Modeling and Understanding Communities in Online Social Media using Probabilistic Methods

THÈSE N^{O} 5059 (2011)

PRÉSENTÉE LE 1ER JUIN 2011

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE L'IDIAP
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Radu Andrei NEGOESCU

acceptée sur proposition du jury:

Prof. P. Dillenbourg, président du jury
Dr D. Gatica-Perez, directeur de thèse
Prof. S. Boll, rapporteur
Prof. J. D. R. Millan Ruiz, rapporteur
Dr R. van Zwol, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Abstract

The amount of multimedia content is on a constant increase, and people interact with each other and with content on a daily basis through social media systems. The goal of this thesis was to model and understand emerging online communities that revolve around multimedia content, more specifically photos, by using large-scale data and probabilistic models in a quantitative approach. The disertation has four contributions. First, using data from two online photo management systems, this thesis examined different aspects of the behavior of users of these systems pertaining to the uploading and sharing of photos with other users and online groups. Second, probabilistic topic models were used to model online entities, such as users and groups of users, and the new proposed representations were shown to be useful for further understanding such entities, as well as to have practical applications in search and recommendation scenarios. Third, by jointly modeling users from two different social photo systems, it was shown that differences at the level of vocabulary exist, and different sharing behaviors can be observed. Finally, by modeling online user groups as entities in a topic-based model, hyper-communities were discovered in an automatic fashion based on various topic-based representations. These hyper-communities were shown, both through an objective and a subjective evaluation with a number of users, to be generally homogeneous, and therefore likely to constitute a viable exploration technique for online communities.

Keywords : social media, online communities, probabilistic modeling, human factors

Résumé

La quantité de données multimedia est en croissance constante, et les gens interagissent constamment pas seulement avec d'autres personnes, mais aussi avec le contenu, dans des systèmes ainsi appelés du "social media." Le but de cette thèse est de modeler et comprendre des communautés virtuelles qui émergent autour du contenu multimedia, et plus précisément autour des photos, en utilisant des données de large échelle et des modèles probabilistiques, dans une approche quantitative. La dissertation apporte quatre contributions. Premièrement, en utilisant des données en provenance de deux systèmes de gestion des photos, on a analysé différents aspects du comportement des utilisateurs qui concernent le téléchargement et le partage des photos avec d'autres utilisateurs, ainsi que des groupes sociaux. Deuxièmement, des modèles probabilistiques ont été employés pour la modélisation des entités virtuelles, comme des utilisateurs et des groupes d'utilisateurs, et on a montré que les nouvelles représentations proposées sont utiles pour la caractérisation plus profonde de telles entités, ainsi que pour des applications plus directes dans des scénarios de recherche et recommandation. Troisièmement, en modelant ensemble les utilisateurs en provenance de deux systèmes de partage de photos différentes, on a montré que des différences au niveau du vocabulaire existent, et que des comportements de partage différentes peuvent aussi être observés. Finalement, en modelant des groupes virtuelles d'utilisateurs comme entités dans un modèle probabilistique basé sur des thèmes abstraits, on a découvert automatiquement des hyper-communautés homogènes, à partir de la représentation proposée dans le modèle.

Mots-clés : social media, communautés virtuelles, modélisation probabilistique, facteurs humains

In a world that is constantly changing, there is no one subject or set of subjects that will serve you for the foreseeable future, let alone for the rest of your life. The most important skill to acquire now is learning how to learn.

John Naisbitt

Acknowledgements

In the wake of my private defence, before even fixing the parts that needed fixing, I set to writing the Acknowledgments section, eager to pour out the feelings of gratitude that still ricocheted inside myself.

I'll start with the person closest to this thesis than anyone else apart from myself: my supervisor and, I dare say, friend, Daniel. An example of condensed social skills and intellect, he's the one that made my life as a doctoral student not only bearable, but enjoyable. It is because of him that the "normal" cycle of excitement and depression in the life of a PhD student was rather disrupted, and in a positive sense that is. Lucky the student who finds a role model in his mentor.

I was also blessed with a jury of incredible value, and I thank them from the bottom of my heart for agreeing to be the ones to bring me to shore. Their comments during the defence were eye-opening and the discussion we had was worth the four years' wait. Pierre Dillenbourg was a great jury president, and made sure everything went smoothly. Roelof van Zwol, highly learned opponent as the Dutch say, was, strange as it may seem, a reassuring presence, to whom I also owe a constant ringing at the back of my brain who prompts me to ask myself, when considering any new idea, "how would Roelof push this to Yahoo!?" Speaking of brain, José del R. Millán was everything I expected him to be, and gave me invaluable insights about ways of improving my work. Last, and by no means least, Susanne Boll, who made an extra effort to be part of this jury, and brought along what must be the youngest juror ever. I am humbled and eternally grateful for their presence and help during this trial.

A big thank you to Ingrid Le Duc, whose seminar on preparing the thesis defence has put me ahead of the game and reduced my stress level to nearly zero (on a scale accordingly offset...), and who would not believe that despite all she taught me I had nearly two slides per minute.

Several other people had a contribution towards the scientific content of this thesis, and they receive well deserved thanks for their patience and commitment in working with me from the other side of the world: Svetha Venkatesh, Alexander Loui, Dinh Phung and Brett Adams.

Our parents are the foundation on which we build our accomplishments. Time and again have I given silent thanks to my mom for being the smart woman that she is, and for having opened countless windows of opportunity in my youth towards foreign languages and intellectual jousting – some of my friends probably regret the latter. It is therefore time to thank my mom openly and in writing. Thank you mom! And also a big "thank you!" to my adoptive dad, who was of great support during these years, making sure I had nothing to worry about back home.

Far from home, there was an emptiness to be filled, and there to do just that were scores of friends and colleagues with whom I've crossed paths during these years. A few ones stuck closely to my heart, and with the risk of forgetting some so deeply inserted in there that I don't even see them anymore, I'll thank them for being my friends, and putting up with my pre-doctoral manic syndromes. Paco, who brought me a sombrero all the way from Mexico. Dayra, who managed to survive as my flat mate for two years (and counting). Elisa, who would never say "yes", at least not easily, no matter what the question was. Tatiana who doesn't like my pasta, but is always open for a bear hug. Joan, whom I thought I'd kill after our first five minutes together (actually, I'm still considering it), but who was a most amazing travel companion in China and a friend ever since. Francesco, who is the incarnation of bounds and the reason why now my phone is off on weekend mornings. Kate, my first office mate and prefered target of survival questions, and the source of my mustache envies, Dinesh. Chris, who was far from being inobservant of my jactitations as I was struggling to include sesquipedalian words such as antidisestablishmentarianism in my thesis, will forever remain the doctor with the coolest name.

At Idiap, Nadine and Sylvie, our guardian angels, and candy providers, were the best way to alleviate Monday morning's depression. Well, any morning's depression. Florent, who kick-started me when I arrived at Idiap, and has luckily been there anytime I had questions. Olivier, whose extension anyone should know by heart if they have questions without answers.

We don't often take into account the dangers that research into social media may pose to one's personal life, hobbies, and budget. Working with Flickr has transformed me into a passionate photographer, and if I say photography, I must say Basilio. Sometimes called my evil twin for no good reason, for I'm certainly better looking and undoubtedly meaner, he is an example of generosity and creativity. I honestly don't know anyone else who can show me "something cool" (that is actually "awesomely cool") every time I drop by his office. Another Elisa was my most honest critic during my growth as a photographer, and I hope she'll continue on that trend.

Of course none of this would have been possible without the financial support of the Swiss National Science Foundation, through the MULTI and HAI projects, as well as Idiap's.

A special thought for my very dear friend George, his wife Adina, and their kid Alex, who visited me often and will always be in my thoughts.

To all of you who know me, and who touched me in one way or another, thank you!

Contents

List of Figures

List of Tables

Glossary

AP	Affinity Propagation
API	Application Programming Interface
ATM	Author-Topic Model
BM-LDA	Binary membership based Latent Dirichlet Allocation model
FS	Flickr Search
GTM	Group-Topic Model
HDR	High Dynamic Range
HMM	Hidden Markov Model
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MAP	Mean Average Precision
MM-LDA	Multiple membership based Latent Dirichlet Allocation model
PLSA	Probabilistic Latent Semantic Analysis
SM	Social Media
TB-LDA	Tag based Latent Dirichlet Allocation model
TEDS	Tag-Entity Distance Search
TES	Topic Expert Search

1

Introduction

It is nowadays common to hear people say "you can find anything on the internet." While this may not hold true for *really* anything, the amount of information available electronically is staggering. In the past few years we have witnessed an increase in the amount of interaction people have over the web, be it simple communication like email, voice calls, instant messages, or more complex forms like blogs, video logs, or podcasts. Systems such as Twitter and Facebook have hundreds of millions of users worldwide, and they have changed the way information is disseminated. From the earthquakes in Haiti and Chile to the revolutions in Iran and Egypt, from academic research to marketing, *social media* has become as big a medium as traditional media like TV networks and newspapers, with the advantage of nearly-instantaneous propagation of information and little to zero entry barriers for content creation.

Although no particular definition has yet become universally accepted, it is relatively easy to agree that the two main components of social media are the ones referred to in the name itself: social relations, and (multi-)media content.

1.1 Content

In 2008, the George Eastman House Museum released online under a Creative Commons license a few hundred photos from their collection of 1,400 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 comes from regular people. The *William M. Vander Weyde* photoset from the George Eastman House on Flickr 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." [1]

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 the next level of the 1890s revolution of photography, where everybody is a photographer. And because of the new ways in which people interact in this digital era, many of the photos and videos end up being available online.

As the web got more and more popular, and internet connections got more affordable and widespread, content creation was no longer a privilege of traditional content creators, such as magazines, newspapers, television networks, radio networks, or other professionals. We witnessed a democratization of content generation, in which common people become photographers, videographers, and journalists. The transition from "My home page" of the late '90s to "My YouTube channel" of 2010 was gradual, but steady. As of the writing of this thesis, YouTube, arguably the world's most popular video sharing website, was boasting 24 hours worth of video uploads each minute [73]. Flickr on the other hand, one of the most popular photo sharing websites, showed an average of 2000 photos uploaded per minute in September 2010. This average was computed by counting the time it took Flickr users to upload 1 billion photos starting from the 4 billionth one in October 2009 to the 5 billionth one, less than a year later, in September 2010.

Most of the content generated by regular users consists of text (blogs, journals, etc.), images (holiday photos, creative photography, photojournalism, social activism, etc.), and video (musical performances, photojournalism, video blogs, comic shows, etc.). However, one of the more interesting aspects of social media today, apart from the sheer amount of data available online, is the even richer metadata. People create, upload, and then annotate content, be it through tags, regions of interest, or comments. Annotation is just one of the explicit ways through which metadata is created - and the most visible, but many other actions also lead to the generation of metadata: rating, adding photos or videos to favorite lists, organizing content in sets or collections, and even simply viewing content. This type of implicit metadata is also useful for the understanding of the dynamics of social media.

With this extraordinary increase in the amount of data, one of the biggest problems for consumers of content is now finding relevant content. We are witnessing the transition from the “search & find” mindset to the “subscribe & receive” one, in which content gets delivered to users according to their interests and preferences. Therefore, being able to understand and algorithmically model user interests and preferences is becoming a fundamental goal.

1.2 Social relations

Another way in which social media collections can be enriched is through the social sphere itself, like user-to-user relationships, or relationships with online communities, such as Flickr Groups. The success of companies such as Facebook, Foursquare, or Flickr, in which the accent is placed on online interaction between users of the system, is witness to the importance that online social relations have in today’s society.

With the adoption of internet-enabled smart phones, not only the content, but also the number of people who are “connected” at any given moment of the day (or night!) has increased tremendously. Users connect to users through comments, favorite links, shared resources, or simply by belonging to the same communities. Users also connect to communities, be them local or global, charitable, corporatist, or governmental. As mentioned earlier, this fact is in itself a rich source of data, and researchers have started taking advantage of it.

For example, a relatively simple piece of information – the number of groups a user belongs to – helped establish statistically significant evidence to support qualitative research about tagging behavior [54]. In another case, metadata stemming from group membership helped understand that photo popularity is closely related to the social networking behavior of the users [67].

Just as finding content is a major problem for today's content consumers, from a social point of view, the issue of finding relevant communities, or even individuals, is equally important. Thus the ability to understand and model communities, as well as their interests, is another important need of today's society.

We will have a look in the following section at one of the most representative social media outlets that we have already mentioned a few times, namely Flickr.

1.3 Flickr.com

Flickr (flickr.com) is an online photo management website, created in 2004, and later acquired by Yahoo!. In its relatively short existence, Flickr grew at an extremely rapid rate, reaching 3 billion hosted photos in November 2008, 4 billion in October 2009, and 5 billion in September 2010. Its growth and popularity were accompanied, and maybe encouraged, by a philosophy of openness. By default, photos uploaded to Flickr are public, and the company also provides an application programming interface (API) for third party developers, which helped make Flickr popular with the research community.

Part of Flickr's success seems to be related to the way photos are central to the website experience: rather than becoming an online photo album storage site like other existing systems, Flickr from the very beginning encouraged users to share their photos with the rest of the world. Why exactly users share photos online is a question that received tentative answers in several recent studies like those of Van House [66] or Ames and Naaman [5]. Apart from simply needing a place to store their photos online, most users seem driven by social motives like self-expression and self-promotion, and social interaction is indeed one of the key aspects of the Flickr experience. Users can explore random photos automatically ranked by "interestingness", leave comments, add tags, mark-up regions within a photo with notes, list a specific photo as a favorite, or add other users as contacts. All these public displays of interaction have a strong impact on community building.

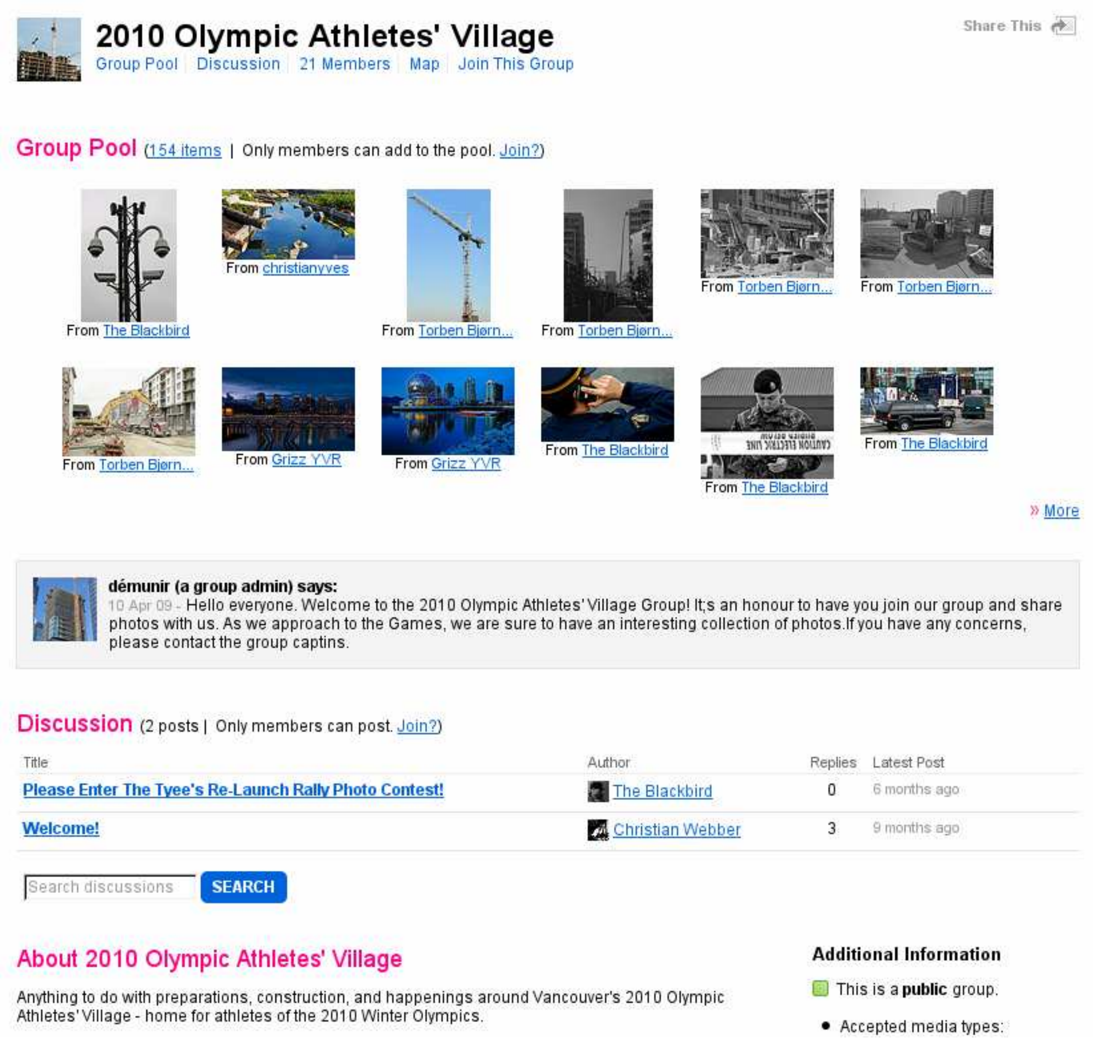
In Flickr, apart from the previously described interaction mechanisms, users can organize themselves in self-managed communities, called *Flickr Groups*. As the name suggests, Flickr groups are sets of users who are brought together by a common feature, and who share photos in the so called *group pool*. There is quite a wide range of interests that bring people together. As a few examples, it may be an interest in a specific kind of photographic technique, leading to groups like *Black and White*, *Closer and Closer Macro Photography*, or *Digital Infrared*. It can also be an interest in a specific geographical location, with groups like *New York City*, *Paris*, or even smaller geographic communities, like the *2010 Olympic Athletes' Village*, shown here in Figure 1.1. Yet again, it may be an interest in promoting one's images, which leads to groups like *Views 7-25* or *Views 1250-1500*, groups in which people share their photos that have reached the respective number of views on Flickr. Within the same category there are groups like *Nature's Finest (Invited Images Only)*, or *Your Best Shot 2009*, shown in Figure 1.2. These are communities of sometimes tens of thousands of people, whose common goal is the gathering and exposure of high quality photography. The list of interests that brings people together in groups does not end here: there are the charity oriented groups, the photojournalism ones, the groups for political activism, and even the groups for corporate marketing.

1.4 Contributions

The overall goal of this thesis is to find ways of modeling and understanding emerging online communities by using large-scale data and probabilistic models for a quantitative approach, as opposed to traditional sociological studies that are qualitative in nature.

Compared to state of the art in online community modeling and understanding, we make in this work several contributions, more specifically:

1. We analyze one of the biggest social media outlets dedicated to photography, namely Flickr.com, from a photo-sharing point of view, showing that users share an important fraction of their photos with groups, online communities centered around a specific topic or activity. On a dataset of more than 22,000 users, 50,000



2010 Olympic Athletes' Village

Group Pool | Discussion | 21 Members | Map | Join This Group

Share This

Group Pool (154 items | Only members can add to the pool. Join?)

From The Blackbird

From christianyves

From Torben Bjørn...

From Torben Bjørn...

From Torben Bjørn...

From Torben Bjørn...

From Torben Bjørn...

From Grizz YVR

From Grizz YVR

From The Blackbird

From The Blackbird

From The Blackbird

» More

démunir (a group admin) says:
10 Apr 09 - Hello everyone. Welcome to the 2010 Olympic Athletes' Village Group! It;s an honour to have you join our group and share photos with us. As we approach to the Games, we are sure to have an interesting collection of photos.If you have any concerns, please contact the group captins.

Discussion (2 posts | Only members can post. Join?)

Title	Author	Replies	Latest Post
Please Enter The Tyee's Re-Launch Rally Photo Contest!	The Blackbird	0	6 months ago
Welcome!	Christian Webber	3	9 months ago

Search discussions SEARCH

About 2010 Olympic Athletes' Village

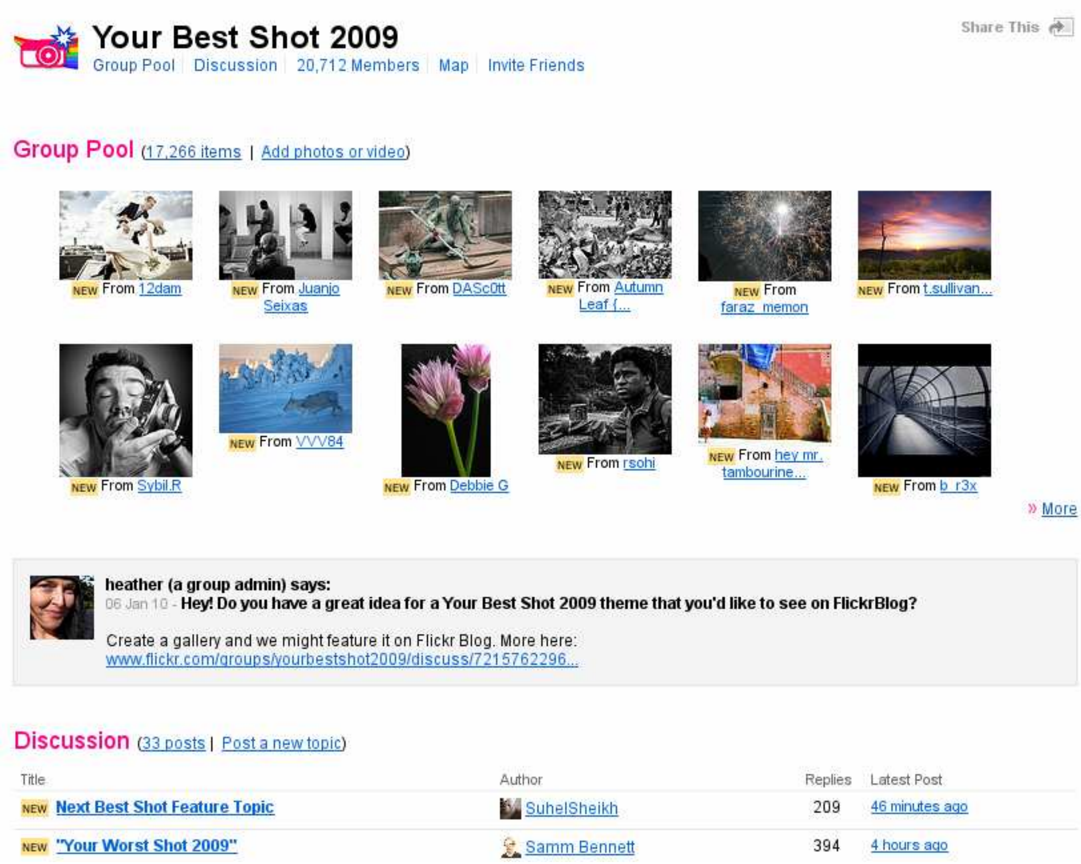
Anything to do with preparations, construction, and happenings around Vancouver's 2010 Olympic Athletes' Village - home for athletes of the 2010 Winter Olympics.

Additional Information

🟩 This is a **public** group.

● Accepted media types:

Figure 1.1: ***2010 Olympic Village*** - Home page of the group *2010 Olympic Athletes' Village*, a very small group with 21 members and just over 150 images. This group is clearly a special interest group, focused on the 2010 Olympic Village.



heather (a group admin) says:
06 Jan 10 - **Hey! Do you have a great idea for a Your Best Shot 2009 theme that you'd like to see on FlickrBlog?**

Create a gallery and we might feature it on Flickr Blog. More here:
www.flickr.com/groups/yourbestshot2009/discuss/7215762296...

Title	Author	Replies	Latest Post
NEW **Next Best Shot Feature Topic**	SuhelSheikh	209	46 minutes ago
NEW **"Your Worst Shot 2009"**	Samm Bennett	394	4 hours ago

Figure 1.2: *Your Best Shot 2009* - Home page of the group *Your Best Shot 2009*, a group with over 20,000 members and more than 17,000 images. The focus in this group is on social interaction via exposure of high quality photography rather than on a specific subject.

groups, 7 million photos, and 2 million unique tags, we examine in detail the sharing behaviors of users. We find consistent patterns for users regardless of their membership status (paying or free), and we show that about half of the users use groups as exposure vehicles, and thus participate actively in creating content pools. We also perform a comparative analysis of two large photo management systems, Flickr and Kodak Gallery, with a combined dataset of roughly 30,000 users and 5 million photos, and find that despite some differences, we can also find similarities at the level of vocabulary. We present this in-depth analysis in Chapter 2.

2. We design a novel joint topic-based representation for users and communities of users, starting from bags-of-tags representations and using a Probabilistic Latent Semantic Analysis (PLSA) model, and we show this to be useful in comparing and finding similar entities at a more abstract level. Furthermore, the new joint topic-based representation of groups and users allows us to directly compare any of these entities to each other from a content perspective, which leads to the possibility to discover similar entities. We also propose a few applications based on our joint representation, allowing entity discovery, or search-through-topic extensions of traditional search-by-keyword paradigms. This work is presented in Chapter 3.

3. We perform a novel human-centered comparative analysis of two major social media outlets, Flickr and Kodak Gallery. Our analysis involves the use of a topic-based representation, this time based on a latent Dirichlet allocation (LDA) model. We show that the effects of the users' motivations and needs can be strongly observed in this large-scale data, in the form of what we call Kodak Moments and Flickr Diamonds. We believe that this study also points out one of the great potentials that social computing holds for future research. Large-scale studies are nowadays much easier to perform, unlike ethnographic studies which are usually small-scale, given their time and subject-effort intensive nature. Large-scale analysis may therefore become the first step in the research process, with in-depth ethnographic studies as a second step once a preliminary hypothesis has been chosen for verification. This analysis is presented in Chapter 4.

4. We propose a framework to detect communities of communities (that we call *hypergroups*), and we show through user studies that this technique may be reasonably well suited in community recommendation scenarios. We use our topic-based

representation of Flickr groups based on an LDA model in order to discover hypergroups. We propose three different models for the topic-based representation and employ a state of the art clustering algorithm to automatically determine the number and composition of the hypergroups. We compare these models numerically, and we also select two of them for a user study with eight annotators, showing that automatic methods applied to large scale data may be feasible for community understanding and social recommendation. We present these results in Chapter 5.

The contributions presented in a unified form in this thesis have been mostly disseminated in the research community through the following publications:

1. R. A. Negoescu and D. Gatica-Perez, "Analyzing Flickr Groups," in CIVR '08: Proc. of the Intl. Conf. on Image and Video Retrieval, 2008.
2. R. A. Negoescu and D. Gatica-Perez, "Topickr: Flickr Groups and Users Reloaded," in MM '08: Proc. of the 16th ACM Intl. Conf. on Multimedia, 2008.
3. R. A. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez, "Flickr Hypergroups," in MM '09: Proc. of the 17th ACM Intl. Conf. on Multimedia, 2009.
4. R. A. Negoescu, A. C. Loui, and D. Gatica-Perez, "Kodak Moments and Flickr Diamonds: how Users Shape Large-scale Media," in MM'10: Proc. of the 18th ACM Intl. Conf. on Multimedia, 2010.
5. R. A. Negoescu and D. Gatica-Perez, "Modeling Flickr Communities Through Probabilistic Topic-Based Analysis," IEEE Transactions on Multimedia, vol. 12, no. 5, pp. 399-416, 2010.
6. R. A. Negoescu and D. Gatica-Perez, "Flickr Groups: Multimedia Communities for Multimedia Analysis," in: Internet Multimedia Search and Mining, Bentham Science Publishers, in press.

2

Analyzing and understanding online photo communities

When faced with new phenomena, the first steps involve almost always attentive observation and analysis, a search for structure, and a desire to better understand the system under scrutiny. As people start using online social media systems to upload, annotate, share, and discover digital content, new behavioral patterns emerge, and so does the desire and often also the need to understand these patterns. Hugely successful online companies such as Flickr or Facebook, with millions of users, need to understand usage patterns in order to deploy their hardware and bandwidth resources accordingly, while online advertisers, brands, and policy makers desire to understand these new patterns in order to exploit them to their own advantage.

In this chapter we take a look at two of the most popular photo management online systems, Flickr and Kodak Gallery, and analyze the behavior of their users, the content, and metadata they generate. Our main goal is to understand what are the factors that influence sharing behavior in such systems, and what the sharing behavior actually is. We are also interested in understanding how design choices of a system impact the vocabulary and sharing behaviors of its users. We find through our analysis of a large-scale dataset collected from Flickr that users who participate in social scenes share on average roughly 30% of their online photo collections with online communities organized as Flickr groups. We also show that system design choices, such as default privacy levels and tagging system implementation have a qualitative impact on the emerging vocabulary. The first part of the work described in this chapter was done in 2007, while the comparative analysis was performed in 2010, and they appeared in modified form in [46, 50].

The structure of the chapter is as follows: we first review related work, then we perform an in-depth analysis of photo sharing behaviors in Flickr in Section 2.2, and we present a large-scale comparative analysis of Flickr and Kodak Gallery in Section 2.3.

2.1 Related work

Flickr data has started to be used in the context of classic content-based image retrieval research [38]. However, one of the most interesting aspects of Flickr, apart from the sheer size of its data, is the plethora of metadata associated with photos, in the form of tags, notes, number of views, comments, number of people who mark the photo as a favorite, and even geographical location data. Recent studies have used notes [64], combinations of tags, geolocation and visual data in order to improve retrieval [6, 56], visualization, and summarization techniques for large databases either over time or over a geographic area [3, 23, 30, 33], to automatically extract place and event semantics [60], or to induce tag ontologies [63] and disambiguate tags [71].

