IDIAP RESEARCH REPORT

Analyzing Flickr Groups

Radu-Andrei Negoescu [1]
Daniel Gatica-Perez [1]

IDIAP–RR 08-03

April 2008

to appear in
Proceedings of CIVR'08

1 IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, {negora,gatica}@idiap.ch

IDIAP Research Report 08-03

Analyzing Flickr Groups

Radu-Andrei Negoescu Daniel Gatica-Perez

April 2008

to appear in
Proceedings of CIVR'08

Abstract. There is an explosion of community-generated multimedia content available online. In particular, Flickr constitutes a 200-million photo sharing system where users participate following a variety of social motivations and themes. Flickr groups are increasingly used to facilitate the explicit definition of communities sharing common interests, which translates into large amounts of content (e.g. pictures and associated tags) about specific subjects. However, to our knowledge, an in-depth analysis of user behavior in Flickr groups remains open, as does the existence of effective tools to find relevant groups. Using a sample of about 7 million user-photos and about 51000 Flickr groups, we present a novel statistical group analysis that highlights relevant patterns of photo-to-group sharing practices. Furthermore, we propose a novel topic-based representation model for groups, computed from aggregated group tags. Groups are represented as multinomial distributions over semantically meaningful latent topics learned via unsupervised probabilistic topic modeling. We show this representation to be useful for automatically discovering groups of groups and topic expert-groups, for designing new group-search strategies, and for obtaining new insights of the semantic structure of Flickr groups.

1 Introduction

Social media repositories involving images, video, text, etc. constitute an emerging challenge for multimedia information management systems. Users of such repositories interact in a variety of ways with the media, thus creating additional metadata that could be exploited by management systems. As of January 2008, Flickr claims to host over 228 million photos, indexed by over 20 million unique tags [1], making it one of the largest online image repositories, with an incredible amount of associated metadata associated.

Social media in general, and Flickr in particular, are interacting online communities, producing, sharing, viewing and repurposing content while participating in a number of social scenes. The understanding of the complex social aspects of Flickr, including its users' motivations and needs, the social uses of the system features, and the collective behaviors that emerge from the intersection of people and content opens doors to entirely new opportunities for the image retrieval community [14, 21, 18, 9].

In particular, Flickr's social link structure has been analyzed, based on connectivity information, i.e., "who is a contact of whom" [10], in the traditional social network set up. However, to our knowledge, little attention has been paid so far to another social connection feature on Flickr, namely "groups". Groups in Flickr are self-organized communities with declared, common interests, and are explicit instantiations of the "content+relations" feature of social media. Groups are created spontaneously but not randomly: people participate in groups (e.g. by sharing pictures) for specific social reasons, and most groups are defined about specific topics or themes (e.g. an event or a photographic style). Aggregating content and metadata for groups could thus offer insights into both large scale behavioral trends (e.g. photo sharing practices), and also provide robust representations (e.g. at the topic level) to characterize groups by their content (and not only by their connectivity). This could in turn offer viable new alternatives to organize and manage visual content. These are the issues addressed by our work.

Our paper contains two contributions. First, we present an analysis of Flickr groups from the perspective of the photo-sharing practices of their members. Such analysis, to our knowledge, has not been previously attempted. Based on a snapshot of the Flickr collection (involving roughly 7 million images extracted from a sample of users belonging to 51000 groups), our work reveals a number of fundamental patterns with respect to the degree of active participation in groups, group affiliation, group loyalty, photo repurposing, and the effects of certain system design choices (user subscription models) in photo sharing practices. Second, motivated by the current limitations to browse and search for Flickr groups, we propose a novel topic-based group representation, which is learned in a probabilistic, unsupervised manner from the groups' tags. We demonstrate that our topic-based representation facilitates the discovery of topic-related groups of groups, allows the creation of new methods of group-search, and is also useful for further analysis of the group structure of Flickr at a higher semantic level. Our paper therefore contributes both to the understanding of relevant collective behaviors in social media repositories and to the development of potentially useful applications.

The paper is organized as follows. Section 2 summarizes existing related work. Section 3 recalls the concept of group in Flickr. Section 4 presents our analysis of the photo sharing practices in Flickr groups. Section 5 introduces our proposed topic-based group representation and presents a topic-based analysis of Flickr groups, discussing some of its further uses. Conclusions are drawn in Section 7.

2 Related Work

Flickr data has started to be used in the context of classic content-based image retrieval research [13]. However, one of the most interesting aspects of Flickr, apart from the sheer size of its data, is the plethora of metadata associated with photos, in the form of tags, notes, number of views, comments, number of people who mark the photo as a favorite, and even geographical location data. Recent studies have used notes [20], combinations of tags, geolocation and visual data in order to improve retrieval [16, 5], visualization, and summarization techniques for large databases either over time or over a geographic area [3, 6, 9, 8], to automatically extract place and event semantics [18], or to induce tag ontologies [19].

Tagging systems have been analyzed by Marlow et al. [14], and a taxonomy of users' motivations to tag has been proposed by Ames and Naaman in [4]. There have also been some studies analyzing the sharing practices, motivations, and privacy concerns of the users [21, 15, 2]. In particular, Van House [21] discusses the main uses of photo sharing amongst users on Flickr. While these studies provide particularly useful insights into user behavior, none of them explicitly address sharing practices in relation to Flickr groups, as we do here.

In addition to the photo metadata, attention has been given to metadata stemming from the (social) links existing on Flickr [10, 12, 11, 22]. Recent work includes studying user-to-user relations by means of contact bookmarking, a direction explored by Kumar et al. [10], with interesting results regarding the structure of the Flickr social network. Other works have considered user-to-photo relations by means of ownership, favorites, or comments. Van Zwol [22] analyzes the way new photos are discovered by users on Flickr, and finds that most photo views and comments occur in the first two days after the upload, concluding that the social network of the user and photo pooling (i.e. sharing with groups) are two major indicators of a photo's popularity. In a similar study, Lerman and Jones [11] found that the number of views a photo receives correlates strongly with the size of the social network of a user, and more particularly the reverse contacts. Lerman et al. [12] use a user's existing social network and a latent topic model on tags in order to filter tag search results for that specific user. The motivation and specific use of topic models is, however, fundamentally different than ours.

In summary, compared to our work, previous works have either exploited different social link information or targeted different goals. At the same time, some of the findings in [11, 22, 21] provide us with a starting point about the user motivations for using the Flickr group functionality, and in understanding why new representations for groups are needed.

3 What are Flickr Groups?

The word "group" has several definitions in the English language, but we find two of them to be most representative for Flickr groups [17]: (1) "An assemblage of persons or objects gathered or located together"; (2) "A number of individuals or things considered together because of similarities". A group is therefore a collection of persons or objects, who are either in physical proximity or share some abstract characteristics. On Flickr, from a strictly technical point of view, groups are collections of users who freely choose to join such a community. The main purpose of groups is to facilitate the sharing of user photos in what is called the *group pool*. This is a collection of photos shared by any member with the group, and, implicitly, all the tags associated with the photo become part of the group photo pool. One can distinguish between several types of groups, which may sometimes be intertwined. A short, non-exhaustive list could include:

geographical/event groups: groups limited to a geographical region or a specific event (local or global), such as *New York City*, *San Francisco Bay*, *Switzerland*, *Live Music*, *World Events (festivals, protests, etc.)*, *Global Photojournalism*;

content groups: groups primarily oriented towards the visual content being shared, such as *R is for Red*, *Leaves (No Trees Please!)*, *Cats - Small to Big*, *Artistic Child Photography*;

visual style groups: groups that concentrate on a specific photographic technique, for example *Life in Black and White*, *Closer and Closer Macro Photography*;

quality indicator groups: the goal of these groups is the identification and regroupment of (perceived) high quality photography, such as *Blue Ribbon Photography [Invited Images ONLY]*, *Superb Masterpiece - Invited pictures only (Vote Now!)*, *The Best: BRAVO (INVITED images only)*, *Flickrs Best (Better than Explore!) - (Invite or Award Only)*;

catch-all groups: these groups do not seem to have any particular content-oriented rules, but rather

