Modeling Flickr Communities Through Probabilistic Topic-Based Analysis

Radu-Andrei Negoescu and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—With the increased presence of digital imaging devices, there also came an explosion in the amount of multimedia content available online. Users have transformed from passive consumers of media into content creators and have started organizing themselves in and around online communities. Flickr has more than 30 million users and over 3 billion photos, and many of them are tagged and public. One very important aspect in Flickr is the ability of users to organize in self-managed communities called groups. This paper examines an unexplored problem, which is jointly analyzing Flickr groups and users. We show that although users and groups are conceptually different, in practice they can be represented in a similar way via a bag-of-tags derived from their photos, which is amenable for probabilistic topic modeling. We then propose a probabilistic topic model representation learned in an unsupervised manner that allows the discovery of similar users and groups beyond direct tag-based strategies, and we demonstrate that higher-level information such as topics of interest are a viable alternative. On a dataset containing users of 10 000 Flickr groups and over 1 milion photos, we show how this common topic-based representation allows for a novel analysis of the groups-users Flickr ecosystem, which results into new insights about the structure of the entities in this social media source. We demonstrate novel practical applications of our topic-based representation, such as similarity-based exploration of entities, or single and multi-topic tag-based search, which address current limitations in the ways Flickr is used today.**

***Index Terms*—Flickr, probabilistic topic models, social media.**

I. Introduction

RECENTLY, the George Eastman House Museum has released online, under a Creative Commons license, a few hundred photos from their collection of 1400 glass plate negatives. They are not the only institution to have enriched today's digital landscape: the Library of Congress, the Smithsonian Institution, and the Powerhouse Museum are just a few others. However, the great majority of today's digital photos found online come from regular people. The *William M. Vander Weyde* photoset from the George Eastman House carries a text description that seems poised to prove that history is repeating itself:

> "In the 1890s faster films, better lenses, hand cameras and the availability of commercial developing and printing services not only made it much easier to make photographs, but to make photographs that captured a wider range of events of everyday life. This fueled a huge explosion in photographic practice; first by significantly expanding the number of amateur photographers and then by irrevocably altering and expanding the nature and practices of professional photography. A greatly expanded world of images—very different in concept and in form—suddenly became an inextricable part of the visual world."[3]

By reading the above quote and replacing 1890s with 2000s, and then redefining huge to mean "in the order of billions", one can characterize the current state of the world of digital images available online. Digital cameras got smaller, faster, more reliable—they became a commodity. Mobile phones have become more and more powerful as well in terms of photo-taking capabilities. This is really the next level of the 1890s revolution of photography, where everybody is a photographer. In addition, because of the new ways in which people interact in this digital era, most of the photos end up being available online.

Flickr, an online photo management and sharing website (http://www.flickr.com), had reached on November 13, 2007 [1] the second billionth photo milestone after less than four years of existence and the third billionth photo was uploaded on November 3, 2008 [2], just one year later. This gives a staggering average of 2000 photos uploaded per minute throughout the year. Although Flickr's userbase size is not publicly available, a rough estimate of over 30 million usernames could be obtained in November 2008.

This dramatic increase in the amount of multimedia resources available to people brings a need to organize, retrieve, and discover relevant or interesting bits. It is this need that is the driving force behind innovative systems that help people organize their (and others') content, be it news items, websites, blogs, videos, photos, or scientific literature. This is why community-based filtering, recommendation systems, or social aggregators have come to life in the recent past and continue to improve and develop into new forms, shaped by the power of the communities using them.

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 multimedia research.

Manuscript received September 03, 2009; revised December 11, 2009; accepted February 22, 2010. Date of publication May 18, 2010; date of current version July 16, 2010. This work was supported by the Swiss National Science Foundation through the MULTI project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qibin Sun.

The authors are with the Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2050649

1520-9210/$26.00 © 2010 IEEE

Why do people share photos with other people? Why and how do they use tags to describe those photos? What is the impact of online interaction with other people on literacy and learning in general? These are just some of the questions asked by researchers in the near past [12], [19], [24], [32], [37].

Flickr's structure has also been analyzed based on social connectivity information, i.e., who is a contact of whom, in the traditional social network setup [20]. However, to our knowledge, little attention has been paid so far to one of the flagship social connection features, namely *Flickr Groups*. Groups 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 about specific topics or themes (e.g., an event or a photographic style). Aggregating content and metadata for groups could thus offer insights into large-scale behavioral trends (e.g., photo sharing practices) and also provide robust representations (e.g., at the topic level) that 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 some of the issues addressed by our work.

Concretely, in this paper we make the following contributions.

- We explore a novel problem, namely jointly analyzing groups and users in Flickr from aggregated image and tag usage. Although users and groups are fundamental components of Flickr, their interrelations are, to our knowledge, not completely understood or fully exploited. The key concept is the assumption that groups and users in Flickr can be reasonably regarded as if they were equivalent entities and that their direct joint modeling is beneficial despite the complex ways in which Flickr groups are created through users' contributions.
- We propose a principled topic-based representation for users and groups in Flickr, starting from a bag-of-tags representation for both types of entities. Our representation, using a probabilistic topic model, allows to discover similar users or groups beyond direct tag-based strategies in a fully unsupervised manner, and demonstrates that higher-level information such as topics of interest are an attractive alternative. Remarkably, although we make use of the users' and groups' photos, we need not understand their visual content, as the associated metadata offers a rich (and more semantic) source of information about the description of users and groups.
- We perform a novel analysis of the groups-users Flickr ecosystem based on the newly proposed representation, which results into new insights about the structure of groups and users in this social media source.
- We demonstrate novel applications of our topic-based representation, such as similarity-based exploration, and single and multi-topic tag-based search, which address current limitations in the way Flickr is used today.

In the following section, we present a review of the related work. We will then have a closer look at the Flickr "ecosystem" in Section III, and in Section IV we will introduce the probabilistic model on which our topic-based representation of Flickr entities is founded. Once the model is defined, we present a new analysis of Flickr entities based on this model in Section V and applications of the topic-based representation in Section VI.

II. Related Work

Flickr provides access to nearly all their public data through an API. This has spurred an increasingly large body of third-party applications, but also research that makes use of this data in a number of different areas. On one end, ethnographic studies have been conducted on Flickr users, looking at the factors behind sharing, tagging, and managing privacy [4], [6], [12], [25], [30], [37]. Furthermore, social links amongst users using Flickr have also been explored, either from a purely computational perspective of the social graph [20], or by using the social network information as additional input to different content filtering algorithms [21], [22]. On the other end, there is research aimed specifically at analyzing content on Flickr, either in the context of traditional image retrieval, or directed towards new uses for aggregated data, such as better organizing and exploring huge repositories, or detecting events and places automatically [5], [14], [17], [19], [32].

Several studies used Flickr data in order to better understand users and the ways they use Flickr as a whole. Tagging systems have been analyzed by Marlow *et al.* [24] and a taxonomy of users' motivations to tag has been proposed by Ames and Naaman in [6]. Their studies point out that multiple motivations come into play when users tag photos, with a particularly important role played by social motivations. Nov *et al.* [30] took this research a step further, showing through a quantitative study that indeed tagging behavior is positively correlated with social presence indicators such as group memberships and number of contacts a user has on Flickr. There have also been some studies analyzing the sharing practices, motivations to share, and privacy concerns of the users [4], [25], [37]. In particular, Van House [37] discusses the main uses of photo sharing on Flickr and finds evidence that the main motivations are social in nature. The most important are maintaining relationships, self-representation, and self-expression. Miller and Edwards [25] found two distinct categories of photo-sharing practices of a sample Flickr population. They called them *Kodak Culture* and *Snaprs*. The main difference seems to be that the *Kodak Culture* users adopt sharing practices similar to those found in colocated scenarios around a physical photo album, where the story is not found in the photo itself but it is told, while *Snaprs* are users who adopt completely new approaches to photo taking, story telling, and online interaction. While these studies provide particularly useful insights into user behavior, especially at the level of photo sharing and metadata creation, none of them explicitly addresses sharing practices in relation to Flickr groups, as we do here. They do prove, however, that some of the most important incentives for users to participate in such communities are social in nature.

Metadata from the (social) links existing on Flickr [20]–[22], [38] has also been used as data source. Recent work includes studying user-to-user relations by means of contact bookmarking, a direction explored by Kumar *et al.* [20], with interesting results regarding the structure of Flickr's social network. Other works have considered user-to-photo relations by means of ownership, favorites, or comments. Van Zwol [38] 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 photo upload, concluding that both 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 [22] found that the number of views a photo receives correlates strongly with the size of the social network of a user and more particularly with the number of reverse contacts, i.e., the number of people who have bookmarked the user. In a different work, Lerman *et al.* [21] use a user's existing social network and a topic model learned on tags in order to filter tag search results for that specific user. The motivation and specific use of their topic model are, however, fundamentally different from ours. In their work, the focus is on improving precision and recall measures for image retrieval based on the user interests. User interests are extracted from previously used tags and the model is learned on tags collected from the first 4500 images retrieved from single-tag searches for *tiger*, *newborn*, and *beetle*. In contrast, our model is learned on a dataset-wide vocabulary of tags and is then used to represent not only users' interests, but also those of the groups.

Flickr data have also been used in the context of content-based image retrieval research [23]. 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. This comes in the form of tags, notes (areas defined on photos, with associated comments), number of views, comments, number of people who mark the photo as a favorite, and geographical location data. Recent studies have used notes [35], combinations of tags, geolocation, and visual data in order to improve retrieval [7], [31], to visualize and to summarize large datasets either over time or over a geographic area [5], [14], [17], to automatically extract place and event semantics [19], [32], or to induce tag ontologies [34]. All these studies show the potential that large scale data aggregation has for the better understanding of the communities generating it.

There have also been recent works that try to exploit the visual information that can be extracted from photos themselves, such as [41] and [11]. In [41], Wu *et al.* learn a visual model for each of the 1000 words (concepts) in their dataset, as their goal is to compute word-to-word distances. Although the authors report significant improvements over textual features alone, it is not clear if in our scenario this would hold true, as our vocabulary is an order of magnitude larger (10 000 words) and we are dealing with entity-to-entity distances and not word-to-word. Crandall *et al.* [11] use visual features in order to determine the location of a set of photos taken at roughly the same geographical location. The authors find that visual features are roughly as effective as text features, when the photos are taken at roughly the same location (100 m accuracy), but not so much so when the geographical scale is larger (100 km). Although encouraging, these results also show that, unless certain conditions are met, visual features can even increase the uncertainty when used in conjunction with the textual ones.

In summary, compared to our work, previous research has either directly exploited social link information, used different content facets, or targeted different goals. At the same time, some of the findings in [22], [37], and [38] provide a good motivation to investigate ways of representing Flickr inhabitants and their online communities in such a way as to allow not only to better understand them, but also to develop new methods for community discovery and integration. We present here a probabilistic approach that simultaneously models users' and groups' interests based on the textual tags used to describe the photos belonging to them. To our knowledge, this is the first attempt to model users and groups as equivalent entities. The user-group link information is not explicitly used, but it is implicitly taken into account when building groups' bags-of-tags. The main goal of this representation is to allow a simple yet direct comparison between Flickr entities, be them groups or users. This comparison results in ways to facilitate user and group discovery based on meaningful, content-based information, rather than by simply relying on random or social-based exploration, as the current system at Flickr or previous work propose. To this extent, we present several simple applications of our topic-based representation that show the advantages of a common representation of users and groups. A preliminary version of this work appeared in [28].

III. Flickr Ecosystem: A Statistical View

Flickr was created in early 2004 and has quickly become one of the most important photo sharing websites. In less than five years, Flickr has reached the figure of 3 billion photos uploaded on its servers as of November 3, 2008 [2]. The primary functionality of Flickr is that users can upload photos to their online accounts. They can also tag each photo with up to 75 unique tags. The photos can be displayed publicly (the default option), or access to them can be restricted to a closed social circle. A secondary functionality of Flickr is that users can join different interest groups. These groups are self-managed communities whose main purpose is to facilitate sharing of user photos in what is called the *group photo pools*. These pools are therefore collections of photos shared by any member with the group and implicitly, all the tags associated with the photos also become part of the group photo pool.