Tagging systems have been analyzed by Marlow *et al.* [41], and a taxonomy of users' motivations to tag has been proposed by Ames and Naaman in [5]. There have also been some studies analyzing the sharing practices, motivations, and privacy concerns of the users [2, 42, 66]. In particular, Van House [66] discusses the main uses of photo sharing amongst users on Flickr. While these studies provide particularly useful insights into user behavior, none of them explicitly address sharing practices in relation to Flickr groups, as we do here.

In addition to the photo metadata, attention has been given to metadata stemming from the (social) links existing on Flickr [34, 35, 36, 67]. Recent work includes studying user-to-user relations by means of contact bookmarking, a direction explored by Kumar *et al.* [34], with interesting results regarding the structure of the Flickr social network. Other works have considered user-to-photo relations by means of ownership, favorites, or comments. Van Zwol [67] analyzes the way new photos are discovered by users on Flickr, and finds that most photo views and comments occur in the first two days after the upload, concluding that the social network of the user and photo pooling (i.e. sharing with groups) are two major indicators of a photo's popularity. In a similar study, Lerman and Jones [36] found that the number of views a photo receives correlates strongly with the size of the social network of a user, and more particularly the reverse contacts Since the social links are not necessarily bi-directional, reverse contacts are the people who call the target user a contact. Lerman *et al.* [35] use a user's existing social network and a latent topic model on tags in order to filter tag search results for that specific user.

In summary, compared to our work, previous works have either exploited different social link information or targeted different goals. At the same time, some of the findings in [36, 66, 67] provide us with a starting point about the user motivations for using the Flickr group functionality, and in understanding why new representations for groups are needed.

2.2 Analyzing Flickr Groups

2.2.1 What are Flickr groups?

The word "group" has several definitions in the English language, but we find two of them to be most representative for Flickr groups, from the American heritage dictionary of the English language [57]:

1. An assemblage of persons or objects gathered or located together;

2. A number of individuals or things considered together because of similarities.

A group is therefore a collection of persons or objects, who are either in physical proximity or share some abstract characteristics. In Flickr, from a strictly technical point of view, groups are collections of users who freely choose to join such a community. The main purpose of groups is to facilitate the sharing of user photos in what is called the *group pool*. This is a collection of photos shared by any member with the group, and, implicitly, all the tags associated with the photo become part of the group photo pool. One can distinguish between several types of groups, which may sometimes be intertwined. A short, non-exhaustive list could include:

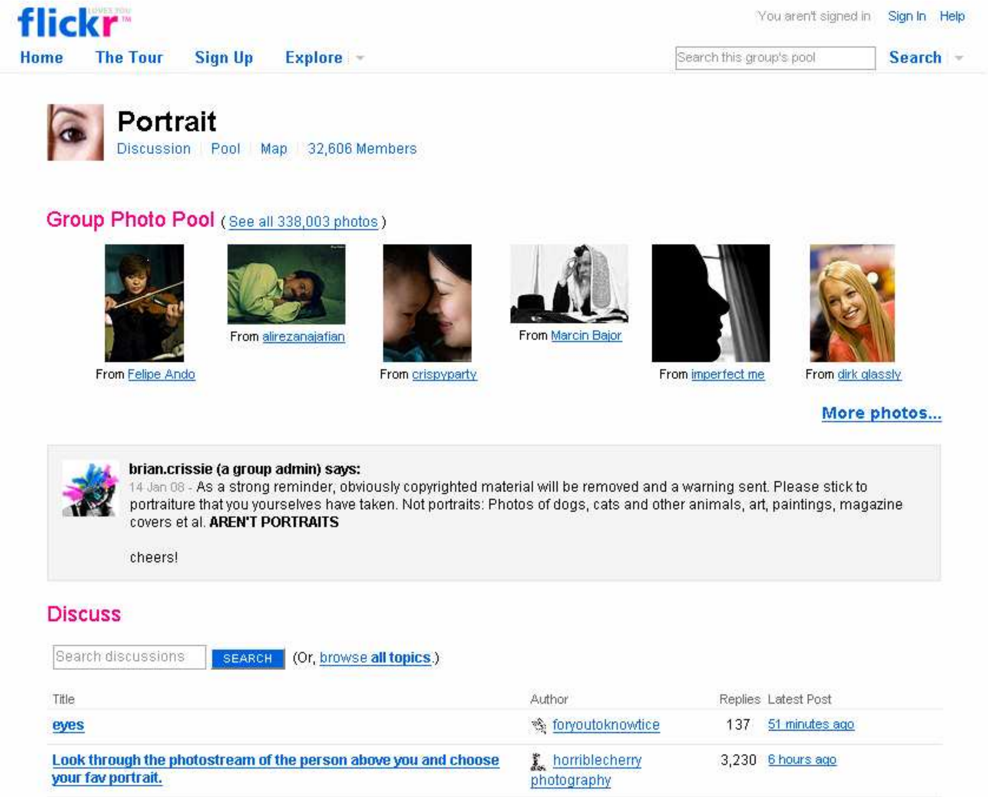
- ***geographical/event groups***: groups limited to a geographical region or a specific event (local or global), such as *New York City*, *San Francisco Bay*, *Switzerland*, *Live Music*, *World Events (festivals, protests, etc.)*, *Global Photojournalism*;
- ***content groups***: groups primarily oriented towards the visual content being shared, such as *R is for Red*, *Leaves (No Trees Please!)*, *Cats - Small to Big*, *Artistic Child Photography*;
- ***visual style groups***: groups that concentrate on a specific photographic technique, for example *Life in Black and White*, *Closer and Closer Macro Photography*;
- ***quality indicator groups***: groups that identify and regroup (perceived) high quality photography, such as *Blue Ribbon Photography [Invited Images ONLY]*, *Superb Masterpiece - Invited pictures only (Vote Now!)*, *The Best: BRAVO (INVITED images only)*, *Flickrs Best (Better than Explore!) - (Invite or Award Only)*;
- ***catch-all groups***: groups that do not seem to have any particular content-oriented rules, but rather they are an invitation for users to share photos in groups. They usually have huge numbers of users and photos: *Flickr Central*, *10 Million Photos*, *The Biggest Group! - Playground for Psychotics!*.

Figure 2.1 shows the home page of a content group, *Portrait*. When users join a group, they can start sharing photos in the group pool. There are three privacy settings for groups: (1) public, meaning that anyone can see the group photo pool, and anyone can join; (2) public, but requiring an invitation from a member; and (3) private, meaning that nobody can find the group, and a user must be invited to join.

2.2.2 Dataset

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

Our dataset consists of approximately 22,000 registered Flickr users, roughly 7 million photos belonging to these users (the most recent 500 photos per user), and about 23 million tags belonging to these photos. We chose to limit the number of photos to the



brian.crissie (a group admin) says:

14 Jan 08 - As a strong reminder, obviously copyrighted material will be removed and a warning sent. Please stick to portraiture that you yourselves have taken. Not portraits: Photos of dogs, cats and other animals, art, paintings, magazine covers et al. **AREN'T PORTRAITS**

cheers!

Title	Author	Replies	Latest Post
eyes	foryoutoknowtice	137	51 minutes ago
Look through the photostream of the person above you and choose your fav portrait.	horriblecherry photography	3,230	6 hours ago

Figure 2.1: Home page of a content group in Flickr, *Portrait*. - Users can see the number of members of the group (in this case more than 30,000!), the number of photos (more than 330,000) and a preview of the latest additions, a short group message from the group admins, and a preview of the group discussion.

most recent 500 primarily to facilitate the data collection process. However, as it has been pointed out in [66], most users see Flickr as a social site, and are mainly interested in the most recent photos (theirs, and their contacts'). The data collection process can be described as follows: we repeatedly retrieved the first approximately 3,900 photos uploaded from a randomly sampled moment t in the interval December 22nd, 2004 - April 2nd, 2007, until information on roughly 187,000 photos was collected. We have thus obtained 22,414 distinct users, the owners of the photos. For each of the users we have retrieved their most recent 500 photos which, in some cases, meant all their photos (and less than 500), for a total of nearly 7 million photos. We have then collected all the tags associated with these photos. Only about 4.7 million photos have at least one tag. In addition to the users, photos, and tags, we have also collected information about the groups the photos belong to, with 1.13 million photos belonging to at least one group. So, although the initial process is random, users' photos and their tags are somewhat complete, giving us insights into the most recent behavior of the users.

Let us formalize the definition of this original dataset (D_O):

– users: $U = \{U_i \mid i = 1...N_U\}$ with $N_U = |U| = 22{,}414$ the number of users;

– groups: $G = \{G_i \mid i = 1...N_G\}$ with $N_G = |G| = 51{,}407$ the number of groups;

– photos: $P = \{P_i \mid i = 1...N_P\}$ with $N_P = |P| = 6{,}926{,}622$ the number of photos;

– tags: $T = \{T_i \mid i = 1...N_T\}$ with $N_T = |T| = 1{,}969{,}813$ the number of distinct tags.

2.2.3 Sharing behaviors

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

– $U_{i,p}$: the total number of photos in user U_i's collection;

– $U_{i,s}$: the total number of photos user U_i shares with groups;

– $U_{i,g}$: the total number of distinct groups in which user U_i shares photos;

– $U_{i,\sigma}$: the total number of sharing instances; this is the count of all photo-group pairs for user U_i.

Using the above notations, we can write the following:

– $\gamma = \frac{U_{i,\sigma}}{U_{i,s}}$: the average number of groups a photo is shared with, for user U_i;

– $\pi = \frac{U_{i,\sigma}}{U_{i,g}}$: the average number of photos shared per group, for user U_i.

With the above definitions, we now set out to address a few key questions about sharing behavior in Flickr groups.

2.2.3.1 To share or not to share?

Figure 2.2 shows the histogram of photos shared with groups by the users in our dataset. Of the 22,414 users in the snapshot, 50.9% share at least one photo with at least one group, 26.4% share more than 50 photos and 9.9% share more than 200 photos. For the full dataset, the average number of photos shared with groups is 54.6. If we only consider the users who actually share any photos with groups, this average is 106.4 photos. Figure 2.3 shows the distribution of the percentages of shared photos for the users who share photos with groups. This is the ratio between the number of shared photos and the total number of the user photos, $\frac{U_{i,s}}{U_{i,p}}$. About a quarter of the users share at least 50.1% of their photos in groups, while almost half share at least 17.2% of their photos. The mean sharing percentage is 29.6%. We consider this to be an indication that sharing photos with groups is an important part of the photo sharing practices of Flickr users. To the best of our knowledge, user motivations for sharing photos with groups have not yet been analyzed, however motivations for tagging photos and uses of personal photography have. Four main uses of personal photography have been observed by van House [66]: (i) *memory, identity, and narrative*, (ii) *maintaining relationships*, (iii) *self representation*, and (iv) *self expression*. We believe that out of these four uses, self expression (or *photo exhibition*) and maintaining relationships are the ones driving users to share photos with groups. Groups ensure a higher exposure of the photos, and it is common practice for thematic groups to require their members to comment on the photo posted in the group pool just before their own. Group photo pools also allow users who have an interest in a specific topic to have a regular photo stream focused on that topic. Some other groups are not thematic, but rather geographically localized, and users sometimes organize offline meetings, creating and maintaining new relationships.

In order to understand whether the size of a user's photo collection influences his or her percentage of shared photos, we have analyzed the relation between these two measures. This is shown in Figure 2.4. The sizes of the photo collections for users who share no photos at all are spread over the entire range of sizes (the thick line overlapping the x axis), and the sharing percentages for the users who have the maximum number of

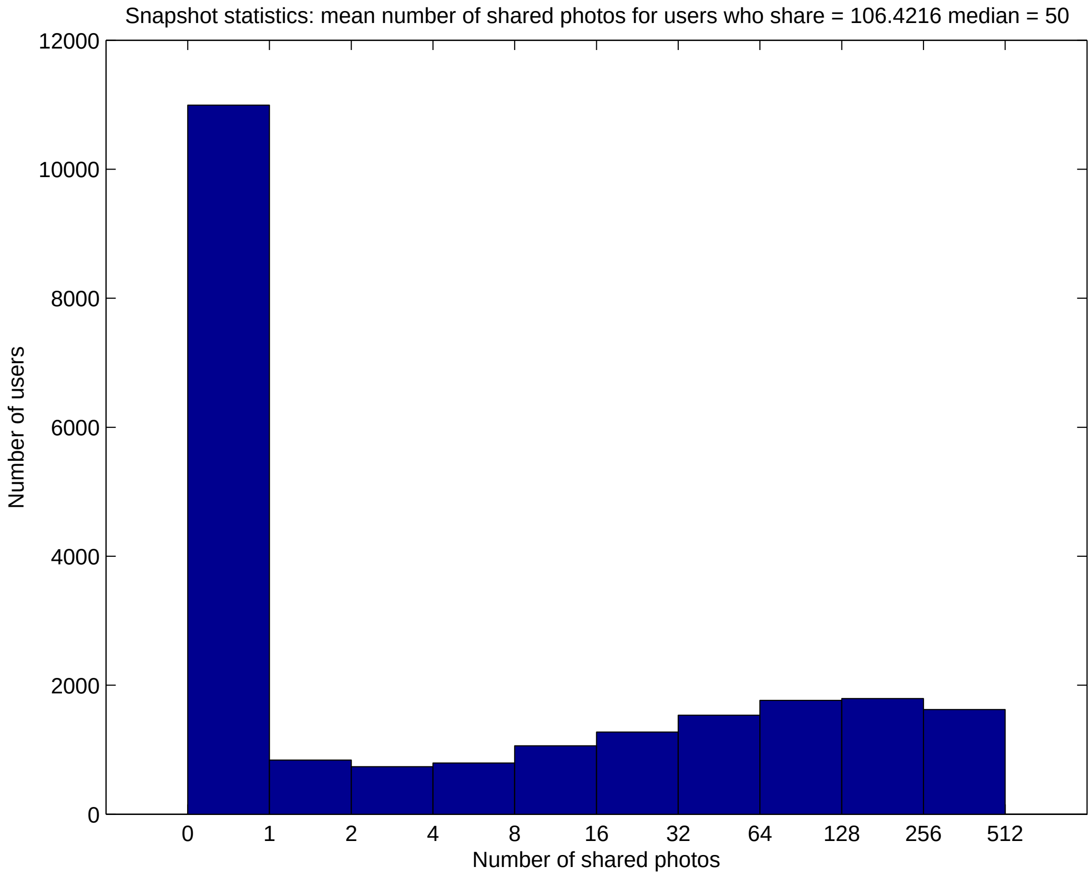


Figure 2.2: Photos shared with groups - Histogram of the number of photos shared with groups $U_{i,s}$, including the users who have not shared any photos. The average number of shared photos is 54.6. The x axis is shown in log-2 scale for displaying reasons.

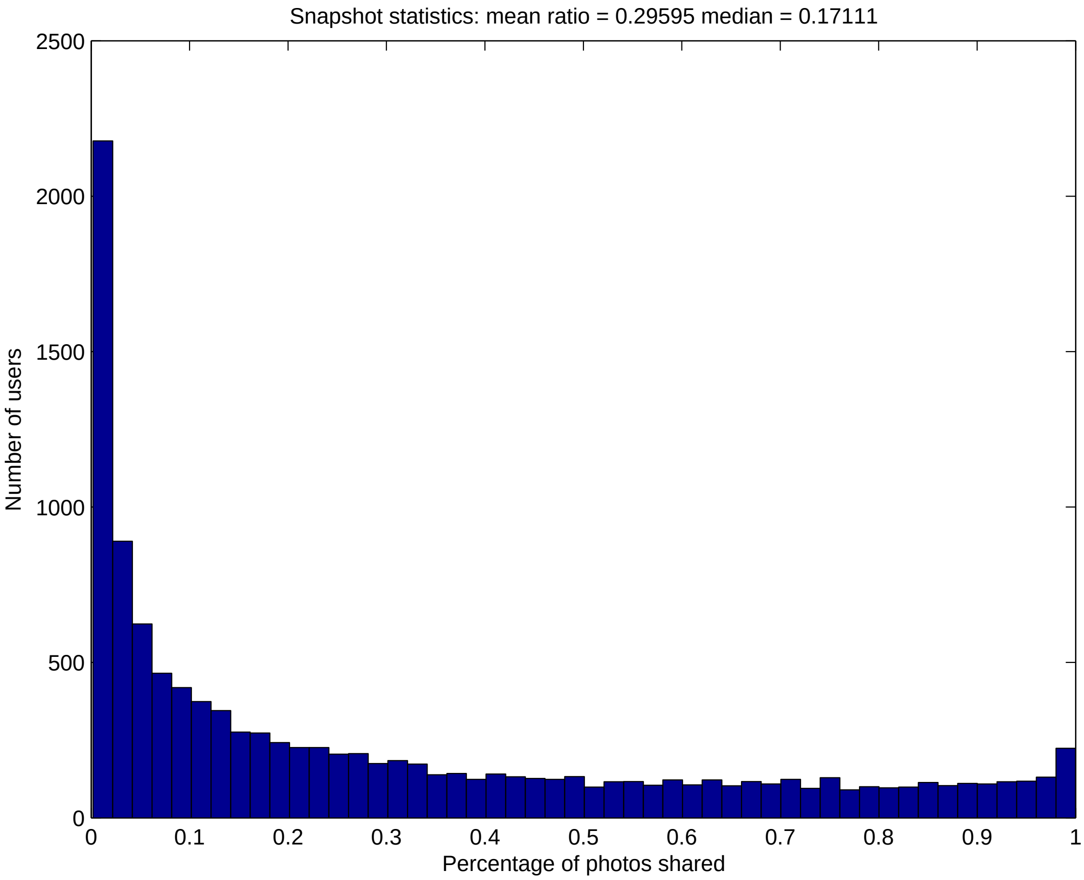


Figure 2.3: Percentage of photos shared with groups - Histogram of $\frac{U_{i,s}}{U_{i,p}}$, the percentage of photos shared with groups, for sharing users. The mean sharing percentage is 29.6%, and the median is 17.1%.

photos allowed in our dataset are also spread over the entire interval $[0, 1]$ (the thick line at $x = 500$). The correlation coefficient between the two measures is 0.1417, indicating a rather weak correlation.

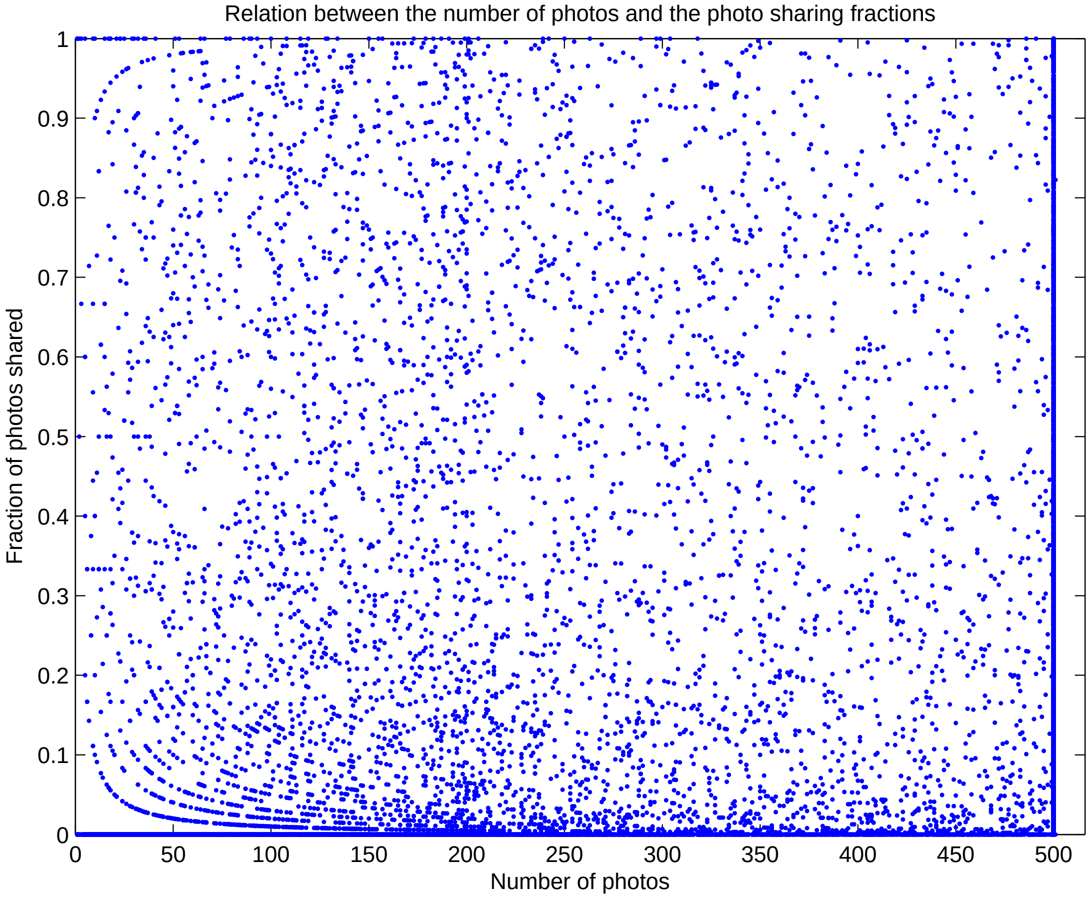


Figure 2.4: Correlation between collection size and group-sharing percentage - The percentage of shared photos (x-axis) vs. the number of photos of each user (the y-axis): the size of the collection of photos for users who do not share any photos at all ($U_{i,s} = 0$) is spread over the entire range of sizes $U_{i,p} \in [1, 500]$; the sharing percentages for users who have the maximum number of photos ($U_{i,p} = 500$) is spread over the full interval $[0, 1]$.

2.2.3.2 Group affiliation through photo sharing: how many groups does a user share photos with?

As pointed out in the previous paragraphs, 50.9% of the users share at least one photo in at least one group. Figure 2.5 shows a histogram of the absolute number of groups users share photos with. For the full dataset, users share their photos with an average of 25.3 distinct groups. If we only consider the users who actually share photos in groups, the average number of groups with which they share photos is 49.6, with a median of 16. Out of all sharing users, 15.1% share their photos with exactly one group, and 45.6% of them share photos with more than 20 groups. A relatively important part of the sharing users, 11.3%, share photos with more than 140 distinct groups.

This highlights two trends: (1) roughly half of the people do not share with groups at all, and (2) half of the users do, and exploit this feature by affiliating with several groups. In the half that shares, several distinct behaviors also emerge: moderate sharers, with fewer than 5 groups, average sharers, and extreme sharers, with hundreds of groups.

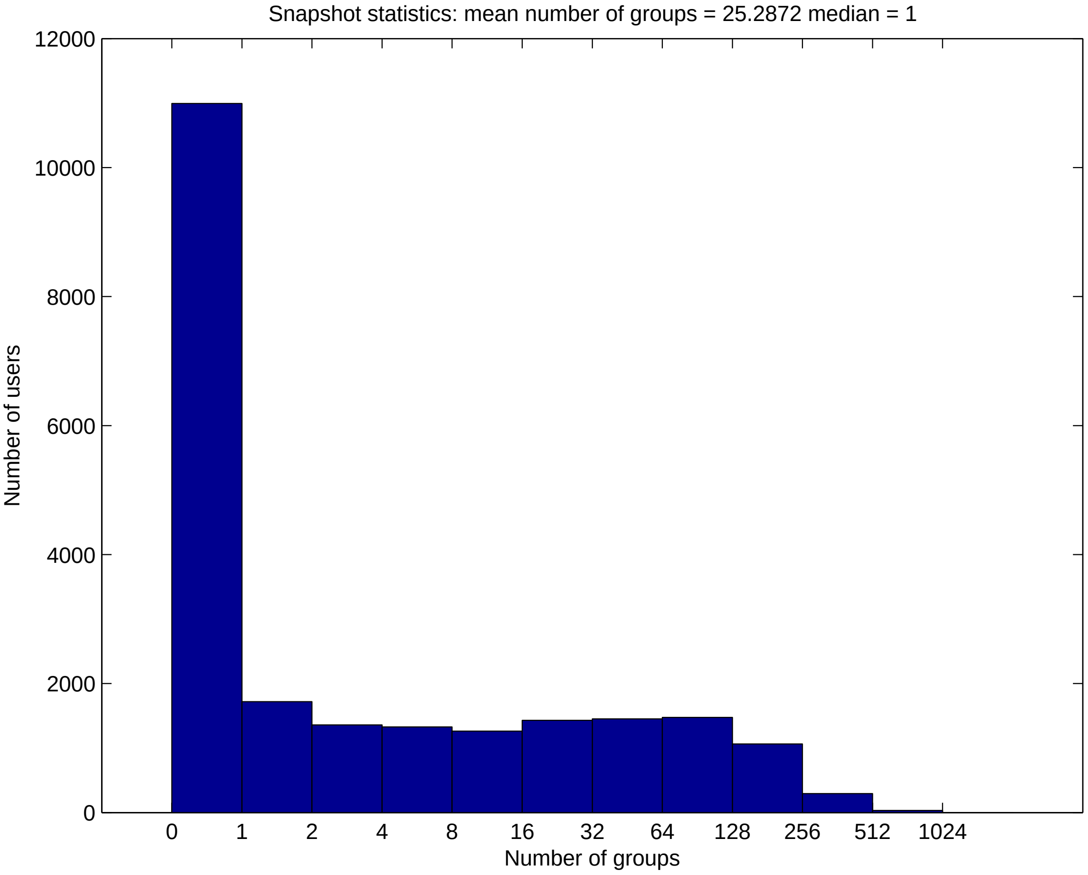


Figure 2.5: Sharing with groups - Histogram of the number of groups photos are shared with per user. The x-axis is in log-2 scale, and the average number of groups is 25.3.

2.2.3.3 Group loyalty: how many photos does a user share with the same group?

Another measure characteristic of the sharing behavior is the average number of photos shared per group, π. For clarity of display, we have plotted the histogram of $\frac{1}{\pi}$ in Figure 2.6. We observe that 9.9% of the users share on average one photo per group ($\frac{1}{\pi} = 1$), and 85.1% of the users share on average less than 15 photos per group ($\frac{1}{\pi} > 0.06$). The mean of the average number of photos shared per group for users who share photos is 9.6, and the median is 5.1. This analysis seems to indicate users tend to share a limited amount of photos with the same group. This could be an effect of the large number of groups on Flickr that share exactly the same theme. For example, searching on Flickr for “black and white” yields about 25,000 results, searching for “sunset” yields about 29,000 groups. Less common words, like for example, “gold”, or “magazine”, get 4,600 and 2,200 results, respectively. Another reason might be the driving force behind sharing with groups: if the motivation is photo exhibition, the users will try to share their photos with many groups, and thus show feeble group loyalty; if the motivation is an interest in a specific theme, they will most likely contribute all their photos belonging to that theme into the same group(s).

2.2.3.4 Photo recycling: how often is the same photo shared with multiple groups?

The ratio between the sharing instances and the number of shared photos effectively represents the average number of groups photos are shared with, γ. Again, for display clarity, we present in Figure 2.7 a histogram of $\frac{1}{\gamma}$. The mean γ value is 3.1, and the median is 1.5. Roughly 27.5% of the users share on average each photo in only 1 group ($\frac{1}{\gamma} = 1$), and only 5.4% of the users share the same photo in more than 10 groups ($\frac{1}{\gamma} < 0.1$). This seems to indicate that most users share the same photos in a rather limited number of groups. How these groups are chosen by the users from the (possibly) hundreds of similar groups with the same theme is open to speculation. Users may either stumble upon a group and not look for other similar ones, or search and select a group out of the search results based on the perceived affinity with the group in terms of content, members, and rules. Existing social links to other users may also play an important role in group discovery and sharing behavior, as users will be exposed to groups their friends are a part of.

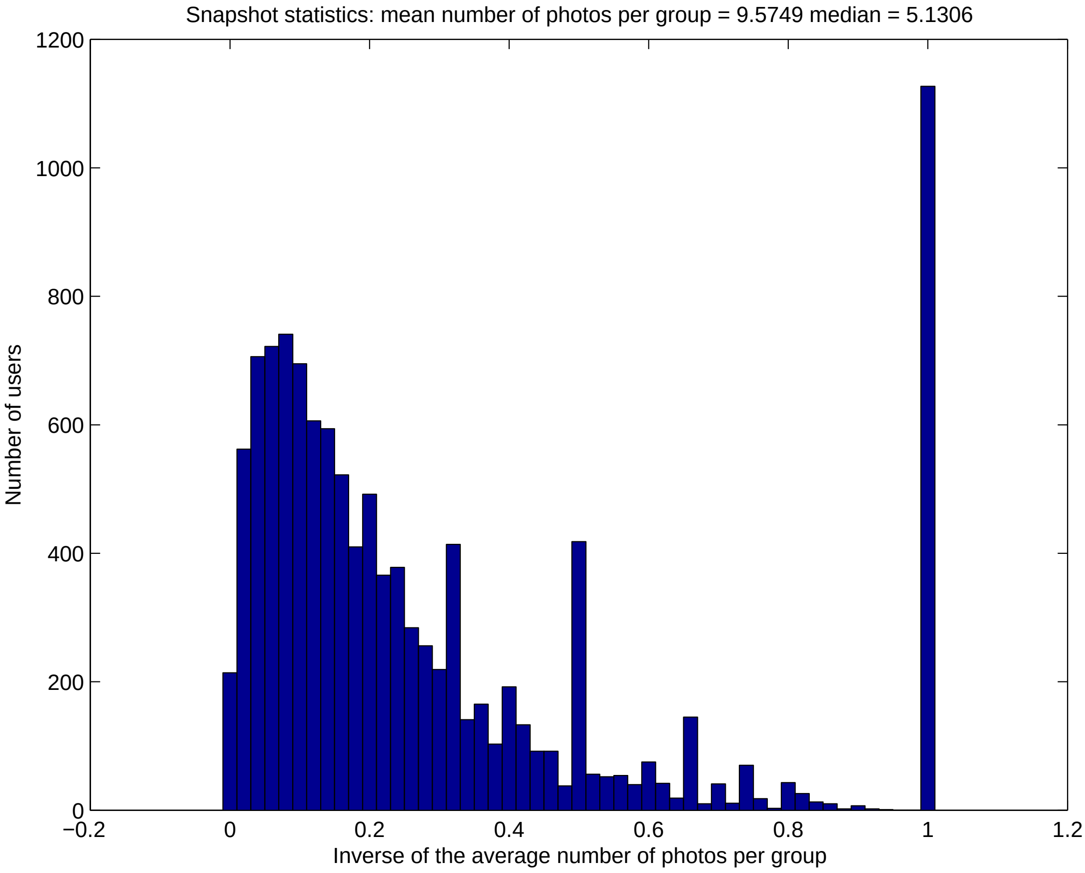


Figure 2.6: Group loyalty - Histogram of $\frac{1}{\pi}$, which is the inverse of the average number of photos shared in the same group. The mean of π over the sharing users is 9.6.

