group of OSNs, random-walk based sampling techniques, such as MRW and RDS, are used [34] [36]. The important issue in measurements based on sampling is that enough samples be collected to assure that samples are representative of the users of the whole network.

The latter method of data collection, crawling, works by having an automated software progressively querying for users to collect their associated information and learn about other users in the network through their friendship links. Unlike sampling, this method exhaustively captures information of all the available users and results in a complete snapshot. However, following challenges should be addressed on crawling based measurements: (*i*) OSNs continuously change over time and crawling may result in a distorted snapshot if it takes a relatively long time for the crawler to capture the complete snapshot [35]. (*ii*) There are some parts of the networks that are unreachable for crawlers, e.g. singletons[5].

[3] Vertices of this graph are users of the OSN and the edges are the interaction between users.

[4] Most of the OSNs create rules that restricts the access of users to their data. For example Twitter only allows sending of 100 quries in each hour and Flickr limits the access to 10 quries per second.

[5] Nodes which are not connected to any other user and no other user is connected to them.

(*iii*) Furthermore, captured snapshot is dependant on the initial seeds if the graph is directional. This problem may occur if there are parts of WCC that are not reachable from the other parts.

Overall, measurement-based characterization of OSNs is not an easy task and in any study based on measurements, challenges introduced in this section, need to be addressed.

C. Importance of Characterizing OSNs

As more users join OSNs, their limitations become more apparent. These limitations can be revealed through characterizations of users and network properties of OSNs. Such characterizations can be helpful in the following areas:

First, we can use insights obtained through characterizations in designing OSNs. By characterizing OSNs, we can gain a better view of user behavior which can lead to better QoS and resource management. As an example, we can consider a characterization on pattern of watching videos on an OSN. Such characterization can be used to discover the influence of user behaviors on each other and it can eventually lead to design of a better recommendation system by system architects. Youtube and Netflix are two major OSNs that are struggling with designing an efficient recommendation system.

Second, findings of user behavior characterizations can also be applied in controlling some potential negative impacts of OSNs on the Internet. For instance, users show a correlation in their interest with the other users in their vicinity (both geographically and network wise). System designers can use this information on how to distribute contents over the servers and it can lead to reduction in network traffic created by OSNs.

Finally, lots of applications and protocols are created everyday to be specifically used either by or through OSNs. Facebook is reporting an ever-growing increase in the number of such applications [20]. Findings in user behavior characterizations are beneficial in design of new applications and protocols for OSNs.

III. Flickr Overview

A. Overview

Flickr, the largest photo-sharing OSN, is widely used by professional and amateur photographers. Flickr has also gained popularity among bloggers as a repository for the images used in their blogs. Ludicorp launched Flickr in Vancouver Canada in 2004. Due to its rapid growth, Yahoo! bought the company in late 2005 and migrated all of its content to servers in the United States.

There are two types of users in Flickr: (*i*) professional, and (*ii*) normal. Professional users can upload photos without any limitation. However, normal users can only upload up to 100MB of photos per month and 200 photos total. For using Flickr one only needs to acquire a Yahoo! ID. Subscription as a normal user is free, but needs additional fee for a professional account.

There is no reliable information about the current population of Flickr. Our estimation suggests that there are more than 25 million registered and about 5 million active users in Flickr. Flickr's popularity comes from the facilities it offers to its users. Flickr allows users to easily manage their content, and is among the first websites supporting folksonomy[6]. Other appealing features include organizer (a web application for organizing photos within Flickr), access control, slide-show and Flickr's API.

B. Organization of Data

Users' data in Flickr has a hierarchical structural as shown in Fig. 1(a). At user level, following information is available about a user: (*i*) profile, (*ii*) contact list, (*iii*) list of photos, and, (*iv*) list of favorite photos. Photos posted by users are in the next level of hierarchy and following information is available for each photo: (*i*) photo profile, (*ii*) list of fans, and (*iii*) list of comments. In this section, we elaborate on this hierarchy of information.

Upon joining Flickr, each user creates a profile and enters some general information, such as full name and age. Flickr assigns some information to users' profiles uppon their arrival, such as a numerical ID and join date. Information added by Flickr to user profiles doesn't change over time. After creating a profile, users may begin to upload photos along with photo's associated information, such as titles and descriptions[7]. Users have control over the access by other users to their photos. Furthermore, Flickr provides some specific information about each photo including a unique photo ID, upload time, and the permanent URL.

After uploading the photos, the owner can organize the photos into different groups called *sets*. Users can add a description of a set in addition to descriptions on individual photos. Grouping of photos into sets allows users to find related photos more easily.

A user can also add other users to her *contact list*[8]. A contact link from user A to B may be a sign of real social friendship, or, may simply signify A's interest in B's content. A contact link in Flickr is directional, such that when user A adds user B as a contact, user B will appear in A's contact list. After adding a user as a contact, she will be notified about that and she may reciprocate the friendship.

Users can group their contacts into three categories: (*i*) friend, (*ii*) family, and (*iii*) normal. Users can restrict access of members of each group to their photos. For this study, we only have access to public photos of users which are available to everyone. In the rest of this report, we use the term photo to refer to publicly available photos.

Flickr allows users to create a list of favorite photos. When users add photos as favorite, they become *fan* of those photos and their name would be added to the list of fans of the photo. When a user logs in, Flickr randomly chooses to display a few photos recently added by the user's contacts as favorite. Through this, the information about the favored

[6] The practice and method of collaboratively creating and managing tags to annotate and categorize content

[7] Flickr recently provided video upload service. This service was not enabled when this research was conducting and it is not considered in this report.

[8] Contact and friend are used interchangeably in other sections of this report.

photos disseminates in Flickr. On the welcome page, users will also be notified about the recent activities of their contacts, such as posting new photos.

Each user can also write comments on any photo to which she has access (including photos of herself). Flickr displays the name of the writer and time of writing under the photos.

Figure 1 shows an overall view of information organization on Flickr.

C. Direct and Indirect Interaction in Flickr

In Flickr, users can interact with each other directly or indirectly. Direct interaction occurs when user A sends a message to user B. No one else except user B is informed about this message. Indirect interaction happens through photos. For this project, we assume that adding a photo as favorite is an instances of indirect interaction. Because information about direct interaction of users is not publicly available, this study only focuses on indirect interactions.

Indirect interaction on Flickr can be demonstrated in two different views: (i) fan-photo-owner view, which emphasizes on the role of photos as the medium of interaction, and, (ii) fan-owner (graph) view, that focuses mainly on interactors.

The first view, demonstrated in fig. 2(a), is called Fan-Photo-Owner view. This view displays three lists: fans, photos and owners. Each photo has one owner and one or more fans. Each fan can have one or more favorite photos but can not become fan of a photo more than once. Owners, on the other hand, can have one or more photos in photo lists. Users may appear in both owners and fans lists. We call a photo of a user which has at least one fan, a *favored photo*. This view helps us consider photos as the main component of interaction and analyze the role of them better.

The second view, Fan-Owner view (also, graph view), can be represented by a weighted graph. Users are the nodes of the graph and appear only once. Edges of the graph represent occurrence of interaction between users. There is a directed edge from user A to user B with the weight w if user A has added w photos of user B as her favorite photos. Fig. 2(b) is the relative Fan-Owner view of Fig. 2(a). This view is user-centric. It eliminates photos from the middle of interactions and focuses on the parties involved in interactions. This view is beneficial in analyzing user behavior.

D. Flickr API

Flickr API constitutes one of the most attractive features of this network. Flickr API supports third party (independent) developers in creating non-commercial applications and expanding services.

Overview: The core functionality of Flickr relies on standard HTML and HTTP features, which enables using different platforms to use available services. Flickr expanded its services by introducing API in late 2005. After getting an API key from Flickr, one can use Flickr features by sending queries to Flickr server and receiving responses. Queries are sent in REST, XML-RPC or SOAP format, while responses can have REST, XML-RPC, SOAP, JSON or PHP format.

Users can develop web and desktop applications using the API. Flickr imposes this limitations a) each user can only apply for one pair of API keys and, b) each pair of keys can only send 10 queries per second. This helps controlling the load on the Flickr server and avoid any malicious attacks, such as DoS.

sample API call: To communicate with Flickr using its API, a user should first acquire authorization token to gain access. One can have *write* access to one's own account, *read* access to friends' accounts and general[9] access to public content of all users. After acquiring the token, users send queries and ask for a service. Flickr server will respond with the proper result if that service is available for the caller user (based on the authentication type) or with an error message otherwise. In Fig. 3(a) a sample API call and its response is depicted. This figure depicts the reply for a query about user profile information. The response is in XML format and the high level element, person, shows that it contains information of a user. The parameters of person element shows high level information about the user, including user id. Inside the person element, other information, such as user-name, real name, location and information about the photos of the user is demonstrated.

IV. Identifying the Interaction Graph

In the previous section, we introduced a detailed view of interaction in Flickr using Figure 2(a). In this chapter we discuss how we use that view to extract data from the hierarchical data structure of Flickr. Throughout the process of data collection, we faced challenges that we discussed in detail in section II. Below, we explain how we dealt with those challenges. After that, we explain about the datasets that we use for this study and go through different properties of them.

A. Data Collection

We begin this section by explaining the data collection process. We found out that exhaustively crawling users and their photos to capture a snapshot of Flickr user information is practically impossible because: (i) list of users in Flickr is not available. Therefore, the only way to discover all the users, is by exhaustively investigating the existence of each ID in ID-space and investigating existence of a user with such an ID (ii) even if there existed a list of user IDs, the API limitation on the number of queries per second wouldn't allow us to extract the information associated to each user in a timely manner.

1) Crawling random users: Based on the reasons mentioned above and the huge amount of data on Flickr, we start with sampling. We have leveraged random users information by generating random IDs based on the specific format of IDs in Flickr[10]. Then we query the server to extract photos and associated information of that user. Using photo-IDs of a

[9] This is the default access. Users are not required to do anything to get this type of access

[10] User IDs in Flickr have a well known format that consists of a six-eleven digit prefix, followed by ”@N0” and a one-digit suffix, e.g. 1234567890@N00.