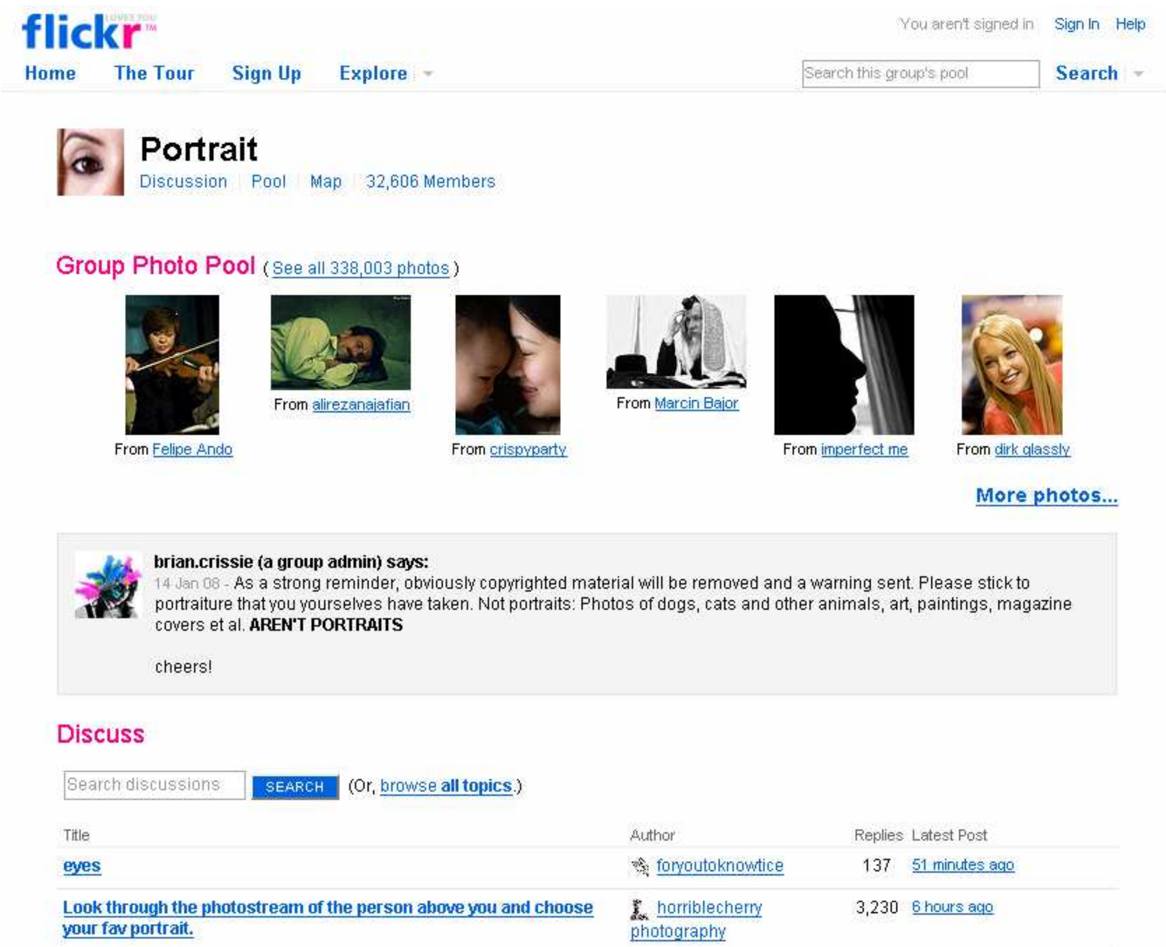


Figure 1: The typical home page of a Flickr group

they are an invitation for users to share photos in groups. They usually have huge numbers of users and photos: *Flickr Central*, *10 Million Photos*, *The Biggest Group! - Playground for Psychotics!*.

Figure 1 shows the home page of a content group, *Portrait*. When users join a group, they can start sharing photos in the group pool. There are three privacy settings for groups: (1) public, anyone can see the group photo pool, and anyone can join; (2) public, requiring an invitation from a member; and (3) private, nobody can find the group, and a user must be invited to join.

4 An Analysis of Flickr Groups

4.1 Datasets

We have collected the data used in this study using Flickr's API. All the information extracted about a particular user is publicly available, and statistics linked to the number of photos may vary if users employ restrictive privacy settings for their photos. This private information was not available to us for this study.

Our dataset consists of approximately 22,000 registered Flickr users, roughly 7 million photos belonging to these users (the most recent 500 photos per user), and about 23 million tags belonging to these photos. We chose to limit the number of photos to the most recent 500 primarily to facilitate the data collection process. As pointed out in [21], most users see Flickr as a social site, and are only interested in the most recent photos (theirs, and their contacts'), which supports the decision of using the most recent 500 photos. The data collection process can be described as follows: repeatedly retrieve the first approximately 3,900 photos uploaded from a randomly sampled moment t in the interval December 22nd, 2004 - April 2nd, 2007, until information on roughly 187,000 photos has been collected. We have thus obtained 22,414 distinct users, the owners of the photos. For each of the users we have retrieved their most recent 500 photos which, in some cases, meant all their photos, for a total of nearly 7 million photos. We have then collected all the tags associated with these photos. Only about 4.7 million photos have at least one tag. In addition to the users, photos and tags, we have also collected information about the groups the photos belong to, with 1.13 million photos belonging to at least one group. Let us formalize the definition of this original dataset (D_O):

- users: $U = \{U_i \mid i = 1...N_U\}$ with $N_U = |U| = 22{,}414$ the number of users
- groups: $G = \{G_i \mid i = 1...N_G\}$ with $N_G = |G| = 51{,}407$ the number of groups
- photos: $P = \{P_i \mid i = 1...N_P\}$ with $N_P = |P| = 6{,}926{,}622$ the number of photos
- tags: $T = \{T_i \mid i = 1...N_T\}$ with $N_T = |T| = 1{,}969{,}813$ the number of distinct tags

4.2 Data Analysis

In this section we analyze the structure of our dataset and posit that, given the random selection process of the users, this structure is characteristic of the Flickr community. Users who do not use Flickr to upload photos will most likely have different usage patterns all together.

In order to understand how users make use of the groups they join, we have also analyzed the statistics of our dataset D_O from a *sharing photos with groups* perspective. Let us define the following notations:

- $U_{i,p}$: the total number of photos in user U_i's collection
- $U_{i,s}$: the total number of photos user U_i shares with groups
- $U_{i,g}$: the total number of distinct groups in which user U_i shares photos
- $U_{i,\sigma}$: the total number of sharing instances; this is the count of all photo-group pairs for user U_i

Using the above notations, we can write the following:

- $\gamma = \frac{U_{i,\sigma}}{U_{i,s}}$: the average number of groups a photo is shared with, for user U_i
- $\pi = \frac{U_{i,\sigma}}{U_{i,g}}$: the average number of photos shared per group, for user U_i

Figure 2 shows histograms of the real number of members and the real number of photos for all the groups in our dataset D_O. These numbers have been retrieved directly from Flickr and represent the real-life sizes of the groups. Both the number of members and the number of photos seem to approximate a log-normal distribution. For the rest of our study we focus on the numbers present in our D_O dataset.

To share or not to share? Figure 3 shows the histogram of photos shared with groups for the users in our dataset. Of the 22,414 users in the snapshot, 50.9% share at least one photo with at least one group, 26.4% share more than 50 photos and 9.9% share more than 200 photos. For the full dataset, the average number of photos shared with groups is 54.6. If we only consider the users who actually share photos with groups, this average is 106.4 photos. Figure 4 shows the distribution of the percentages of shared photos for the users who share photos with groups. This is the ratio between the number of shared photos and the total number of the user photos, $\frac{U_{i,s}}{U_{i,p}}$. About a quarter of the users share at least 50.1% of their photos in groups, while almost half share at least 17.2%

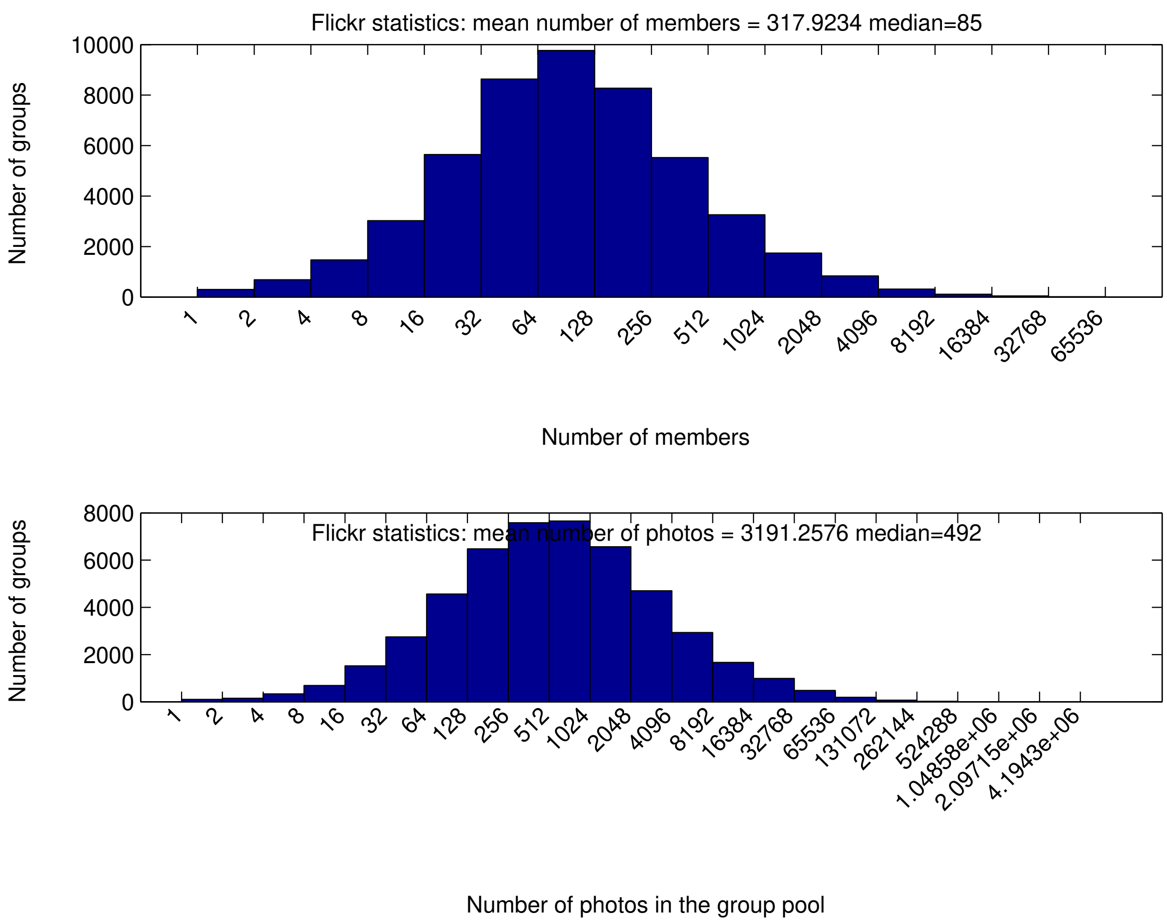


Figure 2: Top: histogram of the number of members per group. The mean number of members per group is 317.9 and median is 85. Bottom: histogram of the number of photos per group photo pool. The mean number of photos is 3191.3 and the median is 492. The x axis are shown in log-2 scale for displaying reasons.