In any case, it appears that important numbers of users in our dataset do not seem to fully profit from the possibility of increasing the visibility of (we hypothesize) their preferred photos, choosing not to recycle their content more often. It should be noted that, at the time of this analysis, the maximum number of groups a photo could be shared with was set by Flickr to be 60 for paying members, and 10 for non-paying members, so this might have played a strong role in user behavior.

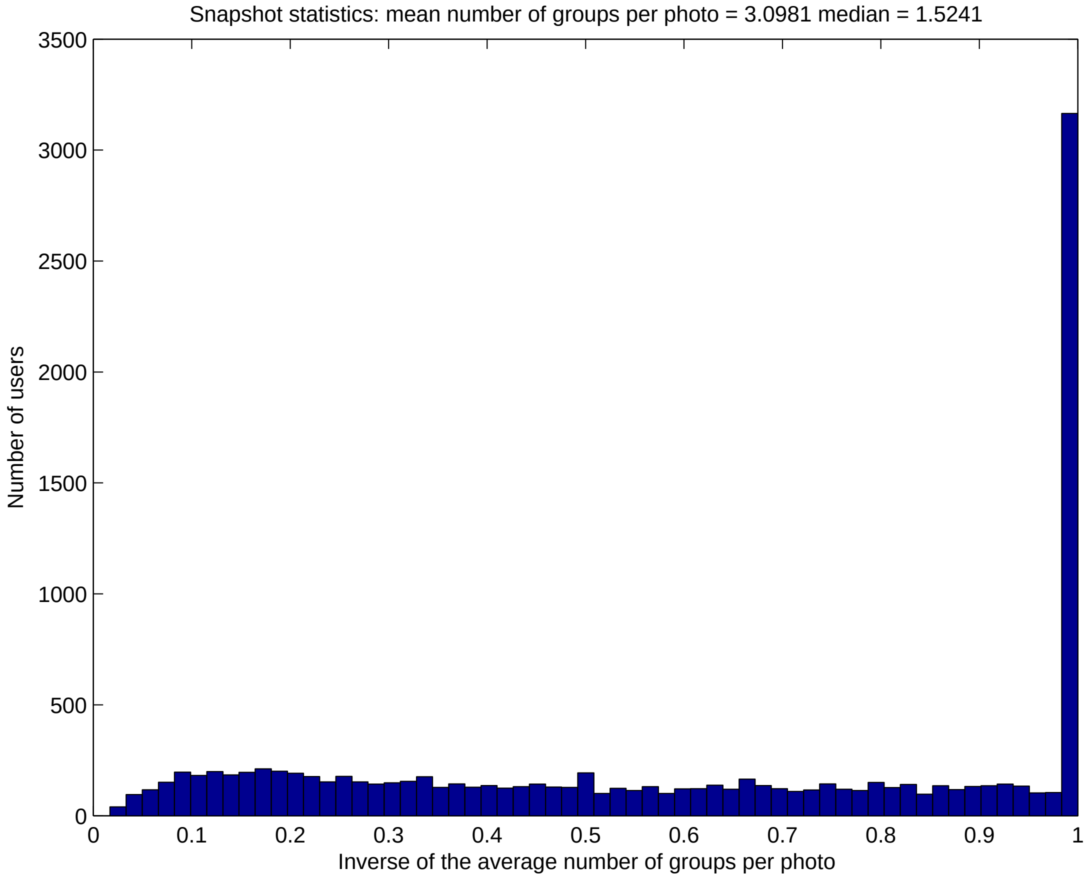


Figure 2.7: Photo recycling - Histogram of $\frac{1}{\gamma}$, which is the inverse of the average number of groups per photo. The mean of γ over the sharing users is 3.1.

In order to determine whether a correlation between the average number of groups per photo and the average number of photos per group exists, we have computed the correlation coefficient between γ and π over the set of users sharing photos. This coefficient is 0.2159, which seems to indicate a relatively weak correlation between the two measures. Figure 2.8 shows that users sharing a large number of photos per group often do so in only a few groups (see the points aligned with the y-axis), while users sharing fewer photos per group often tend to share photos in more groups (see the points aligned with the x-axis). This large variation might suggest that several motivations for sharing photos with groups exist, and these motivations result in different practices for photo sharing. People sharing with many groups might be driven by the *photo exhibition* motivation, while those sharing with only a few groups are probably driven by the more socially anchored motivation of *maintaining relationships* with groups of people either sharing the same passion or interest for a given theme, or being located in the same area.

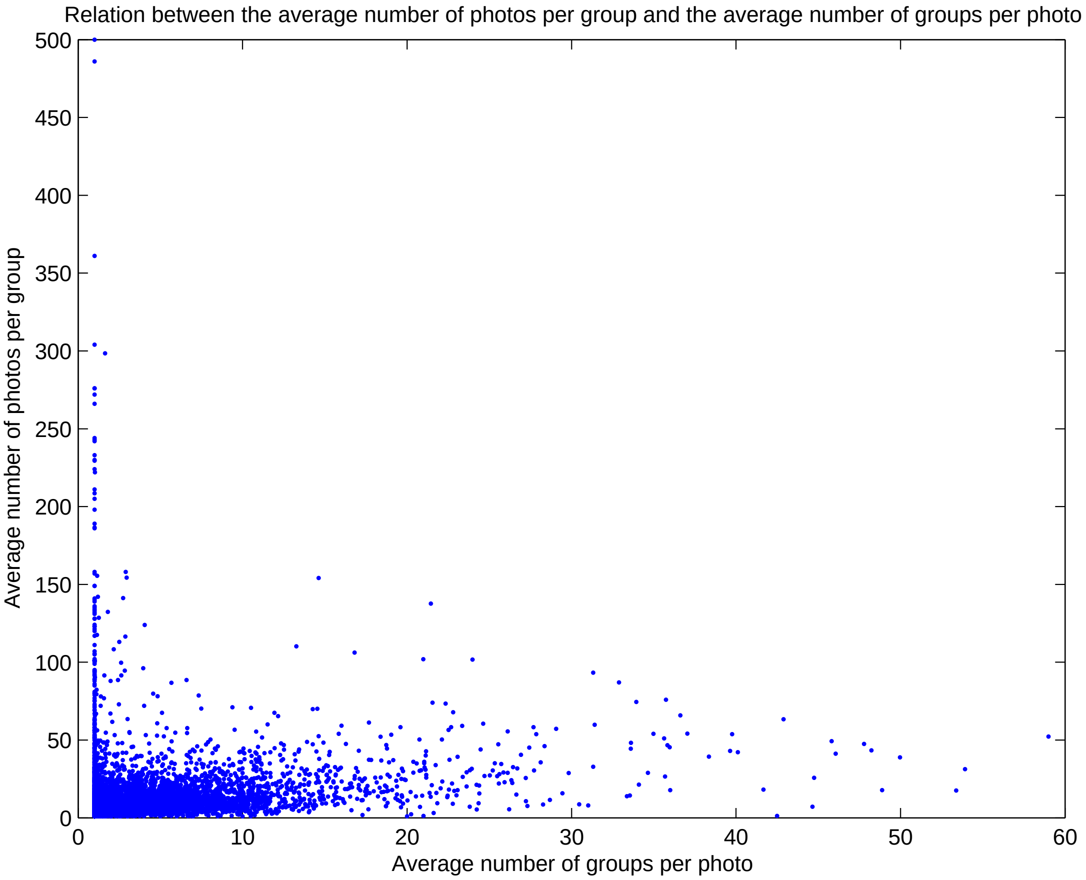


Figure 2.8: Loyalty versus recycling - The average number of groups per photo γ versus the average number of photos per group π, for all users who share photos with groups.

2.2.3.5 Effects of paying membership on sharing behavior

We have not yet discussed some intrinsic limitations of the system on the way users behave. Part of the sharing behavior could likely be influenced by the type of Flickr account a user might have: free accounts allow users to only display the most recent 200 photos from their collection, and to only share a photo with a maximum of 10 groups; paying members (called *pro* members by Flickr) have no limit on the number of photos that are displayed in their account, and can share a photo with a maximum of 60 groups. We take a look in the following paragraphs at the differences in sharing behavior for paying and non-paying members.

In our dataset D_O, the two types of users exist in nearly equal quantities: 51.43% paying users and 48.57% non-paying. The percentages of users who share photos with groups show a significant difference: for paying users, 69.79% share photos with groups, while for the non-paying users, only 31.01% do.

We present in Table 2.1 the most important statistics for the paying users, non-paying users, and both types together. Only users who share photos with groups are taken into account, in order to establish if significant differences exist in sharing behavior. It is clear that the Flickr-imposed maximum limits of 200 visible photos and 10 groups per photo do affect the way non-paying members use their accounts in terms of photos uploaded and groups shared with; however, it is interesting to observe that, although on average pro members upload more and share with more groups (rows $U_{i,p}$, $U_{i,s}$, and $U_{i,g}$ in Table 2.1), the overall sharing ratio is not influenced by their paying or non-paying status (the $\frac{U_{i,s}}{U_{i,p}}$ row). The average sharing measures γ and π also show differences, but at a smaller scale. In conclusion, while sharing volumes may differ, sharing behavior seems consistent across the two categories of paying and non-paying members.

2.2.4 Content-wise comparison of users and groups

When thinking about how groups' photo collections are explicitly formed – they are basically aggregations of user photos – and how groups' tag vocabularies are implicitly formed from those photos in the group photo pool, it could be assumed that groups' tag statistics might be radically different from those of the users.

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

Table 2.1: Statistics for the users who share photos with groups according to their paying status; (μ, m)=(mean, median).

distinct tags	$N_t = 10,236$
users	$N_u = 8,061$
groups	$N_g = 10,838$
photos	$N_p = 1,016,199$

Table 2.2: The reduced dataset D_R.

For this part of the analysis we filtered the original dataset in a number of ways. We concentrated on a smaller vocabulary of the most common 10,236 tags in the dataset, 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 10K 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 10K 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 2.2. 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 20K entities and a total number of photo-tag-group occurrences of roughly 38 million.

In Figure 2.9 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 5,000 unique tags, while on the other hand, there are a number of groups (43) with tag vocabularies of 5 to 10 thousand 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 and 250 bins, the Bhattacharyya distance is 0.2662 and 0.1501 respectively. This distance measure is bounded by the interval [0..1], 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 Figure 2.10, 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 5,000 tag occurrences, while the percentage of groups with less than 5,000 tag occurrences is considerably higher, at 87.3% (as we previously pointed out, groups are by construction smaller, as we only take into account members in our dataset). On the other hand, the dash-dotted-blue and dotted-red curves representing the number of unique tags for groups and users respectively 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.

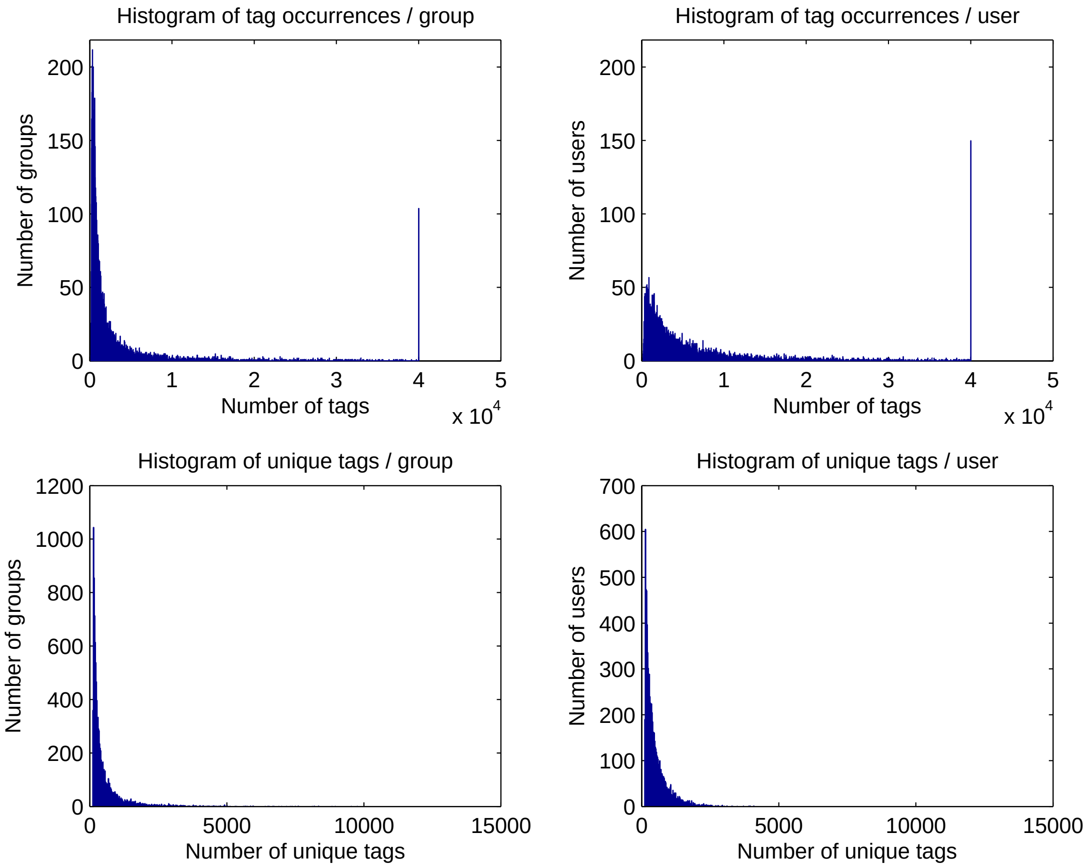


Figure 2.9: User and group tag histograms - Top half: histograms of the total number of tag occurrences per group and per user; bottom half: histograms of the number of unique tags per group and per user.

2.3 Kodak Gallery and Flickr

Among the existing online photo management systems, Kodak Gallery (kodakgallery.com), formerly known as ”Ofoto” and owned since 2001 by Eastman Kodak Company, is one of the leading online digital photo-developing services, operating a number of international sites including the main US site, Canada, and Europe. The company was originally founded in 1999 in Berkeley, California. In 2001, Ofoto, Inc. became a wholly-owned subsidiary of Eastman Kodak Company.

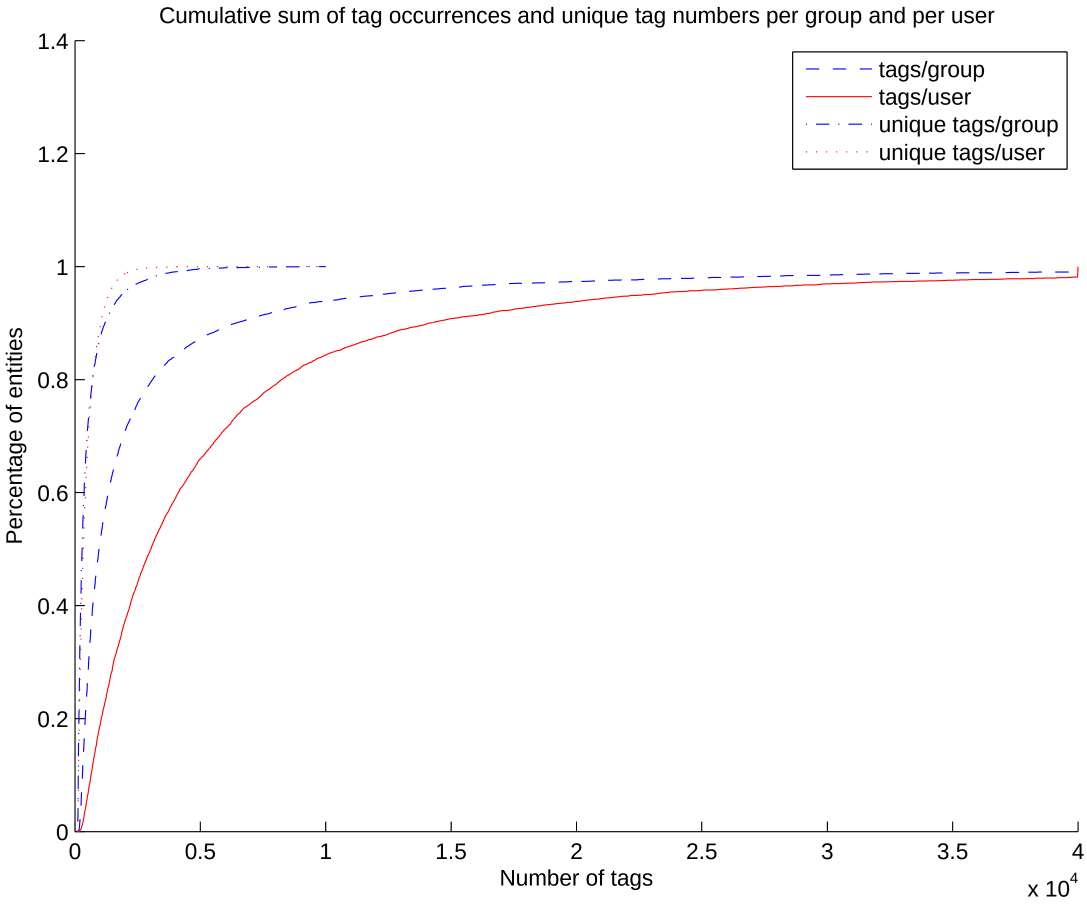


Figure 2.10: Tags and unique tags - Cumulative sums of the total number of tag occurrences and unique tags for groups and users respectively; 66.2% of the users and 87.3% of the groups have less than 5,000 tag occurrences, but in terms of unique tags the two types of entities are very similar.

Apart from Kodak prints of digital pictures, Kodak Gallery offers several additional services around the digital images, such as online photo storage and sharing options, personalized photo gifts, photo books, and mobile phone access to stored photos. It also allows a user to share individual photos or entire albums directly from their Gallery account through social networking sites, such as Facebook. Users are able to provide free form captions of their assets, both at the image level or album level. Kodak Gallery has over 60 million users and storage of billions of images. It is estimated that in 2009 it averaged in excess of 2 million image uploads a day.

In this section we present the first part of a large-scale comparative analysis of these two online photo services, which differ in their design and affordances, and as a result may cater to different needs of their (maybe overlapping) audiences. The analysis will be deepened in Chapter 4.

In Flickr the accent is placed on sharing images with the world, and a real tagging system is employed by users in order to annotate their images and make them searchable to the system. Previous research has shown that the main motivations for tagging on Flickr [5, 47, 66] come from the social involvement within the online community. Many users involve in showcasing high quality photographs, often by joining online Flickr groups, such as *Diamond Stars, Flickr Diamond, The Best of Flickr, Shield of Excellence*, etc.

In Kodak Gallery, on the other hand, the focus is placed on getting physical copies of digital photos, and then on sharing photos mostly with family and friends. The motivations in this case are most likely different, as suggested by results of an ethnographic study [42] with 10 Flickr users. In their study, Miller and Edwards suggested that two distinct categories of users could be found, based on their sharing behavior: *Snaprs* and *Kodak Culture* sharers. The first group takes photos with the primary objective of sharing them with the world, while the second group takes photos to share with a small existing social group, and to archive them. In contrast to ethnographic works which use traditionally a small number of users, we approach this comparison from the other end of the scale, in a study using several orders of magnitude more data.

There is an increasing interest in social media to understand phenomena *across* media sites. For instance, Mislove *et al.* [43] analyzed the connectivity network properties of four major websites, Flickr, YouTube, LiveJournal, and Orkut, while Leskovec *et al.* [37] investigated those of Flickr, del.icio.us, Yahoo! Answers, and LinkedIn. In a work more related to ours, Schifanella *et al.* [62] analyzed Last.fm and Flickr from a social and semantic interplay perspective, showing that a substantial level of local lexical and topical alignment can be observed among users in proximity in the social network. Other works, like Chum *et al.*'s [16] and Philbin *et al.*'s [56], combine image databases from Flickr and other sources to improve performance for image retrieval tasks. However, despite these initial works, the large-scale differences across photo repositories (or social media websites in general) in terms of tagging behavior and tagging content are not yet fully understood.

In this first comparative study, using over 5 million tagged photos from both these sites, we analyze the differences and similarities of Kodak and Flickr users and their tag usage, bearing in mind that two major components are in constant interplay: on one side, from a social study perspective, the impact of the system design on the actual behavior of the users in terms of media usage, and on the other, from a sociological perspective, the impact of the users' motivations and needs on how the systems are actually used. Our results suggest that "Kodak Moments" and "Flickr Diamonds" are indeed two phenomena associated to the large-scale content generation by users.

2.3.1 Datasets

The Kodak dataset is made up of 3,941,463 photos with free-text descriptions, and was provided by Kodak Gallery through collaboration with Dr. Alexander Loui. In total, these photos come from 21,238 different users. A total of 2,681,901 empty captions appear in the dataset, which means more than 65% of the photos have no description at all. Furthermore, almost 697,000 captions contain the camera standard file name as only caption, and 19,337 of them are file names entered by the user, such as *family.jpg, All the grandchildren.jpg,* or *Riding party.jpg*, etc. Since in Kodak Gallery the concept of tags per se does not exist, we preprocessed photo captions, extracting words and using them as tags. As already mentioned, an important number of photos have as caption their filename, and this leads to artifact tags which are quite popular, such as *img, jpg, copy.* In order to get a clearer idea of the actual words employed by users, we have filtered them based on this observation. Additionally, as these tags are extracted from free text, stop-words are quite popular. We have therefore also removed stop-words from the list of tags, using the MySQL list of stop-words [1]. After pre-processing, we

[1] http://dev.mysql.com/tech-resources/articles/full-text-revealed.html#stopwords

Statistic	**Flickr**	**Kodak**
Total photos	4.6M	413,000
Total tag occurrences	13M	900,000
Total users	25,800	5400
Photos / user	157	76
Unique tags / user	81	34

Table 2.3: Statistics of Flickr and Kodak vocabularies.

kept in the Kodak dataset the most popular 50,000 distinct words.

The Flickr dataset, expanded from the one described in the previous section, is composed of 4,794,868 million photos from 32,751 users. These photos are tagged with roughly 23.9 million tags. For our study, we decided to keep an equivalent number of distinct tags for each dataset. We ordered each dataset's tags by popularity (that is, the number of distinct users who employed them), and then kept the most popular 10,000 of them. We present in Table 2.3 some statistics of the two filtered datasets. For Kodak, we have a total of 900,000 tag occurrences from 5,400 users and 413,000 photos. For Flickr, we have a total of 13 million tag occurrences from 25,800 users and 4.6 million photos. We can see that the average number of photos per user is 76 for Kodak and 157 for Flickr, while the average number of unique tags per user is 34 for Kodak and 81 for Flickr.

2.3.2 Speaking the same language?

In order to be able to understand the two vocabularies better, we manually annotated the most popular 200 tags of each vocabulary. We designed a simple taxonomy of 9 categories (*landmark, location, nature, object, action/dynamic, event, time, person, adjective/adverb*), and a 10th catch-all one, labeled *other*.

The distributions of tags over categories for both datasets are shown in Figure 2.11. First we notice that roughly 23% of each vocabulary falls in the *other* category, which is a reflection of the wide variety of subjects. The two vocabularies also show comparable tag frequencies for three other categories, namely *landmark* (with tags *church, bridge, house, building*, etc.), *location* (tags *home, street, museum, city,* etc.), and *adjective/adverb* (*cute, black, green, happy,* etc.). In contrast, the remaining categories display quite important differences between the two vocabularies: *nature* is represented

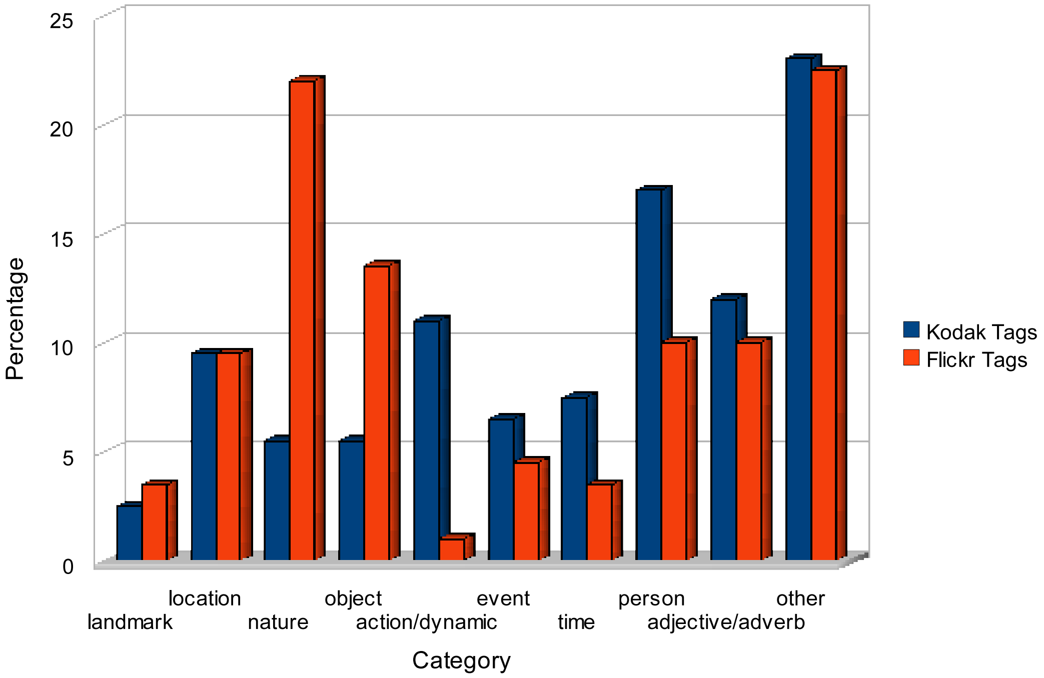


Figure 2.11: Tag categories - Ten-category tag taxonomy and the percentage of the top 200 most popular tags that belong to each category.

4 times more often in the Flickr vocabulary than in the Kodak one, while tags belonging to the *action/dynamic* category appear 5 times more often in the Kodak vocabulary than in the Flickr one. Flickr also shows a higher percentage of *objects*, at around 13%, as opposed to just about 5% in Kodak. Tags belonging to the *time*, *event*, and *person* categories appear much more frequently in the Kodak vocabulary.

While these statistics are computed on only the top 200 tags of each vocabulary, they are likely a good indicator of the inherent differences between the two sets. At the larger scale, Kodak photos are more about *events* and *persons* taking part in them, while many of the Flickr photos seem to be about *nature*. Also, because of the fact that in Flickr tags are used as search keywords, there is a higher number of content descriptive tags, most of which belong to the *object* category. In other words, the "Kodak Moment" concept (family events) and the "Flickr Diamond" one (artistic photos) do show up in the data when taken at large scale. This result therefore backs up the results of Miller and Edwards [42], but with several orders of magnitude more data and users.

As an illustration, we show in Table 2.4 the most popular three tags for our categories. For some categories, the most popular tags are common to Flickr and Kodak users. This is the case for *locations*, *events*, and *nature*. Some differences can be observed for *action/dynamic*, where Flickr has only two tags in the top 200, as well as for *adjective/adverb* where, in contrast to Kodak tags which mainly relate to persons, Flickr tags are dominated by color names.

Going back to the full 10,000 word vocabularies, we are interested in understanding how they compare at the word level. We show in Figure 2.12 the overlap computed at tag-level: the x-axis represents the size of the compared vocabularies ranked by popularity, and the y-axis is the number of common words. Interestingly, we observe a linear relationship between the size of the vocabulary and the amount of overlap, with overlaps between 0 and 58%. The mean overlap is of 54% and the overlap for the full vocabularies is of 56.81%. As with the half-full (or half-empty) glass of water, this shows two things. On one hand, more than half the words are common to the two vocabularies. On the other hand, a significant amount of words is different across the datasets, and their absence may also tell us something about the two vocabularies. A look at the most popular missing tags from each vocabulary shows tags like *macro, selfportrait, blackandwhite, photoshop, insect, flickr, abigfave, impressedbeauty, geotagged* missing from the Kodak vocabulary, while the Flickr one misses tags like *enjoying, lots, put,*

Flickr		Kodak	
Tag	**% users**	**Tag**	**% users**
Category: **landmark**			
bridge	21.5	house	8.6
house	20.7	bridge	3.9
church	19.7	church	3.4
Category: **location**			
park	22.4	home	7.0
street	19.9	park	6.4
garden	18.6	school	5.1
Category: **nature**			
sunset	30.9	beach	7.8
beach	29.8	water	5.4
tree	28.7	tree	5.0
Category: **object**			
flowers	28.1	picture	14.1
flower	28.0	cake	4.6
car	23.4	car	4.2
Category: **action/dynamic**			
work	13.0	ride	4.8
dance	10.8	playing	4.6
		waiting	4.4

Flickr		Kodak	
Tag	**% users**	**Tag**	**% users**
Category: **time**			
night	26.4	day	11.7
winter	19.4	time	7.8
summer	18.1	night	5.9
Category: **person**			
family	22.4	i	14.8
me	22.1	family	10.8
portrait	19.8	mom	10.4
Category: **adjective/adverb**			
red	27.4	big	8.1
blue	26.8	happy	7.2
green	25.5	good	6.7
Category: **event**			
christmas	24.2	birthday	7.8
birthday	21.2	party	7.4
party	21.1	christmas	7.1
Category: **other**			
dog	27.4	view	9.2
cat	25.7	back	8.3
art	23.8	love	7.4

Table 2.4: Top 3 words per category, and percentage of users using them for the two vocabularies.

showing, giving, checking, heading, loved, weeks, visiting, dressed, wearing. The first set of tags represents, more or less, Flickr jargon: photographic techniques (*macro* and *blackandwhite*), Flickr groups' tags (*abigfave, impressedbeauty*), and "modern photographer" related activities, such as *geotagging*. The second set of tags is clearly dominated by *action/dynamic* tags, probably a by-product of the free-text descriptions of the Kodak dataset and the general orientation of the Kodak users towards events and people.