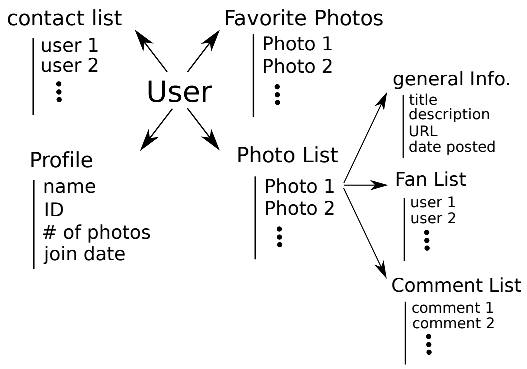
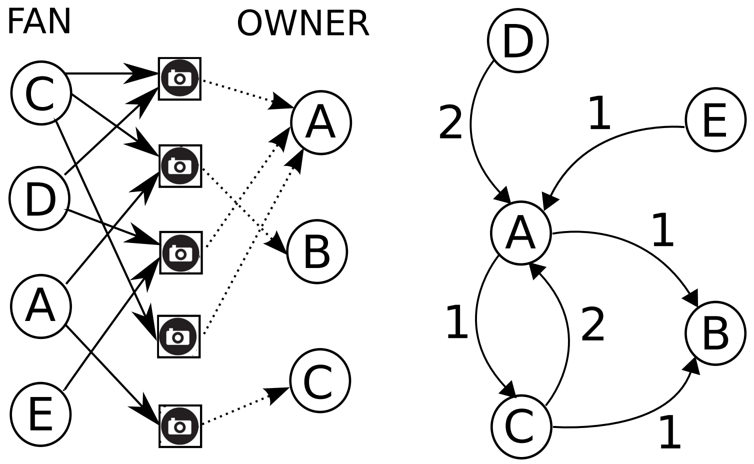


Fig. 1. Information hierarchy in Flickr - Users have profile, list of photos, list of contacts, and, list of favorite photos. Each photo has a profile, list of fans, and, list of comments



(a) fan-photo-owner view

(b) fan-owner (graph) view

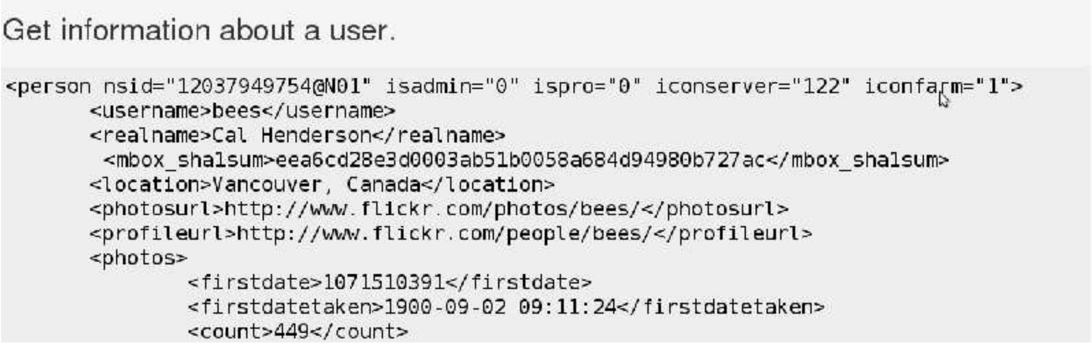
Fig. 2. Views of interaction

user, next, we extract available information of those photos, including timing and fan list. With this method, we extract when a photo is posted and when other users have added the photo as their favorite.

The main drawback of this method is its low speed because (*i*) during the random ID generation phase, most of the randomly generated ID are not assigned to any user, and (*ii*) number of queries needed in this method is in the order of number of photos, while most of the photos don't have any fan, and so are irrelevant to our work. Nevertheless, with this method we collect adequate random samples which are representative of the entire users on Flickr.

2) Crawling favorite photo lists: The second method of data collection in our work, is capturing user interactions by crawling favorite photo lists of known users. If we find a way to have a list of all fans in Flickr, then we can query Flickr for their favorite photo lists and through this indirect method we can collect the information of the photos that have been added as favorite in the network. Using data gathered through the first method, we found out that more than 95% of the interactions in Flickr happen in its WCC and, thus, we can focus on users in WCC in order to efficiently capture interactions. Later, we explain about how we found this out in more depth.

The advantage of this method over the previous method is that the order of needed queries in this method is in the order of number of users and thus, two times less than the previous approach. The main drawback of this method is that we can not get the timing information related to each interaction.



Get information about a user.

```
<person nsid="12037949754@N01" isadmin="0" ispro="0" iconserver="122" iconfarm="1">
        <username>bees</username>
        <realname>Cal Henderson</realname>
         <mbox_sha1sum>eea6cd28e3d0003ab51b0058a684d94980b727ac</mbox_sha1sum>
        <location>Vancouver, Canada</location>
        <photosurl>http://www.flickr.com/photos/bees/</photosurl>
        <profileurl>http://www.flickr.com/people/bees/</profileurl>
        <photos>
                <firstdate>1071510391</firstdate>
                <firstdatetaken>1900-09-02 09:11:24</firstdatetaken>
                <count>449</count>
```

Fig. 3. XML file in response to get-user-info API call

B. Datasets

Based on the two methods that we just discussed, we collected two datasets. Flickr limits the rate with which a user can communicate through API with their server. Such limitations significantly affect the speed of data collection process. Below, We describe datasets that we collected:

Dataset I (random samples): This dataset contains complete information of about 123K users. This information include profile data, list of favorite photos, list of posted photos, photos associated data, and their fan list. Data in this dataset is collected using sampling and is expected to be representative. To validate the data, we repeated the whole process for the second time. The data in the both sets show consistency with each other. Furthermore we compare information of sampled WCC users of Flickr with the information of the entire WCC users and they show more than 98% consistency.

To gain insight about the topological structure of connections between random users and other users, we crawl friendship graph of Flickr OSN using sampled IDs as seeds. Through this, we discover another 4.2 Million users which are tightly connected to each other and make a WCC (donated by MC_f). We believe this is the largest WCC in Flickr, because if there existed any other WCC larger than this, with a very high probability, there were some users of that among our samples and thus we could discover it[11]. We discover that 21K of original 123K users belong to this component and the others are singletons (not connected to anyone)[12].

Based on the proportion of MC_f users in our random dataset, we can speculate that the total population of Flickr is about 6 times the size of MC_f, that is, around 25 million (at the time of crawl). Table I summarize other related information about dataset I.

Dataset II (Interaction in MC_f): In order to capture a more complete snapshot of the fan-owner interactions among the users of Flickr, we crawled all the users that we discovered through random sampling and the friendship graph crawl for their favorite photos. With this crawl, we captured all those interaction edges that the initiator (fan) is in MC_f and of course we missed all those interactions that are initiated by singletons (outside our random dataset). However, these missed edges are expected to be very small. In the next section, we show that these edges consist at most 5% of the edges of the interaction graph. Table II shows the summary statistics for the dataset II.

V. Extent of Fan-Owner Activity

In this section, we focus on the extent of interactions among users in Flickr and investigate uniformity of interaction among users. Due to the large population of Flickr and the limitations on number of queries, it is important to find an efficient way to capture interactions. Hence, we first turn our attention toward dataset I which contains the representative data of the whole network. At this point, we are interested in discovering the topological place of active portion of the photos, owners, and fans in Flickr.

A. posted photos vs. active photos

The randomly selected users in Dataset I have collectively posted 3.5 million photos. Based on the topological crawl that we performed on this dataset, we can distinguish between users who are in MC_f and singletons. Interestingly, we discover that most of the posted photos belong to random users that are located in MC_f, although most of the sampled users are singletons, as shown in table I.

Fig. 4(a) shows another difference between MC_f users and singletons. This figure demonstrates the distribution of file per users for singleton and MC_f users. It shows that only around 20% of singletons post more than one photos while MC_f users are more active and 50% of them post more than one photos[13].

Next, we focus on photos in dataset I, that have fans (we call photos with fan *active photos*). Table I shows that only

[11] As we explained previously, not all the nodes connected to WCC could be discovered. These nodes are having contacts in WCC and are part of them, but no user in the WCC has any link to them. So they can only be captured if we start the crawl from them (having their IDs as seeds).

[12] A negligible number of these users make very small component with a few other nodes. We consider them like other WCC users.

[13] The sudden drop at 200 photos/user is due to the limit that Flickr imposes on the number of photos users with free accounts can post

TABLE I
DATASET I: RANDOMLY SELECTED SAMPLES

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

TABLE II
DATASET II: FAVORITE LIST CRAWL

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

about 145K of 3.5M posted photos have fans. 98% of these photos belong to MC_f users and the rest belong to singletons. Because dataset I is a representative sample set of Flickr, we can conclude that most of the interactions happen on the photos that are posted by MC_f users.

Fig. 4(b) demonstrate the distribution of number of fans per photo. This figure shows that distribution of fans for photos posted by MC_f users is more skewed. Furthermore, it reveals a major difference between photos in MC_f and singletons; photos posted by MC_f users can have up to 10K fans, which is not the case for photos of singletons. We can conclude that most of the highly favored photos are located in MC_f.

B. active owners

We call a user an *active owner* if she has a photo or more that is added by other users as favorite. We call photos of an active owner that have fan(s) *favored photos* of that user.

In this subsection, we concentrate on active owners. Table I shows that MC_f users are more active than singletons. It demonstrate that 23% of MC_f users are active owners while only 1.2% of singletons are active. Moreover, Table I reveals that MC_f users attract two orders of magnitude more fans than singleton users.

Fig. 4(c) shows the distribution of the number of favored photos of active owners in Flickr. This figure shows that active singleton owners have less photos compared to active MC_f owners. Furthermore, it shows that active MC_f owners can have up to a few thousands favored photos, while in most cases active singleton owners don't reach that many favored photos.