Flickr has two types of members: at the time of writing, non-paying members had a monthly upload limit of 100 MB, they could share any given photo with at most ten groups and could publicly display only their most recent 200 photos. Paying members, on the other hand, had no monthly bandwidth limit, no restrictions on the number of photos shown on their photostreams, and could share any given photo with up to 60 groups.

The dataset used in this study has been collected during the spring of 2007 by using the Flickr API. All the information we extracted about a particular user was publicly available; thus, real statistics linked to the number of photos may be different if users employed restrictive privacy settings for their photos. No private information was available for this study. The data collection process can be described as follows: we repeatedly retrieved the first approximately 4000 photos uploaded from a randomly sampled moment t in the interval December 22, 2004–April 2, 2007, until information on roughly 187 000 different photos has been collected. We have thus obtained 22 414 distinct users (the photo owners). For each of these users, we then retrieved their most recent 500 photos, which in some cases meant all of their photos, for a total of nearly 7 million photos. Only about 4.7 million photos have at least one tag and this resulted in roughly 23 million tag occurrences and almost 2 million distinct tags. In addition to the users, photos, and tags, we have also collected information about the groups the photos were shared in, with 1.13 million photos being shared in at least one group. To summarize, this original dataset (D_O) has the following characteristics:

- users: $U = \{U_i | i = 1 \ldots N_U\}$ with $N_U = |U| = 22\,414$;
- groups: $G = \{G_i | i = 1 \ldots N_G\}$ with $N_G = |G| = 65\,474$;
- photos: $P = \{P_i | i = 1 \ldots N_P\}$ with $N_P = |P| = 6\,926\,622$;
- distinct tags: $T = \{T_i | i = 1 \ldots N_T\}$ with $N_T = |T| = 1\,969\,813$.

We observed interesting statistical trends, summarized in the following subsections.

Photo Sharing Through Groups: In our dataset, the two types of users (paying and nonpaying) are almost equally represented (51.4% and 48.6%, respectively). We show in Fig. 1 the relation between the size of the users' photo collections (in number of photos) and the fraction of photos shared in groups. As a first observation, 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). Furthermore, the sharing fractions 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 weak correlation. While the restrictions on free accounts do seem to influence the number of photos users have in their accounts (with an average of around 220 photos for nonpaying members as opposed to 450 for paying members) and also the number of groups they share photos with (on average 60 for paying members, with a median of 23 and an average of 24.7 with a median of 7 for nonpaying members), we found that the ratio of photos shared in groups is similar for both categories of users: paying members in our data share on average 29.4% of their photos (median 17.2%) and nonpaying members share on average 30% (median 17.1%). We have also analyzed in previous work measures of group loyalty (how many photos a user shares with the same group) and photo repurposing (how many groups a photo is shared with) for users who participate in Flickr groups [27]. Our results on the same data showed that, on average, a user shares a small number of photos with each group (mean 9.6, median 5.1) and will share the same photo in multiple groups in even smaller numbers (mean 3.1, median 1.5), with small differences between paying and nonpaying members, despite the large differences in the average number of groups noted above. This is an interesting result, showing that users' group-sharing behavior is not influenced by their paying or nonpaying status, or by the amount of photos they upload.

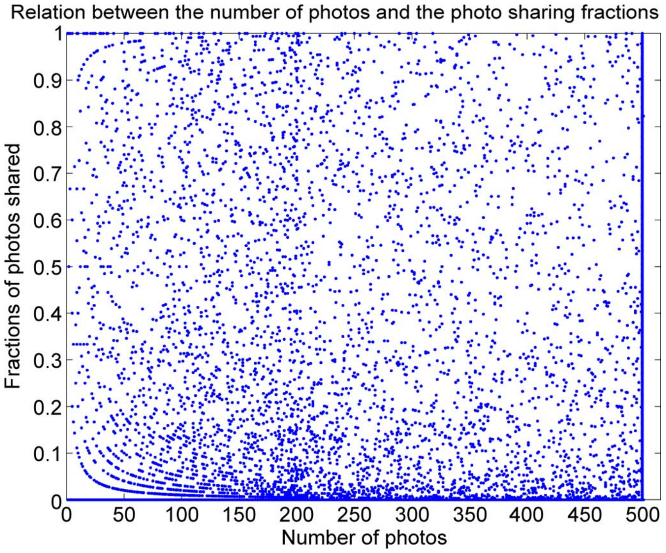


Fig. 1. Fraction of shared photos (y-axis) versus the number of photos of each user (the x-axis): the size of the collection of photos for users who do not share any photos at all is evenly spread over the entire range of sizes [1, 500]; the sharing fractions for users who have the maximum number of photos (500) is evenly spread over the full interval [0, 1].

Overall, the analysis shows that through relatively modest photo repurposing, small but persistent group loyalty, and active participation in groups, Flickr users contribute a significant proportion of their content to communities, which emerge as rich Flickr entities through the aggregation of their members' contributions.

Photo Annotation Through Tags: When thinking about how groups' photo collections are explicitly formed—they are basically aggregations of photos—and how groups' tag pools are implicitly formed from those photos in the group photo pool, it could be hypothesized that groups' tags statistics might be radically different from those of the users'.

For this analysis and subsequent experiments, we filtered the original dataset in a number of ways. We concentrated on a smaller vocabulary of the most common 10 236 tags in our data, obtained by removing tags that contained among others numeric and non-Latin characters, or that were used by less than 100 users. This effectively eliminated the heavy tail of the tag distribution, including among others, dates (*20060401*, *summer2007*), compound tags generally contextual, that only appear once (e.g., *explore22aug2006*, *sustainabilityandsangria*, *jimmyshands*), typos (e.g. *commedians*), and languages other than English that use non-Latin characters (e.g., Arabic, Chinese, or Japanese). An additional constraint was imposed on the groups and users, in order to focus our analysis on groups and users that have a minimum amount of representation in the 10 K vocabulary. More specifically, we kept those entities that have a vocabulary overlap of at least 125 tags (i.e., the group or user vocabulary should contain at least 125 unique tags from the 10 K vocabulary, a mere 1.2% vocabulary overlap). Finally, only users who shared photos with at least one group and groups for which we had at least one member were kept. We can summarize this reduced dataset D_R in Table I.

TABLE I
REDUCED DATASET D_R

distinct tags	$T = \{T_i\}$ with $N_t = 10,236$
users	$U = \{U_i\}$ with $N_u = 8,061$
groups	$G = \{G_i\}$ with $N_g = 10,838$
photos	$P = \{P_i\}$ with $N_p = 1,016,199$

While these filters may seem overreaching, they are likely to insure a more coherent corpus from a semantic point of view. The dataset is still quite large, with almost 20 K entities and a total number of photo-tag-group occurrences of roughly 38 million.

In Fig. 2, we display four histograms, depicting the total number of tag occurrences and the total number of unique tags for groups and users, respectively. We can observe that groups tend to have smaller numbers of overall tag occurrences (on average 3286, with median 972) and just about 100 groups having more than 40 000 tag instances. On the other hand, users tend to have slightly larger tag numbers (a mean of 6414 and a median of 3035), including 150 users with more than 40 000 tag instances. This effect is likely correlated with the fact that the groups' tag pools are only considering tags from the users in our dataset. However, when looking at the number of unique tags, the histograms show a similar distribution. The users' mean vocabulary size is 494, with a median value of 350 unique tags, while the groups' mean vocabulary size is 555, with a median value of 296. One noteworthy aspect, otherwise quite intuitive, is that no user in our dataset has more than 5000 unique tags, while on the other hand, there are a number of groups (43) with tag vocabularies of 5000 to 10 000 tags. One relatively simple way of comparing these two distributions is to compute the Bhattacharyya distance between the histograms of the users' and groups' vocabularies. When binned in 2500 bins, the Bhattacharyya distance is 0.2662 and 0.1501 when binned in 250 bins. This distance measure is bounded by the interval $[0\ldots1]$, and the smaller the distance, the more similar the two distributions are. So although groups' tags collections are constructed from aggregating partial user tag collections, they remain comparable to those of the users in terms of unique tags. We can see this more clearly in Fig. 3, where we show the cumulative sums for tag occurrences and unique tags for both types of entities. The dashed-blue and continuous-red curves show the cumulative sums of tag occurrences for groups and users, respectively. We observe that 66.2% of the users have less than 5000 tag occurrences. The percentage of groups with less than 5000 tag occurrences is much higher, at about 87.3%. On the other hand, the dash-dotted-blue and dotted-red curves represent the number of unique tags for groups and users, respectively, and present a much more similar shape. Overall, users seem to have slightly smaller vocabularies than groups.

These figures support our earlier observations that, although users contribute only a part of their collections to groups, these aggregated contributions create comparable tag vocabularies for groups. This also supports our hypothesis that groups and users may be treated as reasonably comparable entities from a content-based point of view.

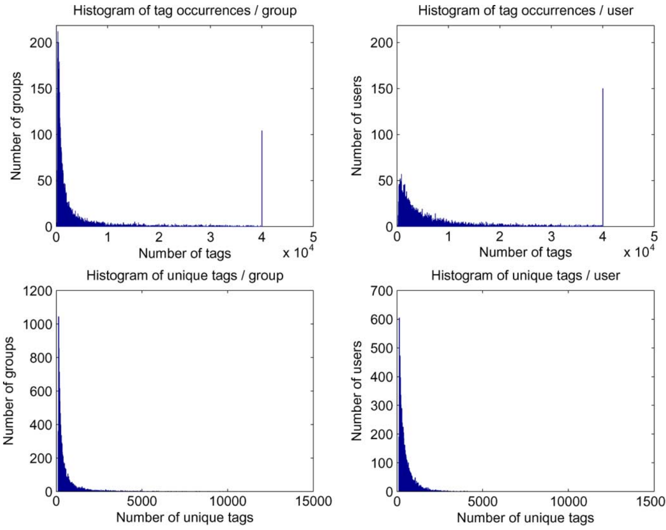


Fig. 2. Top half: histograms of the total number of tag occurrences per group and per user; bottom half: histograms of the number of unique tags per group and per user.

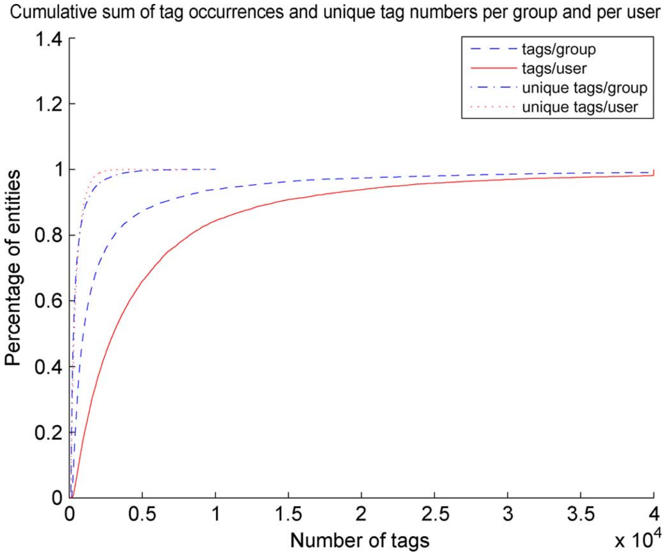


Fig. 3. Cumulative sums of the total number of tag occurrences and unique tags for groups and users, respectively; 66.2% of the users and 87.3% of the groups have less than 5000 tag occurrences, but in terms of unique tags the two types of entities are very similar.

IV. Probabilistic Topic Model for Flickr Users and Groups

A. Flickr Entities and Topics

One can think of groups and users in Flickr primarily as photo collections. From this point of view, they are indeed equivalent entities because, as we have previously shown, they all have a collection of photos with associated tags, and furthermore their vocabularies are quite similar in terms of size. If we consider the full collection of tags for a given entity, we can think of it as a text document, where the words describing the document are the tags associated with that entity's photos, in no particular order.