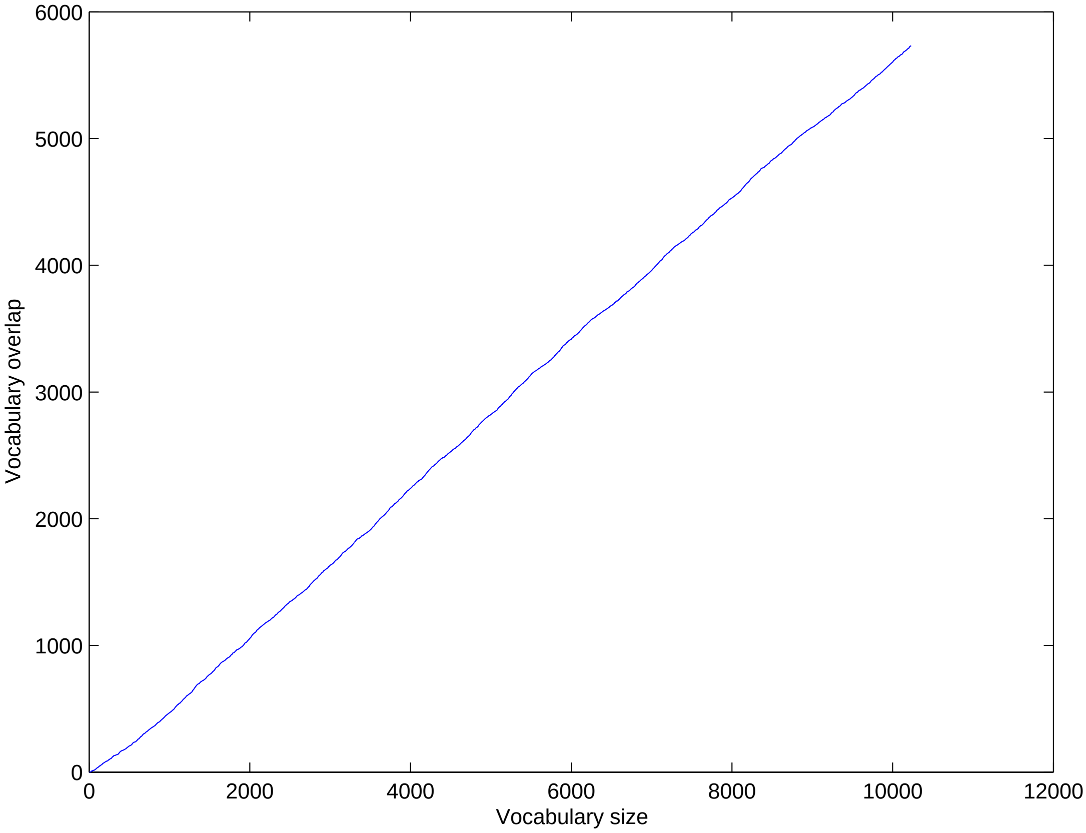


Figure 2.12: Flickr and Kodak vocabularies overlap - Vocabulary overlap computed at every other individual tag rank.

2.4 Conclusions

We have taken a look in this Chapter at two of the most popular photo sharing online systems, Flickr and Kodak Gallery. We tried to understand how users of such systems employ them, and how different system design decisions impact the sharing behaviors and emerging metadata vocabularies.

While the restrictions on free Flickr accounts do seem to influence the number of photos users have in their accounts (with an average of around 220 photos for non-paying 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 non-paying 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 non-paying members share on average 30% (median 17.1%).

Our results also show that, on average, a user shares a relatively 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 non-paying 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 significantly influenced by their paying or non-paying status, or by the amount of photos they upload.

Overall, the analysis shows that through relatively modest photo re-purposing, small but persistent group loyalty, and active participation in groups, Flickr users contribute a significant proportion of their content to communities. These communities emerge as rich aggregated Flickr entities through the integration of their members' contributions, comparable from a content-based point of view with Flickr users.

We have also observed that, despite inherent differences induced by the underlying users of the two different systems (Flickr and Kodak Gallery), by users' motivations and their needs, as well as by system design and affordances, certain similarities exist at the raw vocabulary level between Flickr and Kodak Gallery. Thus, for the top 200 most popular tags, tag categories such as *landmark, location*, and *adjective/adverb* show comparable frequencies for both datasets, and more than half the words exist in both vocabularies. At the same time, differences also exist: the Kodak vocabulary contains substantially more *action/dynamic* tags, which are missing in Flickr, while the Flickr vocabulary contains photography related tags that do not appear in the Kodak one.

Some limitations arise from the intrinsic nature of the two datasets. On one hand, Kodak Gallery does not use tags as a metadata construct, but free text descriptions.

On the other hand, the ratio of photos with a valid description in our Kodak dataset is relatively small, and is also smaller than the number of tagged photos in the Flickr dataset. However, we believe that by pre-processing the captions and discarding stop-words, the tags obtained for the Kodak dataset are a reasonable approximation of what a real tagging vocabulary would have been, and, while the number of photos is smaller than the one obtained for the Flickr dataset, it is still a rather large amount in absolute value.

3

Modeling online communities - jointly modeling Flickr users and groups

We have taken in the previous chapter a close look at the Flickr ecosystem, and the way users contribute to the emergence of groups as comparable entities, from a content-based point of view. The next step is to work under this assumption towards a unified model for users and groups, that would allow us to compare any given user and any given group from a more abstract, almost semantic, perspective. Thus, we present in this chapter a unified probabilistic topic model for both Flickr groups and users. We show that, despite intrinsic differences pertaining to the nature of the two types of entities, at raw vocabulary level they are similar enough to make joint modeling viable. Our model allows us direct comparison between any two entities in the system, and we derive a topic-based similarity measure which can then be used in various application scenarios. The material presented in this chapter was produced in 2008-2009 and it was originally published in [48].

The structure of this chapter is as follows: we first present a review of related literature in Section 3.1, then we describe the probabilistic topic model in Section 3.2, and then we show some of the direct insights obtained from the topic-based representation in Section 3.3. We conclude with some possible applications that make use of the topic-based representation and the similarity measure in Section 3.4 and an analysis of the generalization properties of the model to content-poor entities in Section 3.5.

3.1 Related work

Several studies used Flickr data in order to better understand users and the ways they use Flickr as a whole. From a research perspective, Flickr Groups are interesting for several reasons. First and foremost, many of the groups act like content funnels, gathering in a single place - the group pool - references to photos that match a specific criterion, be it photographic technique, photographic subject, semantic subject, or aesthetic quality. Group administrators, and sometimes even regular group members, act as content filters. This is a great resource to tap into for the research community, and several studies have already used groups as a starting point for data collection [35, 42, 58].

Secondly, groups are natural paradigms of the "content+relations" model of social media. That means that not only there is content that is in one way or another consistent, but there is also information about relations between users. Several research groups have started to take advantage of this information in recent studies [25, 36, 59, 65].

Finally, membership in a specific group also brings additional metadata if information about the group itself is included in the dataset, such as the group name, group size, or group type. Group names can be seen as commonly accepted tags by their respective members, or they can be used as groundtruth during evaluation of automatic analysis methods. This kind of metadata has also been exploited in several works [15, 18, 19, 35].

Tagging systems have been analyzed by Marlow *et al.* [41] and a taxonomy of users' motivations to tag has been proposed by Ames and Naaman in [5]. 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.* [54] 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.

Metadata from the (social) links existing on Flickr [34, 35, 36, 67] 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.* [34], 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 [67] 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 [35] 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.* [36] 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 than ours, as 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 4,500 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.

In a study using Flickr groups' data, De Choudhury [18] modeled group activity over time using a probabilistic approach. The author proposes a continuous HMM to model the activity of each user with respect to a group, whether it is a photo upload, comment on an existing group photo, or marking a photo as a favorite. The output of the HMM over all members of a group is then averaged and used to obtain a measure of *significance* for a given group at a specific time moment. Experiments with a dataset of 200 groups seem to indicate that the proposed model may predict group activity on four variables (new members, uploads, favorites, and comments) reasonably well. In contrast to our work, this study does not make use of the textual content of users and groups, and concentrates specifically on predicting group related activities.

In a related piece of work, De Choudhury *et al.* [19] used image features, tag features, and user activity features to characterize photos in Flickr, and then used these features for group recommendation. They used roughly 15,000 images from 925 groups, and learned a probabilistic topic model over the set of groups, considered as bags-of-features. They then used this model to predict the most likely groups for a given test image, and compared their method to a k-nearest neighbors approach, with better results. Compared to this work, our model is learned jointly on a set of groups almost one order of magnitude larger as well as a set of users, and can be used for recommendation of groups to users based on example users or groups, not individual photos. Furthermore, we advocate a simpler representation of groups and users in our model, that of a bag-of-tags. Nevertheless, adding a measure of user activity in a certain group in the form of comments on photos from the group pool is an interesting avenue to explore in the future.

3.2 A probabilistic topic model for Flickr users and groups

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. Although users' contributions to the groups may be seen as a data-replication mechanism, we believe that this is not really the case. Groups are independent entities and the data they contain (photos and tags), although referencing user data, is rightfully a representation of the groups themselves.

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 [10, 29]. One such model is Probabilistic Latent Semantic Analysis (PLSA), which was introduced by Hofmann [29], as a probabilistic extension of Latent Semantic Analysis (LSA) [20]. Other topic models may have been used, but we chose to favor lower computational complexity and a simpler representation, at the risk of lower generative power.

3.2.1 Probabilistic latent semantic analysis

PLSA is a generative probabilistic model, which assumes the existence of a latent topic variable in the generative process of each word in a document.

For our purposes, we represent each entity (Flickr group or user) E_i as a bag-of-tags, i.e. a vector $\mathbf{t_i} = (t_{i1}, ..., t_{ij}, ..., t_{iN_t})$ 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 naturally 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 $k \in 1, ..., N_z$ and N_z representing the a-priori fixed number of topics for a corpus of documents. The tags, denoted by t_j, with $j \in 1, ..., N_t$, 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 $i \in 1, ..., N_E$, 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 between 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:

$$P(t_j, z_k, E_i) = P(E_i)P(z_k \mid E_i)P(t_j \mid z_k). \tag{3.1}$$

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

$$P(t_j, E_i) = P(E_i) \sum_{k=1}^{N_z} P(z_k \mid E_i)P(t_j \mid z_k). \tag{3.2}$$

This is equivalent 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 $P(z \mid E_i)$. Given topic z_k, a tag t_j is selected based on the conditional probability distribution $P(t \mid z_k)$.

The conditional probability distributions $P(t \mid z_k)$ and $P(z \mid E_i)$ 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 $P(t \mid z_k)$. We denote this matrix by $P(t \mid z)$. Likewise, we denote by $P(z \mid E)$ the matrix storing the parameters of the multinomial distributions $P(z \mid E_i)$ that describe the training documents.

The parameters of these multinomial distributions are estimated by the Expectation-Maximization (EM) algorithm [29], derived from the likelihood of the observed training data:

$$\mathcal{L} = \prod_{i=1}^{N_E} \prod_{j=1}^{N_t} P(E_i) \sum_{k=1}^{N_z} P(z_k \mid E_i) P(t_j \mid z_k)^{n(E_i, t_j)}, \tag{3.3}$$

where $n(E_i, t_j)$, as mentioned earlier, 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:

$$P(z_k \mid E_i, t_j) = \frac{P(t_j \mid z_k) P(z_k \mid E_i)}{\sum_{k=1}^{N_z} P(t_j \mid z_k) P(z_k \mid E_i)}. \tag{3.4}$$

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

$$P(t_j \mid z_k) = \frac{\sum_{i=1}^{N_E} n(E_i, t_j) P(z_k \mid E_i, t_j)}{\sum_{j=1}^{N_t} \sum_{i=1}^{N_E} n(E_i, t_j) P(z_k \mid E_i, t_j)}, \tag{3.5}$$

$$P(z_k \mid E_i) = \frac{\sum_{j=1}^{N_t} n(E_i, t_j) P(z_k \mid E_i, t_j)}{n(E_i)}, \tag{3.6}$$

where $n(E_i)$ is the number of tag occurrences in entity E_i's bag-of-tags. The distributions $P(t \mid z_k)$ describe each topic z_k and are also valid for documents outside the training set. This is however not true for the matrix $P(z \mid E)$ which stores the parameters of the N_E multinomial distributions $P(z \mid E_i)$ and is thus relative to each of the N_E training entities. For unseen documents the distributions over topics can be inferred through a *folding-in* procedure, as proposed in [29]. This method maximizes the likelihood of the unseen documents using a partial version of the EM algorithm described above: $P(t \mid z)$ is obtained from training and *kept fixed*, thus not updated on each M-step. As such, $P(z \mid E_{unseen})$ maximizes the likelihood of entity E_{unseen} with respect to previously learned parameters. Overfitting is largely reduced 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 [29] and annotated images [44].

3.2.2 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 non-trivial 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. Perplexity has been previously used in the topic-modeling literature [14, 61]. Although not a reliable indicator of the semantic quality of the topics themselves, perplexity can be used as an indicator of the "goodness" of a topic model with respect to the data it is learned from.

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, described in Section 2.2.4, 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:

$$per(D_T) = exp[-\frac{\sum_{i=1}^{N_d}\sum_{j=1}^{N_t} n(E_i, t_j) log(p(t_j \mid E_i))}{\sum_{i=1}^{N_d}\sum_{j=1}^{N_t} n(E_i, t_j)}], \tag{3.7}$$

where $p(t_j \mid E_i)$ 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 [29].

We show perplexity values for each of the six different models in Figure 3.1. As previously found in the topic model literature [10, 29], 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 10K 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 counter-balanced 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 higher than 100.

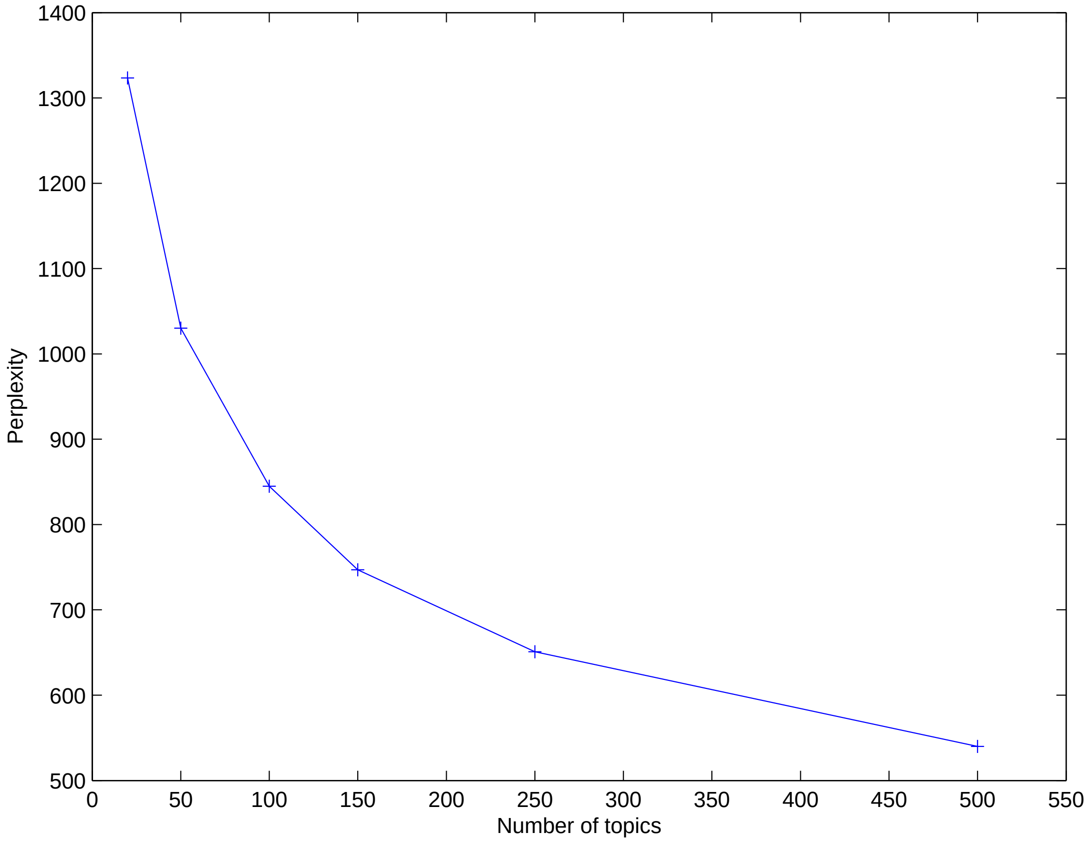


Figure 3.1: Perplexity variation - Variation of the perplexity with respect to the number of topics learned by the model. The + markers show the perplexity values for each N_z in the set $\{20, 50, 100, 150, 250, 500\}$. Perplexity decreases with the increase in number of topics.

3.2.3 Topic-based representation of entities

One of the outputs of the PLSA model for a given entity are the multinomial distributions $P(t \mid z)$, in other words the probability distribution of tags over each topic. The model also outputs, for each entity, the distribution $P(z \mid E_i)$, 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% of the topics as having high semantic consistency in the latter two cases, with slightly more "confused" topics in the case of the 50-topic model. Topics and their most relevant Flickr groups for models learned with 50, 100, 150, and 250 topics can be found online at `http://www.idiap.ch/~negora/flickrcommunities/`.

We show in Tables 3.1, 3.2, 3.3, and 3.4 some of the topics learned by PLSA, 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 3.5 and 3.6 we show some photos from various group pools found in the chosen topics for exemplification.

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 $p(z \mid E)$. 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 the distribution over topics for a Flickr group (*Candid Camera*) in the upper part of Figure 3.2, and the two most probable topics for the group in the lower half. Topic 38 could easily be described as *street portraits* and topic 90 as *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 users and groups exists. To answer this question, we

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)

Table 3.1: Some of the topics learned by the model, characterized by their most probable tags (ranked by $P(t \mid z)$). We also present the most probable entities (ranked by $P(z \mid E)$). Numerical ID entities (such as 56939004@N00) represent Flickr users, while the rest are Flickr groups.

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

Table 3.2: Some of the topics learned by the model, characterized by their most probable tags (ranked by $P(t \mid z)$). We also present the most probable entities (ranked by $P(z \mid E)$). Numerical ID entities (such as 87075398@N00) represent Flickr users, while the rest are Flickr groups.

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

Table 3.3: Some of the topics learned by the model, characterized by their most probable tags (ranked by $P(t \mid z)$). We also present the most probable entities (ranked by $P(z \mid E)$). Numerical ID entities (such as 59289953@N00) represent Flickr users, while the rest are Flickr groups.

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

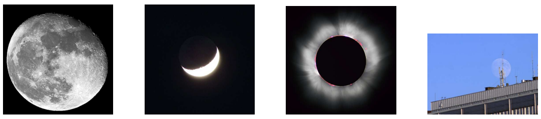
Topic 61	
$P(t \mid z)$	**Tag**
0.2044	london
0.1527	uk
0.1249	england
0.0205	unitedkingdom
0.0119	britain
0.0116	yorkshire
0.0080	brighton
0.0077	thames
0.0076	birmingham
0.0074	kent
0.0073	cornwall
0.0063	oxford
0.0060	manchester
0.0059	norfolk
0.0054	bath
0.0054	park
0.0049	sussex
0.0040	pub

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)

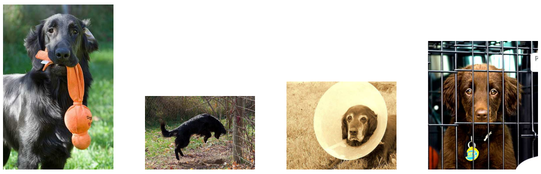
Topic 65	
$P(t \mid z)$	**Tag**
0.0945	portrait
0.0550	woman
0.0515	girl
0.0234	face
0.0185	sexy
0.0179	people
0.0169	beautiful
0.0166	female
0.0152	model
0.0144	beauty
0.0132	man
0.0126	eyes
0.0119	girls
0.0110	women
0.0101	smile
0.0100	pretty
0.0099	hair
0.0086	fashion

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

Table 3.4: Some of the topics learned by the model, characterized by their most probable tags (ranked by $P(t \mid z)$). We also present the most probable entities (ranked by $P(z \mid E)$). Numerical ID entities (such as 89388861@N00) represent Flickr users, while the rest are Flickr groups.

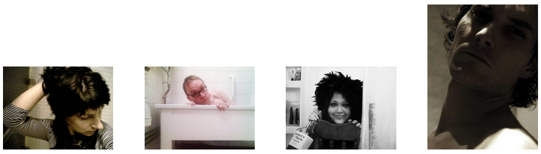


Photos from group *Lunatics*, by *saturn h*, *oceandesetoiles*, *Luc Viatour* ©*GFDL*, and *Steffe*

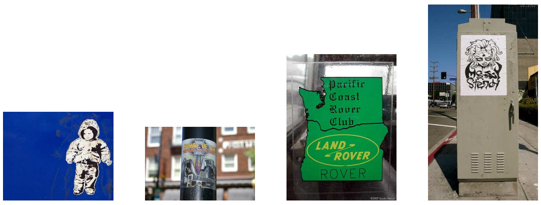


Photos from group *Flatcoats*, by *Wabana* (1,2), *MontanaRaven* (3), and *black dog_brown dog* (4)

Table 3.5: Example photos from pools of groups that are highly probable for topics 3 (left) and 13 (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)

Table 3.6: Example photos from pools of groups that are highly probable for topics 43 (left) and 45 (right).

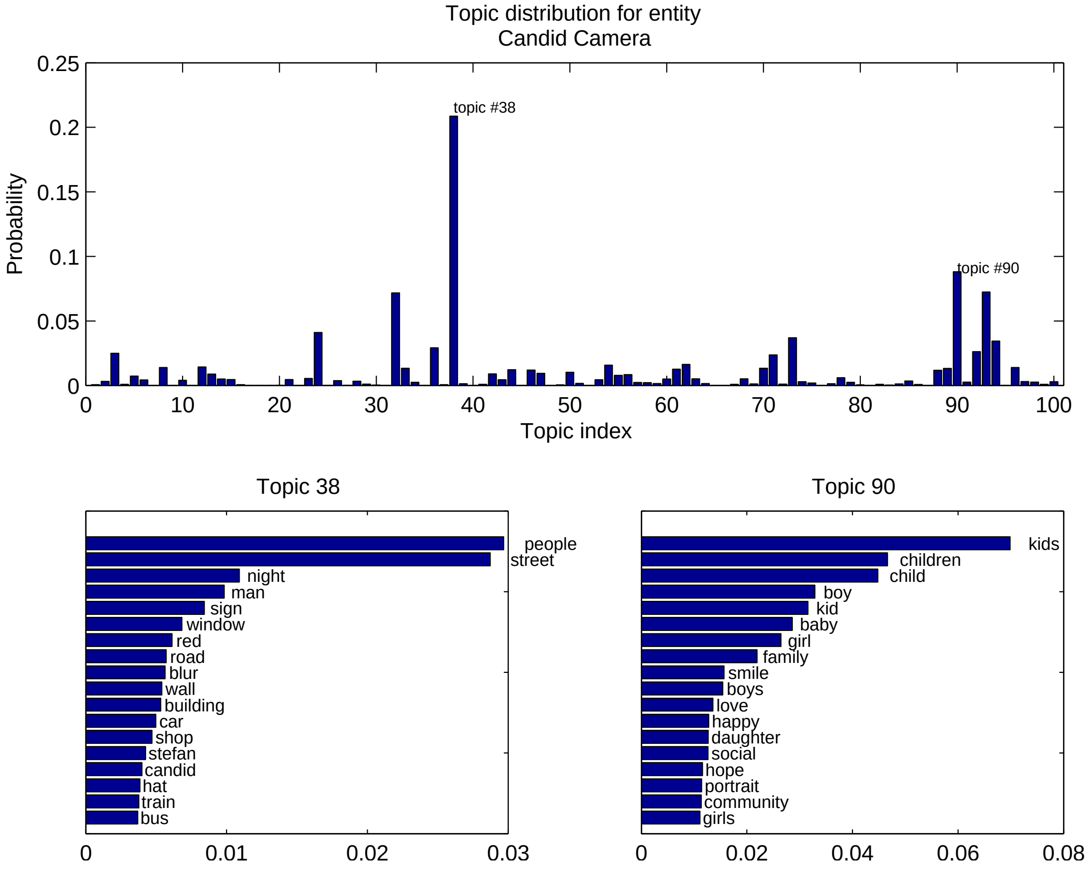


Figure 3.2: Candid Camera - Topic distribution for the entity *Candid Camera* (a Flickr group). In the lower part of the figure, the two most relevant topics, described by their top most probable tags. Topic 38 could be described by the concept “street portraits” and topic 90 by “children”.

have generated the histograms of the number of *relevant topics* for each type of entity in Figure 3.3. 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. After computing the relevant topics, we can observe two main differences, as shown in Figure 3.4:

- on one hand, a higher percentage of groups as opposed to users seem to be focused on fewer topics. For instance 10% of the groups are about 1 or 2 topics, compared to just 4.8% of the users, and 25% of the groups have 4 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. 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 vs. social) by inspecting the number of relevant topics in their topic distributions. Obviously a clear-cut distinction between groups and users cannot however be solely made based on the topic representation.

3.3 Insights into entity and community structures

The main advantage 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

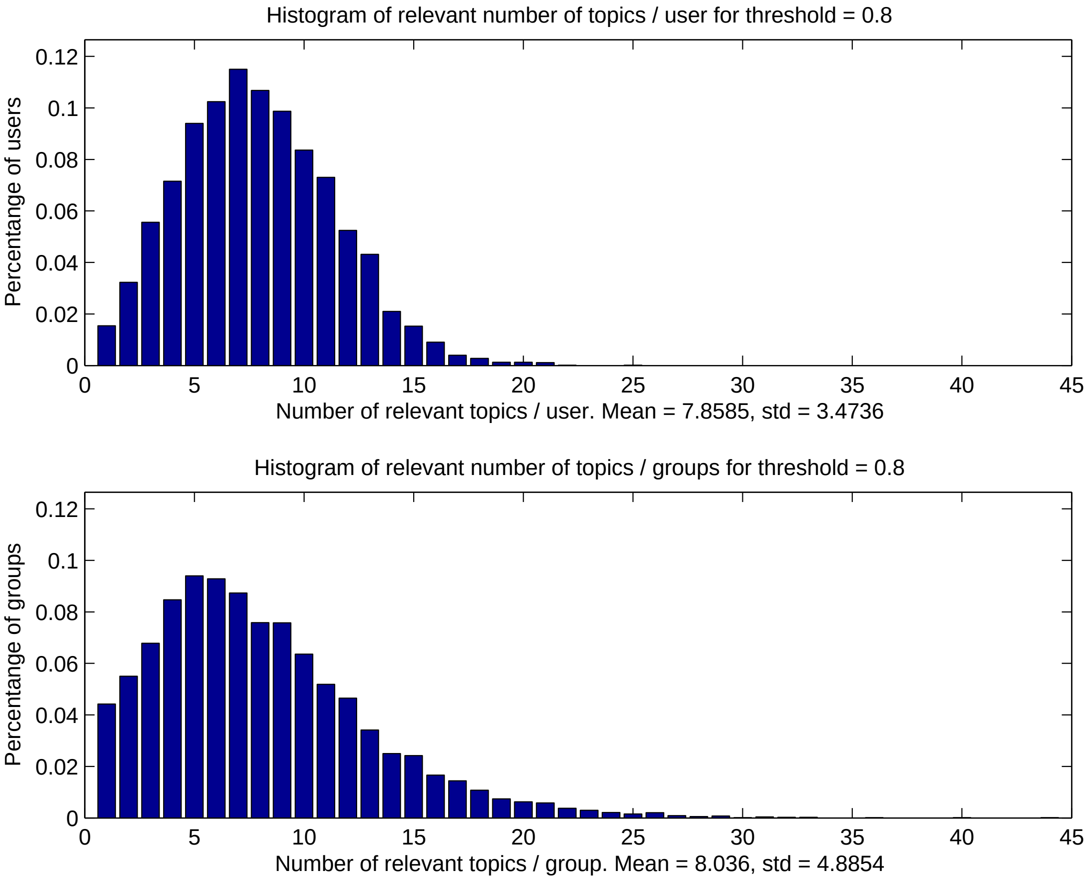


Figure 3.3: Relevant topics per entity - Comparison of the number of relevant topics for groups and users. For ease of comparison, we normalized the histograms and display on the y axis the percentage of users and groups respectively.