C. active fans

We name users that initiate fan-owner interactions by adding another user's photo to their favorite photo list, *fans*. From table I, we notice that MC_f fans are more active than singletons. 2,638 (2.6%) of singletons and 4,053 (18.4%) of MC_f users are fans. Table I also shows that 96% of the total interactions that has been initiated by randomly selected users, are initiated by MC_f users and the remaining 4% by singletons.

Figure 4(d) depicts the distribution of favorite photos among singletons and MC_f users. It shows that MC_f fans have more favorite photos than singleton fans and the tail of MC_f line (very active fans) for MC_f users has the value as large as 3K, while very active singletons don't have that many favorite photo.

Overall, results in this section, show that interactions in Flickr are mostly initiated by MC_f users and they mostly happen on photos associated to MC_f users. Thus, in order to capture interactions efficiently without losing a great portion of interactions, we can focus on MC_f users.

VI. CENTRALITY OF INTERACTIONS IN MC_F

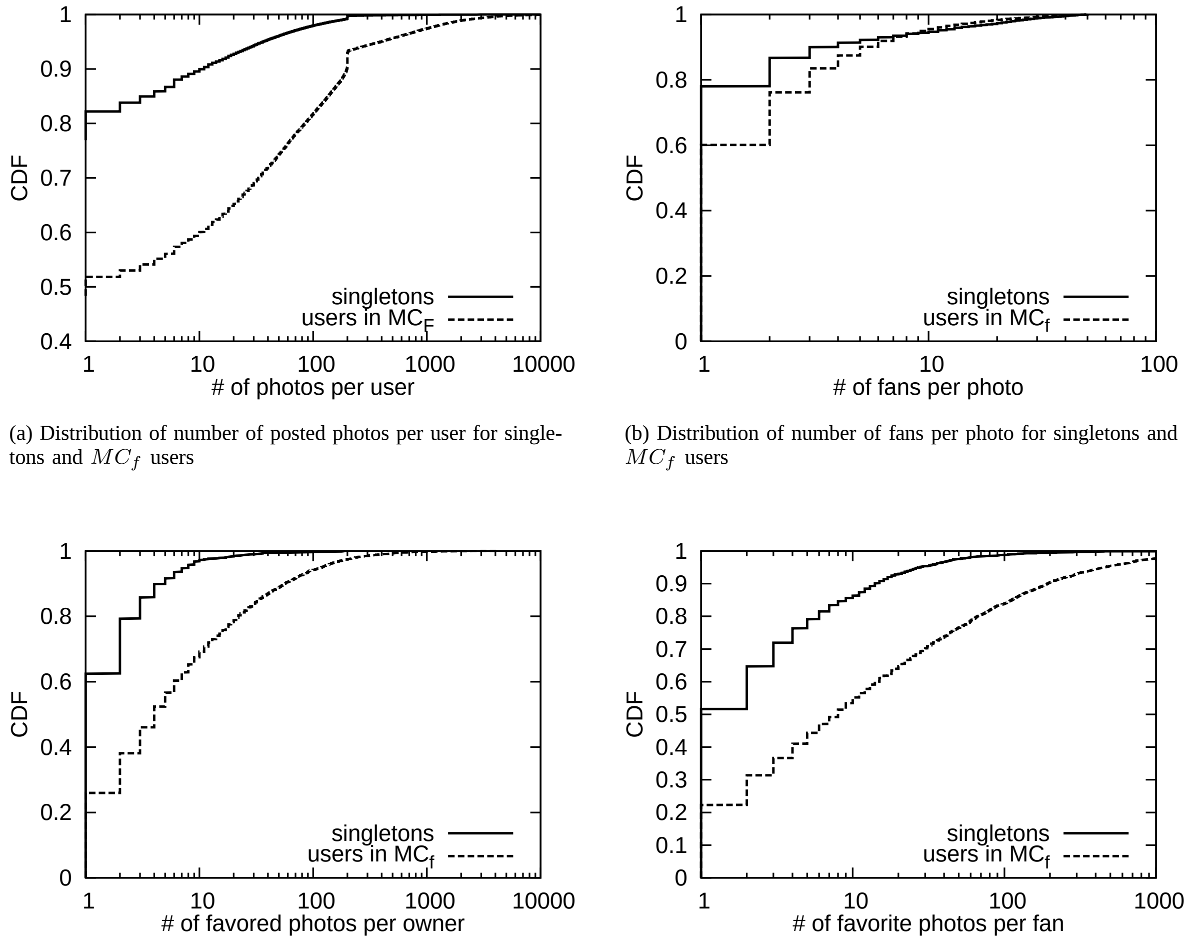
Given that almost all of the interactions happen among MC_f users, one can ask "how these active owners and fans interact with each other?" and "whether the inferred interaction graph has a core (a very dense subgraph)?". To asnwer the questions, we first focus on interaction at user level. Next we investigate pairwise interactions. In the end, we explore reciprocation between users and examine existence of a core for interaction graph.

A. Interaction Centrality

To explore the interaction centrality at user level, we present our results using fab-photo-owner view from two different perspectives: (*i*) ranking, and (*ii*) overlap.

Ranking: To quantify the nature of fan-owner relationship, Fig. 5(a) depicts the number of fan-owner interactions associated with the top active owners, fans and photos. This figure shows that 10% of active owners and fans cover 90% and 80% of interactions, respectively. However, interactions on photos is not as dense; 10% of photos with most fans cover only about 55% of interactions. There are two reasons for this fact: (i) the number of photos with fans are an order of magnitude larger than the number of active owners and fans, i.e., total number of favored photos, active owners and fans are 30M, 1M, and, 800K respectively. (ii) The range of values for contribution among fans and owners is two orders of magnitude larger than the range of popularity of photos. Overall, these two reasons make interactions at a user level more centralized than interaction at photo level.

Overlap: Interaction-wise, based on Fan-Photo-Owner view, each user can have two roles; owner role and fan role. Hence, users can appear both in the owner list and in fan list. To explore the extent of activity for users in each role, Figure 5(b) demonstrates the percentage of overlap between top x



(a) Distribution of number of posted photos per user for singletons and MC_f users

(b) Distribution of number of fans per photo for singletons and MC_f users

(c) Distribution of number of favored photos per user for singletons and MC_f users

(d) Distribution of number of favorite photos per user for singletons and MC_f users

Fig. 4. Characteristics of MC_f users versus singletons (Dataset I)

active owners and fans. It shows that the overlap between top 1K active owners and fans is about 30% and it monotically increases as it reaches its heights at about 60% for top 200K and has a slight drop afterward.

To examine correlation between activities as an owner and as a fan, figure 5(c) plots the distribution of the number of favored photos across three groups of MC_f users with different number of favorite photos: (i) weakly active (number of favorite photos between 0 and 10) (ii) moderately active (number of favorite photos between 10 and 100) (iii) Highly active (number of favorite photos between 100 and 1000). This figure illustrates a significant correlation between activities of a users as a fan and as an owner.

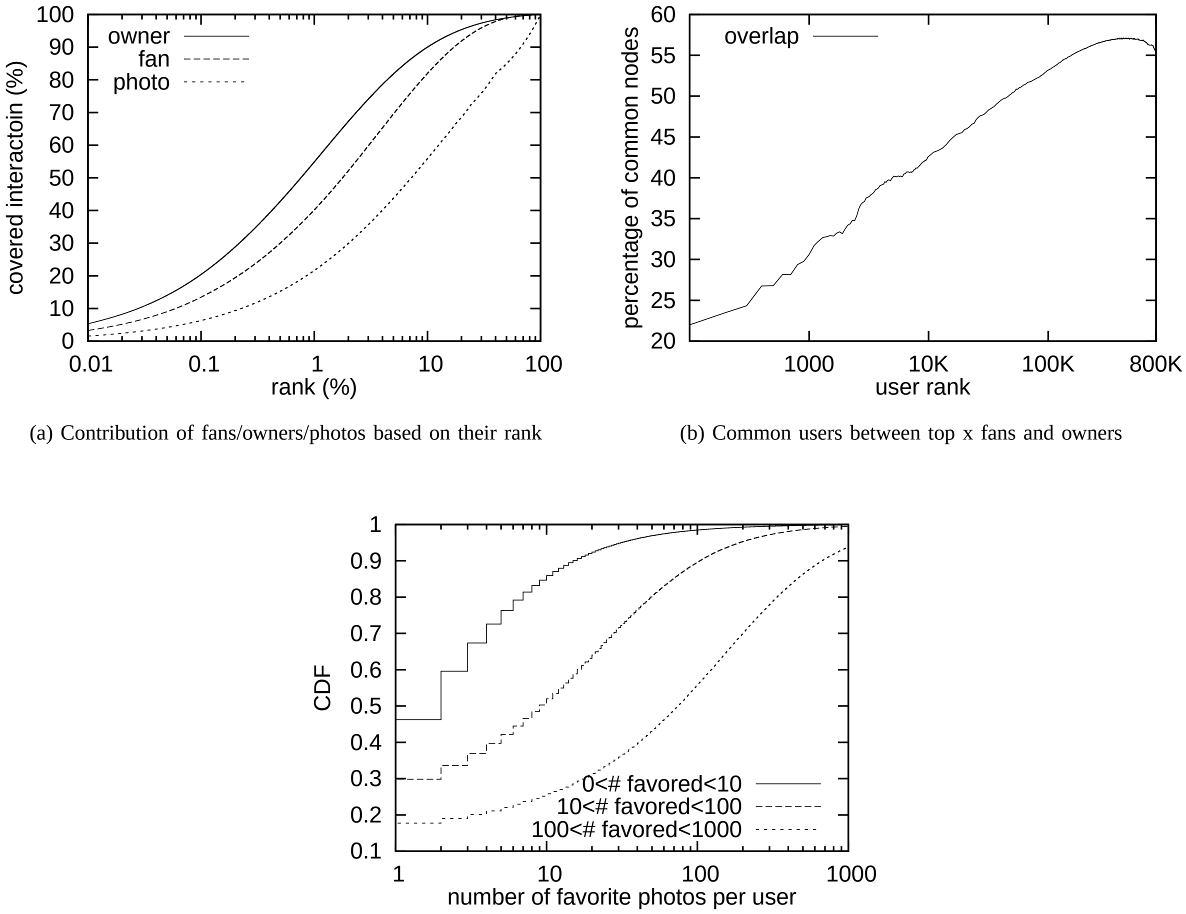
The results in this section show that not only interactions mostly happen in MC_f but also they happen through a smaller portion of MC_f users. They also reveal that there is a correlation in activities of a user in owner and fan roles.