An intuitive way to describe a text document is by considering the different topics it talks about. These topics are not always explicit but can be derived from the document and represent an accurate and compact summary of the original content. Several probabilistic models have been proposed for the extraction of *latent topics* in the context of text corpora [9], [15]. One such model is probabilistic latent semantic analysis (PLSA), which was introduced by Hofmann [15], as a probabilistic extension of latent semantic analysis (LSA) [13]. This model assumes the existence of a latent topic variable in the generative process of each word in a document.

In our work, we represent each entity E_i as a bag-of-tags, i.e., a vector $\mathbf{t} = (t_{i1}, \ldots, t_{ij}, \ldots, t_{iN_T})$ of size N_T (the number of distinct tags in the corpus). Here t_{ij} is the shortcut notation for $n(E_i, t_j)$ and represents the number of times tag j occurs in entity E_i's bag-of-tags. It is worth noting that in our scenario the entities are natural bags-of-tags, as there is no predefined order for the tags in an entity's pool of tags. The PLSA model described below is trained on the bag-of-tags representations of groups and users regardless of their type.

Let z_k represent the latent topics, with $k \in 1, \ldots, N_z$ and N_z representing the *a priori* fixed number of topics for a corpus of documents. The tags, denoted by t_j, with $j \in 1, \ldots, N_T$, make up the words vocabulary, with N_T denoting the total number of distinct words in the corpus. Finally, documents, denoted by E_i, with $i \in 1, \ldots, N_E$, are made up of words from this vocabulary and N_E denotes the total number of documents in the corpus. Introducing the latent topics effectively breaks the conditional dependence of the words and the documents, that is to say, each occurrence of a word t_j is conditionally independent from the document E_i it belongs to, but it is on the other hand dependent on the topics the document is about, the latent variables z_k.

Formally, this corresponds to the joint probability

$$P(t_j, z_k, E_i) = P(E_i) P(z_k|E_i) P(t_j|z_k). \tag{1}$$

The joint probability of the observed variables (words and documents) is the marginalization over all the N_z latent topics

$$P(t_j, E_i) = P(E_i) \sum_{k=1}^{N_z} P(z_k|E_i) P(t_j|z_k). \tag{2}$$

This is equivalent in our model to the following generative process: an entity E_i is selected with probability $P(E_i)$, then a hidden topic z_k is sampled from the conditional probability distribution $P(z|E_i)$. Given topic z_k, a tag t_j is selected based on the conditional probability distribution $P(t|z_k)$.

The conditional probability distributions $P(t|z_k)$ and $P(z|E_i)$ are multinomial, given that both z and t are discrete random variables. For an entity collection with vocabulary of size N_T, a N_T-by-N_z matrix stores the parameters of the multinomial distributions $P(t|z_k)$. We denote this matrix by $P(t|z)$. Likewise, we denote by $P(z|E)$ the matrix storing the parameters of the multinomial distributions $P(z|E_i)$ that describe the training documents.

The parameters of these multinomial distributions are estimated by the expectation-maximization (EM) algorithm [15], derived from the likelihood of the observed training data

$$\mathcal{L} = \prod_{i=1}^{N_E} \prod_{j=1}^{N_t} P(E_i) \sum_{k=1}^{N_z} P(z_k|E_i) P(t_j|z_k)^{n(E_i,t_j)} \tag{3}$$

where $n(E_i, t_j)$ is the number of occurrences of tag t_j in entity E_i.

The algorithm has two steps.

- **Expectation-step**: the conditional probability distribution of the latent topic z_k given the observation pair (E, t) is computed from the previous estimate of the model parameters

$$P(z_k|E_i, t_j) = \frac{P(t_j|z_k) P(z_k|E_i)}{\sum_{k=1}^{N_z} P(t_j|z_k) P(z_k|E_i)}. \tag{4}$$

- **Maximization-step**: the parameters of the multinomial distributions $P(t|z)$ and $P(z|E)$ are updated with the new expected values $P(z|E, t)$

$$P(t_j|z_k) = \frac{\sum_{i=1}^{N_E} n(E_i, t_j) P(z_k|E_i, t_j)}{\sum_{j=1}^{N_t} \sum_{i=1}^{N_E} n(E_i, t_j) P(z_k|E_i, t_j)} \tag{5}$$

$$P(z_k|E_i) = \frac{\sum_{j=1}^{N_t} n(E_i, t_j) P(z_k|E_i, t_j)}{n(E_i)}. \tag{6}$$

The distributions $P(t|z_k)$ describe each topic z_k and are also valid for documents outside the training set. This is, however, not true for the matrix $P(z|E)$ which stores the parameters of the N_E multinomial distributions $P(z|E_i)$ and is thus relative to the N_E training entities. For unseen documents, the distributions over topics can be inferred through a *folding-in* procedure, as proposed in [15]. This method maximizes the likelihood of the unseen documents using a partial version of the EM algorithm described above: $P(t|z)$ is obtained from training and *kept fixed*, thus not updated on each M-step. As such, $P(z|E_{unseen})$ maximizes the likelihood of entity E_{unseen} with respect to previously learned parameters. Overfitting is prevented by early stopping based on the folding-in likelihood of a validation set. This procedure has proven successful in several uses of PLSA, including work on text corpora [15] and annotated images [26].

B. Relation With Other Topic Models

Other topic-based formulations that involve (implicitly or explicitly) the existence of individuals and groups characterized by their content have been proposed in the text modeling literature [9], [33], [39].

Latent dirichlet allocation (LDA) is a fully generative probabilistic model [9]. It works under the same assumption as PLSA, namely that documents are mixtures of latent topics, which in turn are characterized by distributions over words. LDA is said to better the PLSA model in the way it generalizes to unseen documents. Although appealing in theory, in practice LDA has been reported to produce mixed results when compared to PLSA [16], [36], specifically in the case of multimedia data (images and tags).

The author-topic model (ATM) is an extension of LDA that includes authorship information in modeling text documents [33]. ATM uses a topic-based representation to model simultaneously the content of documents and interests of authors in the context of scientific articles and it assumes multiple authors for each document. The special case of one author per document is equivalent to the LDA model. Applied to Flickr, one could consider Flickr groups as being the documents in the model, with multiple authors, namely the group members. While this could be an alternative worth exploring, the ATM model would lose the potential of comparing users and groups directly.

The group-topic model (GTM) clusters entities based on their mutual relations, as well as on attributes of those relations [39]. This work does not explicitly take into account groups as existing entities, but rather tries to discover *latent* groups of people, specifically in the context of legislative voting patterns. Trying to apply GTM onto our problem, one could attempt using the users' representations for discovering latent groups. However, a way of taking into account *existing* groups is not straightforward.

To our knowledge, none of these options have been investigated to model Flickr groups and users and their content. While these models are potentially interesting, the complexity of some of them is higher and their applicability (e.g., in the case of GTM) might not be straightforward given the type of user-to-group membership evidence that is assumed.

In contrast, although several other models could be feasible, we advocate for a simpler computational modeling option (PLSA) that is nevertheless powerful. The key ideas are that groups and users in Flickr can be reasonably modeled as if they were comparable entities and that their direct joint modeling is beneficial despite the complex ways in which Flickr groups are created.

C. Learning the PLSA Model

The number of topics in the PLSA model is not known in advance, and learning it from the corpus itself is a nontrivial task. However, given the very nature of the corpus, we can assume that the accuracy of this number is not of extreme importance. We have thus approached this problem with the intention of finding a relative optimum, by analyzing the variation of the perplexity of the model with respect to the number of learned topics.

For this analysis, we have trained six different models, varying the number of topics N_z between the values in the set {20, 50, 100, 150, 250, 500}. We have trained the models on the dataset D_R split in a 9 to 1 ratio for training and testing, respectively. For each model, we have then computed perplexity, which is one of the standard measures for the performance estimation of a probabilistic model for a text collection. Given our probabilistic model and a set of test entities D_T, the perplexity of the model is computed as

$$per(D_t) = exp\left[-\frac{\sum_{i=1}^{N_d}\sum_{j=1}^{N_t} n(E_i,t_j) log\left(p(t_j|E_i)\right)}{\sum_{i=1}^{N_d}\sum_{j=1}^{N_t} n(E_i,t_j)}\right] \quad (7)$$

where $p(t_j|E_i)$ is the probability of tag t_j given entity E_i from the test data, N_d denotes the number of testing documents, N_T denotes the size of the vocabulary, and $n(E_i, t_j)$ denotes the count of tag t_j in entity E_i's bag-of-tags [15].

We show in Fig. 4 perplexity values for each of the six different models. As previously found in the topic model literature [9], [15], perplexity decreases with the number of topics. It appears that fixing a number of topics in the order of a few hundred is an adequate choice. For the experiments described in the rest of the paper, a value of $N_z = 100$ was used. This number represents a 100-times dimensionality reduction from the original 10 K tag vocabulary and facilitates both the manual inspection of the discovered topics and the visualization of the overall results. Larger values of N_z (e.g., 250 or 500) bring a decrease in perplexity; however, this is counterbalanced by the complexity of manually inspecting the model. We have experimented with other values of N_z, but omit their discussion at length for space reasons. In a nutshell, larger values of N_z (e.g., in the order of 200–500) tend to result in more "specialized" topics at the cost of a lower reduction in dimensionality. For this case, the main qualitative results (i.e., the consistent extraction of meaningful topics and their ability to be used for comparison between users and groups) do not change. On the other hand, smaller values of N_z (e.g., less than 50) result in topics that are more and more "general", becoming too broad (e.g., merging too many different actual topics) if N_z decreases substantially. For a realistic system, the number of topics would most likely be slightly higher than 100.

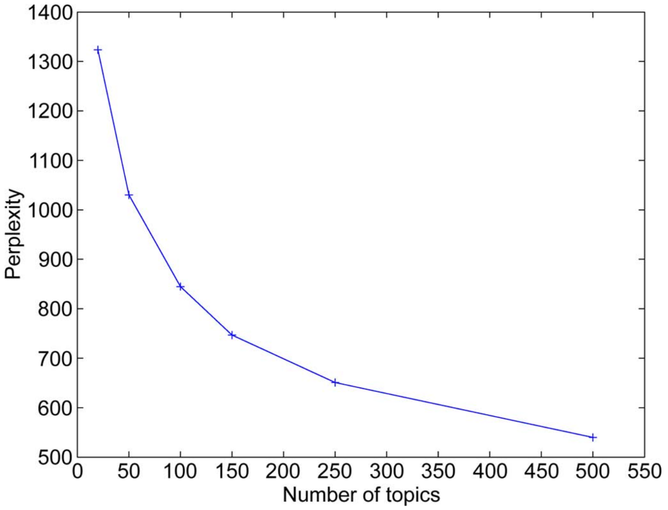


Fig. 4. Variation of the perplexity with respect to the number of topics learned by the model. The + markers show the perplexity values for each N_z in the set {20, 50, 100, 150, 250, 500}. Perplexity decreases with the increase in number of topics.

V. Using Topic Models for Community Understanding and Exploration

Unsupervised topic models trained on a corpus of documents output for each document a topic representation. In this case, a PLSA model is trained on the bags-of-tags representations of all users and groups.

A. Topic-Based Representation of Entities

One of the outputs of the PLSA model for a given entity are the multinomial distributions $P(t|z)$, in other words the probability distribution of tags over all topics. The model also outputs, for each entity, the distribution $P(z|E_i)$, or otherwise put, the probability distribution of the topics for that particular entity. Most of the topics in the model appear to be semantically consistent. We performed a subjective evaluation of a few models with different numbers of topics (50, 100, and 150), and we identified roughly 70% topics with high semantic consistency in the latter two cases, with slightly more "confused" topics in the case of the 50 topics model. Topics and relevant Flickr groups for 50, 100, 150, and 250 topics models can be found online at http://www.idiap.ch/~negora/flickrcommunities/.