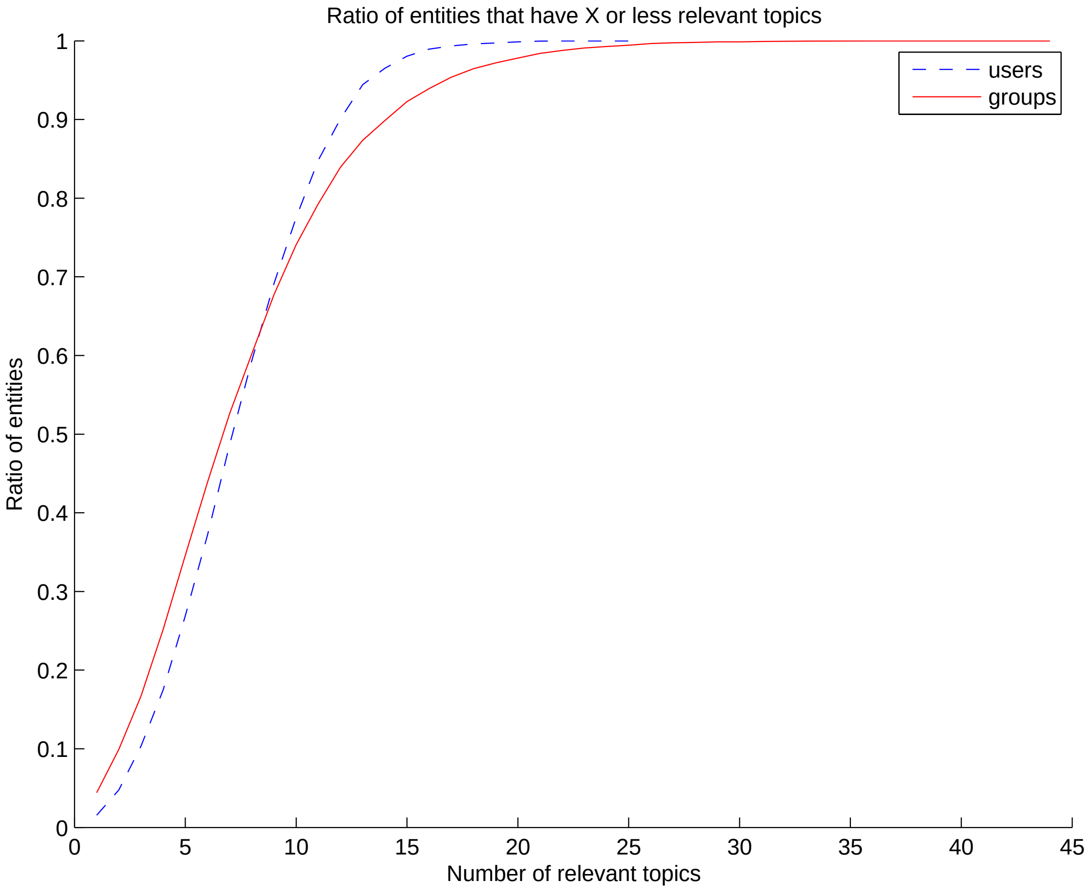


Figure 3.4: Topicality of entities - Ratio of either type of entities that are about x or less topics. For example, 60.2% of groups and 59.3% of users are about at most 8 topics.

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 that also has the properties of a full metric, we finally adopted the Hellinger distance, as used in [17], which is based on the Bhattacharyya coefficient. In the case of discrete probability distributions, the Bhattacharyya coefficient is defined as:

$$BC = \sum_{x} \sqrt{p(x)q(x)}. \tag{3.8}$$

The similarity metric is then the distance given by:

$$\rho(p, q) = \sqrt{1 - BC}. \tag{3.9}$$

This distance has the advantage of being a true metric: 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 [31]. It also has the advantage of being confined to the interval [0..1].

For each entity in our dataset we have thus computed the distance ρ to all other entities in the dataset, resulting in a $N_E \times N_E$ distance matrix. With this new information, we explore new ways of understanding communities' structure.

3.3.1 Group similarities

First, we started by looking at the distribution of the mean distances between groups. As pointed out in the earlier analysis of Flickr in Section 2.2, on average users share any given photo with about 3 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 Figure 3.5 the two histograms 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 that groups which share at least one member are more likely to be similar. The null hypothesis that the two distributions have the same mean is rejected by a two-tailed t-test at $\alpha = 0.01$.

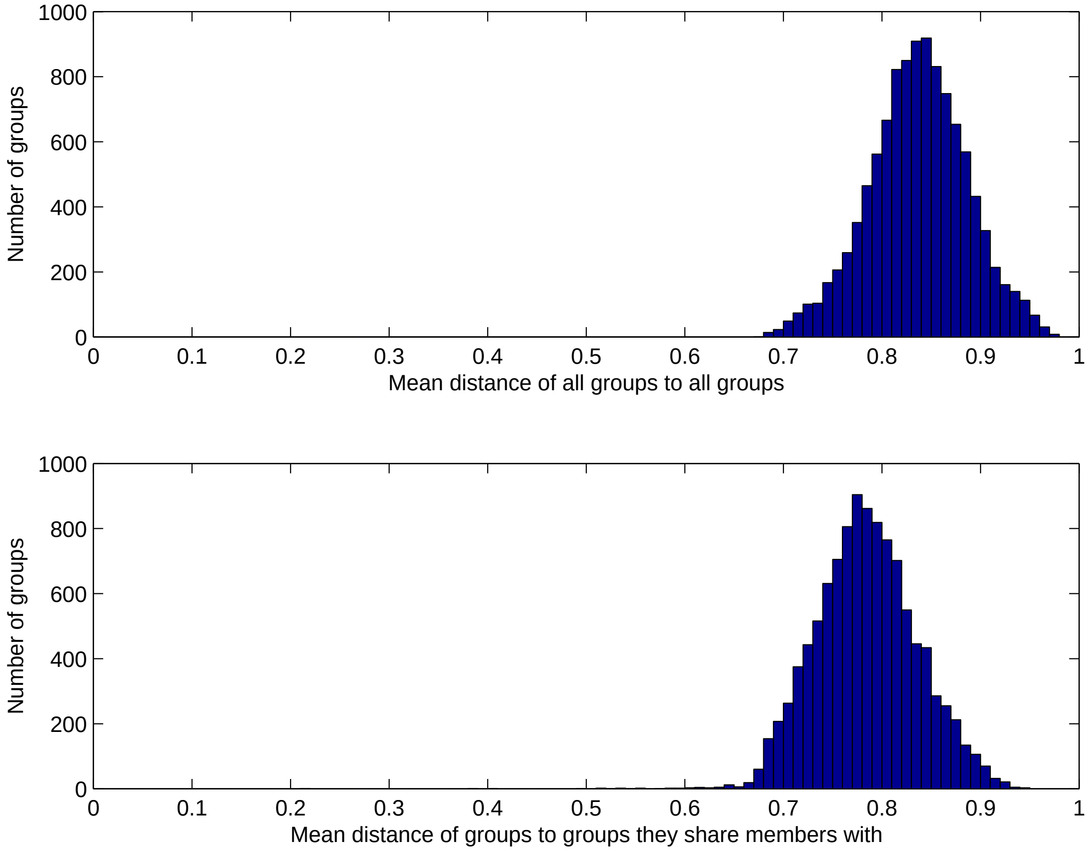


Figure 3.5: Group similarities - Distributions of the mean distances from each group to all groups (top), and to only groups who share at least one member (bottom). We observe a significant shift in distances when only "overlapping" groups are considered.

3.3.2 User similarities

Second, we analyzed the distances between users. As previously for groups, we have constructed two histograms, shown in Figure 3.6: 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 as pronounced as that observed earlier for groups, however we can still observe a clear shift towards lower values when only users who belong to common groups are taken into account. A 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 whom they share no groups at all. One can also observe that the histograms in Figure 3.5 have a larger variance than the histograms in Figure 3.6, which again indicates that groups might be a more variable construct.

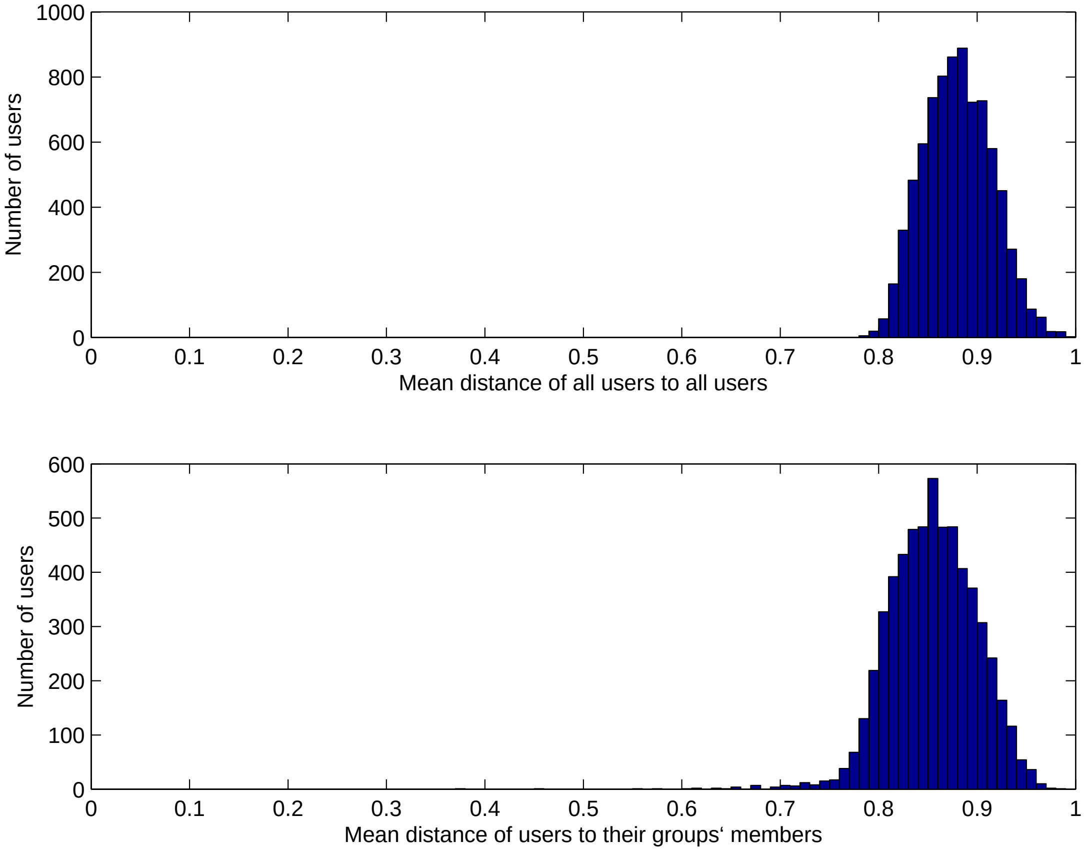


Figure 3.6: User similarities - Distribution of the mean distances from all users to all users (top), and to only users whom they share at least one group with (bottom). We observe a clear difference between the two histograms, the distances between users who are part of the same groups are smaller on average than those between all users.

3.3.3 Group-user similarities

Finally, in Figure 3.7 we plot 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 0.8, while the mean distances from groups to just their members are generally lower than 0.8. 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 seems 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 to some degree the semantic similarity of users to either the groups they belong to or to users with whom they share the same groups, and it might also be partially a consequence of the way groups' bag-of-tags representations are built, starting from their members. It is nevertheless an indication that topic-based similarity can be an useful measure for recommendation of groups or users.

3.4 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, could be 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 the site, 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 decomposition.

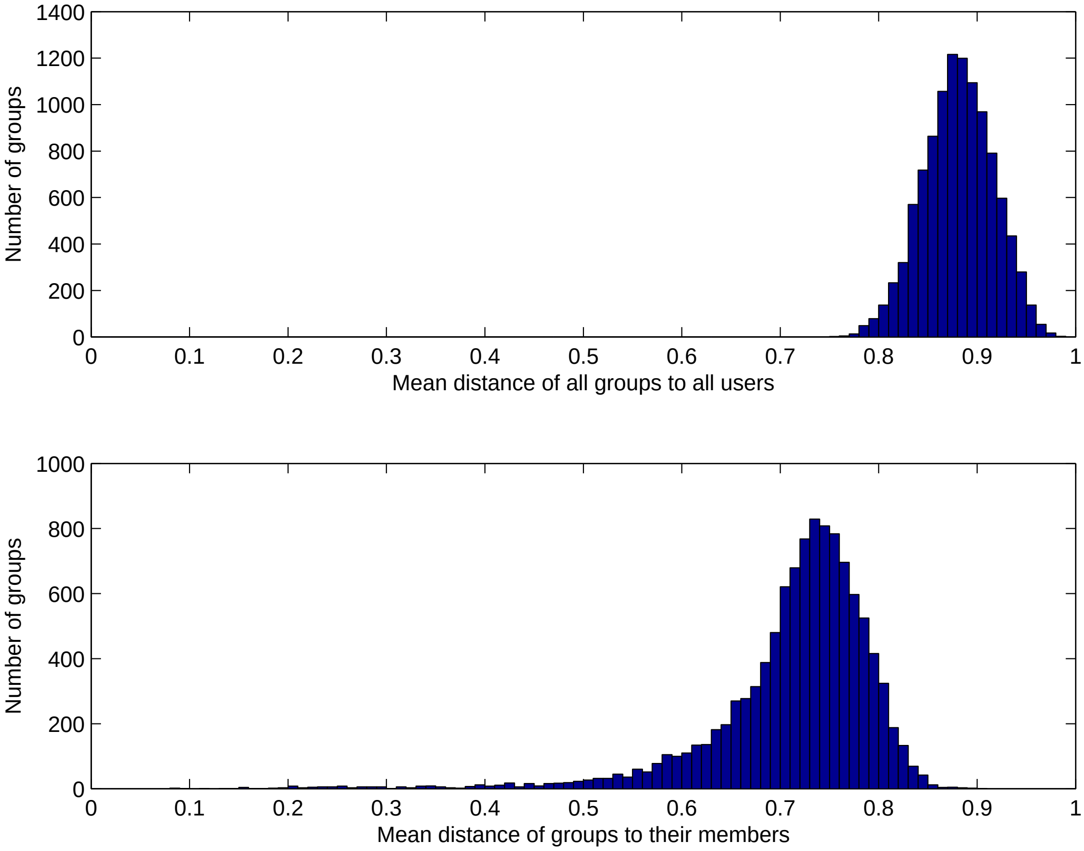


Figure 3.7: Group-user similarities - Distribution of the mean distances from all groups to all users (top), and to only member users (bottom). We observe a distinct difference, explained by the fact that members' representations are closer to that of the group they belong to than those of users who do not belong to the group.

3.4.1 Topickr: an interest-based entity exploration tool

The exploration mechanism can be very well used with our topic-based representation. We can envisage three scenarios.

First, 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 Equation 3.9.

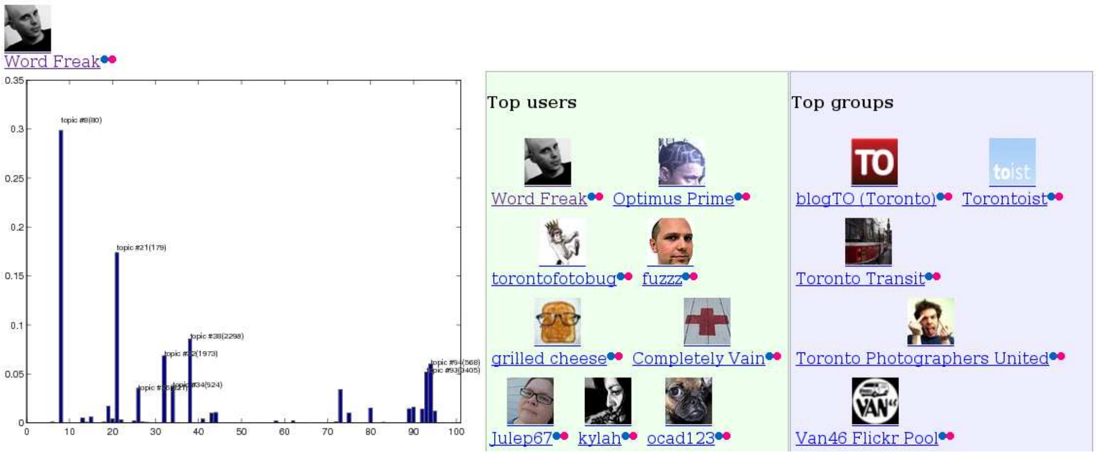


Figure 3.8: Topickr - An exploration application that uses similarity of the topic-based representations in order to present the most similar users and groups for a given entity. On the left, the topic representation of the given entity (user *Word Freak* in this case), and on the right the top most similar users and most similar groups.

Our *Topickr*[1] application, of which a snapshot is presented in Figure 3.8, 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 they particularly like. This is not straightforward for a human observer, but in our model this can be 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 $P(E \mid z) \propto P(z \mid E)$, we can rank entities based on their probabilities given this starting topic. As we have seen in Section 3.2.3, some entities have spiky topic distributions, with a single topic in their representation. We call these entities

[1] see demo at http://www.idiap.ch/˜negora/acmmm08

topic-experts. We show in Figure 3.9 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.

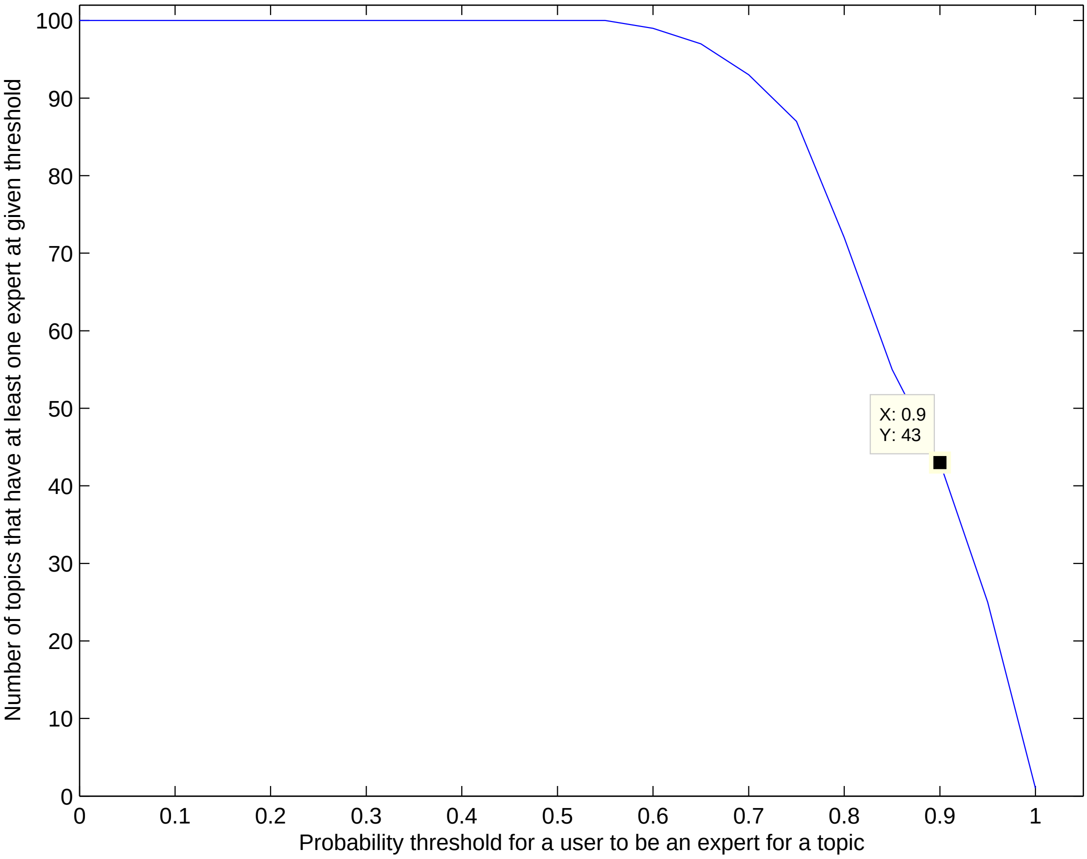


Figure 3.9: Topic experts - The number of topics that have at least one topic-expert, varying with the topic-expert's probability for the given topic. In this model, 93 topics out of 100 have at least one entity whose probability for that topic is higher than 0.7 and 43 topics out of 100 have at least one entity whose probability is higher than 0.9.

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.

3.4.1.1 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 relevance.

We compare three similarity measures in two retrieval scenarios. The first similarity measure is the previously denoted ρ distance in Eq. 3.9, 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 Figure 3.10 and 3.11 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 denotes 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, user retrieval from groups, we retrieve the most similar users for each group. In this case (Figure 3.10, 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 all 10,000 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

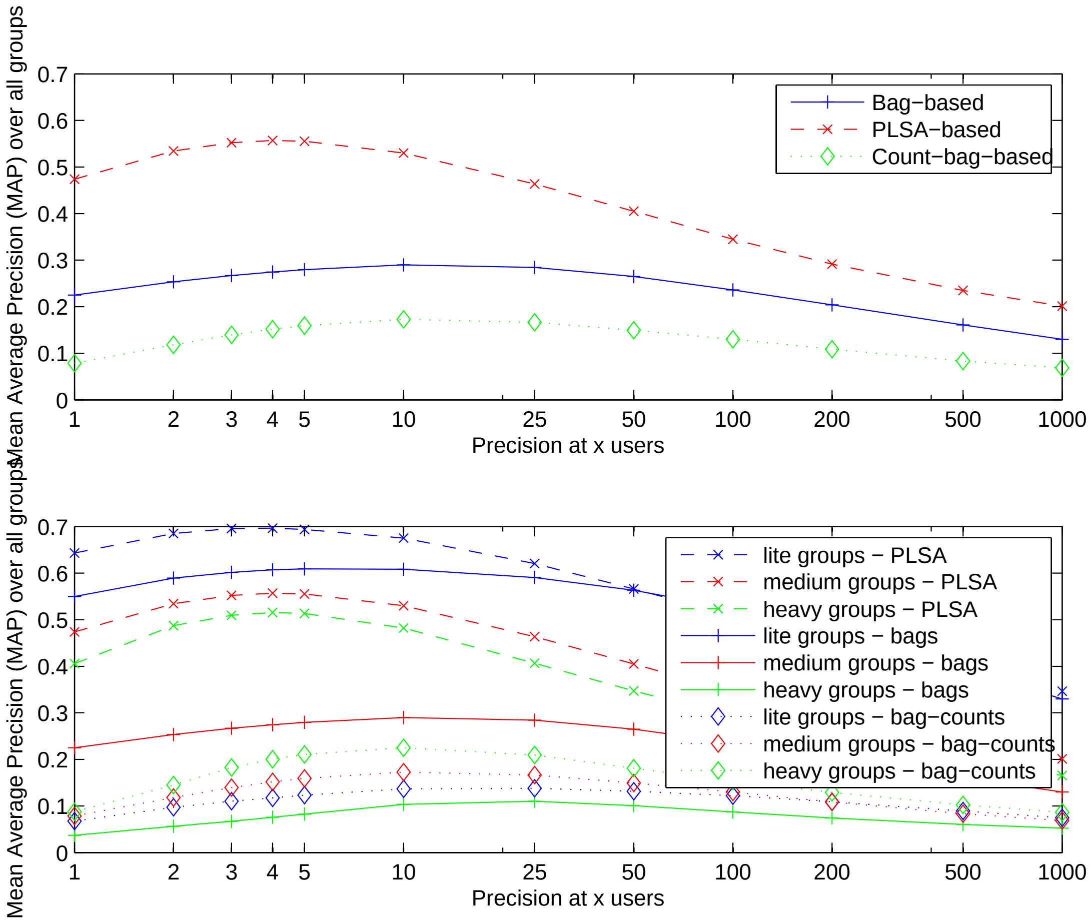


Figure 3.10: User retrieval MAP - Mean Average Precision for the user retrieval experiment, computed for all data (top) and separately for lite, medium and heavy groups in terms of user memberships (bottom). Lite groups are in the first quartile (less than 12 users), medium groups in the second and third quartiles (12 to 49 users) and heavy groups in the forth quartile (more than 49 users).

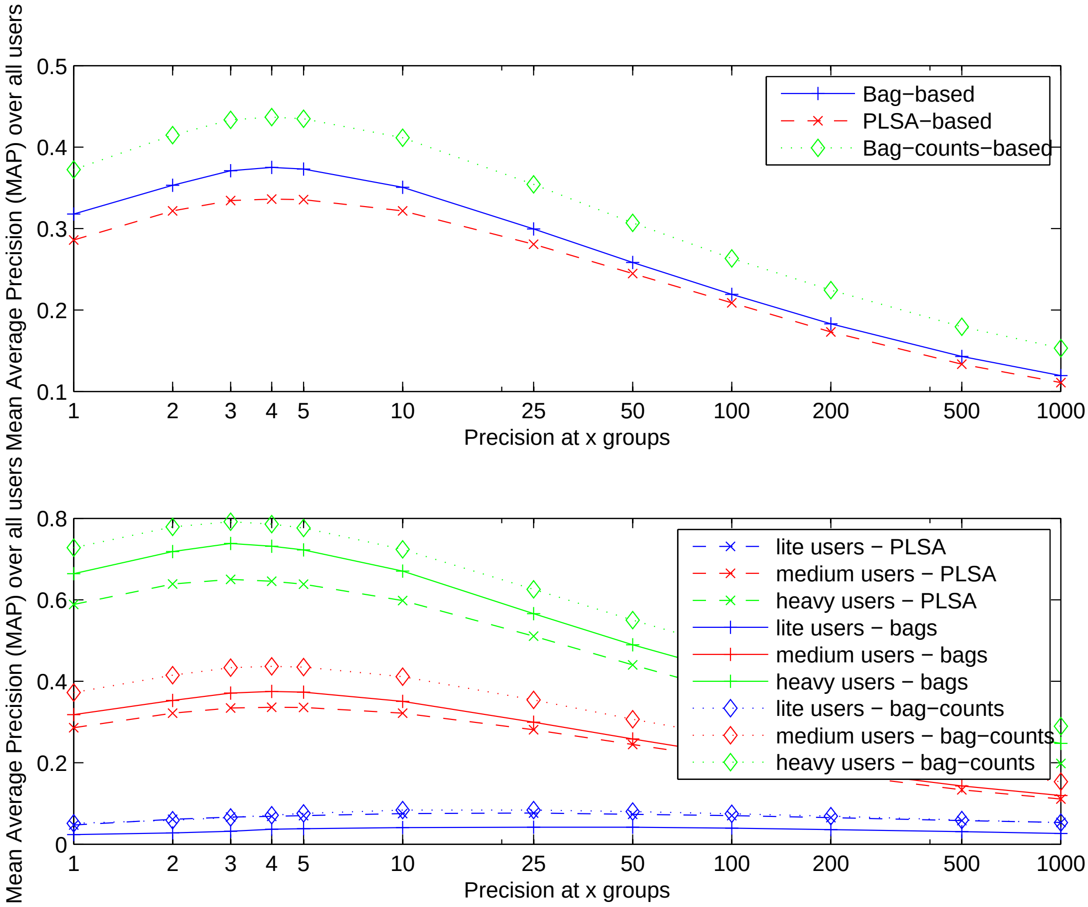


Figure 3.11: Group retrieval MAP - Mean Average Precision for the group retrieval experiment, computed for all data (top), and separately for lite, medium and heavy users in terms of group membership (bottom). Lite users are in the first quartile (less than 10 groups), medium users in the second and third quartiles (10 to 91 groups) and heavy users in the forth quartile (more than 92 groups).

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 3,091 for the topic-based method, 5,446 for the bag-based one and 35,693 for the bag-count-based similarity.

For the second experiment, group retrieval from users (Figure 3.11, 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-binary-based similarity measure retrieves only 12 different groups for the 8,000 users (which additionally correspond to the largest groups in terms of members, with an average of 894 users per group), while the bag-counts-based measure retrieves 85 different groups (also the largest groups as well as some medium sized ones, on average (respectively median) 327 members per group (respectively 113)). In contrast, the topic-based similarity measure retrieves 3137 distinct groups, with an average of 25 members (and median 12 members) per group. This indicates that the bag-based similarity measures are heavily biased towards the largest and so the most popular groups, while the topic-based representation is able to return less popular groups, which may be desirable in the exploration scenario, as the user might access groups he might not otherwise. 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 forth 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 Figures 3.10 and 3.11 we show the breakdown by user and group types, respectively. For the group retrieval scenario (Figure 3.11 bottom), all three similarity measures perform similarly when exposed to all three types of users, preserving their relative ranking to each other. For lite users (sparse information) all three measures perform the worst, and their best performance is for heavy users (plenty of information). For the user retrieval scenario on the other hand (Figure 3.10 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 degrades as the size of the groups increases for the topic-based and bag-based measures, unlike the bag-counts-based measure, which performs the worst across all three types of groups, but it works better as the groups get larger.

3.4.2 Single and multi topic-based keyword search

As mentioned in the previous chapters, finding relevant groups in Flickr at the original time of writing (2008) was not a particularly easy task. Unless the group uses the searched keyword in its name, description, or in the group discussions, direct tag-matching against the group photo pool was not possible.

By using the topic model we can effectively transform the keyword into relevant topics using the $P(t \mid z)$ matrix. We select those topics and then retrieve the most likely entities for each individual topic, using the $P(z \mid E)$ 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 $P(z \mid t)$ for the given tag, we can then compute the distance ρ from the full 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 ten results for the tag *guitar* in Table 3.7 using the current Flickr search method (FS), TES and TEDS.

The search for the keyword *guitar* on Flickr yields about six thousand 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.

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 3.7: Flickr search (FS), topic-expert search (TES) and tag-entity distance search (TEDS) results for the tag *guitar*.

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 3.8: Flickr search (FS), topic-expert search (TES) and tag-entity distance search (TEDS) results for the tag *artist*.

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

Table 3.9: Flickr search (FS), topic-expert search (TES) and tag-entity distance search (TEDS) results for the tag *airplane*.