B. Interaction Degree

To investigate interactions in details, leveraging the Fan-Owner view, we focus on the interactions at edges level (pairwise interaction) in this subsection. Figure 6(a) shows the distribution of weight of interactions between users. This figure shows that only 30% of edges have weight more than one; therefore, most of the users interact with other users just once.

To explore the impact of top-weighted edges on the total interactions happened, Figure 6(b) demonstrates the number of fan-owner interactions associated to top-weighted edges. This figure reveals that the 30% of edges that have weight 2 or more, cover 70% of interactions happened in Flickr.

Results presented in this section show that there is a centrality among interaction edges; meaning that a small portion of edges (30%) cover most of the interactions (70%), although most of the edges in Flickr are spread everywhere with weight one.



(a) Contribution of fans/owners/photos based on their rank

(b) Common users between top x fans and owners

(c) Distribution of number of favored photos across users with certain number of favorite photos

Fig. 5. Characteristics of interaction in Flickr - User level (Dataset I)

C. Reciprocation

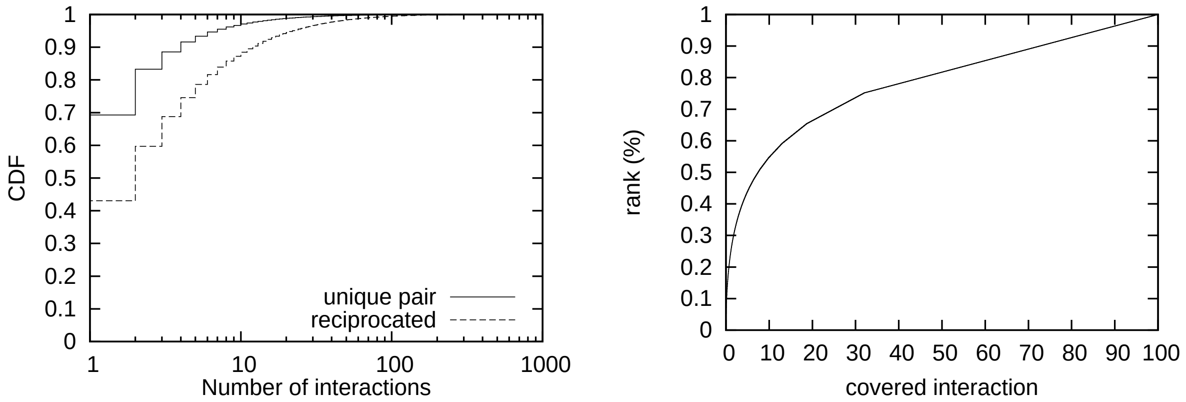
The results we have in this section so far, show that a great portion of interactions happen through a small portion of users and edges. However, one can ask ”whether highly active user interact with each other or with weakly active users”? In this subsection using Fan-Owner view, we focus on reciprocation of interactions among users to answer the posed question. Note that, we call an interaction reciprocated if there is a bidirectional edge between two users.

We start by raising this question that ”are the reciprocated edges different than the uni-directional edges”? Figure 6(a) compares reciprocated and other edges in terms of their weights. It plot the distribution of weights for reciprocated and other edges and shows that reciprocated edges have higher weights.

To explore reciprocation among different users (in terms of activity), Fig. 7(a) demonstrate the distribution of percentage of reciprocated edges of users. This Figure plots different lines for top 1%, top 10%, and all active users. The figure reveals that more users among highly active users tend to reciprocate interactions. It shows that more than 85% of top 1% of highly active users have reciprocated their edges, however this number among top 10% is 60% and for all users it decreases to 15%. Interestingly, Fig 7(a) reveals that mostly, percentage of reciprocated edges among top 1% and top 10% active users do not go beyond 10%. In conclusion, This figure shows that the highly active users reciprocate more edges but at the same time their are selective about the users they interact with.

To investigate which group of users interactions, highly active users tend to reciprocate, Figure 7(b) demonstrate reciprocation for different subgraphs of interaction graph. The x-axis in this figure, is the size of the subgraph (consisted of top active users) and the y-axis shows the percentage of reciprocation in that subgraph. It shows that the reciprocation significantly decreases with higher values of x which means that active users reciprocate more among themselves than the others.

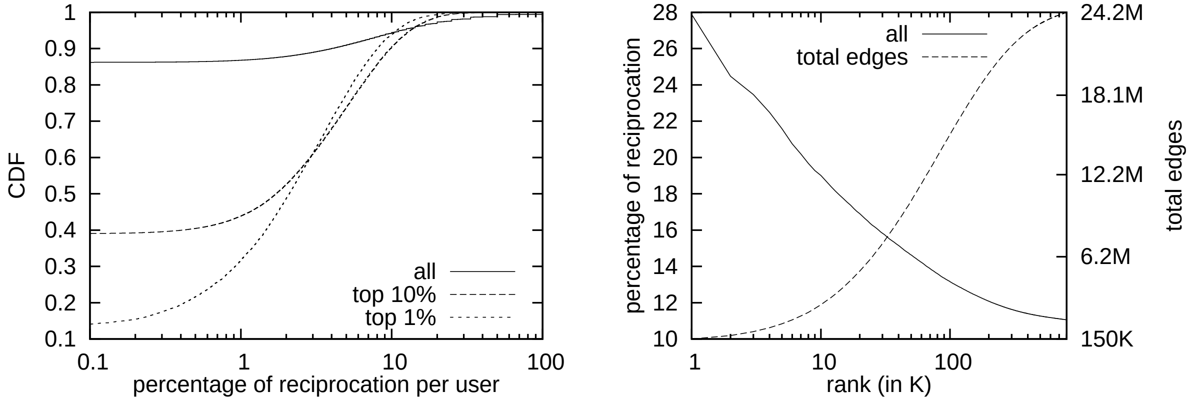
Results in this section suggest existence of a core for interaction graph. They revealed both ownership and fan-ship



(a) Distribution of weight of interactions across all and reciprocated edges

(b) Contribution of edges based on their rank

Fig. 6. Characteristics of interactions in Flickr - edge level (Dataset II)



(a) Distribution of percentage of reciprocated edges per user across users with different ranks

(b) Percentage of reciprocation in different subgraphs of interaction graph

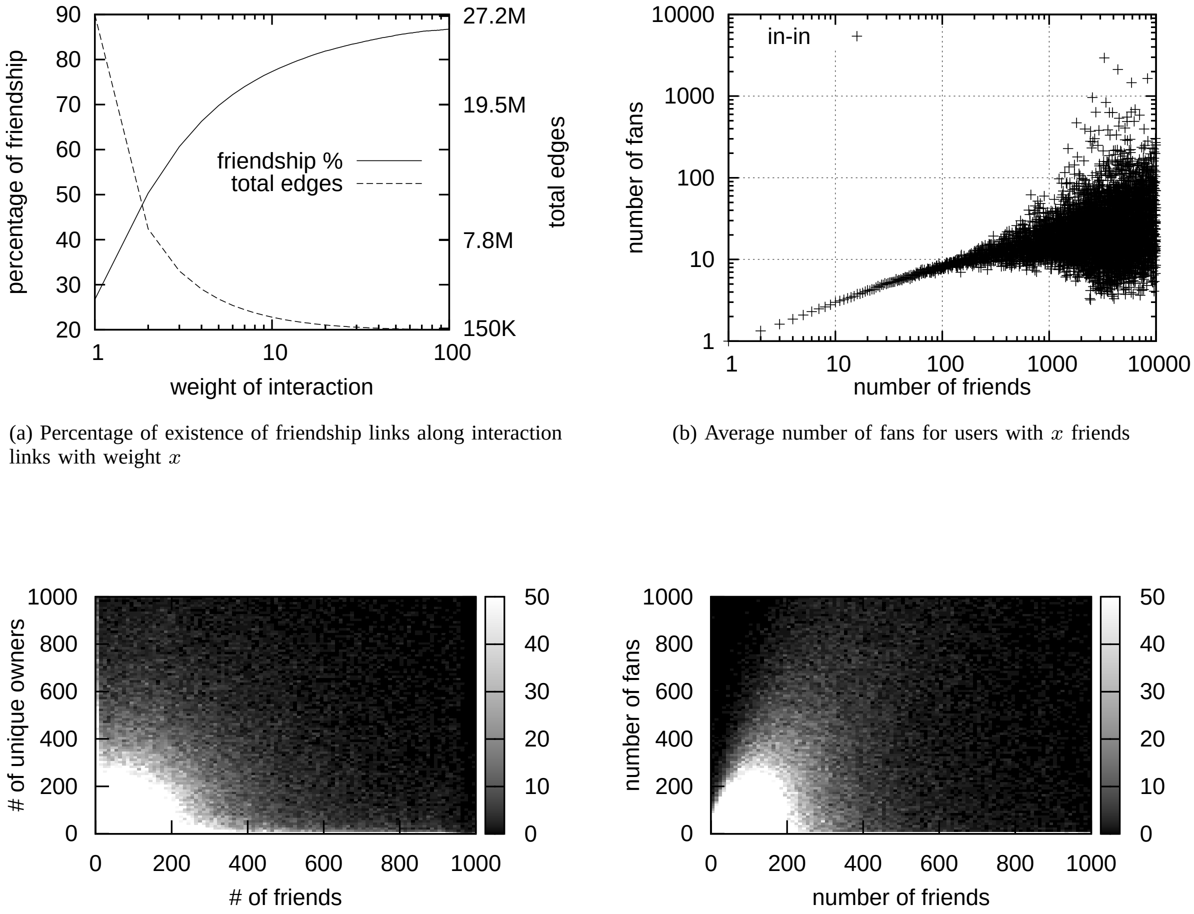
Fig. 7. Interaction reciprocation in Flickr

behavior for highly active users and greater weights of interaction on the edges attached to these users. To complete the last piece of existence of a core, we showed that highly active users tend to reciprocate interactions with higher weights among each other.