TABLE II
SOME OF THE TOPICS LEARNED BY THE MODEL, CHARACTERIZED BY THEIR MOST PROBABLE TAGS (RANKED BY $P(t|z)$). WE ALSO PRESENT THE MOST PROBABLE ENTITIES (RANKED BY $P(z|E)$)

Topic 3	
$P(t\mid z)$	**Tag**
0.0856	sky
0.0608	sunset
0.0546	clouds
0.0493	night
0.0379	light
0.0290	sun
0.0268	blue
0.0160	lights
0.0159	water
0.0131	silhouette
0.0121	sunrise
0.0117	longexposure
0.0117	sea
0.0112	cloud
0.0105	orange
0.0097	moon
0.0089	reflection
0.0089	beach

Topic 3	
$P(z\mid E)$	**Entity**
0.6968	Lunatics
0.6793	moon
0.6697	The Moon
0.6629	Out The Window
0.6587	Lightning
0.6539	MOON Shots
0.6518	capture the sky
0.6505	Lightstream
0.6322	!orange sky
0.6272	Sunburst Specialty

Topic 13	
$P(t\mid z)$	**Tag**
0.1854	dog
0.0648	dogs
0.0382	puppy
0.0340	pet
0.0195	pets
0.0130	retriever
0.0125	cute
0.0122	pug
0.0115	dachshund
0.0083	chihuahua
0.0070	terrier
0.0070	animals
0.0069	mutt
0.0068	black
0.0067	la
0.0066	puppies
0.0064	canine
0.0064	animal

Topic 13	
$P(z\mid E)$	**Entity**
0.9516	For the love of dogs
0.9514	Love Of The K-9
0.9457	Flatcoats
0.9379	Just Puppies!
0.9334	Dogs, Dogs, and More Dogs...
0.9316	Retrievers
0.9231	Gentle Giants - An Extra Large Dog Group
0.9199	Crazy Canines
0.9137	Small cute doggies
0.9025	56939004@N00

Topic 18	
$P(t\mid z)$	**Tag**
0.1937	art
0.0339	painting
0.0278	drawing
0.0178	sculpture
0.0169	collage
0.0150	design
0.0144	illustration
0.0130	sketch
0.0121	artist
0.0104	gallery
0.0092	ink
0.0087	museum
0.0078	artwork
0.0076	paper
0.0072	paintings
0.0068	toys
0.0066	draw
0.0065	exhibition

Topic 18	
$P(z\mid E)$	**Entity**
0.9618	Obsessive Drawing
0.9237	Doodle Art
0.9120	Paper Museum
0.9075	Dragon's Den of Paintings and Other Art
0.9057	Art Critique - Non Photography
0.8907	Art Journal
0.8763	Moleskine: One Page at a Time.
0.8729	Notebookism
0.8679	Line Drawings
0.8634	ALL FEMALE ARTIST(ALFA FEM)

Topic 19	
$P(t\mid z)$	**Tag**
0.0383	handmade
0.0330	craft
0.0276	knitting
0.0247	prague
0.0220	vintage
0.0176	praha
0.0173	czechrepublic
0.0150	diy
0.0133	cute
0.0132	knit
0.0124	pink
0.0121	yarn
0.0112	eu
0.0111	etsy
0.0108	crafts
0.0088	sewing
0.0079	fabric
0.0079	bunny

Topic 19	
$P(z\mid E)$	**Entity**
0.9706	tezukuri life!
0.9691	Do It Yourself
0.9676	The Bag Blog
0.9596	Handmade Jewelry
0.9591	handbags
0.9573	Sewing
0.9564	Do It Yourselfers
0.9542	Cut Out + Keep
0.9441	83373306@N00
0.9437	MADE for the HOLIDAYS!

Topic 21	
$P(t\mid z)$	**Tag**
0.1615	music
0.0935	concert
0.0622	band
0.0569	live
0.0399	livemusic
0.0399	rock
0.0354	show
0.0221	gig
0.0197	dance
0.0185	guitar
0.0184	performance
0.0145	festival
0.0134	jazz
0.0121	bands
0.0117	musician
0.0109	concerts
0.0087	gigs
0.0084	stage

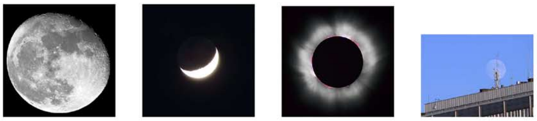
Topic 21	
$P(z\mid E)$	**Entity**
0.9030	livemusic
0.8940	11289325@N00
0.8832	Gigs Pool
0.8818	Support Local Music
0.8234	LIVE in CONCERT
0.8130	87075398@N00
0.7948	Live Music Photography
0.7786	SINGERS SING!
0.7557	Live Music Photographs
0.7224	Rock and Roll : live shows only please

Topic 22	
$P(t\mid z)$	**Tag**
0.0806	bird
0.0589	birds
0.0586	nature
0.0561	animal
0.0494	animals
0.0295	wildlife
0.0218	featheryfriday
0.0174	ilovenature
0.0170	natureza
0.0163	aves
0.0161	ave
0.0155	naturaleza
0.0136	out
0.0134	colors
0.0126	colorful
0.0126	color
0.0122	cores
0.0112	brazilian

Topic 22	
$P(z\mid E)$	**Entity**
0.9840	Birds of the world
0.9786	Birds Special Interest Group
0.9664	Birds and Bees and More
0.9660	Aves - Birds
0.9653	For Love Of Birds
0.9616	Garden Birds
0.9557	Wildlife Watch
0.9485	Free As A Bird
0.9388	Birds From Around The World
0.9355	Birding & Butterfly Enthusiasts

We show in Tables II and IV some of the topics learned by our model, described by their most probable tags, as well as their most probable entities. In these tables, when an entity is represented by just the Flickr ID (e.g., 56939004@N00), it represents a user; otherwise, it represents a group (e.g., *Lunatics*). In Tables III and V we also show some of the photos

TABLE III
EXAMPLE PHOTOS FROM POOLS OF GROUPS THAT ARE HIGHLY PROBABLE FOR TOPICS 3 (LEFT) AND 13 (RIGHT)

 Photos from group *Lunatics*, by *saturn h*, *oceandese-toiles*, *Luc Viatour* ©*GFDL*, and *Steffe*	Photos from group *Flatcoats*, by *Wabana* (1,2), *MontanaRaven* (3), and *black dog_brown dog* (4)

appearing in the most probable groups for topics 3, 13, 43, and 45.

Most topics are about places (e.g., topics mainly about *The Netherlands*, *Germany*, *Italy*, *Canada*, *UK*, *Spain*, or *France*), others about specific types of photography or photographical subjects (e.g., *black and white portrait photography*, *flowers*, *art*, *cats*, and *dogs*), while other topics are about events (e.g., *party*, *wedding*, or *music concerts*). For some of the topics (e.g., topics 13, 19, or 22), many of the top entities are very much about that specific topic, with very high values for the probabilities $p(z|E)$. We observe also that some topics' top entities are dominated by groups (e.g., topics 3, 18, or 22), while others are dominated by users (e.g., topic 61).

We also show in the upper part of Fig. 5 the distribution over topics for a Flickr group (*Candid Camera*) and the two most probable topics for the group in the lower half. Topic 38 could easily be described by *street portraits* and topic 90 by *children*. The next two most probable topics, 32 and 93, are about *black and white portraits* and *women portraits*, respectively.

Once these topic distributions are known for each entity, we are interested in knowing whether a difference between the two types of entities exists. To answer this question, we have generated the histograms of the number of *relevant topics* for each type of entity in Fig. 6. By relevant topics, we mean the highest ranked topics that account together for at least 80% of the probability mass in a given entity's topic probability distribution. We can observe two main differences.

- On one hand, a higher percentage of groups as opposed to users seem to be about fewer topics. For instance, 10% of the groups are about one or two topics, compared to just 4.8% of the users and 25% of the groups have four or less relevant topics compared to just 17% of the users. This is explained by the presence of a large number of *specific thematic groups* like *North New Jersey*, *Wildlife Watch*, or *Knitted Textile Art*, where the emphasis is placed on a specific geographical location, photo subject, or photographic technique, and as such, there is a high concentration in just a few topics of interest. People who belong to these groups contribute to the group pool just those photos that are relevant to the specific group interest theme, but they may have a wider range of interests themselves.
- On the other hand, certain groups are about more topics than any of the users. For example, 12.6% of the groups are about more than 13 topics, compared to only 5.6% of the users, as shown in Fig. 7. This is explained by the presence of *social groups* like *What's the Story?*, *Photos of people taking photos*, or *FlickrCentral*, where the emphasis is placed on social interaction. In these groups, there are (nearly) no restrictions on the kind of content members may submit to the group pool, and this results in all content types being shared in the group, even if individual members may have very specific photographic interests.

This is an interesting result, showing that we can distinguish between these two different types of groups (thematic versus social) just by looking at the number of relevant topics in their topic distributions. A clear-cut distinction between groups and users cannot, however, be solely made based on the topic representation.

B. Insights into Entity and Community Structures

The main purpose of having a common representation for groups and users is, of course, the ability to compare all these entities directly. This direct comparison would allow us, for example, to recommend groups and users to people based on their own topics of interest. Alternatively, a query-by-example scenario can also be envisaged, where a user would want to see all groups and users similar to a given entity of his or her choosing. Once a distribution over topics is obtained for each entity, by simply measuring the distance between any such two distributions, we should be able to tell if user X is more similar to user Y or user Z, or if user X is more similar to group A or group B. A few methods have been widely used to compute the similarity between distributions, such as the Kullback-Leibler (KL) divergence, Jensen-Shannon divergence, histogram intersection, or Bhattacharyya distance. As we were interested in a symmetrical distance, we have explored two of the above-mentioned methods: a variation of the KL distance which is symmetrical and the Bhattacharyya distance. However, as none of these two is actually a true metric, we adopted the distance used in [10], which is based on the Bhattacharyya coefficient. In the case of discrete probability distributions, the Bhattacharyya coefficient is defined as

$$BC = \sum_{x} \sqrt{p(x)q(x)}. \tag{8}$$

Our similarity metric is then the distance given by

$$\rho(p,q) = \sqrt{1 - BC}. \tag{9}$$

This distance has the advantage of being a true metric, as it is non-negative, it is zero if and only if the two distributions are identical, it is symmetric, and it obeys the triangle inequality [18]. It also has the advantage of being confined to the interval $[0 \ldots 1]$.