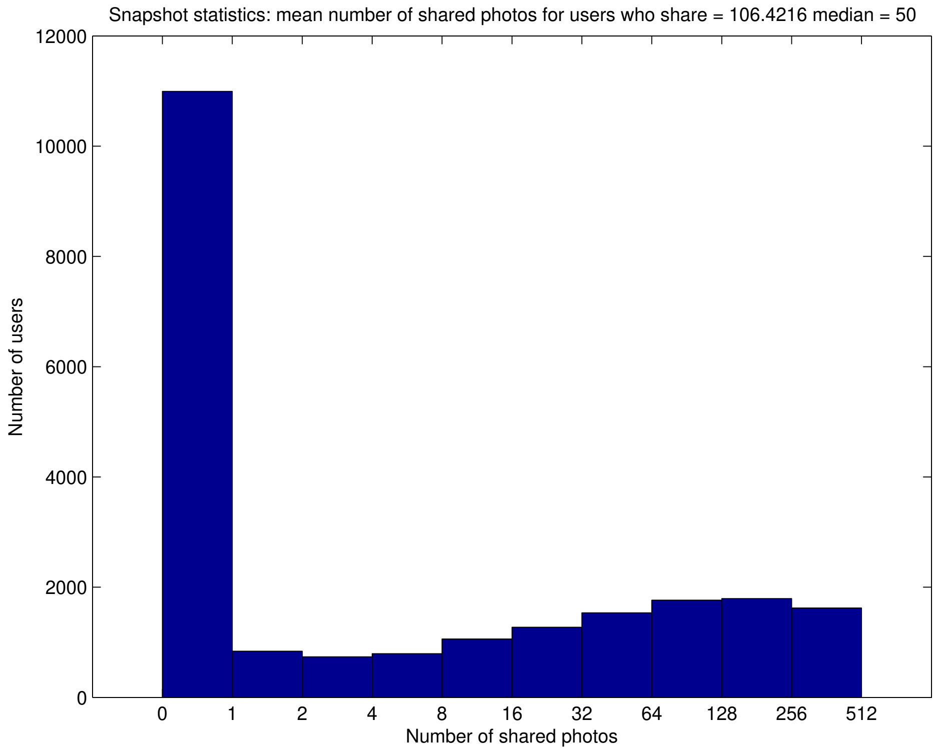


Figure 3: Histogram of the number of photos shared with groups $U_{i,s}$, including the users who have not shared any photos. The average number of shared photos is 54.6. The x axis is shown in log-2 scale for displaying reasons.

of their photos. The mean sharing percentage is 29.6%. We consider this to be an indication that sharing photos with groups is an important part of the photo sharing practices of Flickr users. To the best of our knowledge, user motivations for sharing photos with groups have not yet been analyzed, however motivations for tagging photos and uses of personal photography have been. Four main uses of personal photography have been noted in [21]: *memory, identity, and narrative*, *maintaining relationships*, *self representation*, and *self expression*. We believe that out of these four uses, self expression (or *photo exhibition*) and maintaining relationships are the ones driving users to share photos with groups. Groups ensure a higher exposure of the photos, and it is common practice for thematic groups to require their members to comment on the most recent photo posted before their own. Group photo pools also allow users who have an interest in a specific topic to have a regular photo stream focused on that topic. Some other groups are not thematic, but rather geographically localized, and users sometimes organize offline meetings, creating and maintaining new relationships.

In order to understand whether the size of a user's photo collection influences his or her percentage of shared photos, we have analyzed the relation between these two measures. This is shown in Figure 5. The sizes of the photo collections for users who share no photos at all are evenly spread over the entire range of sizes (the thick line overlapping the x axis), and the sharing percentages for the users who have the maximum number of photos allowed in our dataset are also evenly spread over the entire interval $[0, 1]$ (the thick line at $x = 500$). The correlation coefficient between the two measures is 0.1417, indicating a rather weak correlation.

Group affiliation through photo sharing: how many groups does a user share photos with? As pointed out earlier, 50.9% of the users share at least one photo in at least one group. Figure 6 shows a histogram of the absolute number of groups users share photos with. For the full dataset, users share photos with an average of 25.3 distinct groups. If we only consider the users who actually

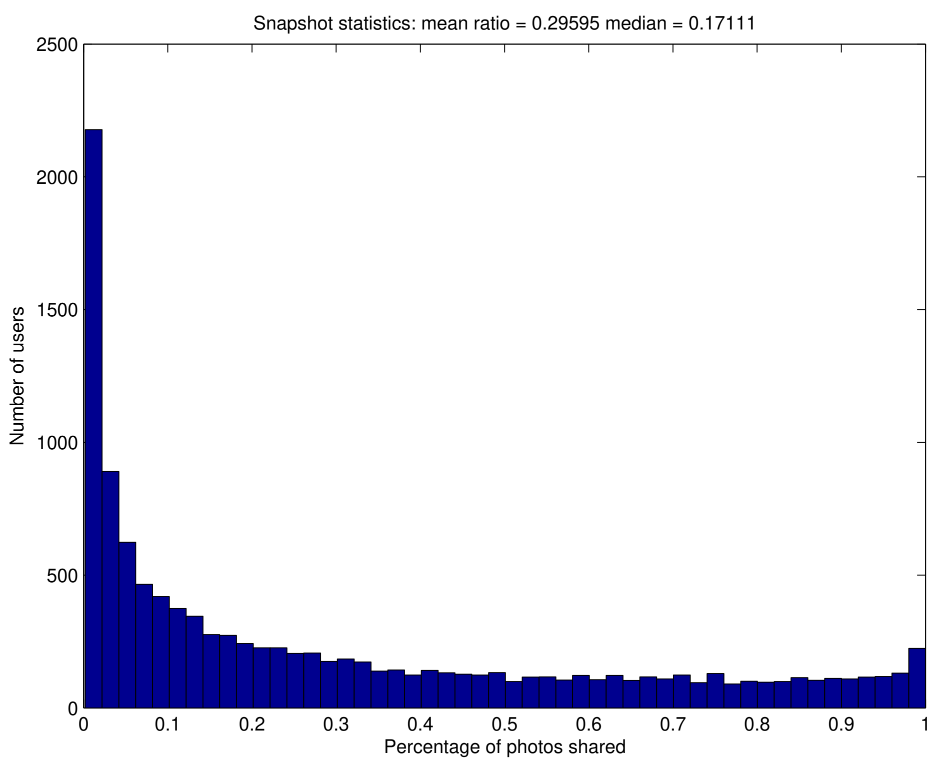


Figure 4: Histogram of $\frac{U_{i,s}}{U_{i,p}}$, the percentage of photos shared with groups, for sharing users. The mean sharing percentage is 29.6%, and the median is 17.1%.

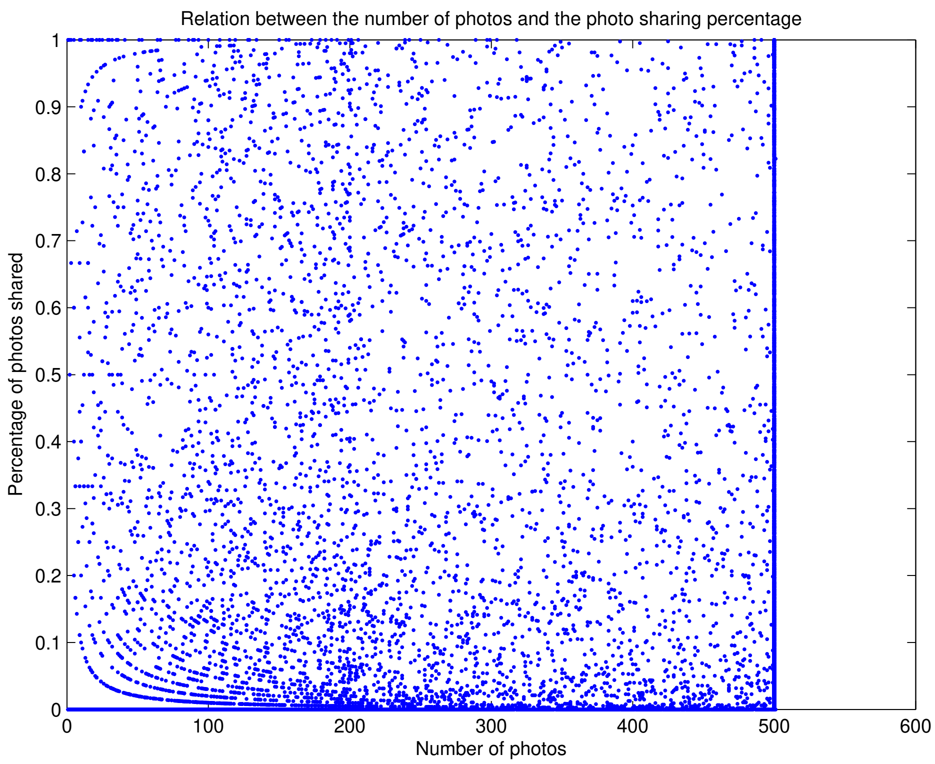


Figure 5: The percentage of shared photos (x-axis) vs. the number of photos of each user (the y-axis): the size of the collection of photos for users who do not share any photos at all ($U_{i,s} = 0$) is evenly spread over the entire range of sizes $U_{i,p} \in [1, 500]$; the sharing percentages for users who have the maximum number of photos ($U_{i,p} = 500$) is evenly spread over the full interval $[0, 1]$.