FS	TES	TEDS
Ericaceae	Only pink flowers	Azaleas and Rhododendrons
PLANT	Flower Petal Macro ... Petal Art	Lilies (3/day)
Plant Taxonomy	WEEDS - SO MISUNDERSTOOD!	Purple Flora
UBCBG Botany Photo of the Day	Flowers of Passion	Daffodil World
Guide to Oregon Wildflowers	Flowers with Rain Drops	Botany
CaliFlora	flowerhearts - 3 pics a day!	Daisy Chain....
Orman Gülü Çiçek Grubu Fotoğraflar, Resimler ve Videolar Pa	Daisy Chain....	Iris Flowers
Azaleas and Rhododendrons	Nature Photo Close-Up - 3 per day	Orchids
1001 Gardens You Must See...	Pollen Swapping	Colorful Flowers
Vancouver Island Wildflowers	FlowereZ	Pollen Swapping

Table 3.10: Flickr search (FS), topic-expert search (TES), and tag-entity distance search (TEDS) results for the tag *ericaceae* (a plant family comprising cranberries, blueberries, azaleas, and rhododendrons, amongst others).

Another interesting example is the search for the tag *artist*, presented for the three methods in Table 3.8. The topic-based searches retrieve mostly groups about drawing and painting that, 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 and forth example for the tags *airplane* and *ericaceae* are shown in Tables 3.9 and 3.10, where we can also see the quality of the returned search results for TES and TEDS compared to that of FS.

From these examples it is clear that 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 topic model.

3.5 Model generalization

In constructing our reduced dataset D_R, discussed in Section 2.2.4 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 10K vocabulary. This threshold gives us roughly 30K groups and 10K users for a total of 40K 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 Figure 3.12. In this case, on the left, the entity is a group, *Arabic Weddings*, with a vocabulary of only 3 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 10K vocabulary seemed 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 Figure 3.13.

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 8 relevant topics in Figure 3.3 to about 3 relevant topics for both users and groups in Figure 3.13, 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 2 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

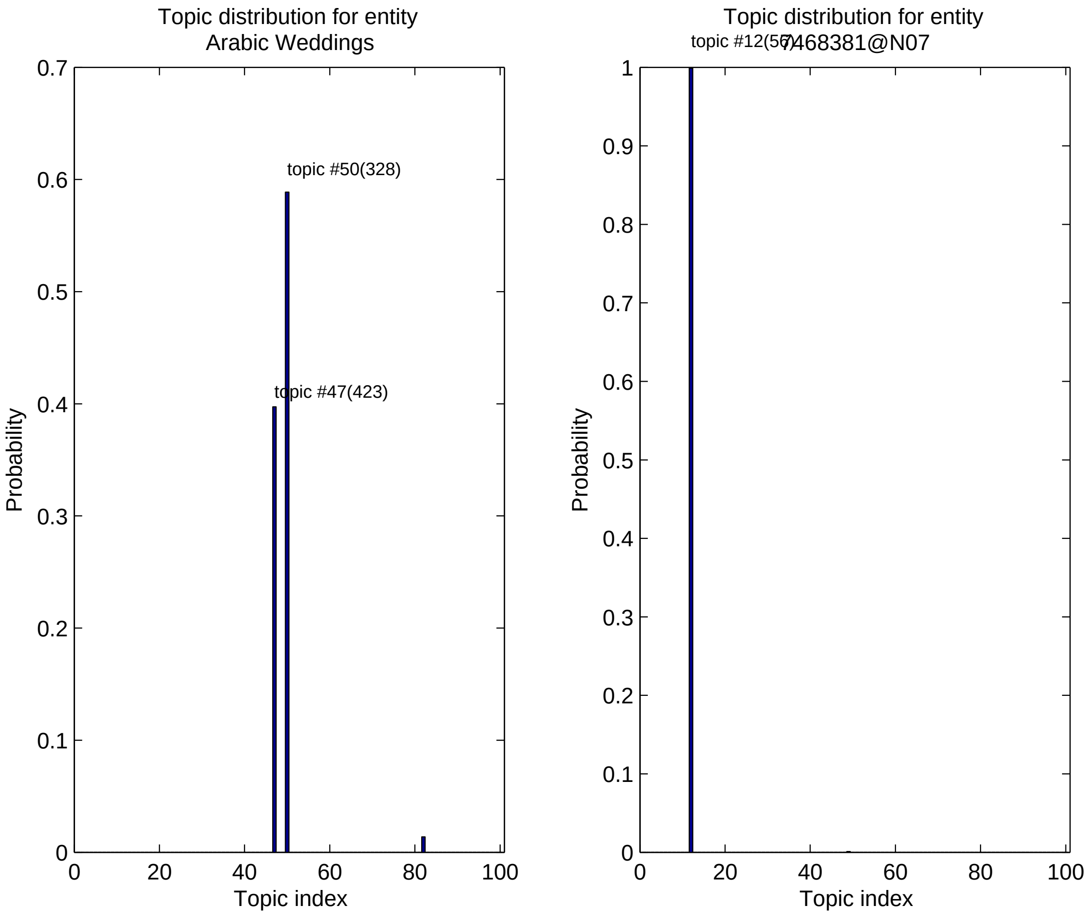


Figure 3.12: Content-poor entitites - On the left, the topic representation of an entity (the group *Arab Weddings*) with only 3 unique tags in its vocabulary: *john*, *dancing*, *wedding*. The relevant topics 47 and 50 are mainly about *parties and friends* and *weddings and proper names* respectively. On the right, the topic representation of an entity (user 7468381@N07) with only 1 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.

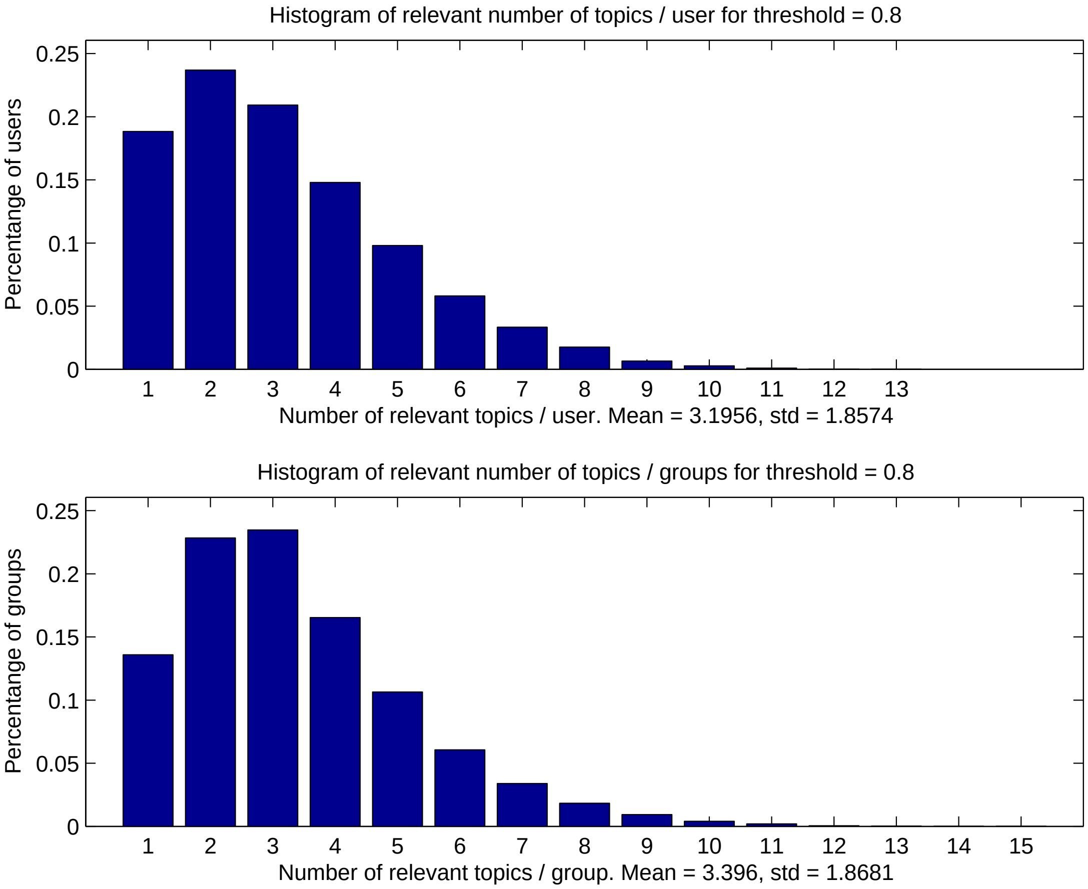


Figure 3.13: Content-poor dataset - Comparison of the number of relevant topics for groups and users with at most 50 unique tags in their vocabularies. For ease of comparison, we normalized the histograms and display on the y axis the percentage of users and groups respectively.

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 “content-poor” entities. This shall be an open issue for 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 hours on a IntelCore2 CPU 6700 machine with 3GB RAM, running at 2.66GHz. On a new document, inference takes in the order of 2 seconds. Learning the full topic model in principle can be sped up through a number of strategies, discussed for instance in a number of recent works including [51], or [70]. 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 is 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 [23]), which would be another relevant direction to pursue in the future.

3.6 Conclusions

Social media repositories such as Flickr constitute an emerging challenge for multimedia information management systems. We have analyzed in this chapter an unexplored issue, that is jointly modeling Flickr users and groups. Our analysis in Chapter 2 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. 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 [9]. Hierarchical topic models may also be a viable alternative to be explored in the future.

We have also 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 one way to offset the sparsity.

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. Evaluation of the model performance in prediction scenarios, with part of the data held out for validation, and membership used as ground truth may be another way of assessing the quality of topic models for community modeling. 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, as we have shown in the second part of Chapter 2, the different interaction modalities available on each site result in different ”annotation qualities” and as such a bag-of-words model can be a good representation of users, if the vocabularies are comparable. Some of these issues are investigated in the next chapter. 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. Lastly, there is the obvious 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.

4

Kodak Moments and Flickr Diamonds - jointly modeling disjoint communities

We examined in Chapter 2 some of the vocabulary characteristics of Flickr and Kodak Gallery. We have seen that, despite inherent differences induced by their users' motivations and their needs, as well as differences induced by system design and affordances, certain similarities at the vocabulary level exist. This encourages us to attempt a joint modeling of Kodak and Flickr users.

We present in this chapter a probabilistic topic model for jointly representing Flickr and Kodak Gallery users, in very much the same way as we previously modeled Flickr groups and users. Using a large-scale dataset from both systems, we show that two distinct sharing behaviors can be observed, in line with results from an ethnographic study by Miller and Edwards [42]. The material in this chapter was written in 2010 and appeared in part in [50].

We start by discussing the related work, then we discuss the topic model in Section 4.2, and then we present in Section 4.3 a topic-based analysis of the differences between users coming from the two systems, Flickr and Kodak Gallery. We conclude with some considerations on the characteristics of the model itself in Section 4.3.2.

4.1 Related work

In recent years, as more and more online systems obtain and use user information, there has been an increasing desire to be able to share user profiles between such disjoint systems[4, 8, 12, 21, 27, 55]. Although privacy and ethical considerations are of critical importance, they are outside the scope of this discussion, and we will concentrate purely on the interoperability and cross-system modeling literature.

If an internet user belongs to two different online communities, it is reasonable to assume that the two systems serve different needs of the user. Differences may exist either at the level of the functionalities, the type of content, the community to which the user has access, or indeed any combination of the above.

Two main approaches to user modeling across different systems emerge in the related literature, and are mainly related to recommender systems. One is based on standardized ontologies and/or unified user models, and the other on mediation between different models. Attempts to propose standards for user modeling started in the early 90's[4], and followed with approaches that were mainly focused on the reusability of a user's profile accross different systems, for example in e-learning scenarios [21]. By deffinition, these proposed approaches were quite rigid, and have mostly failed to be adopted as feasible solutions in cross-system modeling.

The second main approach to user modeling across systems abandons the idea of a unified model and tries instead to mediate, or map different user models to one another by a set of mapping rules, or by using meta-models. A formal definition of *mediation* is given by Berkovsky *et al.* [8] as "a process of importing the user modeling data collected by other [...] systems, integrating them and generating an integrated user model for a specific goal within a specific context", and is extended in [7] to explain the integration part as a "set of techniques aimed at resolving the heterogeneities and inconsistencies in the obtained data". Other authors, such as Gonzalez *et al.* [27], used a multi-agent approach to split a "smart user model" into objective, subjective, and emotional features of the users. In two different studies, Carmagnola *et al.* [12, 13] looked at two different issues: on one hand, at the model level, an exchange of information based on tag-enriched profiles, while on the other, at the system level, identification of users across systems.

In contrast, in our work we do not need nor wish to identify users who use both systems, and our cross-system model is particularly useful from the content description perspective. The question of how this unified model might actually be implemented in practice is still open.

For a broader look into cross-system user modeling we refer the reader to a recent survey of Viviani *et al.*[68].

4.2 A probabilistic model for Kodak and Flickr users

Using the same analogy as in Chapter 3, Flickr and Kodak users can be seen as text documents composed of the tags associated with their photos, in no particular order. In the same way a text document can be more succintly described by a small number of subjects it treats, a user can be described by a few recurring themes or topics of interest. These interests are not always explicit, but they can be inferred from the complete user collection of photos, or tags, and as such are likely to be discovered by a probabilistic topic model such as the one we used for Flickr groups and users. We will describe in the following section a probabilistic model that represents an improvement over PLSA, called Latent Dirichlet allocation (LDA), first proposed by Blei *et al.* [10].

4.2.1 Latent Dirichlet allocation

LDA [10] is a generative model that assumes, like PLSA, that documents in a corpus are a low-dimensional mixture of hidden topics of interest. With respect to PLSA, LDA is fully generative, and this may be an advantage in cross-system scenarios such as the one we study here, where being able to infer topic distributions for new users of either system may be important.

The model learns, in an unsupervised way, a word-topic and a topic-document distribution from the corpus. The basic LDA model, where the topic-words distributions are smoothed according to a Dirichlet distribution conditioned by the parameter β, is shown in graphical representation in Figure 4.1. The shaded node represents words, and is the only observed variable, while the unshaded nodes are all unobserved variables. α and β are corpus level parameters of symmetrical priors of Dirichlet distributions, and are assumed to be fixed while generating the corpus. The plates show repeated sampling for the variables they enclose: the outer plate denoted D shows the number of documents that are sampled, the plate denoted T shows the number of topics sampled, and N_d shows the number of words per document that are sampled when generating a document. The variables θ are document-related, and they are sampled once per document, ϕ are topic-related and they are sampled once per topic, while z and w are word-related, and are sampled once for each word in each document.

LDA assumes the following generative process for each document (in our case each user U):

1. Choose $\theta^{(d)} \sim Dirichlet(\alpha)$, where d is the user index and $d \in \{1, .., D\}$;
2. Choose $\phi^{(z)} \sim Dirichlet(\beta)$, where z is the topic index, and $z \in \{1, ..., T\}$;
3. For each of the N_d words w in a document U:
   i) Choose a topic $z \sim Multinomial(\theta^{(d)})$;
   ii) Choose a word $w \sim Multinomial(\phi^{(z)})$.

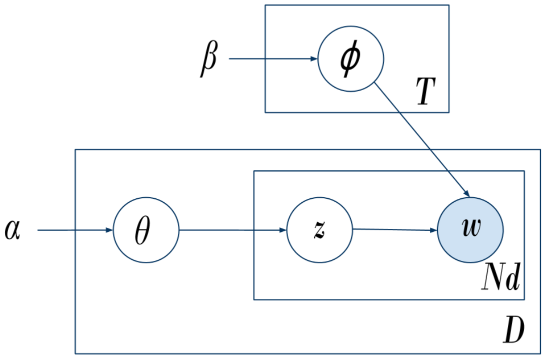


Figure 4.1: LDA - Graphical model representation of LDA. α and β are the parameters of the prior Dirichlet distributions, θ represents the probability distributions of documents over topics, ϕ the probability distributions of topics over words, z are the latent topics, and w, the only observable variable, are the words in each document. The plate notation shows repeated nodes, with N_d showing the repeated choice of words and topics in a document, T denoting the number of topics, and D representing the number of documents in the corpus, in our case, users.

4.2.2 Learning the model

Because exact inference in LDA is known to be intractable, we used collapsed Gibbs sampling with 5000 iterations, as proposed in [28]. Unlike in other approaches, θ and ϕ are not considered parameters to be estimated, but instead they are considered fixed.

Estimates of θ and ϕ are then obtained by examining the posterior distribution over the assignments of words to topics, which we denote by $P(\mathbf{z} \mid \mathbf{w})$, where $\mathbf{z}$ and $\mathbf{w}$ represent the corpus-level topics and words respectively. In order to evaluate $P(\mathbf{z} \mid \mathbf{w})$, a Markov chain Monte Carlo procedure is employed. A Markov chain is constructed such that it converges to the target distribution, and samples are taken from it. Each state of the chain is an assignment of values to the variables being sampled, in our case $\mathbf{z}$. A simple rule is used for transitions between states. The next state is reached by sequentially sampling each variable from its distribution when conditioned on the current values of all other variables and the data. To apply this algorithm, the full conditional distribution $P(z_i \mid \mathbf{z}_{-i}, \mathbf{w})$ is needed, which is expressed as:

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}, \tag{4.1}$$

where W is the vocabulary size, T is the number of topics, the notations $n_j^{(w)}$ and $n_j^{(d)}$ are counts of the number of times word w has been assigned to topic j, and the number of times a word from document d has been assigned to topic j respectively. $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment of z_i. In other words, the negative subscript $-i$ indicates exclusion of the i^{th} component. Equation 4.1 represents an intuitive result: the first ratio is the probability of the word w_i under topic j, while the second ratio is the probability of topic j in document d_i, with α and β acting as smoothing parameters of these counts. The fact that these counts are the only information needed for computing the full conditional distribution is what allows this algorithm to be implemented efficiently.

Once the full conditional probability is obtained, the Monte Carlo algorithm initializes the z_i variables to values in $\{1, ..., T\}$, thus determining the initial state of the Markov chain. An online version of the Gibbs sampler is used to perform this operation, using Equation 4.1 to assign words to topics, but with counts computed from the words seen so far, and not the full data. A number of iterations is run on the Markov chain, each time finding a new state by sampling each z_i as specified above. The chain converges to the target distribution after a number of iterations, and from that point forward samples of the z_i variables are recorded every several iterations, introducing an appropriate lag to ensure that auto-correlation is low. The last sample is then used to compute the word-topic and topic-document distributions, given by:

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}, \tag{4.2}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \tag{4.3}$$

In our case documents are Flickr and Kodak users, represented as bags-of-tags. As we have seen previously, the two vocabularies only share 56% of the tags, therefore the joint vocabulary is composed of almost 15,000 words. In order to avoid an unbalanced dataset due to the much larger number of Flickr users, we randomly sample 5,400 users from Flickr to match the 5,400 users from Kodak, for a total of 10,800 users.

In Figure 4.2 we show the histograms of tag occurrences per user, split by their original dataset. Kodak users have a median of 34 tags, and standard deviation 525, while Flickr users have a median of 124 tags, with standard deviation 867. These statistics show there is a high variability even within each of the two datasets. In terms of vocabulary size per user (number of unique tags), Kodak and Flickr users have a median of 8 and 15 tags respectively, with standard deviations of 94 and 148.

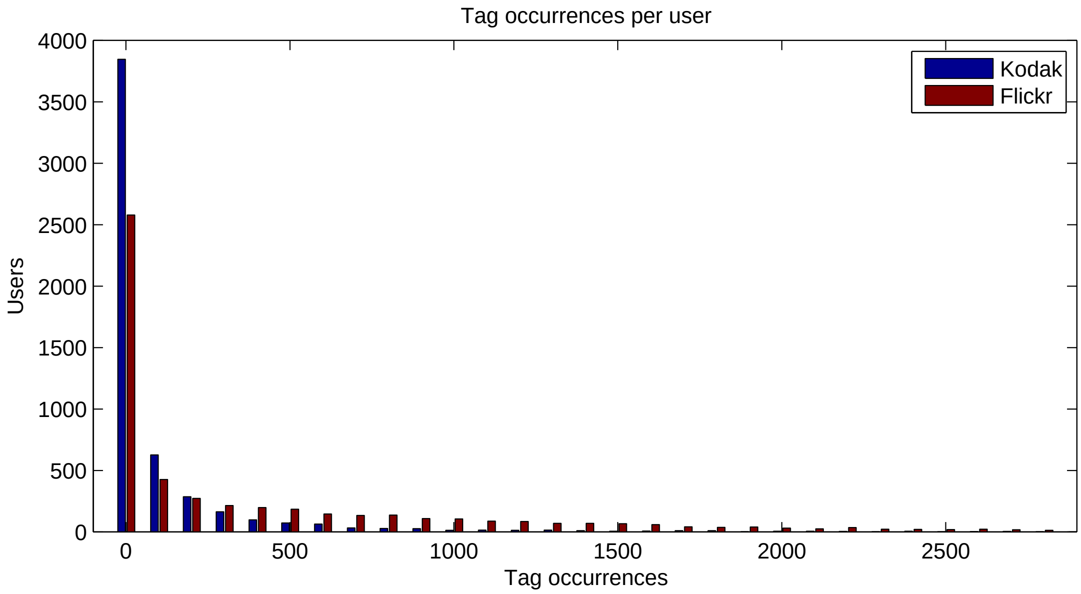


Figure 4.2: Vocabulary statistics - Number of tag occurrences by user split by their belonging to either the Kodak or Flickr dataset.

We trained the LDA model using a bag-of-tags representation for all users, counting for each of them the number of times any given tag was used. The model parameters are the number of topics $T = 200$, the parameters of the per-user topic distribution $\alpha = 50/T$, the parameter for the per-topic word distribution $\beta = 0.01$, and $G = 5000$ the number of iterations for the Gibbs sampling algorithm. Training time on a machine with Dual Core2 CPUs 6700 @ 2.66Ghz and with 3.5GB RAM is 15 hours, using the Matlab Topic Modeling Toolbox [28]. Although we train a joint model, we hypothesize that the inherent differences of the two populations should show up at the topic level.

4.3 Topic-based analysis

4.3.1 User analysis

As the output of the LDA model we have the distributions over topics for each user, or $P(\mathbf{z}|u)$, where $\mathbf{z}$ and $\mathbf{u}$ represent the hidden topics and the users respectively, as well as the distributions over words for each topic, denoted by $P(\mathbf{w}|\mathbf{z})$, where w represents the words. For each user it is then possible to compute which are the most relevant topics, by setting an arbitrary threshold τ on the cumulative sum of the most probable topics. We show in Figure 4.3 the histograms of the number of relevant topics per user for each of the two datasets, for $\tau = 0.8$. We observe that Kodak users are more likely to have fewer topics than their counterparts from Flickr. On average Kodak users are about 4.6 topics, while Flickr users are about 5.2 topics. This difference is statistically significant at $\alpha = 0.01$ in a two-tailed t-test.

Another way to assess the differences between the two populations of users at the topic level is by using the entropy of the topic distributions. Here we use entropy as a measure of the diversity of each user's topic distribution: the lower the entropy measure for a user, the less topics of interest he or she has.

We show in Figure 4.4 the distributions of the entropies for the two sets of users, with the distribution for Kodak users in the upper part of the figure, and the one for Flickr users on the lower one. The two distributions are significantly different in a two-tailed test with $\alpha = 0.001$, with a mean of 0.25 for Kodak users and 0.28 for Flickr users. Here we can observe again the higher number of Flickr users with very low entropy, hence single-topic interests, as well as the fact that Flickr users' topic-based

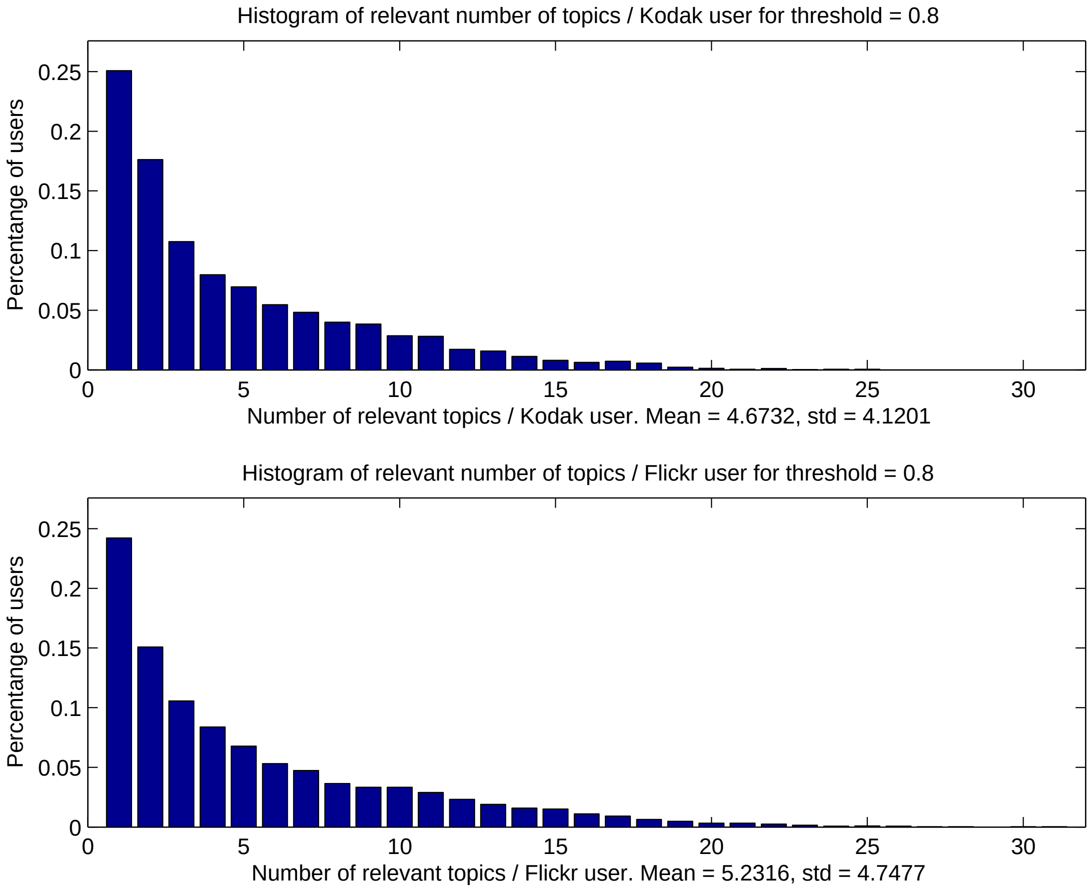


Figure 4.3: Relevant topics - Number of relevant topics per user, for the Kodak (top) and Flickr (bottom) datasets.

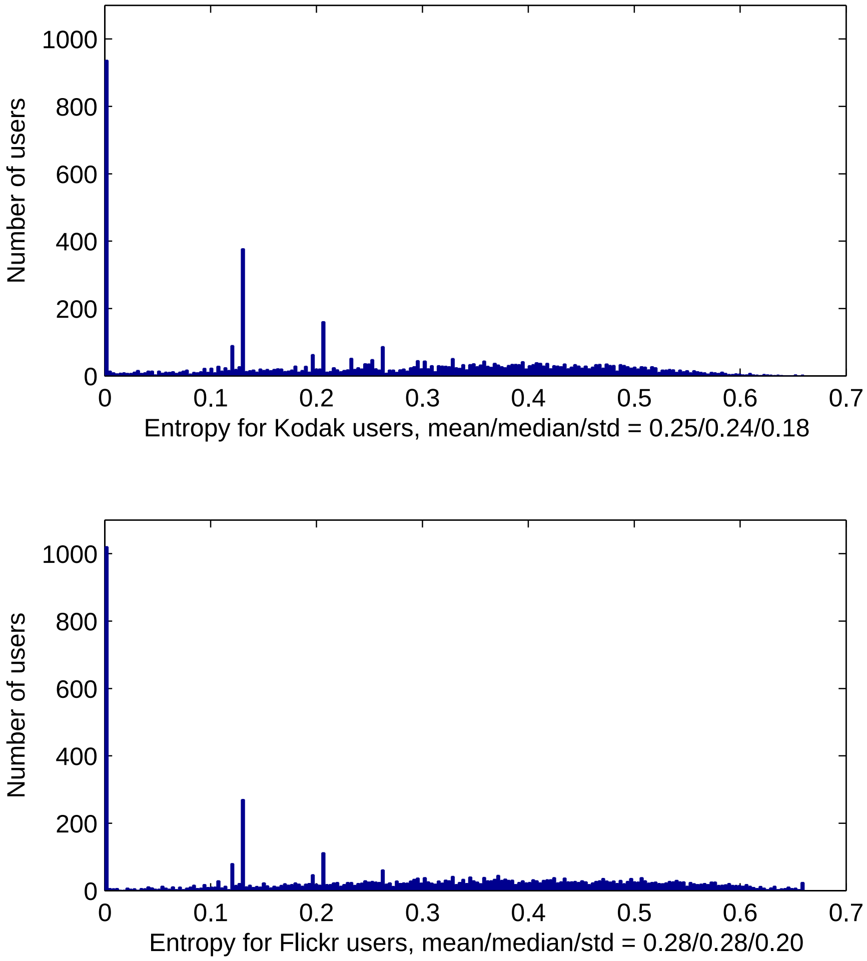


Figure 4.4: User entropies - Entropy distributions for the Kodak (top) and Flickr (bottom) datasets.

representations are in general more spread, which results in higher mean and median entropy values at the population level.

4.3.2 Model analysis

We now turn our attention towards the topic model itself. As the LDA model is learned on the joint vocabulary and all users irrespective of their original data set, we are interested in the differences that can be observed at the topic level. In Figure 4.5 we show a plot of the topic specificity among Flickr and Kodak users.