For each entity in our dataset, we have thus computed the distance ρ to all other entities in the dataset, resulting in an $N_E \times$

TABLE IV
SOME OF THE TOPICS LEARNED BY THE MODEL, CHARACTERIZED BY THEIR MOST PROBABLE TAGS (RANKED BY $P(t|z)$). WE ALSO PRESENT THE MOST PROBABLE ENTITIES (RANKED BY $P(z|E)$)

Topic 26	
$P(t\mid z)$	**Tag**
0.0451	red
0.0397	blue
0.0299	green
0.0259	light
0.0245	yellow
0.0202	white
0.0198	abstract
0.0147	orange
0.0146	wall
0.0132	black
0.0129	shadow
0.0124	glass
0.0123	color
0.0115	window
0.0111	reflection
0.0083	shadows
0.0079	texture
0.0073	metal

Topic 26	
$P(z\mid E)$	**Entity**
0.7116	27986376@N00
0.6957	34204690@N00
0.6956	MAXIMUM minimalism
0.6919	DIGITAL IMAGE
0.6919	Miksang
0.6914	haphazart! Contemporary Abstracts
0.6868	29718473@N00
0.6831	pavement pix: a sequence of images
0.6693	To Inspire Abstract Art.
0.6590	OPTIME GALLERY

Topic 43	
$P(t\mid z)$	**Tag**
0.1721	me
0.0929	selfportrait
0.0500	self
0.0170	bw
0.0148	portrait
0.0127	myself
0.0110	mirror
0.0107	blackandwhite
0.0086	reflection
0.0076	hand
0.0075	home
0.0073	feet
0.0064	face
0.0058	ofme
0.0057	friend
0.0054	hair
0.0048	eye
0.0048	red

Topic 43	
$P(z\mid E)$	**Entity**
0.8317	alter ego
0.7484	Toilet Vanity
0.7240	International (TBA) Week
0.7200	365 Days: Rejects
0.7119	ME
0.7010	lights & skin
0.6998	365 Days Crybaby Edition
0.6959	It's Friday, so put your feet up and take a break! ? FUTAB!
0.6906	365 Days
0.6824	My Self Portrait

Topic 45	
$P(t\mid z)$	**Tag**
0.1255	losangeles
0.0998	graffiti
0.0920	streetart
0.0573	la
0.0410	art
0.0217	california
0.0203	hollywood
0.0200	street
0.0180	santamonica
0.0172	stencil
0.0146	sticker
0.0137	socal
0.0133	urban
0.0101	stickers
0.0098	mural
0.0098	angeles
0.0094	los
0.0093	russia

Topic 45	
$P(z\mid E)$	**Entity**
0.9080	STICKER
0.8268	Street Stickers
0.7929	stickerart
0.7755	City Stickers
0.7616	Stickers & Decals
0.7510	59289953@N00
0.7159	Los Angeles Street Art
0.7119	Street Stickers and Stencils
0.7034	66115732@N00
0.6771	Suburban the streetart magazine

Topic 57	
$P(t\mid z)$	**Tag**
0.1129	car
0.0475	cars
0.0431	auto
0.0192	ford
0.0167	automobile
0.0141	vw
0.0136	classic
0.0126	truck
0.0121	show
0.0111	carshow
0.0102	motorcycle
0.0100	bmw
0.0100	chevrolet
0.0088	classiccar
0.0088	volkswagen
0.0082	vintage
0.0078	honda
0.0073	mercedes

Topic 57	
$P(z\mid E)$	**Entity**
0.8595	BadAss CaRZ TrucKZ N BikEZ
0.8534	76713602@N00
0.8427	89388861@N00
0.7200	Antique, Vintage, Classic Cars and Trucks
0.6997	CHEVROLET
0.6904	US Cars
0.6901	1,000,000 Car Photos
0.6900	Porsche
0.6806	Car Parts and Details
0.6783	Classic Cars

Topic 61	
$P(t\mid z)$	**Tag**
0.2044	london
0.1527	uk
0.1249	england
0.0205	unitedkingdom
0.0119	britain
0.0116	yorkshire
0.0080	brighton
0.0077	thames
0.0076	birmingham
0.0074	kent
0.0073	cornwall
0.0063	oxford
0.0060	manchester
0.0059	norfolk
0.0054	bath
0.0054	park
0.0049	sussex
0.0040	pub

Topic 61	
$P(z\mid E)$	**Entity**
0.7960	49612551@N00
0.7824	86881049@N00
0.7771	82078478@N00
0.7505	29328061@N00
0.7498	84806883@N00
0.7373	15179025@N00
0.7076	85696534@N00
0.7020	Norwich UK
0.7006	49767717@N00
0.6795	LONDRA by ITALIANI (LONDON)

Topic 65	
$P(t\mid z)$	**Tag**
0.0945	portrait
0.0550	woman
0.0515	girl
0.0234	face
0.0185	sexy
0.0179	people
0.0169	beautiful
0.0166	female
0.0152	model
0.0144	beauty
0.0132	man
0.0126	eyes
0.0119	girls
0.0110	women
0.0101	smile
0.0100	pretty
0.0099	hair
0.0086	fashion

Topic 65	
$P(z\mid E)$	**Entity**
0.9529	5,000+ Views
0.9487	Views 5000
0.9470	Views 8000
0.9405	5000+ Views (3 per day)
0.9386	4,000+ Views
0.9353	3000 Views
0.9333	Over 10000
0.9230	Views 3000
0.9205	5000 VIEWS
0.8962	Views 4000

N_E distance matrix. With this new information, we explore new ways of understanding communities' structure.

First, we started by looking at the distribution of the mean distances between groups. As pointed out in the earlier anal-

TABLE V
EXAMPLE PHOTOS FROM POOLS OF GROUPS THAT ARE HIGHLY PROBABLE FOR TOPICS 43 (LEFT) AND 45 (RIGHT)

Photos from group *Toilet Vanity*, by *gretchi2000*, *ugglan*, *jamelah*, and *phil h*	Photos from group *STICKER*, by *sbluerock* (1,2), *smenzel* (3), and *Lush.i.ous* (4)

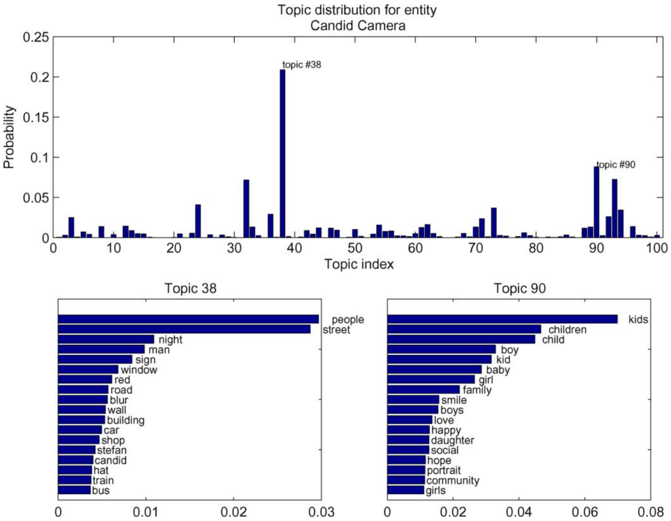


Fig. 5. Topic distribution for the entity *Candid Camera* (a Flickr group). In the lower part of the figure, the two most relevant topics are described by their top most probable tags. Topic 38 could be described by the concept “street portraits” and topic 90 by “children”.

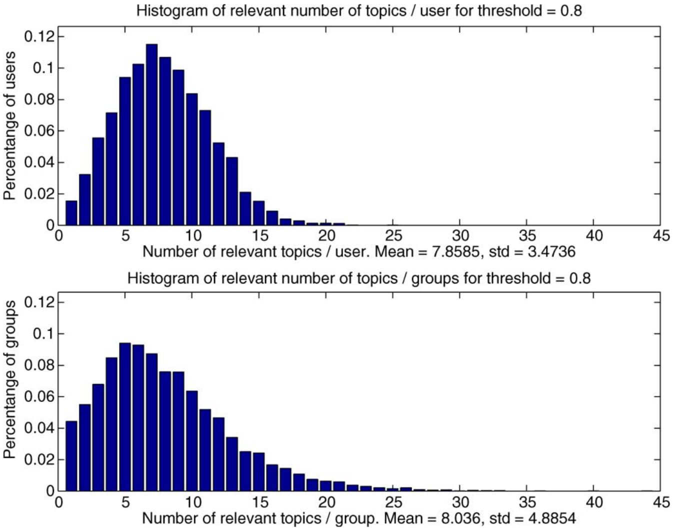


Fig. 6. Comparison of the number of relevant topics for groups and users. For ease of comparison, we normalized the histograms and display on the y axis the percentage of users and groups, respectively.

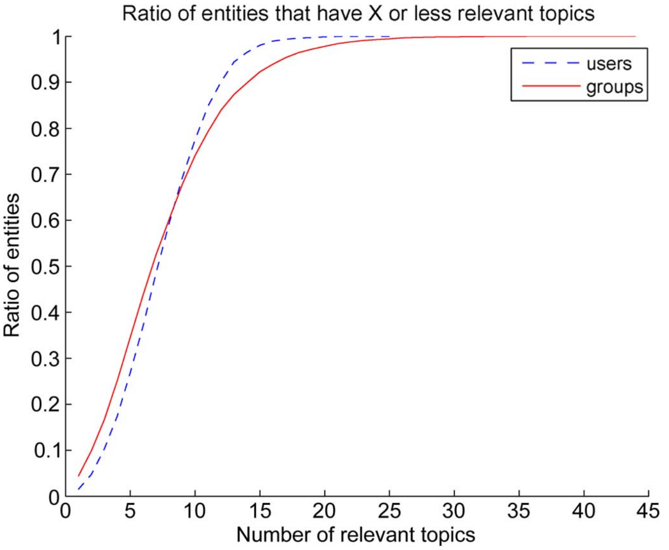


Fig. 7. Ratio of either type of entities that are about x or less topics. For example, 60.2% of groups and 59.3% of users are about at most eight topics.

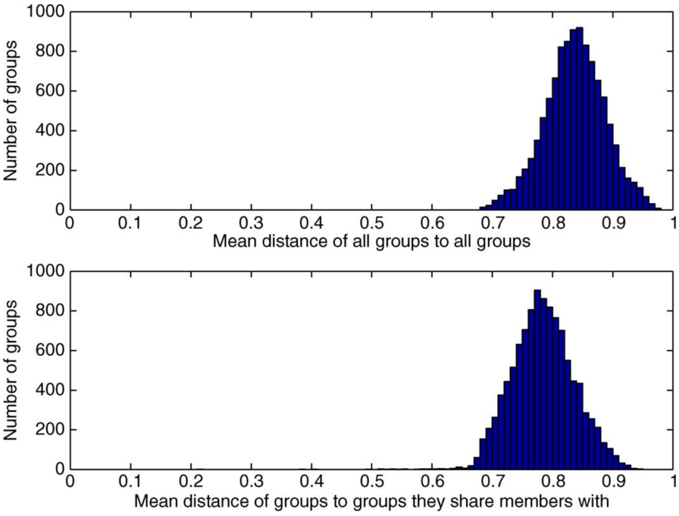


Fig. 8. Distributions of the mean distances from each group to all groups, or to just groups who share at least one member. We observe a significant shift in distances when only “overlapping” groups are considered.

ysis of Flickr in Section III, on average users share any given photo with about three groups. For this reason, we compute the average group-to-group distance for two cases: first, from all groups to all other groups in the dataset; second, from all groups to only all other *overlapping* groups in the dataset—i.e., groups with which they share at least one member. Our hypothesis is that in the second case, distances should be smaller as the members themselves “validate” the similarity of the groups by joining both of them. We present in Fig. 8 the two histograms

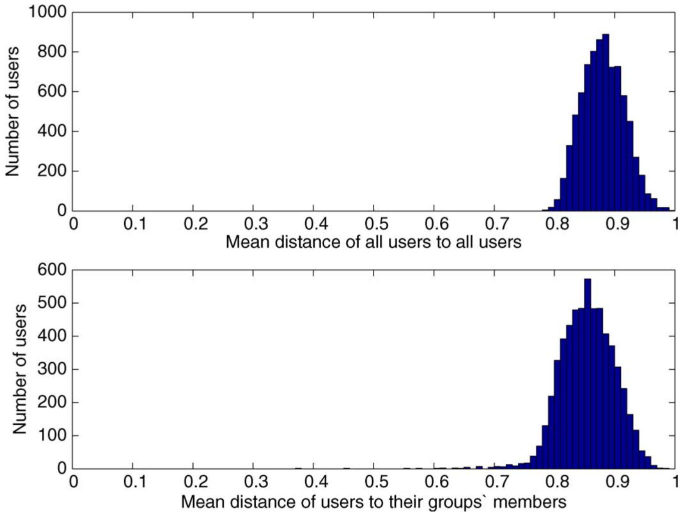


Fig. 9. Distribution of the mean distances from all users to all users, or to just users with whom they share at least one group. We observe a clear difference between the two histograms; the distances between users who are part of the same groups are smaller on average than those between all users.