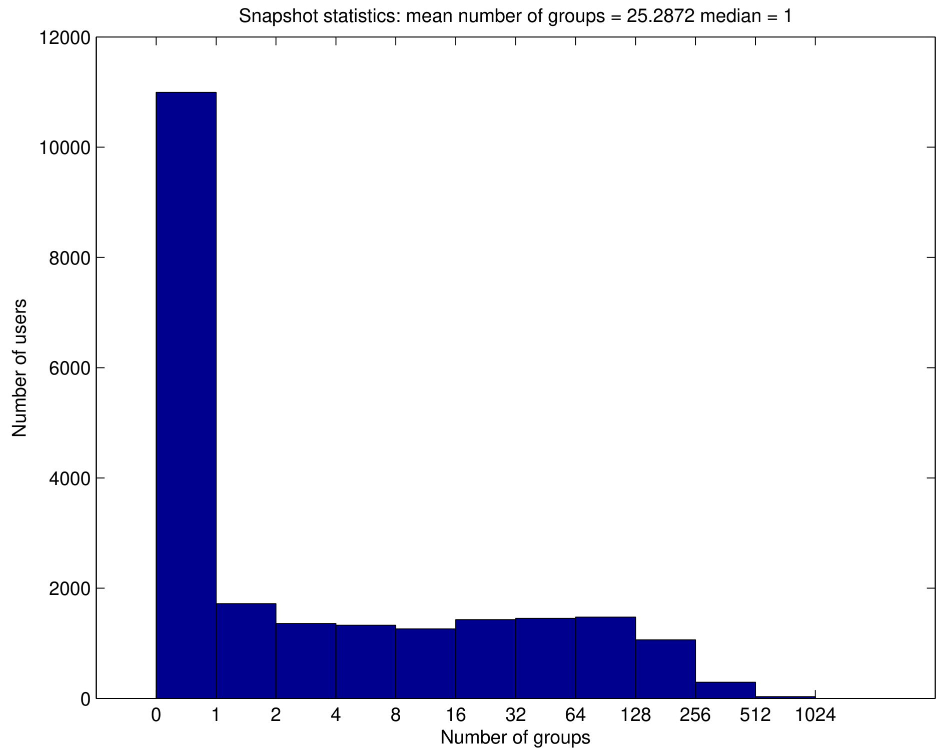


Figure 6: Histogram of the number of groups photos are shared with per user. The average number of groups is 25.3.

share photos, the average number of groups with which they share photos is 49.6, with a median of 16. 15.1% of the sharing users share their photos with exactly one group, and 45.6% of them share photos with more than 20 groups. 11.3% of the sharing users actually share photos with more than 140 groups. This highlights two trends: (1) roughly half of the people do not share with groups at all, and (2) half of the users do, and exploit this feature affiliating to several groups. In the half that shares, several distinct behaviors also emerge: moderate sharers, with fewer than 5 groups, average sharers, and extreme sharers, with hundreds of groups.

Group loyalty: how many photos does a user share with the same group? Another measure characteristic of the sharing behavior is the average number of photos shared per group, π. For clarity of display, we have plotted the histogram of $\frac{1}{\pi}$ in Figure 7. 9.9% of the users share on average one photo per group, and 85.1% of the users share on average less than 15 photos per group. The mean of the average number of photos shared per group for users who share photos is 9.6, and the median is 5.1. This analysis seems to indicate users tend to share a limited amount of photos with the same group. This could be an effect of the large number of groups on Flickr that share the same theme. For example, searching on Flickr for “black and white” yields about 25,000 results, searching for “sunset” yields about 29,000 groups. Less common words, like for example, “gold”, or “magazine”, get 4,600 and 2,200 results, respectively. Another reason might be the driving force behind sharing with groups: if the motivation is photo exhibition, the users will try to share their photos with many groups, and thus show feeble group loyalty; if the motivation is an interest in a specific theme, they will most likely contribute all their photos belonging to that theme into the same group(s).

Photo recycling: how often is the same photo shared with multiple groups? The ratio between the sharing instances and the number of shared photos effectively represents the average number of groups photos are shared with, γ. Again, for display clarity, we present in Figure 8 a histogram of $\frac{1}{\gamma}$. The mean γ value is 3.1, and the median is 1.5. 27.5% of the users share on average

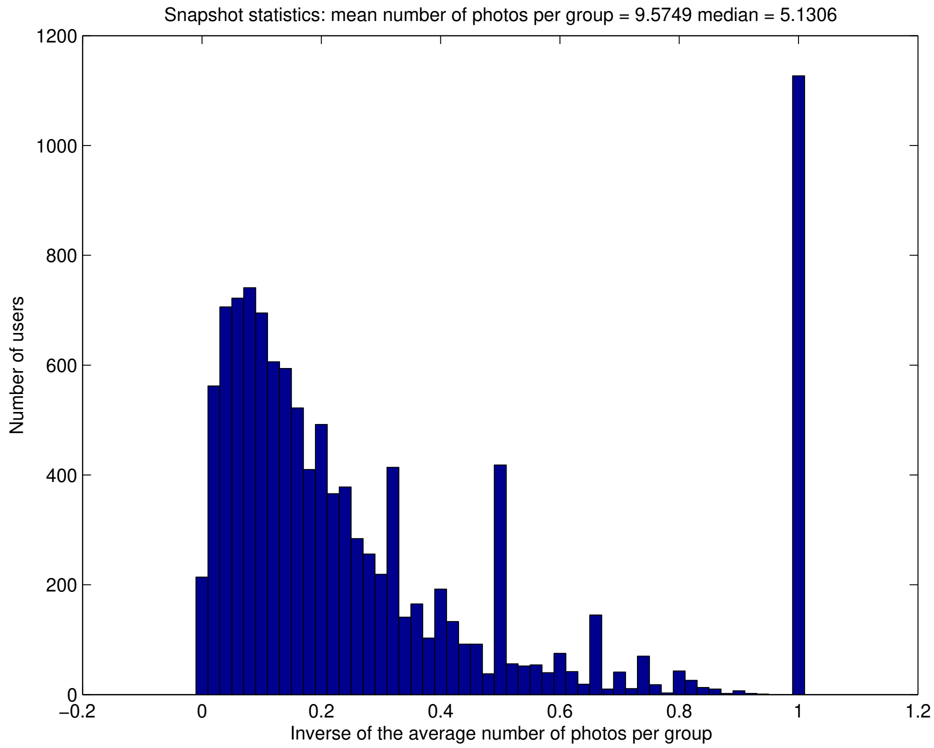


Figure 7: Histogram of $\frac{1}{\pi}$, which is the inverse of the average number of photos shared in the same group. The mean of π over the sharing users is 9.6.

each photo in only 1 group, and only 5.4% of the users share the same photo in more than 10 groups. This seems to indicate that most users share the same photos in a rather limited number of groups. How these groups are chosen by the users from the (possibly) hundreds of similar groups with the same theme is open to speculation. Users may either stumble upon a group and not look for other similar ones, or search and select a group out of the search results based on the perceived affinity with the group in terms of content, members, and rules. In any case, it appears that important numbers of users in our dataset do not seem to fully profit from the possibility of increasing the visibility of (we hypothesize) their preferred photos, choosing not to recycle their content. It should be noted that, at the time of this analysis, the maximum number of groups a photo could be shared with was set by Flickr to be 60 for paying members, and 10 for non-paying members.

In order to determine whether a correlation between the average number of groups per photo and the average number of photos per group exists, we have computed the correlation coefficient between γ and π over the set of users sharing photos. This coefficient is 0.2159, which seems to indicate a relatively weak correlation between the two measures. Figure 9 shows that users sharing a large number of photos per group often do so in only a few groups, while users sharing fewer photos per group often tend to share photos in more groups. This large variation might suggest that several motivations for sharing photos with groups exist, and these motivations result in different practices for photo sharing. People sharing with many groups might be driven by the *photo exhibition* motivation, while those sharing with only a few groups are probably driven by the more socially anchored motivation of *maintaining relationships* with groups of people either sharing the same passion or interest for a given theme, or being located in the same area.