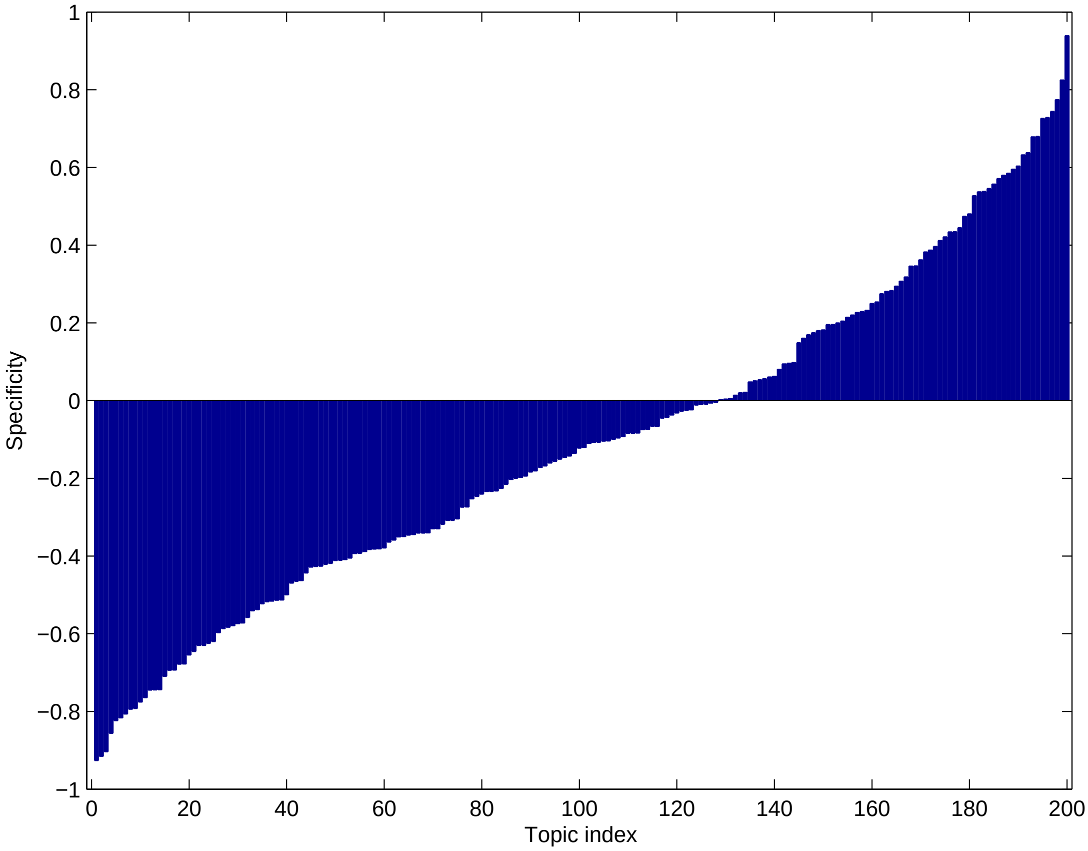


Figure 4.5: Topic specificity - Topic specificity among the two communities. Specificity is computed as the ratio of the difference between Kodak and Flickr users for which that topic is relevant, and the total number of users for which the topic is relevant. Positive values of specificity thus imply that a topic is relevant for more Kodak users than Flickr users, while negative values imply that the topic is relevant for more Flickr users than Kodak.

We define topic specificity as the ratio of the difference between the number of Kodak and Flickr users for which that topic is relevant and the total number of users for which the topic is relevant. This quantity is therefore bounded between -1 and 1. Positive values of specificity thus imply that a topic is relevant for more Kodak users than Flickr users, while negative values imply that the topic is relevant for more Flickr users than Kodak. In this figure, the topics are ordered by their specificity in order to improve trend readability. We can see that about 64% of the topics are more specific to Flickr users. In fact, almost 20% of the topics are relevant for twice as many Flickr users than Kodak ones (specificity below -0.5), while on the other side, 10% of the topics are relevant for twice as many Kodak users than Flickr ones. This is most likely a direct result of the relative imbalance at the bag-of-tags level, as users of the two systems are equally represented in our data.

We also show in Tables 4.1 and 4.2 some examples of the topics discovered by our model. On the left, we show the most probable words for that topic, effectively giving an understanding of what the topic is about. On the right, we show the most probable users for the same topic, identified by their IDs. Flickr users' IDs start with the letter *F*, while Kodak users' IDs with the letter *K*. Some of the topics are dominated by Kodak users, some others by Flickr users, and there are also topics where the top entities are a mix of Kodak and Flickr users, like topics #19, #24, and #77. From the distributions $P(\mathbf{w}|\mathbf{z})$ we can extract the most probable 10 words for some of these "special" topics.

We show in Table 4.3 the most probable seven words for topics taken from the three regions of topic specificity: Flickr specific, cross-over topics, and Kodak specific. One of the main Flickr topics (specificity below -0.80) is characterized by words such as *abigfave, explore, bravo*, and *impressedbeauty*, a vocabulary found exclusively in Flickr, and related to photo-exposure activities within Flickr, these words being tags attached to photos that are invited to specific Flickr groups. Another Flickr specific topic, having the highest specificity at -0.90, is characterized by words such as *lomo, lca, xpro, crossprocessed, film, analog*, and *fisheye*, terms very specific to a particular photography technique and aparatus.

In the middle of the table we find examples of topics specific as much to Flickr users as to Kodak users, such as those defined by the words *marathon, newyearseve, ottawa, karaoke, george*, and *nightlife*, or *wedding, reception, media, jen, dinner*, and *bachelorette*, topics that describe mostly priate or public events, and places.

Topic 19	
$P(t \mid z)$	**Tag**
0.4109	concert
0.0521	concerts
0.0329	musicians
0.0279	drag
0.0254	dragqueen
0.0221	patrick
0.0220	harrypotter
0.0217	music
0.0184	bbq
0.0168	shows
0.0166	sheila
0.0153	russell
0.0114	beyonce
0.0113	icons
0.0109	story

Topic 19	
$P(z \mid E)$	**Entity**
1.0000	K-10271111312
1.0000	K-139702247112
1.0000	K-160533531112
1.0000	K-215429381112
1.0000	F-88588822
1.0000	F-11704188
1.0000	F-60267693
1.0000	F-51368284
0.9811	K-240588927112
0.8696	F-61555160

Topic 24	
$P(t \mid z)$	**Tag**
0.1673	garden
0.0610	flowers
0.0452	spring
0.0296	flower
0.0157	backyard
0.0152	leaves
0.0125	tree
0.0118	gardening
0.0107	pond
0.0101	plants
0.0099	wildflowers
0.0091	tulips
0.0087	purple
0.0084	pink
0.0079	green

Topic 24	
$P(z \mid E)$	**Entity**
1.0000	K-198702771112
1.0000	F-36009846
1.0000	F-51318048
1.0000	F-94249917
1.0000	F-70422559
0.7826	F-13244772
0.7083	K-177727010112
0.6667	F-33802167
0.6571	F-32928279
0.5103	F-40351040

Topic 52	
$P(t \mid z)$	**Tag**
0.2954	cat
0.1233	cats
0.0520	kitten
0.0353	kitty
0.0188	tabby
0.0154	feline
0.0121	modeling
0.0107	cute
0.0100	barn
0.0098	kittens
0.0094	basketball
0.0092	kitties
0.0082	pussy
0.0080	dirt
0.0075	fireworks

Topic 52	
$P(z \mid E)$	**Entity**
1.0000	F-12554088
1.0000	F-34397348
1.0000	F-45291351
1.0000	F-60263124
1.0000	F-51356455
1.0000	F-38758195
1.0000	F-7656004
1.0000	F-71785859
1.0000	F-81734161
1.0000	F-70059713

Table 4.1: Example topics from the model: the top most probable 15 words for each topic and the most probable 10 users. Kodak user IDs start with the letter **K**, and Flickr user IDs start with the letter **F**.

Topic 60	
$P(t \mid z)$	**Tag**
0.1213	mom
0.0739	dad
0.0284	emily
0.0195	grandma
0.0179	jim
0.0173	joe
0.0151	jeff
0.0142	dave
0.0133	uncle
0.0120	grandpa
0.0105	aunt
0.0105	michael
0.0103	mike
0.0093	dads
0.0091	ron

Topic 60	
$P(z \mid E)$	**Entity**
1.0000	K-10108342712
1.0000	K-108613447112
1.0000	K-113767892112
1.0000	K-137469224112
1.0000	K-190045142112
1.0000	K-225624955112
1.0000	K-229511657112
1.0000	K-236383727112
1.0000	K-246507276112
1.0000	K-250349310112

Topic 77	
$P(t \mid z)$	**Tag**
0.4055	sanfrancisco
0.0436	pix
0.0399	tahoe
0.0243	monterey
0.0214	sausalito
0.0205	goldengate
0.0177	goldengatepark
0.0166	j
0.0158	alcatraz
0.0124	reno
0.0116	marin
0.0113	oakland
0.0111	goldengatebridge
0.0089	gaypride
0.0089	mission

Topic 77	
$P(z \mid E)$	**Entity**
1.0000	K-241644021112
1.0000	F-20872613
1.0000	F-91416511
0.9444	F-67839131
0.9058	K-146725060112
0.9054	F-60365458
0.8019	F-62118219
0.6889	K-254026706112
0.6667	F-50858610
0.6667	F-86708053

Topic 79	
$P(t \mid z)$	**Tag**
0.7777	picture
0.1595	pictures
0.0428	edited
0.0019	png
0.0016	mackinaw
0.0015	robbie
0.0013	bass
0.0010	sauce
0.0009	round
0.0008	barbie
0.0007	aimg
0.0007	dot
0.0006	project
0.0006	confirmation
0.0006	smilebox

Topic 79	
$P(z \mid E)$	**Entity**
1.0000	K-10080125512
1.0000	K-10188098812
1.0000	K-10222937512
1.0000	K-10236932612
1.0000	K-10264394212
1.0000	K-10318458312
1.0000	K-10322057712
1.0000	K-10705003912
1.0000	K-10736235612
1.0000	K-10766279112

Table 4.2: Example topics from the model: the top most probable 15 words for each topic and the most probable 10 users. Kodak user IDs start with the letter **K**, and Flickr user IDs start with the letter **F**.

Most probable words for Flickr specific topics (specificity between -0.90 and -0.77)

lomo	lca	xpro	crossprocessed	film	analog	fisheye
selfportrait	photoshop	self	wow	collage	dark	portrait
nyc	manhattan	tibet	centralpark	soho	subway	brooklyn
japan	tokyo	kyoto	ricoh	osaka	sakura	temple
abigfave	explore	bravo	impressedbeauty	ben	melissa	e
film	holga	mediumformat	kodak	leica	toycamera	blur
nikon	dslr	nikkor	coolpix	micro	water	portrait
cameraphone	treo	motorola	office	torino	working	stella
bw	color	blackwhite	portrait	origami	blackandwhite	sepia
nature	flora	photography	indiana	ilovenature	ivy	county

Most probable words for cross-over topics (specificity between -0.02 and 0.03)

ontario	panasonic	lumix	jon	scott	ye	connecticut
vancouver	chris	morocco	mountainbiking	victoria	whistler	northshore
tn	sigma	fiesta	childhood	saopaulo	cincinnati	balloon
marathon	newyearseve	big	ottawa	karaoke	george	nightlife
florida	miami	meetup	vermont	photowalk	tampa	keywest
europe	disneyland	jpeg	castle	hearst	lg	fam
christmas	bike	bicycle	emily	engagement	cycling	presents
hongkong	cottage	town	va	richmond	tacoma	isle
atlanta	happy	georgia	kurt	musicians	concerts	costume
graduation	gay	pride	college	parade	rally	protest
wedding	reception	media	jen	dinner	rich	bachelorette

Most probable words for Kodak specific topics (specificity between 0.70 and 0.95)

bday	bmp	rugby	tigers	final	cam	raiders
mom	dad	grandma	michelle	joe	brother	grandpa
view	room	bldg	kitchen	front	bedroom	master
dcp	tim	becky	w	thomas	sweet	size
dec	feb	nov	oct	pix	sept	j
view	back	top	road	river	front	bay
m	e	s	t	d	high	c
grad	rosa	bible	sun	wed	dec	omaha
liam	al	jim	taylor	mary	dylan	riley
i	daddy	time	im	big	mommy	good
picture	png	raj	mackinaw	robbie	bass	round

Table 4.3: Examples of topics grouped by their specificity to Flickr users (top), Kodak users (bottom), or neither (middle). The most probable 7 words for each topic are listed on each table row.

At the other extreme, amongst the Kodak specific topics, we find topics described by words such as *mom, dad, grandma, brother*, and *grandpa*, or by words such as *view, back, top, road, river, front*, and *bay*, mostly related to family and vacations. We are obviously looking on one side at the nature-loving and exposure-seeking users of Flickr, and the family oriented users of Kodak Gallery on the other, with an interesting meeting zone in the middle of the specificity graph.

4.4 Conclusions

We presented in this chapter a large-scale cross-site topic modeling of two large online photo-sharing communities, Flickr and Kodak Gallery, which so far have not been jointly analyzed. The choice of the two communities is not arbitrary, but partly dictated by the pragmatic fact of data availability. Not only are Flickr and Kodak Gallery two of the most important photo-sharing communities, one for wide audiences, and the other for closed social circles, but data from both systems could be obtained relatively easy compared to other such websites.

The results we presented in Chapter 2, coupled with the joint topic modeling of the two datasets shown here provide strong support for the observations reported by Miller and Edwards in their ethnographic study with 10 users [42]. This time on a joint dataset with 5 million images and more than 10,000 users, we show two types of emerging phenomena: *Flickr Diamonds*, the product of *Snaprs* who tend to take photos and share them with the world, and whose main concern is very likely artistic, and *Kodak Moments*, a product of the *Kodak Culture* users, who take photos mostly at family events, and mainly share them within their existing social circle.

We believe that this study also points out the potential for large-scale studies, which are nowadays much easier to perform, unlike ethnographic studies which are usually small-scale, given their time and subject-effort intensive nature. Large-scale analysis might become the first step in the research process, with in-depth ethnographic studies as a second step once a preliminary hypothesis has been chosen for verification.

A drawback of a cross-site modeling like the one we presented here is the lack of objective evaluation methods. While our model allows us to differentiate between two different types of behaviors, corresponding in general to the two types of users found in the analyzed systems, there is no straightforward way of assessing the “goodness” of the model.

Future work could investigate the use of additional metadata, such as gender and occupation, as well as the impact of having access to the full user photo collection (including private photos). Also, while the results suggest that system design has a definite impact on the data being created by users, we have not addressed the problem of using the findings of the study to provide specific guidelines for system designers, as an interpretation of our results in this direction is not straightforward. This could be the subject of cross-field research in the future.

5

Beyond groups: Flickr hyper-communities

In this chapter we propose a novel method to discover *hypergroups* in Flickr, that is, communities consisting of groups of Flickr groups. Our hypothesis is that groups that are similar probably host the same kind of content (in terms of images and associated tags), and depending on their popularity, they may also share an important number of members. Based on this observation, our work has four major contributions.

First, starting from almost 11,000 groups, we propose to use these two sources of information, content (through photo tags) and relations (through group memberships) in a bag-of-words model to represent groups in Flickr. In particular, we propose a novel angle to modeling relations. While traditional approaches to social networks have mainly examined a user's explicit contacts, participation in the same groups can also be viewed as an implicit social link; this is how we will approach relations in this work.

Second, using a probabilistic topic model, we build three comparable topic-based representations, one based on content, one based on binary membership links, and a hybrid, based on membership links weighted by the content-wise contributions of the user to the group.

Third, we employ a state-of-the-art clustering algorithm that discovers cohesive hypergroups, and analyze and compare the three models from several viewpoints.

Finally, we develop an annotation tool that allows us to perform a subjective evaluation of the quality of the hypergroups in a user study with 8 users.

Overall, our approach provides a prototype solution to the problem of how users can find interesting groups, as it allows users to find potentially unknown groups, that are still relevant to their search, based on how similar the target groups are to an example group a user would provide.

The part of this work related to hypergroup discovery was done in collaboration with a group of researchers from Curtin University of Australia in 2009, and it was published in Negoescu *et al.* [45].

Our approach is described in detail in Section 5.1, with an overview of related work in (hyper-)community detection following it in Section 5.2. The analysis of our method's results is presented in Section 5.3. Finally, we conclude in Section 5.5.

5.1 Hypergroup discovery

Finding groups in Flickr is relatively easy for popular groups whose names and/or descriptions include the keywords used for searching. However, when these keywords are not present, or when the group is not very popular, finding groups can be problematic. We propose as solution to this problem through the automatic discovery of hypergroups, or groups of groups, a process that allows a user to find interesting groups starting from one group he or she considers relevant.

The conceptual workflow for hypergroup discovery is illustrated in Figure 5.1. Groups can be seen as entities containing three types of data: users, through the memberships to the group; photos, through the photos contributed to the group pool by its members; and tags, associated with the photos in the group pool. Thus we start by creating three different bag representations for the groups, one based on a binary membership feature, one based on memberships and weighted by the tag contribution of the given user, and one on tags alone. Each of these bags is then used to learn a probabilistic topic model. Finding hypergroups is then cast as a clustering problem where the number of clusters is unknown. We will describe these steps in more detail in the following subsections.

5.1.1 Latent Dirichlet allocation

As presented in Chapter 4, LDA [10] is a probabilistic topic model, which is fully generative. When applied to a corpus of documents, it assumes these documents can

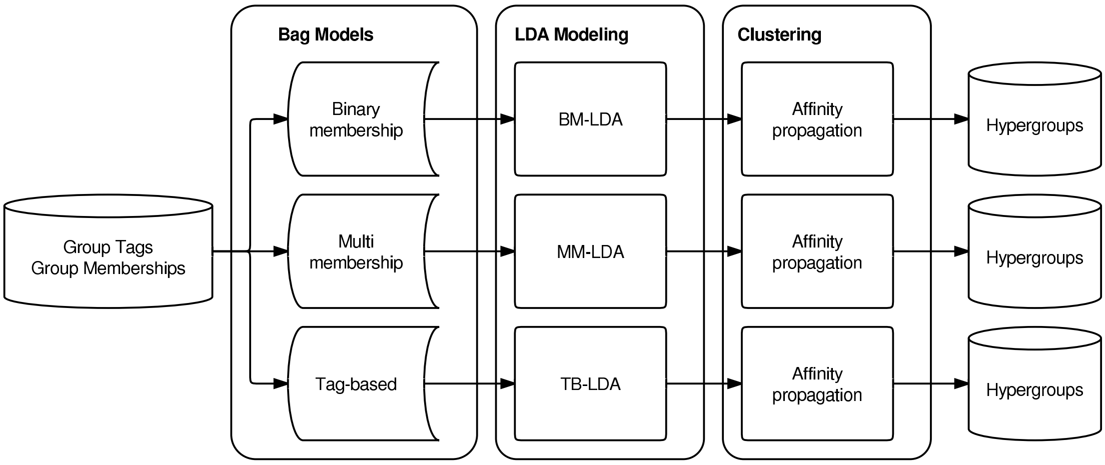


Figure 5.1: Hypergroup discovery workflow - From group content and membership we create bags of words, then we learn LDA models for each bag model, and finally we obtain hypergroups through clustering with affinity propagation.

be succintly described through a mixture of latent topics. In our case, we consider the Flickr groups to be the documents in the corpus, and several alternative representations of their "contents" are proposed, in the form of the three bag-of-words models previously described. The LDA model is then learned on each of these three representations separately.

Because exact inference in LDA is known to be intractable, we use again Gibbs sampling as proposed in [28], and described in more detail in Chapter 4.

5.1.2 Bag models

In order to test our hypothesis (similar groups have similar content and/or members), we develop three topic models: one based on a bag-of-tags representation for each group, and the other two on two different bag-of-members representations.

We construct three bag representations for the documents in our corpus, namely the groups:

1. a bag-of-users representation, by counting once each member of a given group; this is a binary-membership bag;
2. a bag-of-weighted-users representation, by counting for each user in a group all the unique tags they contributed to the group; thus this represents a membership bag too, but weighted by content, with multiple occurrences for the same user based on his or her contribution to the group vocabulary;

3. a bag-of-tags representation, by counting all the occurrences of a given tag in a given group's photo pool.

These three representations are just a small subset of the many representations for groups that may be envisaged. They are then used for learning three different topic models, using LDA.

5.1.3 Using LDA to characterize hypergroups

For brevity, we shall name the three models BM-LDA for binary-membership, MM-LDA for multiple-occurrence membership, and TB-LDA for the tag-based representation respectively.

For the membership-based representations we learned the LDA models starting from the two bags described in Section 5.1.2, i.e., binary-membership (BM-LDA), and multiple-occurrence membership (MM-LDA). The words in these two topic models are therefore users. Each topic, given that a word in the bag is a user ID, is characterized by a probability distribution over users, so their meaning is linked to shared memberships of the users in groups. Each group is now characterized by a probability distribution over topics, given by $p(z_u \mid G)$, where z_u is the notation for the user-based topics.

For the content-based representation, each document is also characterized by a distribution over topics, given by $p(z_t \mid G)$, where z_t is the notation for the tag-based topics. In the case of this model, given the words in the bags are tags, the learned topics are mostly topics of interest, described by semantically similar tags. As observed in previous work relying on similar models (PLSA) [35, 47], tag-based topics are likely to be homogeneous, and this is the case for this LDA-based model as well.

5.1.4 Clustering

For clustering we rely on a pairwise measure of similarity S between any two given groups starting from their topic-based representations. A few distribution measures were explored, including Kullback-Leibler divergence, and a parameterized (and thus generally asymmetrical) Jensen-Shannon divergence.

The similarity measure was calculated for every pair of groups, yielding a N_GxN_G similarity matrix, where N_G is the total number of groups. Hypergroup discovery is now cast as a clustering problem on this similarity matrix. A number of clustering algorithms could be used, and we chose the recently proposed Affinity Propagation method (AP) [26].

The algorithm can be summed up as follows. For each group G_i an entity C_i is created to represent its exemplar. A factor graph is then created with function potentials that encode the similarity between data points in addition to enforcing a valid configuration among exemplar nodes, where "valid" means that if a node i is voted as an exemplar by another node j, then i must vote itself as its own exemplar. Max-sum message passing is performed on this graph to minimize the energy function and the result is the desired partitioning of groups into hypergroups (clusters), each with a single exemplar. Intuitively, during clustering, most exemplars cede the right to be exemplars to another exemplar, through "negotiation" with other exemplars, accomplished through the passing of real-valued messages of two types, *responsibilities* and *availabilities*. The winning node (after negotiation) subsumes other exemplars into its own cluster.

AP has good properties for our problem: it is non-parametric, the number of clusters is automatically determined, and it does not assume the similarity function to be neither a metric, nor symmetric. Hence we can use any of the aforementioned similarity measures, some of which model asymmetric relationships in the formation of Flickr groups. For example, preferential attachment leads to nonreciprocating influence – a smaller group may be aware of a large, popular group, and mimic its tagging practices, without the larger group being aware of the smaller.

An additional benefit of AP is the discovery of *exemplars* as a by-product of the clustering process. Exemplars are the "most representative" members of a cluster, and hence provide a ready-made description of a hypergroup.

For a more detailed description of AP, we refer the reader to Frey and Dueck's work [26]. The results of the AP clustering algorithm applied to the three models separately are presented in Section 5.3.

5.2 Related work

With more and more social media systems becoming immensely popular, the problems of community detection and abstractization of content have been tackled by numerous research groups, and from different perspectives. We give in this section a brief overview of related work in the areas of topic models and (hyper-)community detection.

5.2.1 Topic models

Other topic-based models have been proposed in the context of text modeling [10, 61, 69].

In an earlier chapter, we have used Probabilistic Latent Semantic Analysis (PLSA) for modeling Flickr users and groups, a model first proposed by Hoffman in [29]. PLSA is a relatively simple generative probabilistic model, that has given promising results by learning in an unsupervised manner hidden (or latent) topics in the corpus. Its main disadvantage compared to LDA is the lack of ability to generalize easily to unseen documents, but it also has the advantage that it is less computationally expensive.

The Author-Topic Model (ATM) is an extension of LDA that includes authorship information in modeling text documents [61]. 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.

The Group-Topic Model (GTM) clusters entities based on their mutual relations, as well as on attributes of those relations [69]. 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 and would not ultimately allow the discovery of groups of groups, but that of groups of people.

Although each of the models has its own merits, LDA seems to be the more attractive one, as it makes no implicit or explicit assumption about the document representation, and is fully generative.

5.2.2 Community detection

Community discovery has become in recent years an active research domain, prompted by the rapid growth of online social networks, and the integration of social relations in other online content networks, such as photo or video sharing websites. Most community discovery approaches fall into either link analysis [22, 24, 39, 40, 74] or content analysis [32, 53], with a small number of works using a combination of the two [72].

Link analysis is mostly approached through graph theory, in either static or time-dynamic scenarios. In [74], Zhang *et al.* describe an LDA-based hierarchical algorithm, in which they model communities as latent variables with distributions over the social actors' space, and apply it to two datasets of collaborative networks, CiteSeer, and NanoSCI. In their work, the number of communities is defined a priori, and evaluation is performed using the perplexity measure for three different kinds of models, as well as an evaluation from a clustering perspective using a measure of compactedness of the clusters, similar to our measure of homogeneity of hypergroups.

Lin *et al.* [40], on a dataset of around 400 blogs, model the process of mutual awareness expansion using a random-walk algorithm. In their model, the authors extract communities based on an interaction space, and they also track the evolution of these communities. Their algorithm requires the number of discovered communities to be set in advance, and evaluation is done based on conductance, coverage, and entropy of the resulting communities, compared to three base-line algorithms.

In [39], Lin *et al.* propose to jointly analyze the structure as well as the evolution of communities in different synthetic and real networks, dressed as a maximum a posteriori estimation problem. The number of communities, although not necessarily fixed a priori, is determined through a series of simulated partitions, which requires a certain amount of domain knowledge for the appropriate number of communities, as exploring the space of partitions is otherwise expensive. One distinct advantage of their approach is the possibility to perform soft-clustering, which is often a characteristic of modern social networs, with members of one community belonging to several other communities simultaneously.

In another work using the link structure of the network, Du *et al.* [22] proposed an algorithm for the detection of communities that does not require prior knowledge of the number of communities, enumerating all maximal cliques. Their method requires counting all triangle relationships in the network, and finding maximal cliques, which then become cluster kernels, or in other words, disjoint communities. For some of the datasets in this study, evaluation of the resulting communities is done by visual inspection, while for others Newman's network modularity measure Q [52] is used.

From a content-based perspective, Nguyen *et al.* [53] use a blogging dataset with sentiment annotations. The authors propose two representations, one content-based and one sentiment-based. Communities are then found by affinity propagation clustering using these representations, and resulting clusters are assessed by visual inspection.

Kammergruber *et al.* [32] take advantage of user tags to compute similarities between users at the content level, and then use a clustering algorithm without a preset number of clusters for community detection. The clusters they obtain on a del.icio.us dataset with 2270 users seem however quite limited, with 92% of the users assigned to a catch-all giant cluster labeled as noise. In contrast, we use tags both as a feature as well as a weighing factor for user membership and we discover communities of communities.

Finally, in an approach that combines link structure and content, Yang *et al.* [72] propose a discriminative model for community discovery, sustaining that generative models may not accurately capture the real factors leading to community formation in complex systems, such as citation networks. Their approach also requires the number of communities to be set beforehand, but they report significant improvements over state of the art approaches for community detection on benchmark datasets with a number of communities between 2 and 20.

We conclude with one recent work that deals with the lack of macro-structure related to Flickr Groups. At the time of writing, discovering new groups in Flickr was still a matter of searching by keywords or of serendipitous discovery while browsing someone else's photos. There is no hierarchy per se, nor any other kind of classification. In a study using a dataset of 300 groups, Egger *et al.*[25] used a membership-based measure they termed *GroupConnectivity* in order to perform community segmentation. This measure is simple to compute, as the fraction between the number of shared members of two groups and the total number of members of the smallest of the two groups. As such, this ratio is bounded by 0 if two groups have no members in common, and by 1 if all members of the smaller group are also members of a larger group. Taking this a step further the method builds, using the same measure of connectivity, a tree of groups. Their assumption was that larger groups are semantic parents of smaller groups with which they are highly connected. Through their experiments their assumption seemed to be generally confirmed, although some counterexamples were also found. The authors obtained semantically meaningful taxonomies, partially shown here in Fig. 5.2. Every node in the tree is a group, and edges imply dependence. With respect to the computational effort involved, this method of automatically extracting taxonomies of Flickr Groups seems very appealing.