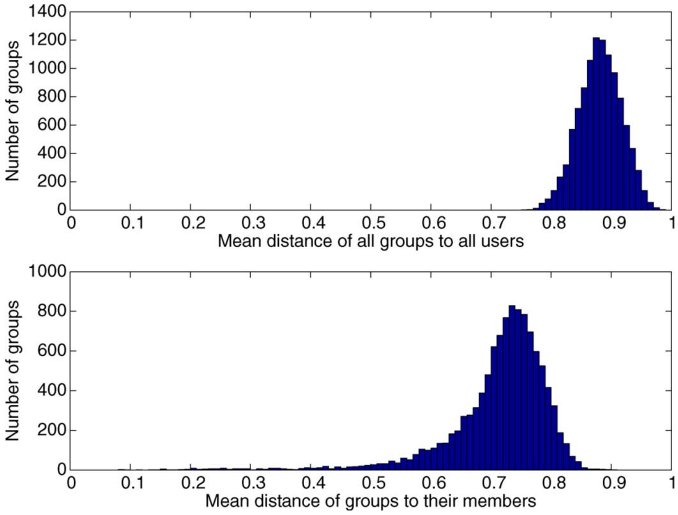


Fig. 10. Distribution of the mean distances from all groups to all users, or to just member users. We observe a distinct difference, explained by the fact that members' representations are closer to that of the group they belong to than those of users who do not belong to the group.

of distances for the two considered cases. We can observe a significant shift in mean distance when only overlapping groups are considered, which seems to confirm our intuition. The null hypothesis that the two distributions have the same mean is rejected by a two-tailed t-test at $\alpha = 0.01$.

Second, we analyzed the distances between users. As previously for groups, we have constructed two histograms, depicted in Fig. 9: in the upper part, mean distances from all users to all users and in the lower part, mean distances from all users to only those users with whom they have at least one group in common. The difference between these two histograms is not very pronounced; however, we can observe a clear shift towards lower values when only users who belong to common groups are taken into account. The two-tailed t-test rejects the same-mean hypothesis at $\alpha = 0.01$. Again this can be explained by the fact that users who participate in the same groups are likely more similar to each other than to users with whom they share no groups at all. One can also observe that the histograms in Fig. 8 have a larger variance than the histograms in Fig. 9, which again indicates that groups might be a more variable construct.

Finally, in Fig. 10, we plotted two histograms: in the upper part, a histogram of the mean distances from all groups to all users in the dataset and in the lower part, a histogram of the mean distances from all groups to only their members. Here we can observe a much more pronounced difference, in means and variances of the two distributions. The mean distances from groups to all users are generally higher than 0.82, while the mean distances from groups to just their members are generally lower than 0.82. This difference in means is statistically significant, confirmed by a two-tailed t-test at $\alpha = 0.01$. Furthermore, in the case of distances to group members, about 30% of the groups have an average distance smaller than 0.7, which would seem to indicate higher homogeneity in terms of topic distributions of their members. These are interesting but not surprising results, as one would expect the topic model to capture the semantic similarity of users to the groups they belong to, or to users they are in the same groups with and it might also be due to the way groups' bag-of-tags representations are built, starting from their members.

VI. Applications of the Topic-Based Model of Flickr Entities

One of Flickr's most addictive features by the account of its members is the opportunity to explore quasi-random photographs through the *Explore* feature of the site. Using a proprietary algorithm that takes into account different meta-parameters of a photo (some of which one may guess, like the number of views, number of comments, or number of times the photo has been marked as a favorite), Flickr provides a ranking measure called *interestingness*, which is then used to display interesting photos from people the user may not necessarily know. Flickr groups are also a very important feature of this community, yet finding groups is limited to keyword-based searching through the group names and group forum discussions. Inspired by these features and shortcomings, we present a concept of two simple applications: one that allows topic-based exploration of Flickr entities rather than photos and another one that allows keyword-based searching of users and groups alike, based on their topic decompositions.

A. *Topickr: An Interest-Based Entity Exploration Tool*

The exploration mechanism can be very well used with our topic-based representation model. Instead of ranking photos based on interestingness as done in Flickr, we rank users and groups with respect to each other based on the inter-entity distances computed previously as per (9). Our *Topickr*[1] application, of which a snapshot is presented in Fig. 11, allows us to explore the topic model visually: starting from any given entity in the model, we present the most similar users and most similar groups. This is in fact a query by example scenario. A user may want to discover entities that are similar to a given user or group

[1] See demo at http://www.idiap.ch/~negora/acmmm08.

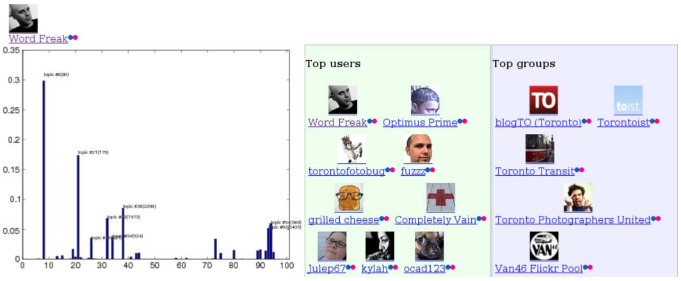


Fig. 11. Topickr: an exploration application that uses similarity of the topic-based representations in order to present the most similar users and groups for a given entity. On the left, the topic representation of the given entity (user *Word Freak* in this case), and on the right, the top most similar users and most similar groups.

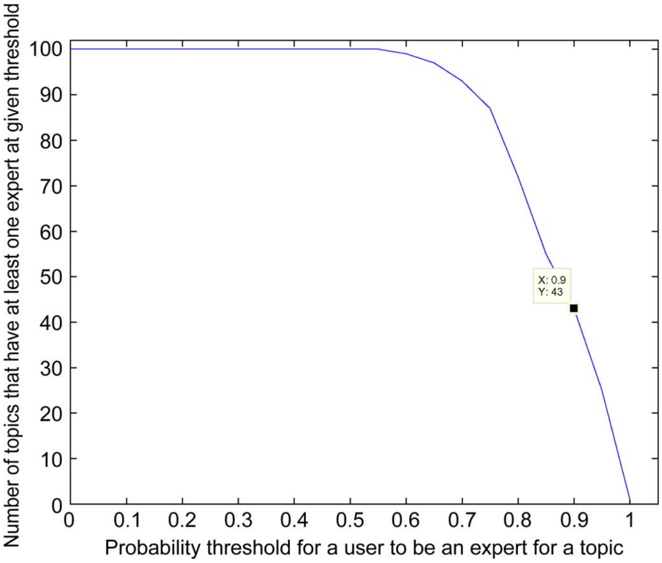


Fig. 12. Number of topics that have at least one topic-expert, varying with the topic-expert's probability for the given topic. In this model, 93 topics out of 100 have at least one entity whose probability for that topic is higher than 0.7 and 43 topics out of 100 have at least one entity whose probability is higher than 0.9.

they particularly like. This is not straightforward for a human observer, but in our model, this can be easily accomplished by ranking all entities with respect to the example provided by the user, based on the distances ρ.

As an alternative starting point, a user may choose any topic learned by the model. Using the fact that $P(E|z) \propto P(z|E)$, we can rank entities based on their probabilities given this starting topic. As we have seen in Section V-A, some entities have spiky topic distributions, with a single topic in their representation. We call these entities *topic-experts*. We show in Fig. 12 the number of topics that have at least one topic expert, depending on the threshold set on the entities' probabilities for the given topic: 93 topics out of 100 have at least one entity whose probability for that topic is higher than 0.7 and 43 topics out of 100 have at least one entity whose probability is higher than 0.9. In all cases, for any given topic, a most probable entity across the entire data set will always exist, even if its probability for that topic is lower. The exploring user may thus start from any of the topics in the model and explore its experts and their most similar entities.

A third exploratory option is a combination of the previous two: we start with an example entity, and, in addition to the most similar entities, we also present the topic-experts for the relevant topics in the distribution of the example.

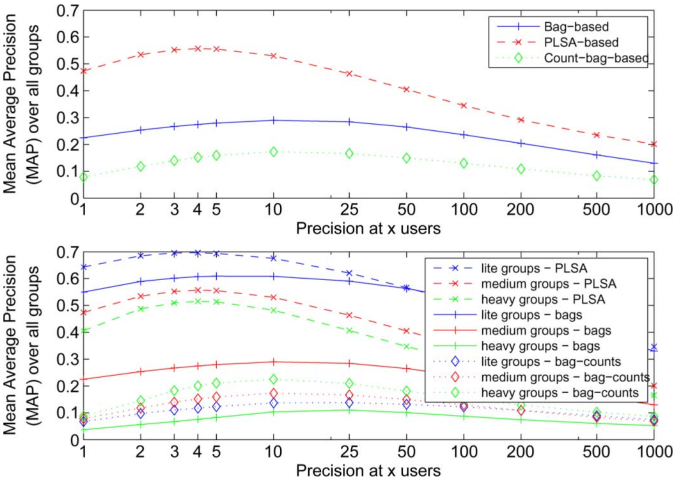


Fig. 13. Mean average precision for the user retrieval experiment, computed separately for lite, medium, and heavy groups in terms of user memberships. Lite groups are in the first quartile (less than 12 users), medium groups in the second and third quartiles (12 to 49 users), and heavy groups in the fourth quartile (more than 49 users).

B. *Evaluation of Topic-Based Exploration*

Although most numerical evaluations are difficult in the context of our data set for lack of ground truth, we can attempt to use the user-group memberships as ground truth for user-group relevancy. We compare three similarity measures in two retrieval scenarios. The first similarity measure is the previously denoted ρ distance in (9), from the topic-based representations. A second measure is based on the raw bag-of-tags representations, namely the distance between two entities is computed as the dot product of the binary bags vectors. Finally, a third measure is also computed as the dot product, but this time between the effective counts of the tags in each bag-of-tags representation of users and groups. We ran two evaluation experiments, one in which we use the full set of groups as queries and rank users by similarity to the query group, and the second one in which we use the full set of users as queries and we rank groups by similarity to the query user. For each of the two experiments, average precision is computed for each query, using the user-group membership information as ground truth. We show in the top halves of Figs. 13 and 14 the mean average precision (MAP) of the two retrieval experiments. In both figures, the blue continuous line shows the MAP for the bag-based similarity measure, the green dotted line the bag-counts-based one, and the red dashed line shows the MAP for the topic-based similarity measure. The x axis is drawn in log scale.

For the first experiment, we retrieve the most similar users for each group. In this case (Fig. 13, top), the best performance in terms of MAP is given by the topic-based similarity measure, with the bag-based measures performing significantly worse. The PLSA-based similarity measure peaks at 56% MAP for the top-5 returned results. The bag-based measures reach their highest MAP for the top-10 returned results, with 29% for the bag-based measure and 17% for the bag-counts-based one. Additionally, a comparison of the top-1 retrieved users for

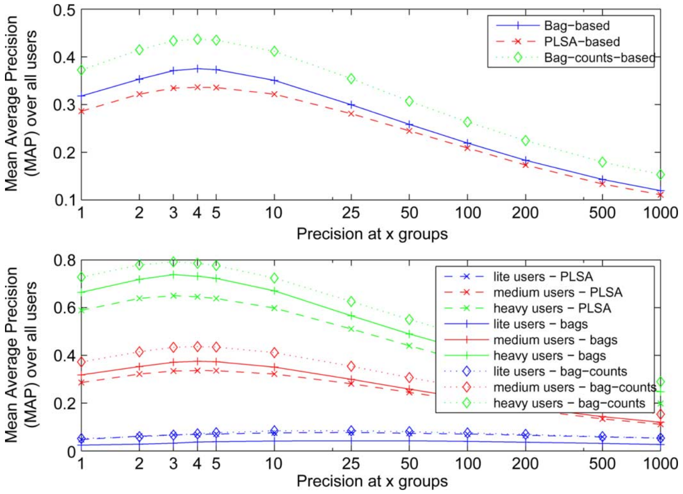


Fig. 14. Mean average precision for the group retrieval experiment, computed separately for lite, medium, and heavy users in terms of group membership. Lite users are in the first quartile (less than 10 groups), medium users in the second and third quartiles (10 to 91 groups), and heavy users in the fourth quartile (more than 92 users).

all groups shows that the bag-based similarity retrieves only 656 distinct users, the bag-counts-based one 236, while the topic-based similarity retrieves 2274 different users. This shows that the topic-based representation is able to retrieve a larger variety of users, which is a good feature for exploration. The first users retrieved by all methods tend to have quite big vocabularies, with a median of 3091 for the topic-based method, 5446 for the bag-based one, and 35 693 for the bag-count-based similarity.