Are you a pro? Part of the sharing behavior might be influenced by the type of Flickr account a user might have: free accounts allow users to only display the most recent 200 photos from their collection, and to only share a photo with a maximum of 10 groups; paying members (called *pro*

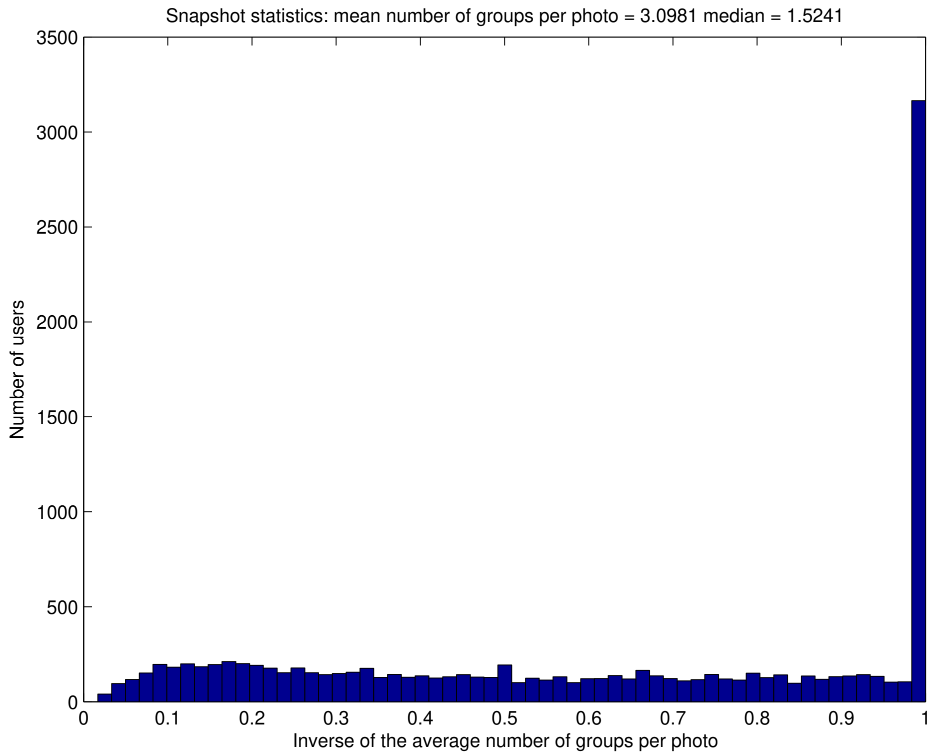


Figure 8: Histogram of $\frac{1}{\gamma}$, which is the inverse of the average number of groups per photo. The mean of γ over the sharing users is 3.1

members by Flickr) have no limit on the number of photos that are displayed in their account, and can share a photo with a maximum of 60 groups. Therefore we have also analyzed the differences in sharing behavior for paying and non-paying members.

In our dataset D_O, the two types of users exist in nearly equal quantities: 51.43% paying users and 48.57% non-paying. The percentages of users who share photos with groups show a significant difference: for paying users, 69.79% share photos with groups, while for the non-paying users, only 31.01% do. We present in Table 1 the most important statistics for the paying users, non-paying users, and the full dataset. Only users who share photos with groups are taken into account, in order to establish if significant differences exist in sharing behavior. It is clear that the Flickr-imposed maximum limits of 200 visible photos and 10 groups per photo do affect the way non-paying members use their accounts in terms of photos uploaded and groups shared with; however, it is interesting to observe that, although on average pro members upload more and share with more groups (rows $U_{i,p}$, $U_{i,s}$, and $U_{i,g}$ in Table 1), the overall sharing ratio is not influenced by their paying or non-paying status (the $\frac{U_{i,s}}{U_{i,p}}$ row). The average sharing measures γ and π also show differences, but at a smaller scale. In conclusion, while sharing volumes may differ, sharing behavior seems consistent across the two categories of paying and non-paying members.

5 Modeling Groups with Topics

We have seen a relatively important interest in sharing photos with groups, and also different types of sharing behaviors. In practice, finding groups on Flickr is relatively cumbersome and does not make use of the plethora of metadata available in the user and groups photo collections. We believe topic modeling is a good way of representing groups in a principled, unsupervised manner, and we will show

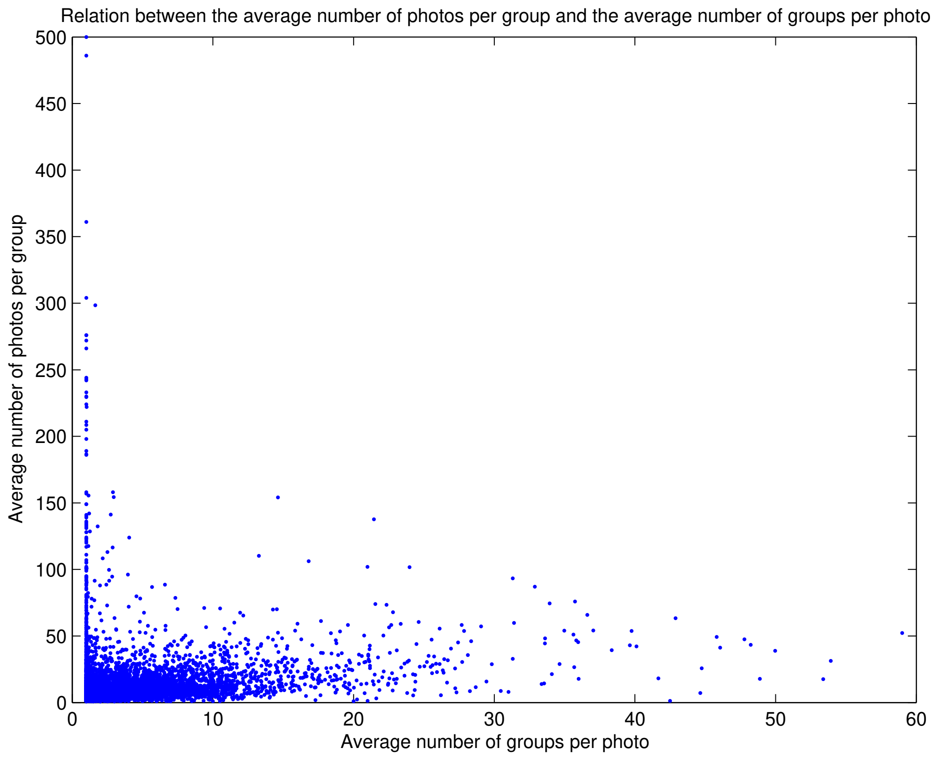


Figure 9: Plot of the average number of groups per photo γ versus the average number of photos per group π.

	Paying (μ, m)	**Non-paying** (μ, m)	**All** (μ, m)
$U_{i,p}$	450.1, 500	220.3, 181	382.2, 500
$U_{i,s}$	127.3, 71	56.75, 25	106.4, 50
$U_{i,g}$	60.07, 23	24.74, 6	49.62, 16
$\frac{U_{i,s}}{U_{i,p}}$	29.4%, 17.2%	30.0%, 17.1%	29.6%, 17.1%
γ	3.3, 1.7	2.5, 1.3	3.1, 1.5
π	9.9, 5.4	8.7, 4.5	9.6, 5.1

Table 1: Statistics for the users who share photos with groups according to their paying status; (μ, m)=(mean, median)

that the topic model we propose is useful (1) to discover groups, and groups of groups, and (2) to do further analysis of the structure of Flickr.

Flickr groups have two main components: their members and the photos shared by the members with the group. Starting from the second component, we propose a representation based on the photos added to the group pool, more specifically on the tags those photos bring into the group.

5.1 Probabilistic Latent Semantic Analysis

We can think of Flickr groups as being a collection of text documents, and the content of these documents are the tags associated with the group photos. In general, an intuitive way to describe a text document is by considering the different topics it is about. These topics are not explicit, but can be derived from the document, and represent an accurate and compact summary of the original content.

PLSA [7] assumes the existence of a latent topic z_k $(k \in 1, ..., N_z)$ in the generative process of each tag t_j $(j \in 1, ..., N_T)$ in a group G_i $(i \in 1, ..., N_G)$. Each occurrence t_j is independent from the document it belongs to given the latent variable z_k, which corresponds to the joint probability expressed by:

$$P(t_j, z_k, G_i) = P(G_i)P(z_k \mid G_i)P(t_j \mid z_k). \tag{1}$$

The joint probability of the observed variables is the marginalization over the N_z latent topics z_k as expressed by:

$$P(t_j, G_i) = P(G_i)\sum_{k}^{N_z} P(z_k \mid G_i)P(t_j \mid z_k). \tag{2}$$

In our model, this is equivalent to the following generative process: a group G is selected, then a hidden topic z_k is sampled from $P(z \mid G)$. Given topic z_k, a tag t_j is selected based on $P(t \mid z_k)$.

5.1.1 Model parameters

The conditional probability distributions $P(t \mid z_k)$ and $P(z \mid G_i)$ are multinomial given that both z and t are discrete random variables. The parameters of these distributions are estimated by the Expectation-Maximization algorithm [7].

5.1.2 Learning

An Expectation-Maximization algorithm can be used to derive from the likelihood of the observed data (Eq.3) the parameters of the distributions $P(t \mid z)$ and $P(z \mid G)$.

$$\mathcal{L} = \prod_{i}^{N_G}\prod_{j}^{N_T} P(G_i)\sum_{k}^{N_z} P(z_k \mid G_i)P(t_j \mid z_k)^{n(G_i,t_j)}, \tag{3}$$

where $n(G_i, t_j)$ is the count of element t_j in document G_i. The two steps of the EM algorithm are the following:
E-step: the conditional probability distribution of the latent topic z_k given the observation pair (G_i, t_j) is computed from the previous estimate of the model parameters.
M-step: The parameters of the multinomial distribution $P(t \mid z)$ and $P(z \mid G)$ are updated with the new expected values $P(z \mid G, t)$.