VII. Correlation interactiongraph and friendship graph

Given that only a portion of nodes in MC_f interact and the inferred interaction graph has a core which is mostly consisted of the high degree (in terms of number of interactions) nodes with lots of reciprocated edges, the next natural question is ”whether there is any correlation between the interactions and the friendship links?”. To answer this question, ion this section we explore the relationships between interaction graph and friendship at node level and edge level. For convenience, throughout this section, we refer to interaction and friendship graphs as i- and f-graph, respectively. All the results presented in this section are from dataset II and they should be interpreted throughg the Fan-Owner (graph) view that was introduced by Fig. 2(b).

We now focus on the correaltion of weight of edges in i-graph and the existence of the same edge in f-graph. Figure 8(a) demonstrates the percentage of existence of friendship edges (y-axis) between users that have interacted with each other x times. This figure shows that the correlation between interaction edges and friendship edges sharply increases as the weights of interaction links increase. This Figure also reveals that less than 30% of interactors who have interacted only once are friends. This percentage experiences a dramatic increase until the value 5 (about 70%) and after that it continues its increase with a slower rate. Basically, this figure reveals that



(a) Percentage of existence of friendship links along interaction links with weight x

(b) Average number of fans for users with x friends

(c) Grid plot of out-degree in friendship graph vs. out-degree in interaction graph

(d) Grid plot of in-degree in friendship graph vs. in-degree in interaction graph

Fig. 8. Comparison of interaction graph and friendship graph

(i) there is no friendship link between most of the users who have interacted with each other (Fig. 6(a) shows that edges with weigh one or 2 consist 80% of total edges in Flickr i-graph), (ii) the more two users interact, the more probable it is for them to make friendship. However, we can not say much about the precedence of interactions or friendship.

To explore any potential relationship between the two graphs at node level, Fig. 8(b) depicts the correlation between in-degree of users in both graphs. The y-axis in this figure indicates the average number of fans of users with x friends. This figure demonstrates a direct correlation between number of friends (in-degree in f-graph) and number of fans (in-degree in i-graph). The noisy part of the end of the graph (for high degree nodes) is due to lack of samples of high in-degree. We see such correlations between number of friends and interaction in other networks as well, such as [14].

In order to further investigate the correlation between interactions and friendships at node level, Figures 8(c) and 8(d) show three dimensional scatter plots of node degrees in i- and f-graph. Because both i-graph and f-graph are directional, we investigate in- and out-degree separately. Fig. 8(c) demonstrates the scatter plot of out-degrees. The x-axis is the number of friends of a user and the y-axis is the number of unique owners the relative user is a fan of their photos. The color of each bin of the graph shows the number of users that fall in the bin; the brighter bins contain more users than darker ones; The bins that have 50 users or more are white. If we put users with small number of friends and low social activity aside, i.e. users with less than 10 owner and 10 contacts, there will be around 450K users which can be put into three groups based on Figure 8(c). The first group are those who lay along the x-axis. 44% of users fall into this group and it seems that they are looking for friendship on Flickr more than other activities. The second group lay along y-axis. 18% users fall into this group. This group of users seems to be interested on the photos on Flickr rather than social features. The rest of the users which consists 38% of users, show both types of activities in Flickr.

Fig. 8(d) shows the correlation of out-degrees in i- and f-graph. Similar to Fig. 8(c), x-axis and y-axis show the degree in f- and i-graph respectively and the color of each bin depicts the number of users that fall into that bin. Interestingly, this figure demonstrate difference structure from Fig. 8(d). The difference is mainly caused because unlike the out-degree, users don't have any control over their in-degree in i- and f-graph[14]. The main difference between the two figures is that there is almost no user with too many fans but no friend (a black triangle is formed along the y-axis). This figure demonstrates an increase in number of friends as the number of fans increases. Furtheremore, it reveals that more than 95% of users fall in the area with x and y less than 200.

Our results in this section show that most of the interaction edges form independent of existence of friendship link between interactors. However, as interactions occur more often between users, those users are more probable to be friend. Our results also demonstrate a correlation between social activity (creating friendship links) and interaction activity.

VIII. Temporal Properties of Interactions

A. Pattern of fan arrival

Given that the most of the interactions happen by only a small percentage of users, we are interested in analyzing the dynamics of these interactions in more depth. The main question we want to answer is ”how popularity of individual photos changes over time”. All the analysis presented in this sections are produced using Dataset I, because that is our only dataset that provides detailed timing information about the interactions.

Intuitively, when a photo is posted, its popularity increase follows a certain pattern until it attracts a majority of its fans. After this period, casual fans may arrive at a slower but constat rate. Essentially, it implies that the older photos have more time to attract fans and thus are more popular than the younger ones. Also we know intuitively that different photos attract fans with different rates. Based on these intuitions, we leveraged these properties for different photos to infer their pattern of fan arrivals: (i) the 10th/50th-/90th-percentile fan arrivals (ii) the duration between first/10th-percentile and last/90th-percentile fan arrival (iii) popularity (total number of fans) (iv) rate of fan arrivals and (v) distribution of fans inter-arrival periods. note that the time between 10th-percentile and 90th-percentile fan arrivals captures how fast a photo attracts its fans without being sensitive to the arrival of first and last few fans.

B. Popularity vs. Age

The first question that we want to answer is ”whether age of a photo affects its popularity?”. Figure 9(a) is a scatter plot of the popularity and the age of individual photos using a log-log scale. It demonstrate that the range of popularity widens as the age of photos increases. But this figure doesn't show whether this is because old photos are more popular in general or because newer photos have not had enough time to become mature, in terms of popularity.

To examine the correlation between age and popularity more closely, next, we focus on distribution of popularity among active photos with different ages. To do this we divide the active photos in Dataset I into different groups based on their age (photos less than 3 days old, between 3 days and 1 week old, etc.) and plot the distribution of popularity for each group in Figure 9(b). Similarly, we also divided photos based on their popularity (photos with less than 10 fans, between 10 and 20 fans, etc.) and plot the distribution of age for each group. Interestingly, these two figures show that age and popularity do not have a strong correlation on each other. Figure 9(c) demonstrates that even distribution of popularity of photos that have been uploaded to Flickr in past few days follows the same pattern of other groups of photos. This property completely contradicts our intuition that the older photos popularity distribution should be more skewed (because they have more time to attract fans) and in the rest of this section we try to find a reason for this observation.

To gain more information about the pattern of fan arrival, we plotted the distributions of fan inter-arrival time (interval between arrival of two consecutive fans) across photos with different popularity and age in Figure 10. The first Fig., 10(a), shows the distribution of fan inter-arrival across photos with different popularity. It demonstrate that interarrival significantly decreases for more popular photos. The second one, figure 10(b) demonstrates that age greatly affect fan inter-arrival time as well. It shows about 70% of fan interarrival times for photos older than a year are more than a week. However, for photos which are between 2 and 4 month old, about 80% of fan inter-arrivals are less than a week.

C. Fan arrival

given that the nature of interactions is very dynamic, the purpose of this section is to explore some aspects of temporal behavior of fan-owner interactions. However, in order to examine interaction patterns, we need to focus on popular photos, i.e. photos that have more than 10 fans. One reason for this is that some of our metrics, such as 10th- and 90th-percentile of fan arrival, are not meaningful defined over unpopular photos. Figures 4(b) and 5(a) show that these photos cover a significant number (about 60%) of interactions on Flickr. For the rest of this section, we only consider this group of photos.

Figures 11(a), 11(b), and 11(c) show the distribution of arrival of 10th-, 50th-, and 90th-percentile of fans for photos with different age. Fig. 11(a) shows that for more than 90% of photos with different ages, 10th-percentile fans arrives within a day except for photos older than a year. There can be two reasons for this: (*i*) it takes longer for some photos to get discovered and these photos are among older photos; thus their 10th-percentile fan arrives later than young photos (*ii*) Continuous arrival of fans pushes the 10th-percentile fans further away from the post time of photos and this effect is harsher for older photos. Fig. 11(b) demonstrates an interesting

[14] A user can not delete any other user from the fan list of her photos; nor can she delete herself from the friend-list of another user.

point that for younger photos, distribution of 50th-percentile fan arrival is very similar to distribution of 10th-percentile fan arrival which was shown in Fig 11(a); for older photos the distribution slowly diverges toward 90th-percentile fan arrival which is depicted in Fig. 11(c). Figure 11(c) reveals that the arrival of 90th-percentile fan is proportional to the age of the photos.

Figures 11(d) and 11(e) demonstrate the distribution of time between 10th-percentile and 90th-percentile of fan arrivals and first to last fan arrivals, respectively. These figures reveal two interesting points. First, they show that distribution of 10th-percentile and 90th-percentile of fan arrivals and first to last fan arrival are almost similar. We speculated that by cutting the first 10 percent of fan arrivals, we eliminate the initiating part that the photos are getting slowly popular (the information of posting of a new beautiful photo is disseminating through the network) and by cutting the last 10 percent of fan arrivals, we eliminate the final phase that fans arrive sparsely. But on contrary, lack of significant difference in these distributions dismisses our speculation. Second, they show that most of the (popular) photos keep receiving photos throughout their lifetime and the distributions show proponsity to the age of the photos.