For the second experiment, group retrieval (Fig. 14, top), the MAP is better for both bag-based similarity measures, with values peaking at 44% for the bag-counts-based measure, 38% for the bag-based one, and 34% for the topic-based measure, all performing best at the top-4 retrieved results. The same observation applies in this case as well: looking at the top-1 retrieved group across all users, we note that the bag-based similarity measure retrieves only 12 different groups for the almost 6000 users (the largest groups in terms of members, on average 894 users per group), the bag-counts-based measure retrieves 85 different groups (also the largest as well as some medium sized ones, on average 327 members per group, with median 113 members per group). In contrast, the topic-based similarity measure retrieves 3137 distinct groups, with on average 25 members (and median 12 members per group). This indicates that the bag-based similarity measures are heavily biased towards the most popular groups, while the topic-based representation is able to return less popular groups, which may be desirable in the exploration scenario. It is also noteworthy that although we designed these experiments as a retrieval scenario where we know the ground truth user-group membership, in practice it is much more interesting to retrieve groups that the user does not already belong to, but to which he or she is similar. This aspect is not accounted for in the experiments.

Another important issue is how these models perform when confronted with different types of entities in terms of size. We defined three categories (lite, medium, and heavy) based on how many groups a user belongs to, or how many members a group has. We then analyzed how the MAP changes with respect to the users' and groups' sizes. Lite users fall within the first quartile of the membership distribution, from 1 to 9 groups, medium users in the second and third quartiles, from 10 to 91 groups and heavy users in the fourth quartile, with more than 92 groups. Similarly, lite groups have between 1 and 11 members, medium groups between 12 and 49 members, and heavy groups more than 50 users. In the bottom halves of Figs. 13 and 14, we show the breakdown by user and group types, respectively. For the group retrieval scenario (Fig. 14 bottom), all three similarity measures perform similarly when exposed to lite and heavy users, preserving their relative ranking to each other. For lite users (sparse information), all three measures perform the worst and they perform the best for heavy users (plenty of information). For the user retrieval scenario on the other hand (Fig. 13 bottom), the results are quite interesting. Lite groups yield the highest MAP for both the topic-based and the binary bag-based measures. MAP performance actually degrades with the size of the groups for the topic-based and bag-based measures, unlike the bag-counts-based measure, which although it performs the worst across all three types of groups, it works better as the groups get larger.

C. Single and Multi-Topic-Based Keyword Search

As mentioned earlier, finding groups in Flickr is not particularly easy today. Unless the group uses the searched keyword in its name, description, or in the group discussions, direct tag-matching against the group photo pool is not possible.

By using the topic model, we can effectively transform the keyword into relevant topics using the $P(t|z)$ matrix. We select those topics and then retrieve the most likely entities for each individual topic, using the $P(z|E)$ distributions. Because we use in each case a single topic for which we retrieve the topic-experts, we call this search method topic-expert search (TES).

Alternatively, by computing the probability distributions $P(z|t)$ for the given tag, we can then compute the distance ρ from all the topic distributions of the entities in the dataset to the search keyword. This allows us to retrieve those entities that have a topic distribution most similar to that of the searched keyword and who are not necessarily topic-experts. We call this search method tag-entity distance search (TEDS).

To illustrate these methods, we present the top-10results for the tag *guitar* in Table VI using the current Flickr search method (FS), TES, and TEDS.

The search for the keyword *guitar* on Flickr yields about 6000 groups that supposedly contain this tag in their names, admin-defined keywords, or their descriptions, although upon manual inspection, the search engine does not seem to work as advertised after the first few pages of results. On the other hand, we observe that the topic-based search methods retrieve groups whose names do not contain (with the exception of the first result for TEDS) the searched keyword but are more related to its context, mostly live music for TES and music in a more general way for TEDS.

Another interesting example is the search for the tag *artist*, presented for the three methods in Table VII. The topic-based searches retrieve mostly groups about drawing and painting that,

TABLE VI
FLICKR SEARCH (FS), TOPIC-EXPERT SEARCH (TES), AND TAG-ENTITY DISTANCE SEARCH (TEDS) RESULTS FOR THE TAG *guitar*

FS	TES	TEDS
Guitar Face	livemusic	Guitar World
Hand Made Guitars	Gigs Pool	Music
Guitar World	Support Local Music	My Love Affair With Music
Teye Guitars	LIVE in CONCERT	Live Music
Fender Guitars	Live Music Photography	musicians
Acoustic Guitar Personages	SINGERS SING!	Band Photography
SCHECTER Guitars	Live Music Photographs	Music Makers
Warmoth Guitars	Rock and Roll : live shows only please	Everything about music
your personal guitar	Band Photography	SINGERS SING!
guitar video	Rock Photography	Rock and Roll : live shows only please

TABLE VII
FLICKR SEARCH (FS), TOPIC-EXPERT SEARCH (TES), AND TAG-ENTITY DISTANCE SEARCH (TEDS) RESULTS FOR THE TAG *artist*

FS	TES	TEDS
Christian Mixed Media & Folk Artists	DRAW!	Obsessive Drawing
Female Self-Portrait Artists' Support Group ;-)	drawing	Doodle Art
Polymer Clay Artists Guild of Etsy (PCAGOE)	Sketchbook	Paper Museum
Artists And Their Art	Artworks on Paper	Dragon's Den of Paintings and Other Art
Etsy Artists Rule: 1 Million Picture Pool	Illustration	Art Critique - Non Photography
Art and Artists.	Doodlegang	Art Journal
Artist Trading Cards	DRAWING (charcoal, pencil, pastel, etc.)	Moleskine: One Page at a Time.
Artist's Hidden World	Sketches	Notebookism
Etsy Glass Artists (EGA)	drawings	Line Drawings
ATC (Artist Trading Cards)	Doodle Art	ALL FEMALE ARTIST(ALFA FEM)

TABLE VIII
FLICKR SEARCH (FS), TOPIC-EXPERT SEARCH (TES), AND TAG-ENTITY DISTANCE SEARCH (TEDS) RESULTS FOR THE TAG *airplane*

FS	TES	TEDS
Airplanes: Classic Airliners	Rocket	Aviation
Airplane Wings	We love planes	Airplanes
Junkers -n- Classics (OLD CARS TRUCKS, TRACTORS, BOATS, AIRPLANES)	Warbirds	Aeronautical
Airplanes: Nose Shots	Air Shows	Military Aviation Photography
Airplanes	Aircraft Spotting	Warbirds
Radio Control Airplanes	Las Vegas Local	Boeing Jetliners
Airplanes and Airports	Aviation	Jet Airplanes
Jet Airplanes	Airportnerds - ”we few, we happy few” :-)	Aircraft
Airplanes: Regional Jets	Military Aviation Photography	Air Shows
. : Airplane Graveyard : .	Pilot's Lounge: Photo Assignment - Biplanes and Triplanes	We love planes

with few exceptions, do not contain the search keyword in their name. It is however quite clear that these groups are highly relevant to the *artist* concept. A third example for the tag *airplane* is shown in Table VIII.

What we are proposing is not replacing the search-by-tag paradigm, because tags are essentially the finest granularity of concepts that we may obtain and the most straightforward way for information retrieval. Rather, we advocate improving search-by-tags by taking advantage of higher-level concepts, like the ones discovered with our model. Clearly, one open issue is model complexity, that is the number of topics with respect to the corpus that is being modeled. Too many topics will make the model intractable, while too few topics will not provide enough concept granularity. This is an active research field [8].

VII. MODEL GENERALIZATION

In constructing our reduced dataset D_R, discussed in Section III, we have set a minimum threshold of tags present in the vocabularies of the entities. This was done in order to ensure that the topic model was learned on good quality data, but it leaves us with several open questions. How does the learned model perform for entities which have small bags-of-tags (and thus are potentially poorly represented)? Is there a difference between the topic representations of entities with smaller bags-of-tags and entities with larger bags-of-tags?

To answer these questions, we tested the model on entities with bags composed of 50 or less unique tags from our 10 K vocabulary. This threshold gives us roughly 30 K groups and 10 K users for a total of 40 K entities, with an average of 15.3 unique tags for users and 15.8 unique tags for groups.

Two examples of typical topic distributions for entities in this set are shown in Fig. 15. In this case, on the left, the entity is a group, *Arabic Weddings*, with a vocabulary of only three unique tags: *john*, *dancing*, and *wedding*. The two relevant topics, 47 and 50, are mainly about *parties and friends* and *weddings and proper names*. While in this particular case, the entity tags seem to have been discriminant enough to determine the correct topics, in other cases, like the one presented on the right of the same figure, this is no longer true. The only tag in the entity's bag (user 7468381@N07) is the tag *bo*. The topic with the highest probability in this case is topic 12, which is mainly about *cats and kittens*. However, for this specific entity, *bo* has nothing to do with cats and, for lack of better information provided by other tags, the inference is poor. At a first glance, the presence alone of the tag *bo* in our 10 K vocabulary seemed

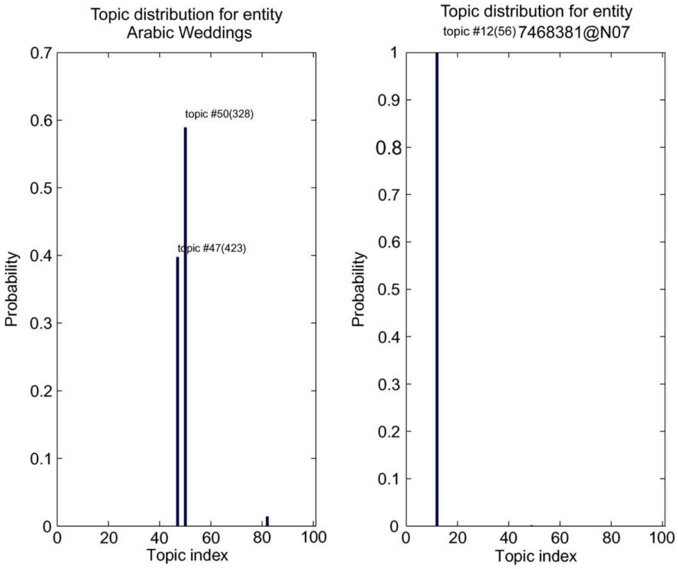


Fig. 15. On the left, the topic representation of an entity (the group *Arab Weddings*) with only three unique tags in its vocabulary: *john*, *dancing*, and *wedding*. The relevant topics 47 and 50 are mainly about *parties and friends* and *weddings and proper names*, respectively. On the right, the topic representation of an entity (user 7468381@N07) with only one unique tag in its vocabulary: *bo*. However, the relevant topic 12 is mainly about *cats and kittens*, which does not correspond to the usage of the tag employed by this user.

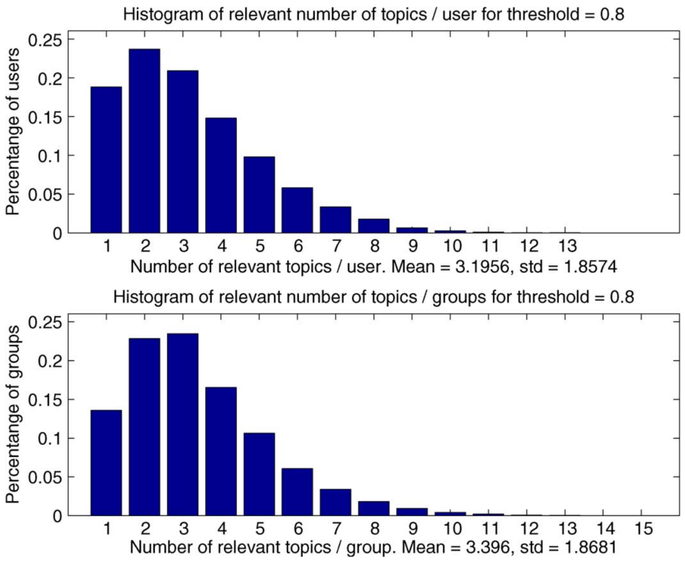


Fig. 16. Comparison of the number of relevant topics for groups and users with at most 50 unique tags in their vocabularies. For ease of comparison, we normalized the histograms and display on the y axis the percentage of users and groups, respectively.

surprising; however, on inspection of the data, it turned out that *bo* is quite a popular name, in particular in the pet world, which also explains why topic 12 is the most probable one for this tag.