5.2 A Topic-Based Group Representation

Each group G_i is represented as a bag-of-tags, i.e. a vector $t_i = (t_{i1}, ..., t_{ij}, ..., t_{iN_t})$ of size N_t (the number of distinct tags), where t_{ij} represents the number of times tag j occurs in group G_i. The PLSA model described in Section 5.1 is trained on the bag-of-tags representation of groups.

For this more in-depth study we filtered our dataset in a number of ways. We have concentrated on a vocabulary of the most common 10,236 tags, by removing tags that contained, among others, numeric characters (e.g. dates or years), or that were being used by only one user. Further constraints were imposed on the groups, more specifically, a vocabulary overlap of at least 150 tags (i.e. the group bag of tags should contain at least 150 unique tags from the vocabulary, a mere 1.5% vocabulary overlap). We can summarize this reduced dataset (D_R) as follows: **tags**: $T = \{T_i\}$ with $N_t = |T| =$ 10,236; **users**: $U = \{U_i\}$ with $N_u = |U| =$ 6,144; **groups**: $G = \{G_i\}$ with $N_g = |G| =$ 7,614; **photos**: $P = \{P_i\}$ with $N_p = |P| =$ 766,056.

Table 2 illustrates some of the topics learned by the model by displaying the most probable 18 tags extracted from the distribution $P(t \mid z)$. We also display the groups most likely to generate these specific topics from $P(z \mid G)$. For the experiments discussed in the rest of the section, we have used a number of hidden topics $N_z = 50$, which is relatively small, but meaningful and convenient for illustration and analysis. We experimented with larger numbers of topics, finding similar behavior of the model, but do not discuss the specific results for space reasons. The experiments show that the PLSA model captures truly meaningful information and give us strong reasons to believe that this representation can have useful applications.

6 Uses of Group Representations

As we have seen, the number of groups on Flickr is far from negligible. At the time of writing, the only method to find a group related to a specific theme was to perform a search by keywords against the group names and descriptions, or against the group discussions (i.e. online message boards where group members can exchange messages). Our topic-based representation for groups allows us to analyze groups from the tag content point of view: the decomposition over topics allows us to find "experts" on a topic, or groups of groups centered around the same theme, or combination of themes. Probably the most important advantage is that this representation also allows search by keywords to be performed indirectly on the group content.

6.1 Finding Topic Experts

The topic representation of groups can help automatically discover groups of groups around single topics, in other words, *topic experts*, by ranking $P(z \mid G)$, with no further computation after the model has been learned. We show in Table 2 some of the topics represented by the ten most probable groups, and in Table 3 we show some of the photos present in the group pools for a few of those groups. The topics are also represented by their most relevant tags, ranked in descending order by their $P(t \mid z)$ probabilities. Many of the topics seem to be quite meaningful: for example, topic 1 primarily relates to flower photography, topic 13 relates to the Netherlands, topic 18 relates to live music performances, topic 24 relates to self portrait photography, and so on. The top group names are mostly self evident.

6.2 Searching by Tags Using the Topic Model

As already mentioned, in practice, finding groups on Flickr is quite difficult, as the search by keyword feature uses only the group names and descriptions. While generally group names are descriptive, they may not necessarily use the same keywords as the user searching for them. It is why we believe a topic representation for groups might be a step forward in group discovery. Keyword search could be transformed into a two step process: the keyword could be first used to recover its most probable topics; then, for each of the topics, their most probable groups could be fetched and the user could then browse search results within each topic. This is different from direct tag search, as it would, in principle, be able to also offer disambiguation information for polysemy and synonymy of the search keyword. Let us illustrate the described method with the results for the term "portrait". Figure 10 shows a histogram of the topics' probabilities for the tag "portrait", and the first five topic-expert

Topic 1	
$P(t \mid z)$	**Tag**
0.0766	flower
0.0555	flowers
0.0550	nature
0.0431	ilovenature
0.0323	spring
0.0295	garden
0.0243	green
0.0221	yellow
0.0212	macro
0.0204	pink
0.0168	white
0.0136	plant
0.0126	blue
0.0122	purple
0.0112	red
0.0110	flora
0.0109	canon
0.0095	rose

Topic 1	
$P(z \mid G)$	**Group**
0.9715	1-Plants World
0.9456	Flickr Gardens
0.8783	In my garden
0.8718	My Garden
0.8347	Daffodil World
0.8337	What plant is that?
0.8214	Gardening for Fun
0.8102	Garden Flowers
0.7993	grow
0.7377	Backyard Nature

Topic 2	
$P(t \mid z)$	**Tag**
0.0957	canada
0.0397	bc
0.0343	snow
0.0334	vancouver
0.0240	britishcolumbia
0.0213	ontario
0.0210	winter
0.0129	water
0.0128	mountain
0.0127	ice
0.0111	alberta
0.0102	tree
0.0099	trees
0.0097	mountains
0.0085	colorado
0.0083	sky
0.0071	cold
0.0070	vancouverisland

Topic 2	
$P(z \mid G)$	**Group**
0.9978	BC Peaks & Mountains
0.9971	A S C E N T - (how you get to the top)
0.9937	British Columbia Provincial Parks
0.9922	Climbing Photography
0.9809	Rock Climbing
0.9667	Climbing lifestyle
0.9650	Climbing
0.9632	Where am I in BC
0.9510	ROCKCLIMBING
0.9421	Alpinism

Topic 13	
$P(t \mid z)$	**Tag**
0.0846	holland
0.0613	netherlands
0.0458	nederland
0.0255	thenetherlands
0.0210	amsterdam
0.0182	denhaag
0.0148	bike
0.0141	dutch
0.0136	bw
0.0119	rotterdam
0.0111	people
0.0105	candid
0.0101	bicycle
0.0084	taco
0.0071	love
0.0070	horse
0.0064	canon
0.0063	song

Topic 13	
$P(z \mid G)$	**Group**
0.9599	Den Haag (The Hague)
0.9591	Den Haag / The Hague, The Netherlands
0.8626	goingdutch
0.8202	1-2-3 Nederland
0.7831	Nederland/The Netherlands
0.7679	Made in Holland
0.7671	Dutch
0.7665	horses
0.7639	Dutch skylines
0.7566	Amsterdam today

Topic 18	
$P(t \mid z)$	**Tag**
0.0478	music
0.0175	rock
0.0171	concert
0.0156	live
0.0131	band
0.0127	party
0.0124	florida
0.0123	guitar
0.0104	friends
0.0088	label
0.0086	show
0.0074	livemusic
0.0070	wii
0.0067	framed
0.0065	fun
0.0064	dance
0.0062	miami
0.0058	singer

Topic 18	
$P(z \mid G)$	**Group**
0.9917	**LIVE in CONCERT**
0.9783	Vinyl Junkie
0.9730	BUSH-IT Artist
0.9512	REHNQUIST RETIRES THE WAR BEGINS
0.9386	Rock and Roll : live shows only please
0.9307	Concerts
0.9234	Rock in Paris
0.9171	Live Music Photography
0.9135	SINGERS SING! (4 pics at any one time)
0.9088	Concerts!!

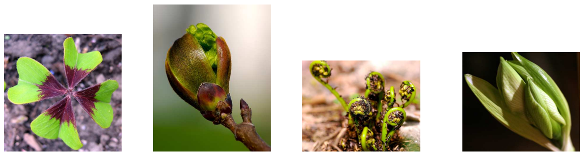
Topic 19	
$P(t \mid z)$	**Tag**
0.0229	handmade
0.0190	craft
0.0167	pink
0.0159	christmas
0.0154	art
0.0149	cute
0.0143	vintage
0.0126	etsy
0.0119	portland
0.0109	blue
0.0105	red
0.0100	paper
0.0096	design
0.0086	green
0.0085	fdsflickrtoys
0.0083	shop
0.0073	collage
0.0068	flower

Topic 19	
$P(z \mid G)$	**Group**
1.0000	Pregadeiras/Pins
1.0000	tezukuri life!
1.0000	Do It Yourselfers
1.0000	MADE for the HOLIDAYS!
0.9999	Crafters HQ
0.9979	crafting world
0.9977	DIY
0.9972	Contemporary Textile Art
0.9967	quilts and quilting
0.9961	The Bag Blog

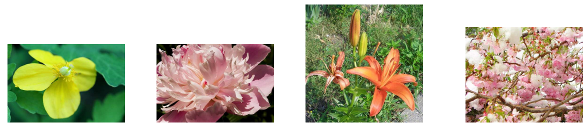
Topic 24	
$P(t \mid z)$	**Tag**
0.0598	me
0.0581	selfportrait
0.0365	portrait
0.0310	woman
0.0262	self
0.0217	face
0.0152	girl
0.0129	eyes
0.0100	female
0.0097	hair
0.0093	light
0.0090	red
0.0082	myself
0.0080	bw
0.0077	lips
0.0066	hand
0.0066	blue
0.0061	skin

Topic 24	
$P(z \mid G)$	**Group**
0.9921	...and god created woman
0.9884	be bad not sad!
0.9706	Beautiful Blue Eyes
0.9664	.Cropped Faces [
0.9210	arm,leg,finger,shoulder...
0.9144	zelfportretten / selfportraits / auto-portraits
0.9142	Blondes Have More Fun (or so they say)
0.9140	Everyday men models...WOOF!
0.9054	My Self Portrait
0.8899	Lighting & Posing Styles

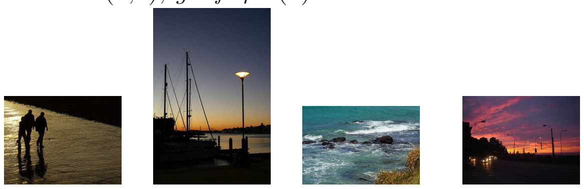
Table 2: Some of the topics in the PLSA model, characterized by their most probable tags (ranked by $P(t \mid z)$), and by their most probable groups (ranked by $P(z \mid G)$).



photos from group *grow*, by *docman*(1), *Ben McLeod* (2,3), *gailf548* (4)



photos from group *Flickr Gardens*, by *Lorika13*, *egg.*, *annethelibrarian*, *Somerslea*



photos from group *Beaches & Sunset*, by *Mallmus*, *marj k*, *The Life of Bryan*, *cakecosas*

photos from group *Sea*, by *Martin Burns*, *mnadi*, *carf*, *Ennor*

Table 3: Example photos from group pools, that are highly probable for topics 1 (top row) and 12 (bottom row).

groups for the top four most probable topics. One can see that different meaningful concepts related to portraits can now be recommended, and that NONE of the group titles contain the word "portrait". Figure 11 shows the search results for the tag "tiger". At a first glance, it would appear that the search for this specific tag does not return relevant results, however, on further observation of the group contents, we have found that topic 23 and its most probable groups are related to the *Tiger F-5E* fighter plane, topic 38 and its groups to various types of toys representing tigers, topic 20 and its groups to pet cats with tiger stripes, and finally, topic 29 and its groups are related to the real feline.