Although these figures do not reveal much about the pattern of fan arrival, they show one interesting point. Fig. 9(b) shows that for various groups of photos with different ages, the distribution of popularity almost follow the same pattern. When we put this fact beside the fact that photos recieve fans all over their lifetime, Figures 11(d) and 11(e), we understand that rate of fan arrival should be higher for newer photos. This fact is demonstrated by Fig. 11(f). This figure plots the distribution of rate of fan arrival across different groups of photos and it shows that rate of fan arrival is significantly higher for photos that are newer in Flickr.

To explore the effect of popularity on fan arrival, Figures 12(a) 12(b), and 12(c) plot the distribution of 10th-, 50th-, and 90th-percentile of fan arrival. They show that across groups with various popularities, the distributions show significant similarity except for the most popular group, i.e. photos with more than 100 fans, which contains less than 1% of total photos.

Unlike the similarity between 10th- and 90th-percentile fan arrivals and first and last fan arrivals of photos grouped by age, figures 11(d) and 11(e), Figures 12(d) and 12(e) that plot the same distributions except that photos are grouped by popularity, show differences especially in the head part of the graphs. This shows that cutting the first and last 10 percent of fan arrivals changes pattern of fan arrival if we classify photos based on their popularity.

Figure 12(f) demonstrate different rate of fan arrivals across groups of photos with different popularities. It reveals that more popular photos have higher rates. As the distribution 10th- to 90th-percentile of fan arrivals for different photos follow similar pattern, we can say that the rate of fan arrival for most photos is proportional to their popularity values.

The results that we discussed in this section, do not provide a complete view on how fans arrive at photos. We can raise this question that ”are the patterns of fan arrival for a photo in different periods of its life the same?” And if the answer is no, ”how different are these patterns?”

To answer the questions raised above, we leverage the rate of fan arrival across photos in different periods. Figure 13(a) shows the distribution of rate of fan arrival for different periods after the arrival of the first fan for all photos that are older than that period. It demonstrate that active photos recieve fans with much higher rate in the first week of their photo-life and then after that, the rate gradually diminishes.

To understand whether this pattern is the same for all photos or not, we investigate this rate across photos with different ages. Fig. 13(b) depict the distribution of rate of fan arrival in the first week for photos with various ages. It shows strong similarity between the two groups and it means that this pattern is homogenous across various photos in Flickr.

Our results in this section show insensitivity concerning the distribution of popularity as the function of age, and age as the function of popularity. Furthermore, we showed that popularity of photos in Flickr experience a sudden pick at the beginning which leads to arrival of most of its fans in a few days. Then fan arrivals decreases over time but photos continue to get fans with a very low rate. We saw that this property holds across all photos and factors like age and popularity of photos do not affect it much.

IX. Related Works

Large-scale graphs have received significant attention in past few years from different areas of studies such as sociology, physics, biology, and computer science. In each area, based on the implications of large scale graphs for that area, different properties of the graphs have been studied. In this section, we briefly review some influential and recent works which are related to this work.

Large scale graphs (LSG) are made by collecting a set of entities and defining an interaction between those entities as the edges. These graphs include, real life social networks, word adjacencies, neural and protein networks, collaboration graphs of film actors, networks of power grid, co-authorship in science writings, citation graphs, and gene network. All the studies on LSGs fall loosely into following categories: (i) static structure of the network (ii) dynamics of the network (how the network topology changes and evolves over time).

Goal of works on former category, static structure of the graphs, is to discover properties of the graphs in order to understand the involving entities (nodes) interactions better and shed light on the nature of those interactions, e.g. in works on protein graphs, the chemical reaction between different groups of protein has been investigated through large scale graphs.

One of the most cited properties on different graphs, in this category of studies, is *small world*. In large-scale graphs with small world property, most of the nodes which are not connected, are within a few hops of each other. This property was first discovered by Milgram [28]. He discovered that on average, there are only 6 hops distance between each two American. Later on, scientist discovered that this property comes with power-law distribution of the node degree and many large-scale graphs have these two properties together including protein network [33], scientific collaboration network [8], web graph [22], and Internet graph [23]. Networks with this property are known as *small world networks*.

The later category of studies focuses on the evolution of graph over time. The main concerns in these studies include how this evolution happen, how new nodes connect to existing nodes, how properties of the graph change over time, and what causes the changes [27] [32].

Works on large scale graphs in computer science, are in both categories. In the following section we consider these works in more depth.

A. Large-scale Graphs in Computer Science

Large-scale graphs attracted attention in computer science in late 90s when Internet and web started to grow explosively. Seminal works in this area were on Web graph and Internet topology. Studies on web graph aimed to improve the performance of search engines [22]. They also helped topic-classification to become more accurate and led to algorithms for enumerating cyber-communities. Researches conducted on the structure of Internet, such as [23], shed light on topology of Internet and this eventually helped to improve the network performance.

Web-pages make the vertices in web graph and hyper links between web-pages constitute edges. For the Internet graph, autonomous systems (AS) are considered as vertices and paths between ASs are represented by the edges of the graph. Internet topology graph is orders of magnitude smaller than web graph.

Works on Internet topology and web graph can be distinguished into two groups: (i) measurement studies (ii) graph generators. Measurement studies try to discover properties of the related graph.

In [5] authors used a BFS search of web graph and they found power-law degree distribution for nodes of web graph. They also investigated shortest path between nodes and found that web graph is a small world network. Broder et al. [10] used Alta-vista search engine (one of the most comprehensive search engines at that time) and collected information of more than 203 million pages and 1,4666 million links between them. Their analysis confirmed power law degree distribution and discovered diameter and WCC size of web graph. Some other properties of web graph are discussed in [22].

[2] [17] [11] were measurement studies conducted on Internet topology. Data used in these studies was gathered by a route server from BGP routing tables of multiple geography distributed routers with BGP connection to the server. These studies also discovered same characteristics for Internet topology such as power law degree distribution and small world property.

Goal of researches conducted in later group was design of algorithms for generating random graphs that have properties discovered in the former group of studies. 6 major methods are recognized, where some are modified versions of the others. In [9] [11] [7] [4] [15] [25] [16] algorithms are discussed thoroughly. Algorithms are distinguished based on the properties that output graph has. These properties are:

- on-line property: nodes can randomly join and leave at any time
- power law degree distribution
- small world
- dense bi-partite subgraph ([22])

B. Related Works on OSN

As Online Social Networks started to grow in past few years, computer scientists started to conduct measurements on different characteristics of them to analyze their impact on Internet. Loosely, works on OSNs fall into 5 categories.

1) empirical characterization of friendship graph: Users attending online social networks usually create a profile and establish connection with their friends on the network. Seminal works on online social networks were all on friendship graph.

Mislove et al. [30] on one of the seminal works captured snapshots of Youtube, Live Journal, Orkut, and Flickr OSNs and found correlation between in-degree and out-degree[15] and a densely connected core for the network.

Ahn et al. [3] calculated some metrics on full graph of Cyworld and random samples of MySpace and Orkut. Authors analyzed degree distribution, clustering coefficient, average shortest path, and degree correlation. In their work, snowball sampling method for OSNs was validated and MySpace, Orkut and Cyworld were compared to each other. They showed common properties between various OSNs.

2) Network Dynamics: Unlike works on the previous section which are focused on the properties of static network of OSNs, works in this section concentrate on formation and evolution of the network. [24] investigates the structure Yahoo! 360 and Flickr networks (two yahoo associated OSNs) and classifies users into three groups: (i) singletons - those who don't have any connection with other users (ii) invitors - who encourage their off-line friends to join the network (iii) linkers - who fully participate in the social evolution of the network. Based on this, authors suggest a model to generate graphs with proportionate number of three groups of users and explain how they should connect to each other.

In [26], authors focus on the evolution of some citation graphs and observe densification of the graph and shrinking of average distance between users in spite of growth of the network in terms of number of nodes. Based on their observation, they suggest a new model for graph generation which is similar to forest fire propagation process.

In [6], instead of considering the OSN as a whole, authors turn their attention toward communities formation and evolutions. Authors used two data sources: friendship links and

[15] OSN friendship graph can be directional or bidirectional. For bidirectional graphs, in-degree equals the out-degree because edges are not directional. On contrary, direction of the edges in directional graphs causes imbalance between in-degree and out-degree of nodes

community membership on Live Journal, and Co-authorship and conference publication in DBLP. They found relationship between propensity of individuals joining communities and underlying network structure.

3) characterization of content: Web 2.0 changed the way users used to interact with websites. Web 2.0 concept enabled users to participate in the process of generation of content. Online social networks fully used this concept and, nowadays, almost all of the OSNs provide services for users to share and use user-generated contents. Some OSNs, such as Youtube and Flickr, focus on one or more types of contents. How user use this content and what are the impact of that on the underlying network are important questions that several works tried to answer.

Youtube is one of the most famous website for uploading and sharing user-produced videos. It is estimated that 10% of traffic of the web is produced by this website. [19] and [12] concentrated on this website and characterized content from different perspectives. Because of abundant number of videos on Youtube each work had its special way to narrow down number of investigated videos. Thus both works lack completeness and results presented in them maybe biased. Results presented in our work are to some extent related these works because they also focused on the contents posted by users and characterized its properties.

Gjoka et al. [20] focused on Facebook and investigated behavior of users on applications. These application are mostly user generated and comparable in some ways to content. They reported the pattern of use, growth of popularity over time, and effect of application category in their work.

4) characterization of interaction: The first work on how people interact with each other on online social networks was [21]. This work is focused on poking and messaging on Facebook and it reports reciprocity, school ties, temporal rhythms, and seasonal variation on how people send messages to each other.

[13] characterized the pattern users add each others' photos as their favorite photos on Flickr and uses that to investigate information dissemination in the system.

In another work on interaction in online social networks, Chun et al. [14] investigated the interaction on Cyworld, the largest OSN in Korea. In their work, the structure of the interaction graph is analyzed and they found out value of properties such as clustering coefficient, degree distribution, network motifs, and disparity for the interaction graph. Next they compared the coherence between interaction graph and friendship graph. At last, they analyzed the time between the time messages were sent and the time they were answered.