The statistics of the topic distributions over this set of entities are shown in Fig. 16. We can clearly observe a shift in the mean number of relevant topics towards lower values compared to the entities in D_R, from around eight relevant topics in Fig. 6 to about three relevant topics for both users and groups in Fig. 16, and also a smaller variance, from 3.4 for users and 4.8 for groups in the case of large bags-of-tags to approximately 1.8 for both types of entities in the case of small bags-of-tags entities. This indicates that the model produces quite sparse topic-based representations, with nearly 50% of the groups and users having at most two relevant topics and almost 19% of users and 14% of groups having one topic only. We have just illustrated that when the topic decomposition is based on very small bags-of-tags, the accuracy of the inference might decrease. This may also cause entities with very few tags to become topic-experts based on very little evidence; clearly it would be more desirable to have as topic-experts entities for which the probability is based on substantial evidence rather than just a few tags. As such, weighting mechanisms should probably be taken into account when dealing with "tag-poor" entities. This will be examined in future work.

One practical issue is that of the computational time of the model. With a non-optimized C implementation, learning the PLSA model on 18 000 entities takes in the order of 2.5 h on a IntelCore2 CPU 6700 machine with 3 GB RAM, running at 2.66 GHz. On a new document, inference takes in the order of 2 s. Learning the full topic model in principle can be sped up through a number of strategies. Refer, for instance, to a number of recent works including [29], or [40]. These works show that using topic models at large scales starts to be a feasible option. Furthermore, for a practical application, in our opinion, the model need not be updated so often once it is learned on a significant amount of data, as often many users tend to remain stable in their main interests about specific topics after some time; the same is even more true for groups. An important issue would be how to detect new topics given an existing model. Overall, a thorough investigation of the dynamics of topic evolution is in itself a very relevant research issue that has not been investigated in enough detail in the Flickr community (an exception to this is [14]), which would be a relevant direction to pursue in the future.

VIII. Conclusion and Future Work

Social media repositories such as Flickr constitute an emerging challenge for multimedia information management systems. We have analyzed in this paper an unexplored issue, that is jointly modeling Flickr users and groups. Our analysis showed that, although the two types of entities are conceptually different, they are also similar enough from a tag point of view to make their joint modeling not only possible but highly beneficial. By modeling tag content at a higher, more abstract level, and without the need to understand the visual content itself, we used groups' and users' photos and their tags to derive a probabilistic topic-based representation of Flickr entities.

On one hand, we showed that having a common representation for Flickr's groups and users allows us to easily compare these entities. On the other hand, we also showed that the representation itself can be a source of information about the characteristics of an entity, like concentration on a specific (photographic) concept, geographical location, or type of social interaction undertaken by or within the entity. Furthermore, we have shown that this common representation allows for new insights about Flickr itself and creates new application opportunities, like similarity-based exploration of the entities using the topic model, as well as single and multi-topic tag-based search.

There are several open issues to be looked at in the future, such as model complexity (balance between number of topics and size of vocabulary), user evaluation, or how to effectively deal with tag-poor entities. We have shown that sparse entities might not provide enough evidence for inference and tend to take over the topic-experts roles. As such, re-ranking mechanisms that take into account the available evidence for a given entity are probably envisageable. Considering the huge size of the databases in use for systems such as Flickr, with billions of photos and their associated tags, the answers to these questions will probably become very important if models such as the one we propose here are to be integrated in large-scale systems. User studies could provide an additional validation mechanism for these methods. Future work may also look at 1) the definition of a subject population of significant size (taken from the actual Flickr users and groups used in our study), 2) a subject recruitment procedure, and 3) an incentive mechanism to encourage users to employ our prototype system to search or browse similar entities.

Another promising avenue to explore in future work is the integration into the model of the visual features from the photos themselves, with the main challenge residing in the feature extraction and selection tasks, often expensive computationally. With an active research field in this area, we are confident this is a realistic future goal.

Finally, an open issue is whether the method presented here could be applicable to other popular photo sites (like Kodak Gallery or fotocommunity.com), which also support tagging or other forms of free-form annotation of individual pictures and image sets. Two basic issues to investigate in this direction are the following. First, the different interaction modalities available on each site likely result in different "annotation qualities" and as such a comparative study of the text sources on each site would be a useful step to figure out if a bag of word model could be a good representation of users. The second direction has to do with the availability of social communities in these other photo sites, analogous to Flickr Groups, so that community models could be built. A comparative study of this particular issue would also be needed. Obviously, there is the technical problem of accessing data from other social media sites, which in Flickr is overcome through a public API, but which is still not a possibility in other sites. All these issues are of clear interest for future work.

References

[1] Holy Moly! at Flickr Blog, Nov. 13, 2007. [Online]. Available: http://blog.flickr.net/en/2007/11/13/holy-moly/.

[2] Billion! at Flickr Blog, Nov. 3, 2008. [Online]. Available: http://blog.flickr.net/en/2008/11/03/3-billion/.

[3] William M. Vander Weyde at George Eastman House Museum, Sep. 2008. [Online]. Available: http://flickr.com/photos/george_eastman_house/sets/72157607377134096/.

[4] 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 *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI '07)*, San Jose, CA, 2007.

[5] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang, "World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections," in *Proc. 2007 Conf. Digital Libraries (JCDL '07)*, Vancouver, BC, Canada, 2007.

[6] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI '07)*, San Jose, CA, 2007.

[7] T. L. Berg and D. Forsyth, Automatic Ranking of Iconic Images, Univ. California Berkeley, 2007, Tech. Rep.

[8] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press, 2004, p. 2003.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[10] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[11] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. World Wide Web Conf.*, Apr. 2009, p. 761. [Online]. Available: http://www2009.eprints.org/77/.

[12] J. Davies, "Display; identity and the everyday: Self-presentation through digital image sharing," *Disc., Stud. Cult. Polit. Educ.*, vol. 28, no. 4, pp. 549–564, Dec. 2007.

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.* vol. 41, no. 6, pp. 391–407, 1990. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7546.

[14] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," in *Proc. 15th Int. Conf. World Wide Web (WWW '06)*, Edinburgh, U.K., 2006.

[15] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, 2001.

[16] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proc. Int. Conf. Image and Video Retrieval (CIVR '07)*, Amsterdam, The Netherlands, Jul. 2007.

[17] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries for large collections of geo-referenced photographs," in *Proc. 15th Int. Conf. World Wide Web (WWW '06)*, Edinburgh, U.K., 2006.

[18] T. Kailath, "The Divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun.Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[19] 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," in *Proc. 15th ACM Int. Conf. Multimedia (MULTIMEDIA '07)*, Augsburg, Germany, 2007.

[20] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data-Mining (KDD '06)*, Philadelphia, PA, 2006.

[21] K. Lerman, A. Plangrasopchok, and C. Wong, "Personalizing results of image search on Flickr," in *Proc. AAAI Workshop Intelligent Techniques for Web Personalization*, Vancouver, BC, Canada, 2007.

[22] K. Lerman and L. Jones, "Social browsing on Flickr," in *Proc. Int. Conf. Weblogs and Social Media (ICWSM)*, Boulder, CO, Mar. 2007.

[23] R. Lienhart and M. Slaney, "PLSA on large scale image databases," in *Proc. 2007 Int. Conf. Acoustics, Speech and Signal Processing (ICASSP '07)*, Honolulu, HI, 2007.

[24] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proc. 17th Conf. Hypertext and Hypermedia (HYPERTEXT '06)*, 2006.

[25] A. D. Miller and W. K. Edwards, "Give and take: A study of consumer photo-sharing culture and practice," in *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI'07)*, San Jose, CA, 2007.

[26] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1802–1817, Oct. 2007.

[27] R. A. Negoescu and D. Gatica-Perez, "Analyzing Flickr Groups," in *Proc. Int. Conf. Image and Video Retrieval (CIVR '08)*, Niagara Falls, ON, Canada, Jul. 2008.

[28] R. A. Negoescu and D. Gatica-Perez, "Topickr: Flickr Groups and users reloaded," in *Proc. 16th ACM Int. Conf. Multimedia (MM '08)*, Vancouver, BC, Canada, Oct. 2008.

[29] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed inference for Latent Dirichlet Allocation," *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 1081–1088, Dec. 2007.

[30] O. Nov, M. Naaman, and C. Ye, "What drives content tagging: The case of photos on Flickr," in *Proc. 26th SIGCHI Conf. Human Factors in Computing Systems (CHI '08)*, Florence, Italy, 2008.

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'07)*, 2007.

[32] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr Tags," in *Proc. 30th Int. Conf. Research and Development in Information Retrieval (SIGIR'07)*, Amsterdam, The Netherlands, 2007.

[33] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty in Artificial Intelligence (Auai '04)*, Arlington, VA, 2004.

[34] P. Schmitz, "Inducing ontology from Flickr Tags," in *Proc. Workshop Collaborative Tagging (WWW '06)*, Edinburgh, U.K., 2006, IW3C2.

[35] P. Schmitz, "Leveraging community annotations for image adaptation to small presentation formats," in *Proc. 14th ACM Int. Conf. Multimedia (MULTIMEDIA '06)*, Santa Barbara, CA, 2006.

[36] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *Proc. Int. Conf. Computer Vision*, 2005.

[37] N. A. Van House, "Flickr and public image-sharing: Distant closeness and photo exhibition," in *Proc. Extended Abstracts on Human Factors in Computing Systems (CHI'07)*, San Jose, CA, 2007.

[38] R. van Zwol, "Flickr: Who is looking," in *Proc. Int. Conf. Web Intelligence (WI '07)*, San Jose, CA, 2007.

[39] X. Wang, N. Mohanty, and A. McCallum, "Group and topic discovery from relations and their attributes," in *Proc. 19th Conf. Advances in Neural Information Processing Systems (NIPS'05)*, Vancouver, BC, Canada, Dec. 2005.

[40] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang, "PLDA: Parallel Latent Dirichlet Allocation for large-scale applications," in *Proc. 5th Int. Conf. Algorithmic Aspects in Information and Management*, 2009.

[41] L. Wu, X. S. Hua, N. Yu, W. Y. Ma, and S. Li, "Flickr distance," in *Proc. 16th ACM Int. Conf. Multimedia (MM '08)*, New York, 2008.

Radu-Andrei Negoescu received the M.Sc. degree in electrical engineering from the Polytechnic University of Bucharest, Bucharest, Romania. After several years spent in the software engineering industry, he has returned towards research, pursuing the Ph.D. degree at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, and Idiap Research Institute, Martigny, Switzerland.

His general interests include artificial intelligence, social computing, and user experience. Currently focusing on what we call social media, he has been researching models that describe large online communities, their multimedia data, and their behavior.

Daniel Gatica-Perez (S'01–M'02) received the B.S. degree in electronic engineering from the University of Puebla, Puebla, Mexico, in 1993, the M.S. degree in electrical engineering from the National University of Mexico, Mexico City, in 1996, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2001.

He is now a Senior Researcher at Idiap Research Institute, Martigny, Swizerland, where he directs the Social Computing Group. His recent work has developed statistical methods to analyze small groups at work in multisensor spaces, populations using cell phones in urban environments, and online communities in social media. He has published over 100 refereed papers in journals, books, and conferences in his research areas.

Dr. Gatica-Perez received the Yang Research Award for his doctoral work. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *Image and Vision Computing*, *Machine Vision and Applications*, and the *Journal of Ambient Intelligence and Smart Environments*.