Although this type of search is quite simplistic, it already shows the potential of a topic-based representation of group content. We intend to explore this direction further in future work, and also explore visual content and text-visual joint modeling.

6.3 Further Group Analysis Based on Topic Decomposition

The topic-based representation can also be useful for gaining further insight into the structure of Flickr groups. We have already seen that the model offers a straightforward way of finding topic experts. By visual inspection of their topic distribution $P(z \mid G)$, such topic experts are mainly about one subject, and their probability to generate the given topic is rather high (see Table 2). However, there are groups for which the distribution over topics is slightly more varied, indicating group interest shared over several themes. Therefore, we can ask ourselves what the group topic "homogeneity" looks like in our dataset. Figure 12 shows the histogram of the number of *relevant* topics per group, where relevant is a term that describes the minimum number of topics (ranked by their probability mass) that account for an amount ϵ of the total probability mass. Obviously the choice of ϵ and N_z will modify the number of relevant topics per group. In Figure 12 we use ϵ=0.8, so all topics with mass at least equal to 0.2 are always considered. Most of the groups have topic distributions that indicate that several themes are present in their representation: 60.5% have between one and five relevant topics, and just under 10% have more than ten relevant topics; 24.9% of the groups are about just one or two topics.

Based on the above analysis, it might therefore be interesting to extend our previous search scenario in such a way as to retrieve those groups which have a similar distribution over topics as that of the search keyword. We will use the tag "portrait" from our previous search in order to exemplify this search alternative. We show in Figure 13 the topic distribution of a group called *Portrait*, along with the four most relevant topics according to the above definition. It would be reasonable, and desirable,

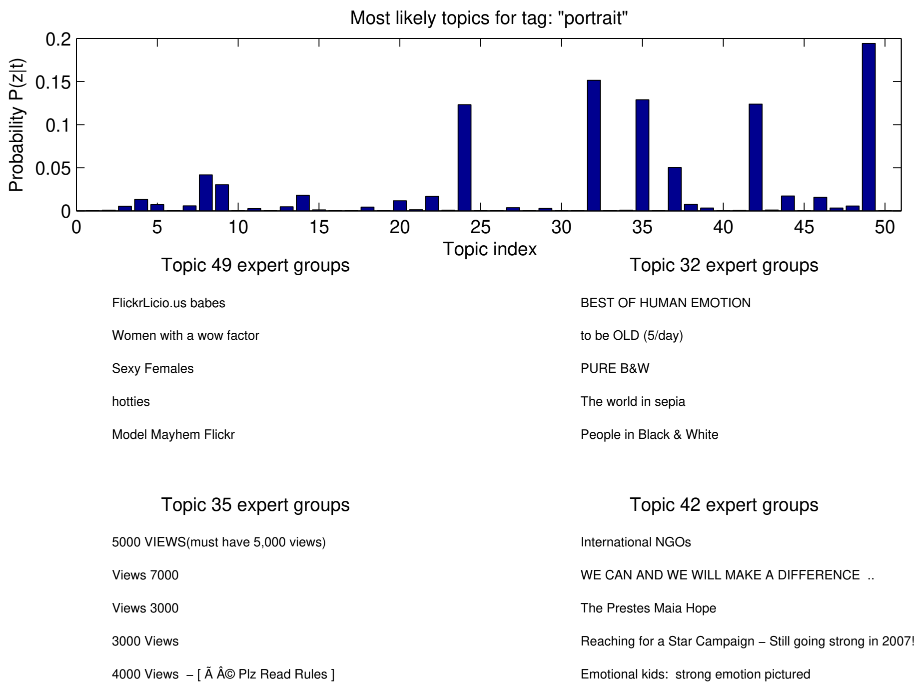


Figure 10: The search results for tag “portrait”: the probabilities of the topics given the tag, and the top 5 expert groups for the most probable 4 topics.

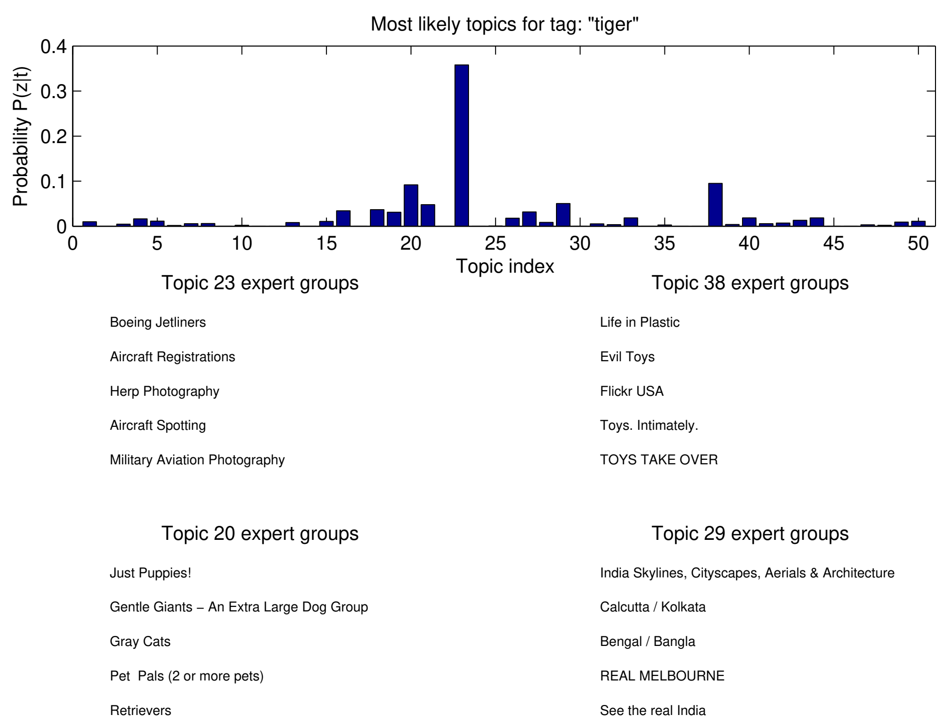


Figure 11: The search results for tag “tiger”: the probabilities of the topics given the tag, and the top 5 expert groups for the most probable 4 topics.

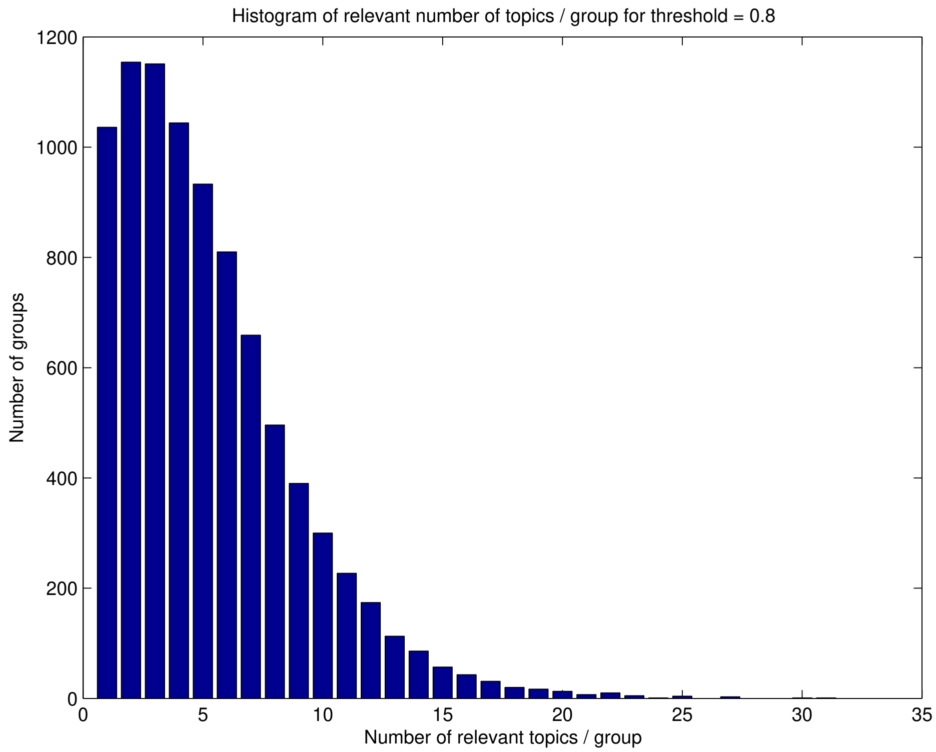


Figure 12: The histogram of the number of relevant topics per group, with threshold $\epsilon = 0.8$.

to expect this group to show as a relevant result for our search term. We can see that the topic distribution is quite similar to that of the tag “portrait” (see Figure 10). Additionally we can note that the four topics have some common ground in the form of the photographic style (portraiture), but differ in granularity, with *female portraits* for topic 49, *self portraits* for topic 24, *child portraits* for topic 42, and *black and white portraits* for topic 32.