Our work has great affinity with works in this section. We go further beyond the basic characterization done in [21] and our work does not have flaws of [14]. The other difference between our work and other works is that our focus is on indirect interaction rather than direct interaction.

5) Embedding OSN features in designing other protocols: Researchers used OSN features, such as friendship links between users, to design new protocols based on assumption that these features are controlled deliberately by users. [37] and [31] uses links in OSNs to create a more secure environment for users. In the former work, links are used against Sybil attacks. In the later one, links are used to protect legitimate users against spammers and promoters.

In [29], Mislove et al. investigated difference in exchange of content in web and in social network and developed an application to exploit feature of social networks for Internet search. They found out that using OSN features can greatly improve the performance of search engines.

REFERENCES

[1] Alexa, Top Sites in United States: http://alexa.com/topsites/countries/US.

[2] L. Adamic and B. Huberman. Scaling behavior of the world wide web. *Science*, 2000.

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

[4] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Symposium on Theory of Computing*, 2000.

[5] R. Albert, A. Barabasi, and H.Jeong. Diameter of the World-wide Web. *Nature*, 1999.

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *KDD*. Cornell University, 2006.

[7] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286, 1999.

[8] A. Barrat, M. Barthlemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *National Academy of Science*, 101, 2004.

[9] A. Bonato. A survey of models of the web graph. In *Combinatorial and Algorithmic Aspects of Networking*, 2004.

[10] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structures in the Web: Experiments and Models. In *WWW*, 2000.

[11] T. Bu and D. Towsley. On Distinguishing between Internet Power Law Topology Generator. In *Infocom*, 2002.

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

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

[14] H. Chun, H. Kwak, Y. ho Eom, Y.-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.

[15] F. Chung and L. Lu. Connected components in random graphs with given degree sequence. In *annals of Combinatorics*, 2002.

[16] C. Cooper, A. M. Frieze, and J. Vera. Random Deletion in a Scale-Free Random Graph Process. *IM*, 2003.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *SIGCOMM*, 1999.

[18] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazieres, and H. Yu. RE: Reliable Email. In *NSDI*, 2006.

[19] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *sigcomm*, 2007.

[20] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking Facebook: Characterization of OSN Applications. In *WOSN*, 2008.

[21] S. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: messaging within a massive online network. In *Third International Conference on Communities and Technologies*, 2007.

[22] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tompkins. The Web as a Graph: Measurements, Models, and Methods. In *WWW*, 1999.

[23] V. Krishnamurthy, J. Sun, M. Faloutsos, and S. Tauro. Sampling Internet Topologies: How Small Can We Go? In *International Conference on Internet Computing*, 2003.

[24] R. Kumar, J. Novak, and A. Tomkins. Structure and the Evolution of Online Social Networks. In *KDD*. Yahoo! Research, 2006.

[25] R. Kumar, P. Raghavan, R. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic Models for Web Graph. In *IEEE symp. on Foundations of Computer Science*, 2000.

[26] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD*, 2005.

[27] D. LibenNowell and J. Kleinberg. The link Prediction Problem for Social Networks. In *ACM International Conference on Information and Knowledge Management (CIKM'03)*, 2003.

[28] S. Milgram. The small world problem. *Psychology Today*, 2, 1967.

[29] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting Social Networks for Internet Search. In *5th Workshop on Hot Topics in Network (HotNets-V)*, 2006.

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

[31] A. Mislove, A. Post, P. Druschel, and K. P. Gummadi. Ostra: Leveraging Trust to Thwart Unwanted Communication. In *NSDI*, 2008.

[32] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review Letters*, 2001.

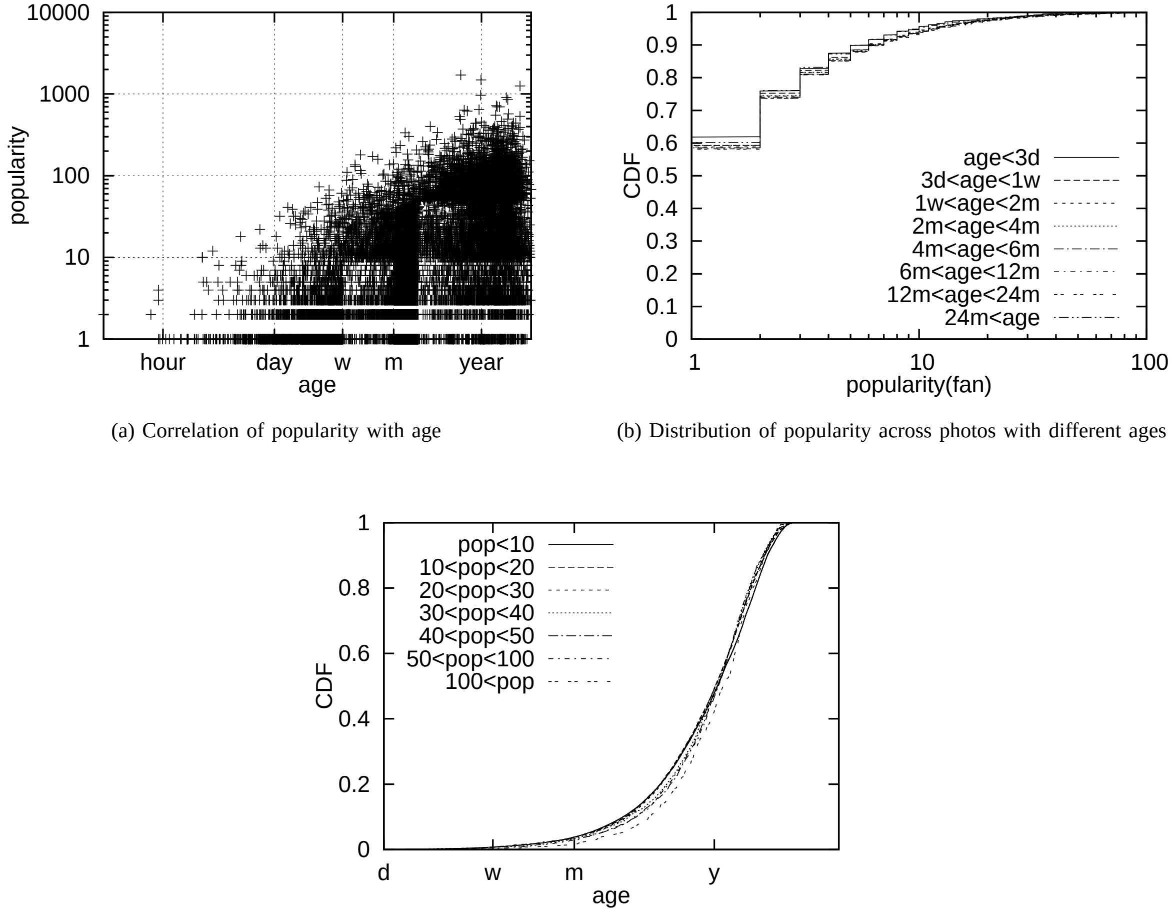
[33] B. P, J. LJ, von Mering C, R. AK, L. I, and M. EM. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3), 2004.

[34] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven Sampling for Characterizing Unstructured Overlays. In *IEEE INFOCOM Mini-conference*, 2009.

[35] D. Stutzbach and R. Rejaie. Understanding Churn in Peer-to-Peer Networks. In *Internet Measurement Conference*, 2006.

[36] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks. Technical Report CIS-TR-06-07, University of Oregon, 2006.

[37] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: defending against sybil attacks via social networks. In *Proceedings of ACM SIGCOMM*, volume 36. ACM Press, 2006.

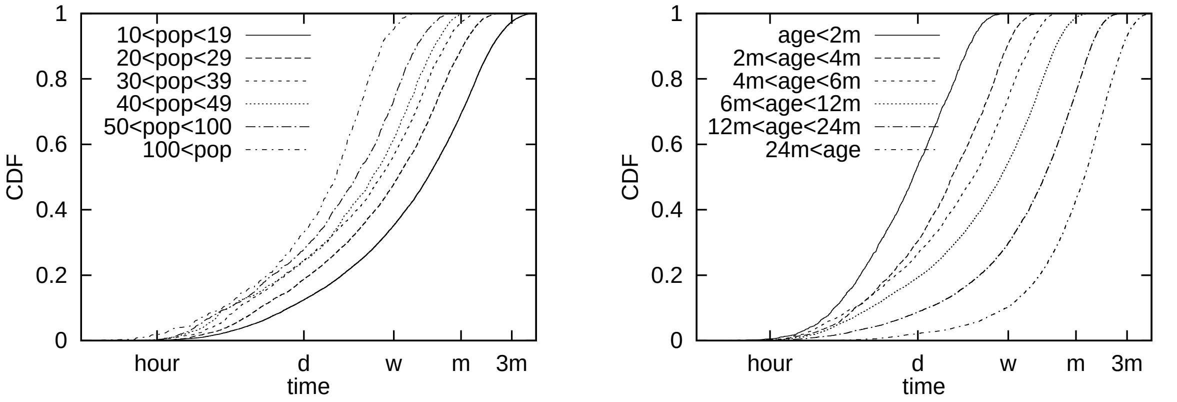


(a) Correlation of popularity with age

(b) Distribution of popularity across photos with different ages

(c) Distribution of age across photos with different popularity

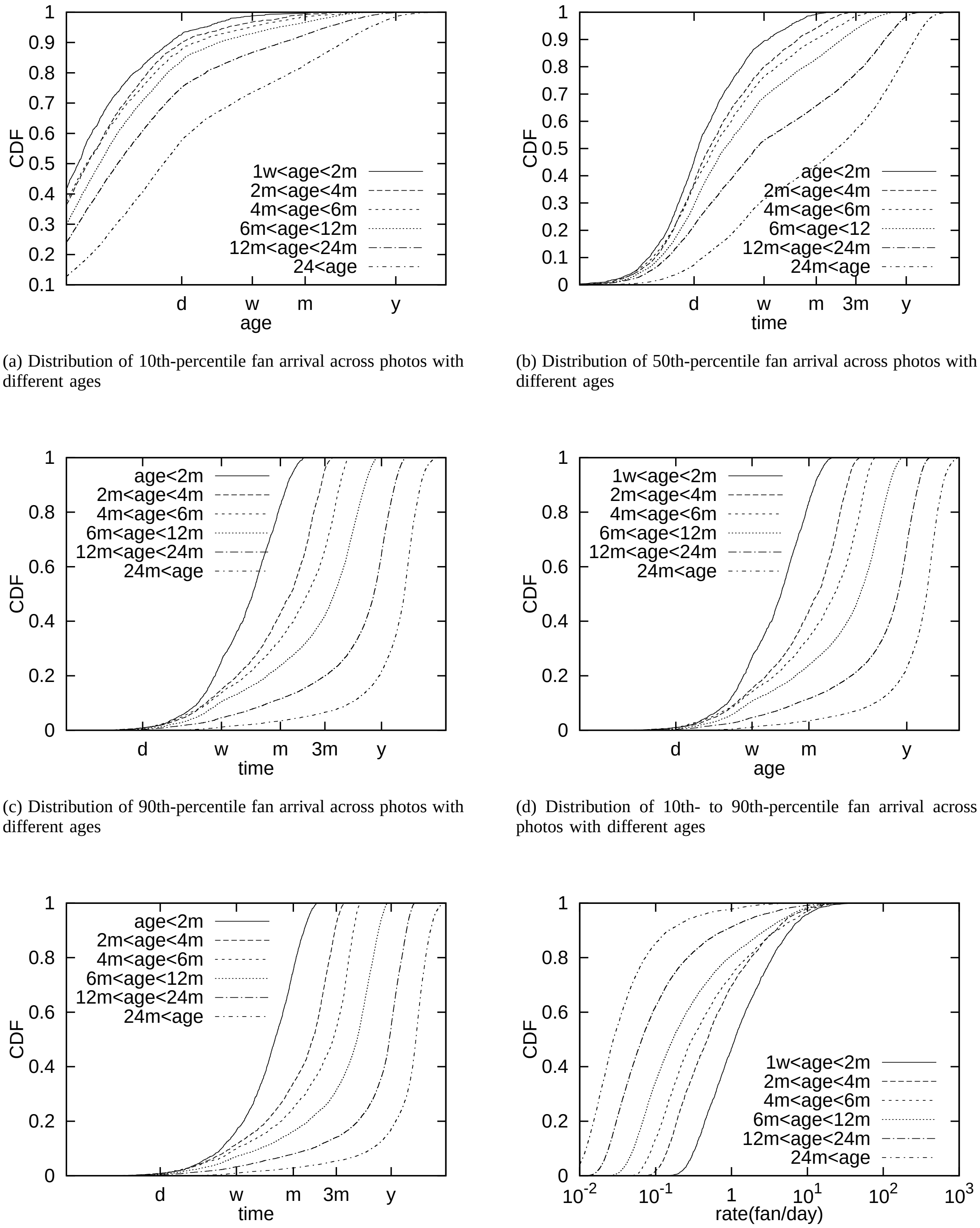
Fig. 9. Relation between popularity and age of photos in Flickr



(a) Distribution of inter-fan-arrival time across photos with different popularity

(b) Distribution of inter-fan-arrival time across photos with different ages

Fig. 10. Effect of age and popularity on interarrival of fans



(a) Distribution of 10th-percentile fan arrival across photos with different ages

(b) Distribution of 50th-percentile fan arrival across photos with different ages

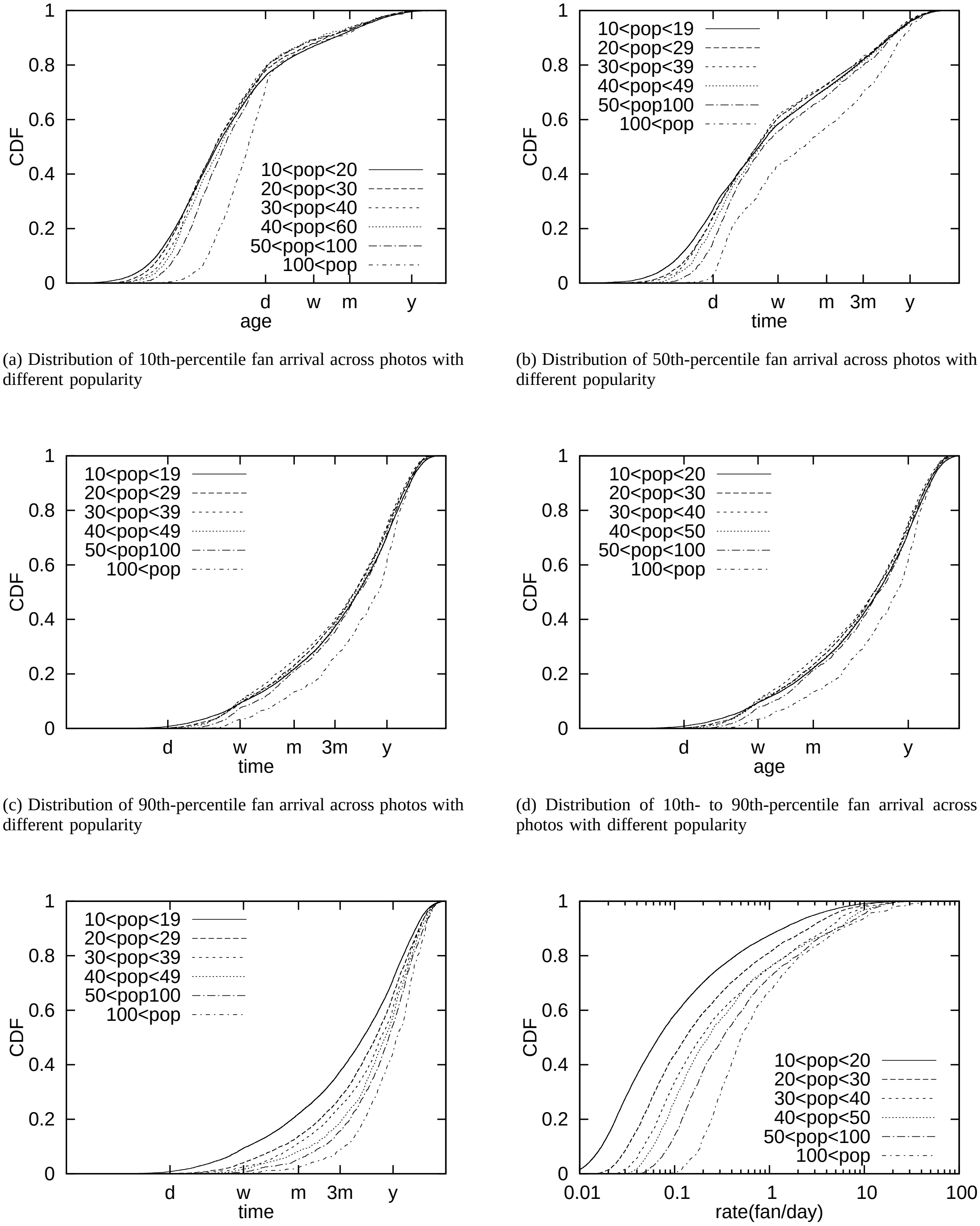
(c) Distribution of 90th-percentile fan arrival across photos with different ages

(d) Distribution of 10th- to 90th-percentile fan arrival across photos with different ages

(e) Distribution of first to last fan arrival across photos with different ages

(f) Distribution of rate of fan arrival across photos with different ages

Fig. 11. Effect of photo age on pattern of fan arrival



(a) Distribution of 10th-percentile fan arrival across photos with different popularity

(b) Distribution of 50th-percentile fan arrival across photos with different popularity

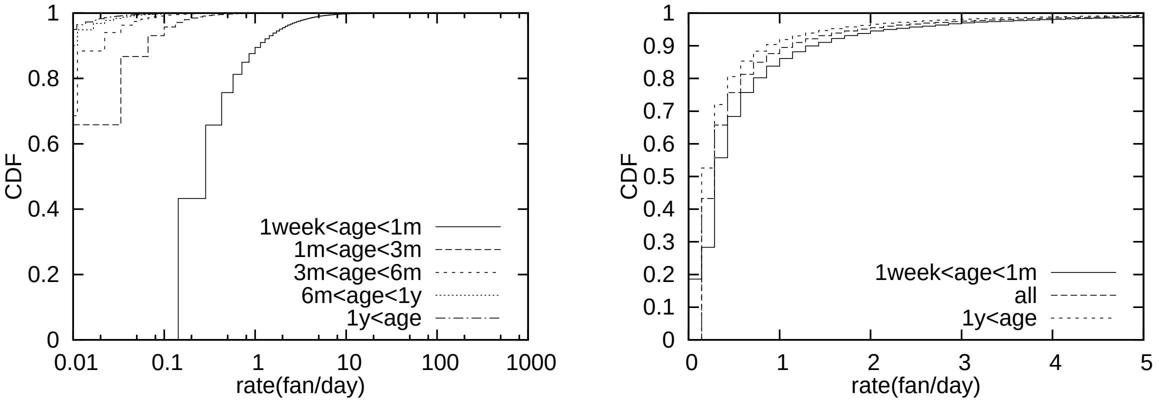
(c) Distribution of 90th-percentile fan arrival across photos with different popularity

(d) Distribution of 10th- to 90th-percentile fan arrival across photos with different popularity

(e) Distribution of first to last fan arrival across photos with different popularity

(f) Distribution of rate of fan arrival across photos with different popularity

Fig. 12. Effect of photo popularity on pattern of fan arrival



(a) Distribution of rate of arrival of fans in different period of lifetime of photos

(b) Distribution of arrival of fans in the first week of lifetime of photos, for photos less than a month and more than a year old

Fig. 13. Effect of different period of life of a photo on rate of fan arrival [believe it or not this caption is the best I could think of!]