In Figure 14 we present for contrast the topic decomposition of one of the “mammoth” groups on Flickr, *FlickrCentral*. This group has roughly 59,546 members and 1,240,939 photos. It can be seen that the distribution over topics shows the nature of the group: a *catch-all* group. While two of the relevant topics for the “portrait” tag are relevant for *FlickrCentral* too, it would be reasonable to rank *FlickrCentral* lower than *Portrait* in the search results, as its range of interests is much wider, and this is very well captured by the difference in the distribution over topics.

7 Conclusions

In this paper, we have analyzed the structure of Flickr groups, highlighting fundamental patterns of photo-to-group sharing practices, with respect to the degree of active participation in groups, group affiliation, group loyalty, and photo repurposing. Our work revealed that a large percentage of users engage in sharing with groups, and that they do so significantly. While the volume of shared photos varies quite a lot, the sharing percentage is on average quite important, with a mean of 30% and a median of 17%. Sharing users can further be categorized as moderate sharers (less than 5 groups), average sharers, and extreme sharers (hundreds of groups). On average, group loyalty is quite low, with users sharing about 9.6 photos in the same group (with median of 5.1). On the other hand, photo repurposing seems to also be surprisingly low, considering the large number of groups, many of which are about the same subjects, with the same photo being shared on average with 3.1 groups (median 1.5). These results leave some open questions, such as whether a correlation between group

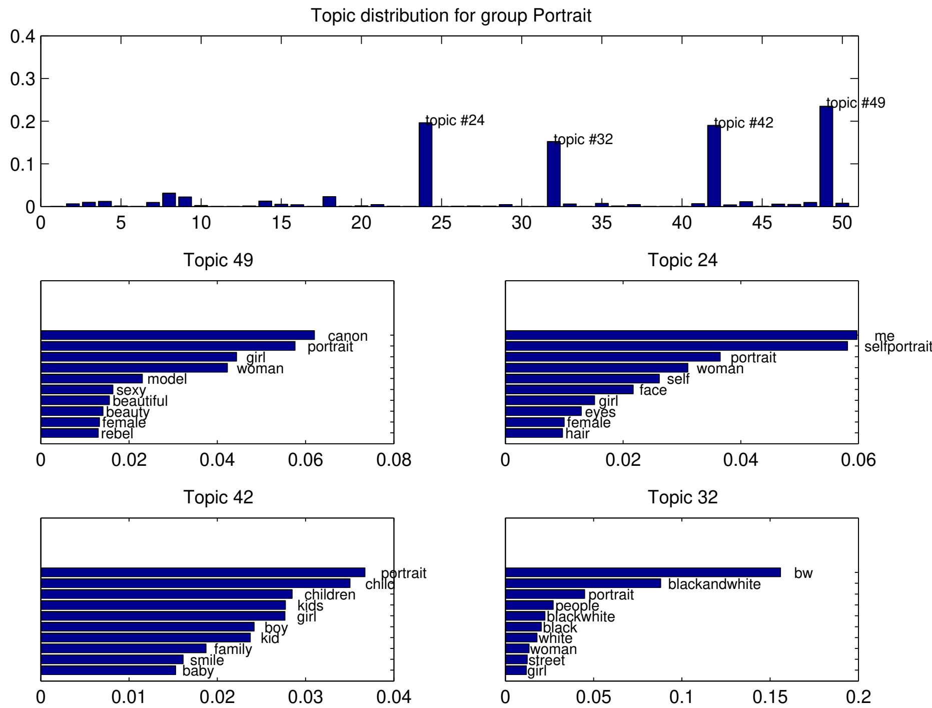


Figure 13: The decomposition over topics of the *Portrait* group.

loyalty and photo repurposing exists at the user-level.

As a second contribution, we have proposed a novel method for group representation, based on latent topics learned via unsupervised probabilistic analysis of group tags. We have shown that this topic-based representation is useful to automatically find topic expert-groups and groups of groups, to facilitate new group search methods, and to obtain further insights into the structure of Flickr groups. We believe the topic-based representation is promising, and brings forward a few questions. One important open issue is to devise a principled way of determining the value of the number of topics. Another research direction is to make use of the topic-based representation for further analysis of group homogeneity, by employing different clustering techniques. Also, a topic-based representation for users might prove very useful in helping devise a similarity measure between users and groups, which would allow implicit group affiliation or group recommendation systems to be implemented. Finally, we intend to investigate models that take into account not only the metadata content, as in this study, but also the visual content, which is widely available. We feel confident that such models shall increase an automated system's ability to help the user annotate his visual content with metadata content. This seems like a next logical step in our attempt to better manage and discover information. These ideas are the goals of our future work.

Acknowledgments

This research has been supported by the Swiss National Science Foundation through the MULTI project.

References

[1] Flickr Blog, Jan. 2008. http://flickr.com/blog.

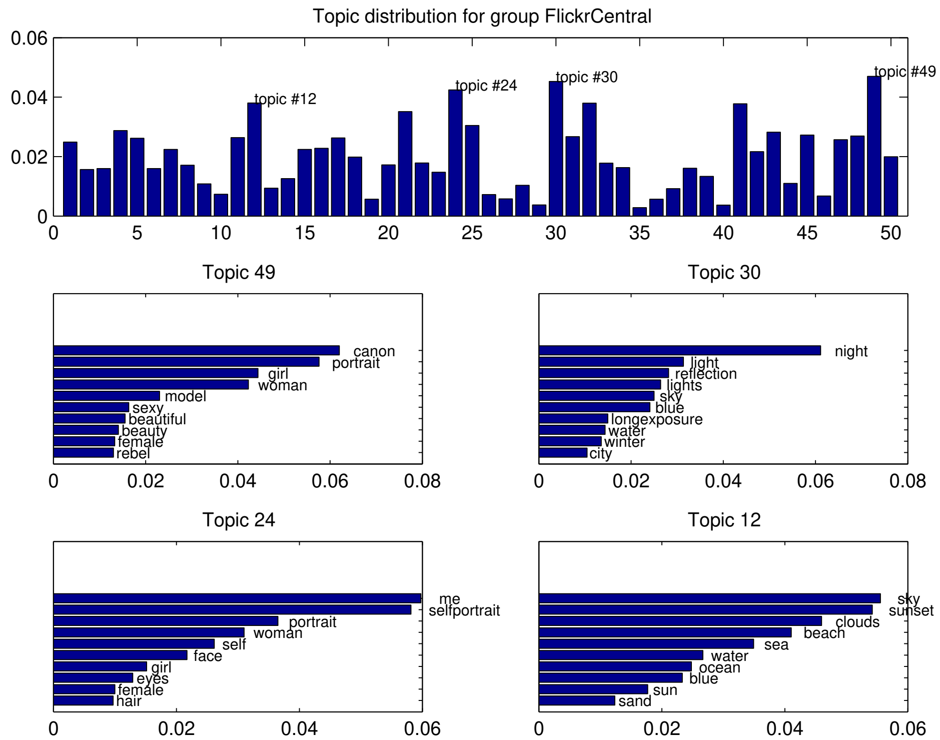


Figure 14: The decomposition over topics of the *FlickrCentral* group.

[2] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007.

[3] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, 2007.

[4] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007.

[5] T. L. Berg and D. Forsyth. Automatic Ranking of Iconic Images. Technical report, U.C.Berkeley, 2007.

[6] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, 2006.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[8] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries for large collections of geo-referenced photographs. In *Proceedings of the Fifteenth International World-Wide Web*, Edinburgh, Scotland, 2006.

[9] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. *Proc. ACM Multimedia*, 2007.

[10] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[11] K. Lerman and L. Jones. Social Browsing on Flickr, Dec 2006.

[12] K. Lerman, A. Plangprasopchok, and C. Wong. Personalizing Image Search Results on Flickr, Apr 2007.

[13] R. Lienhart and M. Slaney. PLSA on Large Scale Image Databases. In *Proceedings of the 2007 International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii*, 2007.

[14] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, 2006.

[15] A. D. Miller and W. K. Edwards. Give and take: a study of consumer photo-sharing culture and practice. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[17] J. P. Pickett, editor. *The American Heritage Dictionary of the English Language*. Houghton Mifflin, January 2000.

[18] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR'07*, 2007.

[19] P. Schmitz. Inducing Ontology from Flickr Tags. In *WWW 2006*. IW3C2, 2006.

[20] P. Schmitz. Leveraging community annotations for image adaptation to small presentation formats. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, 2006.

[21] N. A. Van House. Flickr and public image-sharing: distant closeness and photo exhibition. In *CHI '07 extended abstracts on Human factors in computing systems*, 2007.

[22] R. van Zwol. Flickr: Who is Looking. In *ACM International Conference on Web Intelligence (WI 2007)*, 2007.