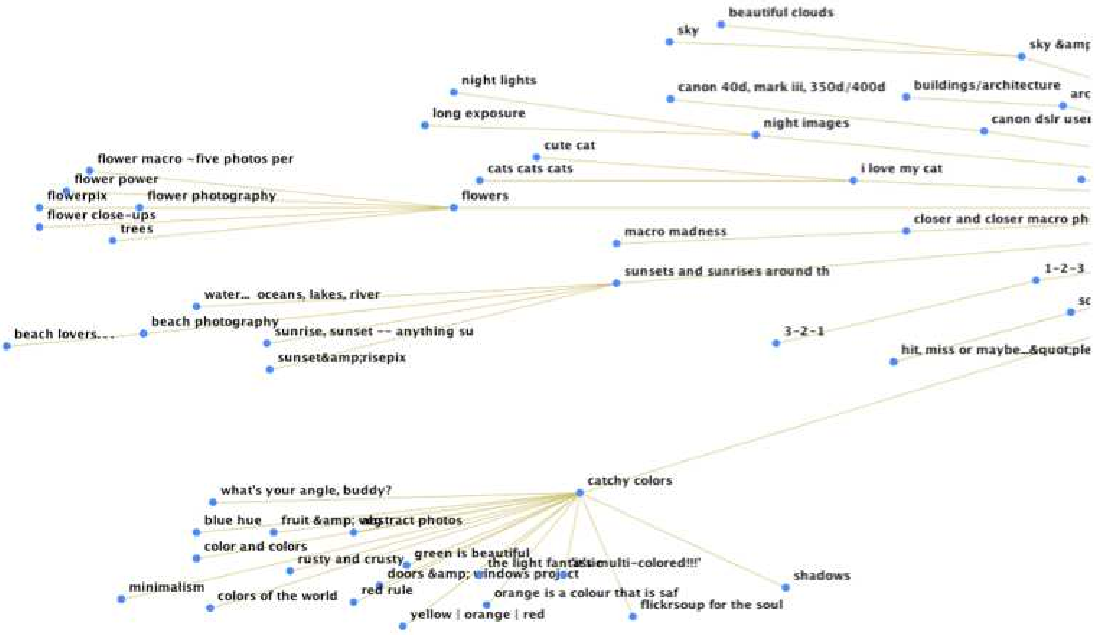


Figure 5.2: Partial view of the automatically discovered taxonomy on 300 Flickr groups - Every node is a group, and edges between groups imply dependence; image courtesy of Egger *et al.* [25].

In our work, we use links, content, and a combination of both through our three models, BM-LDA, TB-LDA, and MM-LDA. We use the same clustering algorithm as in [53], with the number of communities automatically detected during the clustering process, we propose a uniform homogeneity measure for all three models, and we further evaluate two of the models through user studies.

5.3 Experiments and results

Our goal is to find, through clustering, hypergroups that bring together semantically similar groups that do not necessarily have the same keywords in their names or descriptions. But what is a good clustering outcome? Which model performs best? No unique ground truth exists for Flickr groups' similarity, so alternative methods for evaluation need to be designed.

In this section we present the results of hypergroups discovered for each of the three models, then we analyze size and topic-driven statistics for each clustering outcome, and finally we propose and analyze a measure of homogeneity for hypergroups.

The dataset used for these experiments is the same as described in Chapter 2, and consists of 10,800 groups and a sample of their members, for a total of 8,000 users. These users contribute more than 1 million photos to the groups. The total number of tags in the group photo pools is around 38.6 million. Similar to the previous chapter, we have only kept tags that appeared in the list of most popular 10,000 tags.

For all results presented hereafter, the negative Jensen-Shannon (JS) divergence, described in Equation 5.1, was used as the similarity measure: the smaller the value, the more similar two groups are.

$$JS(g_1, g_2) = -\pi D(T(g_1) \parallel T_M) - (1-\pi) D(T(g2) \parallel T_M). \tag{5.1}$$

In Equation 5.1, π is a parameter whose value is set to 0.9. $T(g_1)$ and $T(g_2)$ are the topic distributions of the two compared groups g_1 and g_2, T_M is the mean of the two topic distributions $T_M = \frac{T(g_1)+T(g_2)}{2}$, and $D(X \parallel Y)$ is the Kullback-Leibler divergence between the two distributions X and Y, expressed as:

$$D(X \parallel Y) = \sum_{k=1}^{N} X(k) \log \frac{X(k)}{Y(k)}. \tag{5.2}$$

In Equation 5.2 $X(k)$ and $Y(k)$ represent the k^{th} component of the X and Y topic distributions, and N is the number of topics in the model.

The model parameters are α and β, the parameters of the Dirichlet priors, N, the number of topics, and G the number of iterations for Gibbs sampling. For all three models, the parameter values are chosen equal, with $\alpha = 0.2$, $\beta = 0.01$. The number of iterations was set to $G = 5000$, and the number of topics was set to $N = 100$, an arbitrary choice, with the advantage of easy manual inspection of resulting topics, and a dimensionality reduction of about an order of magnitude with respect to the size of the vocabularies used in each model.

5.3.1 Discovered hypergroups

First, we show in Table 5.1 a couple of examples of hypergroups whose size is around the mean and median of each model's clustering outcome, particularly interesting as they capture related groups that at a first glance have nothing in common, like for example *HDR* and *Photomatix*. HDR stands for High Dynamic Range, and refers to a set of techniques that allow a greater dynamic range of luminance of an image, while Photomatix is a photographic software designed for HDR image processing.

Listed on the first line of each cell and in bold-face is the hypergroup exemplar (the group that defines the hypergroup), and listed under it are the other groups belonging to that hypergroup. We also show the number of members (Mem.) and the size of the vocabulary (Voc.) for each group.

We observe that all three models are able to discover relatively homogeneous hypergroups, with interesting results like the grouping of *RUSTY and CRUSTY* and *Things that Moved* (a group about “Past Tense. Things that moved but don’t anymore. Broken down and retired vehicles. Planes, trains, automobiles, riding mowers, dead weasles, etc, etc, etc.”), or the grouping of *Toysaholic Anonymous* and *Urban Vinyl Fiend* (a group dedicated to “photographs of toys from the designer urban vinyl scene or toys with a flair.”). A cursory inspection of reasonably sized hypergroups confirms this is the case for an important number of hypergroups.

5.3.2 Basic statistics of hypergroups

Second, we look at basic statistics of the discovered hypergroups. The total numbers of hypergroups for each model are 928 for BM-LDA, 1090 for MM-LDA, and 1433 for TB-LDA. In Figure 5.3 we show the histogram of hypergroup sizes for the three models. We observe that MM-LDA and TB-LDA tend to generate more hypergroups of smaller sizes (medians of 4 as opposed to 7 for BM-LDA). Each model also generates a few extra-large clusters, with more than 200 groups, not shown in the figure for scaling reasons. A double tail t-test at $\alpha = 0.01$ for all three models shows that hypergroup sizes for the two membership-based models (BM-LDA and MM-LDA) are likely to have been drawn from the same distribution, while the sizes for the tag-based model (TB-LDA) are significantly different to both other models.

In Figure 5.4 we plot the group size (in members) versus the size of the cluster the group belongs to. The correlation between these two measures is low, with the correlation coefficient equal to 0.167 for BM-LDA, 0.196 for TB-LDA, and lower for MM-LDA, at -0.016, showing that group size and hypergroup size are not significantly correlated.

BM-LDA: hypergroup size median 7, mean 11

Hypergroup #3	Mem.	Voc.
Window seat please	105	656
Aerials	59	621
Cambodia Images	21	329
Central Park	43	321
Bangkok	21	310
Thailand Travel	6	248
Monkeys	37	192

Hypergroup #431	Mem.	Voc.
HDR	275	2750
28mm or wider	113	2383
Photojournalism	101	1953
Photomatix	93	1498
Quality HDR	56	858
TTHDR (True Tone High Dynamic Range)	47	761
HDR Skies (please read the rules!!!!)	47	642
The Moon [*current* photos only]	113	485
Moon/Lua	65	321
HDaRt	10	238
HDR Rides	27	209

Hypergroup #90	Mem.	Voc.
My Everyday Life	26	877
No Mcdonalds	15	640
Healthy Food	28	540
Cookbook - The (un)official Flickr cookbook!	8	394
Cooking (recipe required)	18	352
CookingBloggers	3	329
Macro Sweets.	25	288
Innards	12	258
A cup of tea	38	224
The Coffee Bar	19	155
BACON	21	128

Hypergroup #65	Mem.	Voc.
Nikon D200/D300 Users	134	3430
Nikkor	122	2813
Nikon DSLR Users	123	1737
Sigma Lenses	63	1618
Strobist.com	92	1081
Wedding Photography	101	782
Fast Nikkors (¡= f/2.8) wide open	7	361

Hypergroup #205	Mem.	Voc.
Yorkshire	29	568
North Yorkshire	16	509
York	12	340
National Trust	27	290
North Wales, UK	8	261
York Photographers	3	241
Liverpool	18	205
UK Railways	17	161

Hypergroup #691	Mem.	Voc.
RAW Street Photography	43	599
Travel Photojournalism	43	469
Ethnic	20	466
Portraits Unlimited	40	381
PhotoFixation - Your Fixation Continues ...	16	377
World Families (Family Friendly)	22	270
Challenges and Comments	20	243
My Special Place	16	163
World Community Arts Day	3	146
Digital Gallery Photography Color	17	142
SOMETHING BLUE IN MY LIFE post 1 comment on previous 1	20	138
Empyrean Animals (invite only) - post 1 award 2 in the pool	25	125

Table 5.1: Six examples of hypergroups for the BM-LDA model (with sizes around the mean and median). The top group in each hypergroup (in bold) is the found exemplar. The number of members (Mem.) and the size of the vocabulary (Voc.) for each group in the corresponding hypergroup are also shown.

MM-LDA: hypergroup size median 4, mean 10

Hypergroup #25	Mem.	Voc.
NYC Photobloggers	56	2095
Hello Brooklyn	17	507
Uneasy	9	381
Lonely Moment	8	258

Hypergroup #37	Mem.	Voc.
Toysaholic Anonymous	29	1022
Unbearable Cuteness	20	500
Traveling Toys	16	477
Urban Vinyl Fiend	13	417
Via Alley	5	342
My new Toys and my growing collection	3	246
Little Friends Around the World	4	227
Winnie the Pooh and Friends	3	150
Space-Invaders	25	126

Hypergroup #920	Mem.	Voc.
Original shots	33	960
Minnesota	32	954
Greater Minnesota	11	502
The Great White North	9	455
The Northlands	8	395
Canon S3IS	6	377
Corel Paint Shop Pro	9	285
Beyond Duluth	2	221
Iron Range	1	216
Minnesota Scenery	3	170

Hypergroup #21	Mem.	Voc.
Madras Muffins	6	587
Chennai	7	535
Chennai Photography Club (CPC)	5	300
Metroblogging Chennai	4	286

Hypergroup #135	Mem.	Voc.
Metroblogging Mumbai	8	515
BombayPics	8	406
Maharashtra - India	4	189
The Photography Club of Mumbai	7	132
Indian Roads	5	130

Hypergroup #1043	Mem.	Voc.
Your books	141	845
Churches of Europe	29	597
Gothenburg	7	416
Kirchen / CHURCHES	14	404
Church Furnishings	28	360
Baroque	6	236
churchcrawling	5	205
Hard Men	8	172
Church sculptures	5	160
the stone carvings	12	157
Wesermarsch	1	126

Table 5.2: Six examples of hypergroups for the MM-LDA model (with sizes around the mean and median). The top group in each hypergroup (in bold) is the found exemplar. The number of members (Mem.) and the size of the vocabulary (Voc.) for each group in the corresponding hypergroup are also shown.

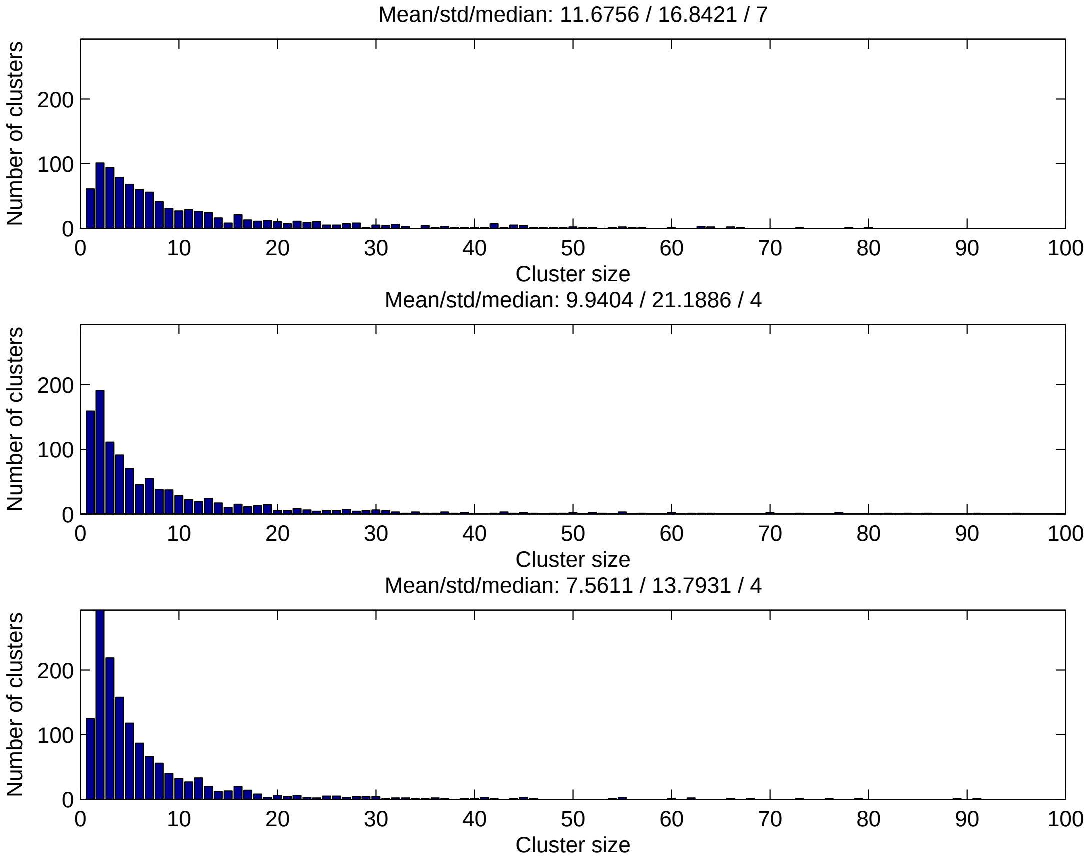


Figure 5.3: Histogram of hypergroup sizes - Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA. The latter two models generate more hypergroups of smaller sizes than the binary-membership model.

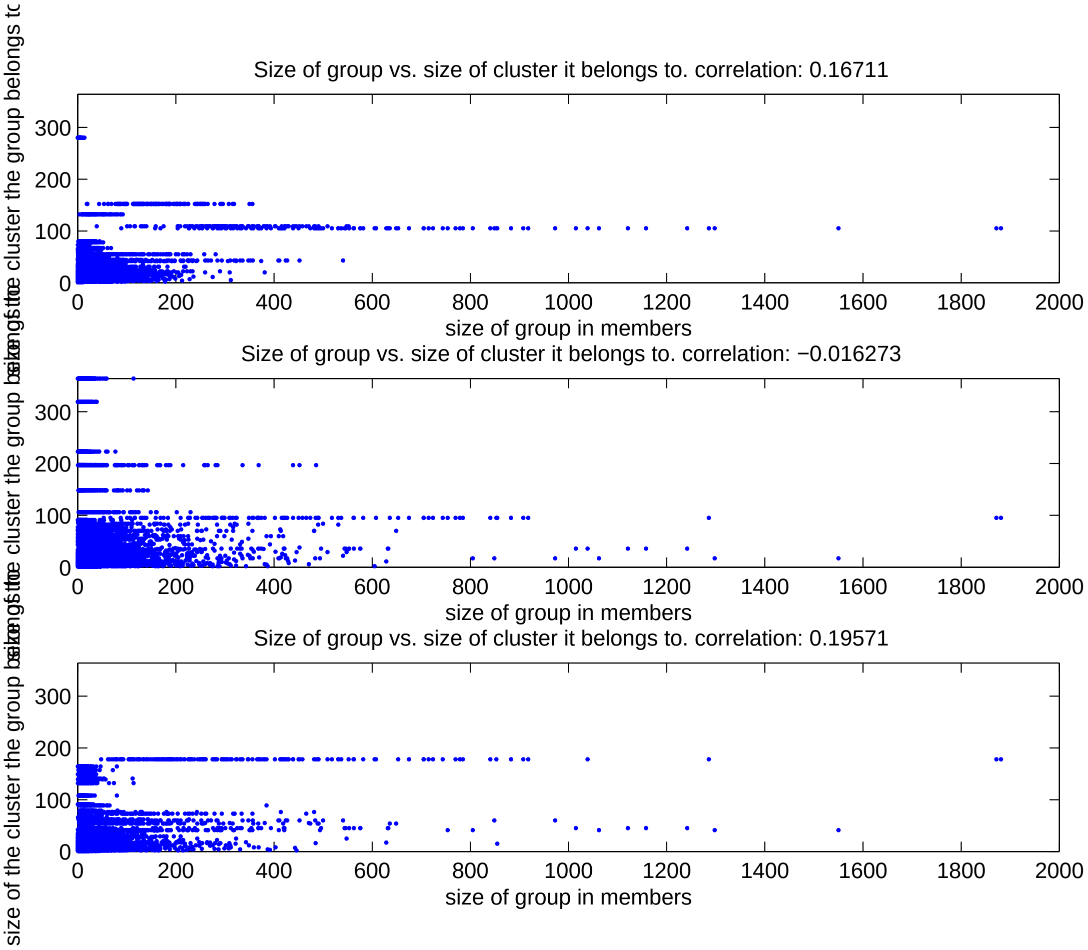


Figure 5.4: Group size versus size of cluster it belongs to - Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA.

TB-LDA: hypergroup size median 4, mean 7

Hypergroup #19	Mem.	Voc.
Patterns and Designs	128	1624
Symmetry	34	1141
Curves vs. Straight Lines	63	1100
A symmetry A	14	362

Hypergroup #567	Mem.	Voc.
RUSTY and CRUSTY	443	2725
Wonders of Oxidation	159	1290
all things rusty	87	904
The Rust Bucket	84	885
Things that Moved	73	689
Rusted	37	516
RUSTY	21	257

Hypergroup #54	Mem.	Voc.
Patterns and Designs	128	1624
Symmetry	34	1141
Curves vs. Straight Lines	63	1100
A symmetry A	14	362

Hypergroup #103	Mem.	Voc.
Chile	44	1588
Latinoamericanos!	35	1282
Free Region de Coquimbo	5	348
Como en el Cine	10	269
Coquimbo	5	217

Hypergroup #85	Mem.	Voc.
People Watching	65	1646
Candid Camera	79	1497
unposed	54	1324
strangers	48	1087
People Watchrs	31	904
Candid	23	662
Strangers & Intimacy	7	280

Hypergroup #266	Mem.	Voc.
Night Shot	43	836
MOON Shots	209	831
Long Exposure Times	45	702
The Moon [*current* photos only]	113	485
Nightscapes - Night Landscapes - *No Cityscapes Thanks*	32	340
Astronomy	38	263
Astrophotography	34	237
Night Sky, The	15	215

Table 5.3: Six examples of hypergroups for the TB-LDA model (with sizes around the mean and median). The top group in each hypergroup (in bold) is the found exemplar. The number of members (Mem.) and the size of the vocabulary (Voc.) for each group in the corresponding hypergroup are also shown.

Starting from the LDA representations, we define *relevant topics* for a group to be those topics that account together for over τ% of the probability mass in its topic-based representation. For the results shown here, the same threshold is used for all three models, $\tau = 80$. We show in Figures 5.5 and 5.6 the histograms of relevant topics for each of the three models, first for individual groups, and then for hypergroups. For a hypergroup, the number of relevant topics is defined as the total number of distinct relevant topics found in its component groups, and it can be seen as a measure of the diversity of the hypergroup topics.

At the group level, MM-LDA appears to generate much more focused topic-based representations, with a mean of around 3 topics per group, as opposed to the BM-LDA and TB-LDA models, which both have means of around 9 topics. This is also observed at the hypergroup level (see Figure 5.6), where aggregating all distinct relevant topics in the hypergroup yields a mean of 6 for the MM-LDA model, while the BM-LDA and TB-LDA have means around 16 and 14 topics respectively. The MM-LDA representation is overall more focused.

In Figure 5.7 we plot the number of relevant topics for each group versus the size of

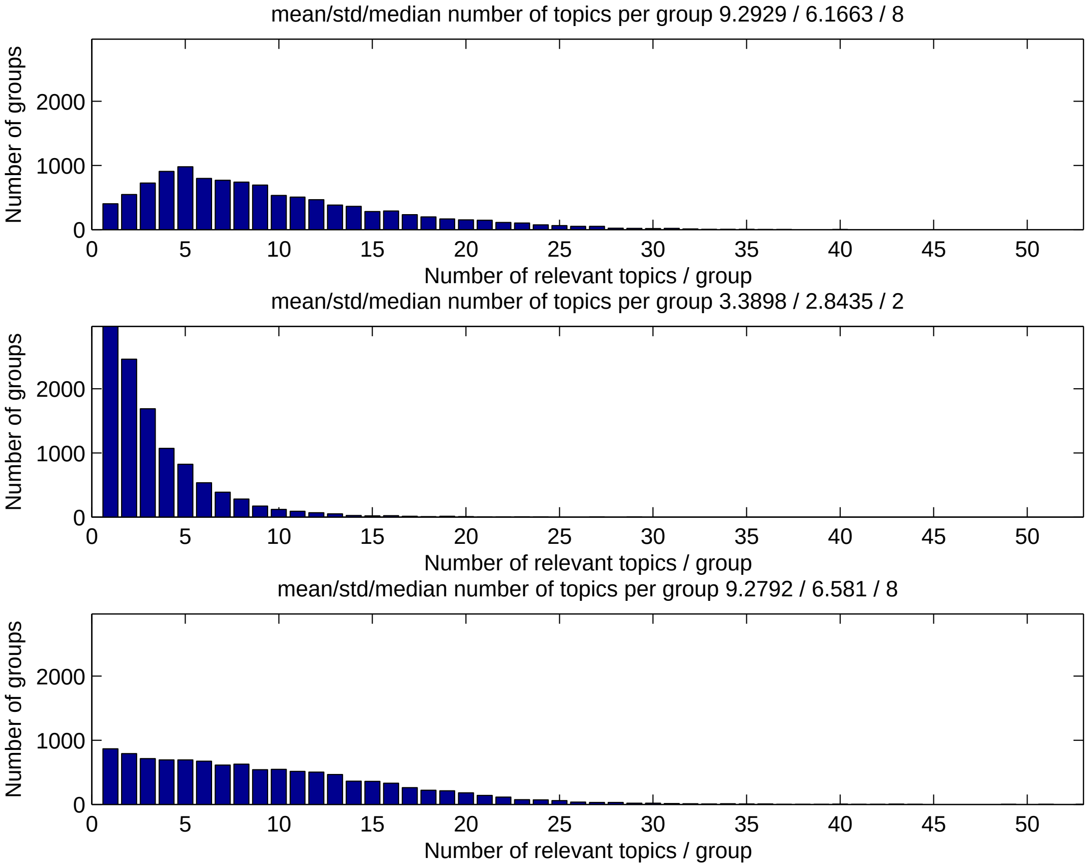


Figure 5.5: Histograms of the number of relevant topics per group for each model - Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA.

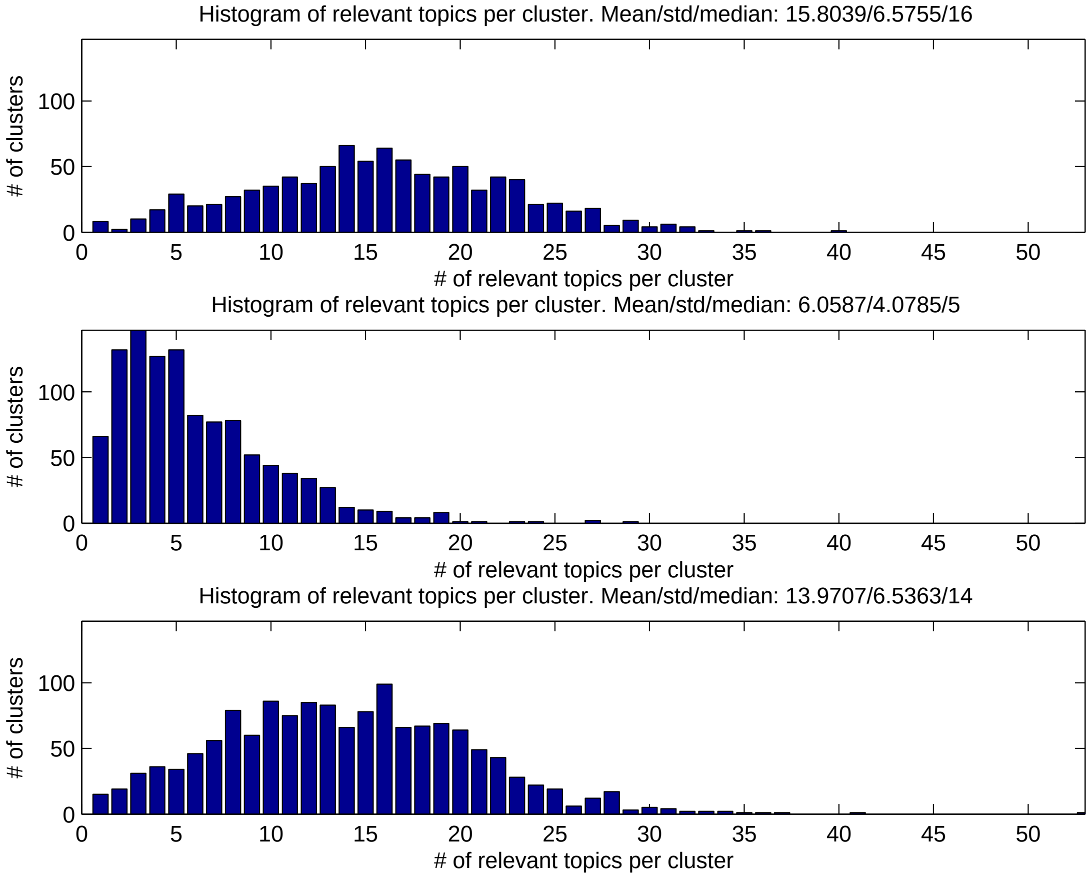


Figure 5.6: Histograms of the number of relevant topics per hypergroup for each model - Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA.

the cluster the group belongs to. A correlation test indicates no significant link between the topic diversity of a group and the size of the cluster it is assigned to by the clustering algorithm.

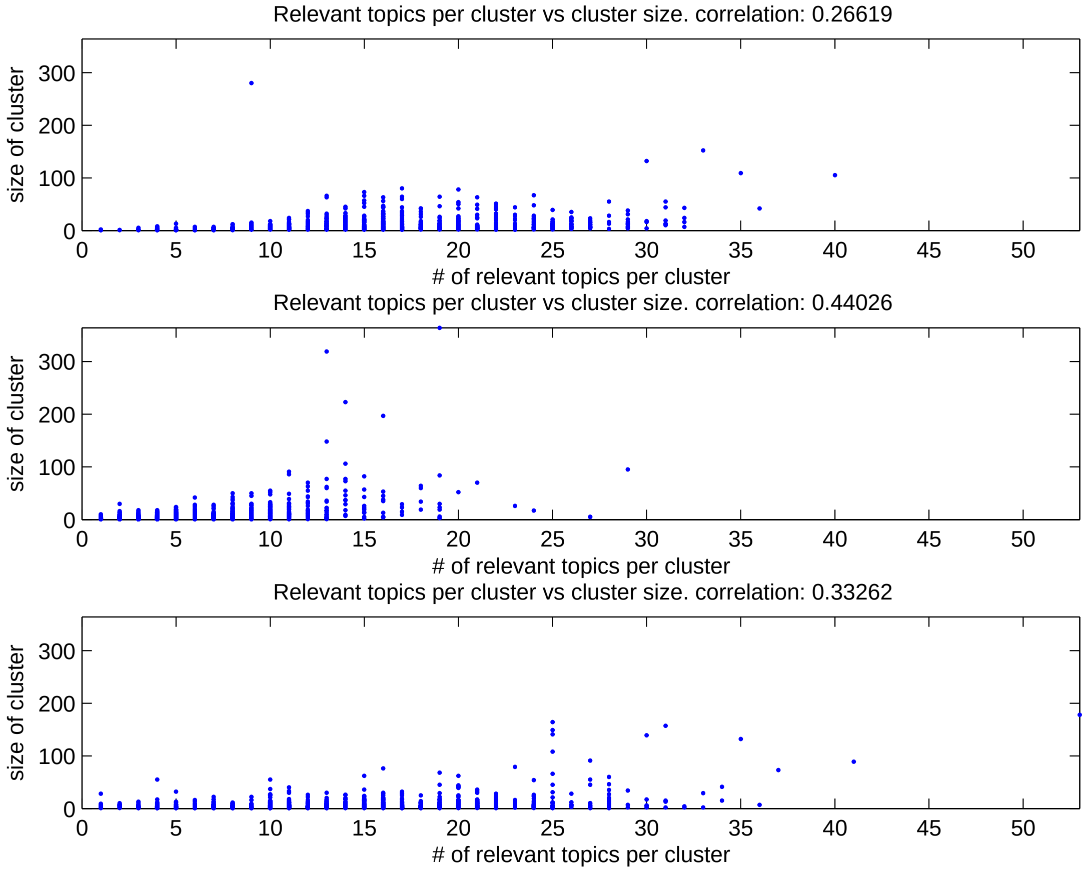


Figure 5.7: Relevant topics per group versus size of cluster the group belongs to - Top: BM-LDA; middle: MM-LDA; bottom: TB-LDA.

5.3.3 Hypergroup homogeneity

Finally, we define a measure of homogeneity for a hypergroup based on the intra-cluster similarity, by averaging the pair-wise similarities for all groups in a hypergroup. For each of the three LDA models we use a Jensen-Shannon similarity measure, dubbed JS-BM, JS-MM, and JS-TB for the similarity derived from each of the three LDA models. These are the same similarities used for the AP clustering algorithm. We then analyzed the effect of each similarity measure on the homogeneity of hypergroups dis-

	BM-LDA	MM-LDA	TB-LDA
JS-BM	0.557 / 0.622	0.491 / 0.547	0.484 / 0.521
JS-MM	0.549 / 0.602	0.372 / 0.420	0.388 / 0.411
JS-TB	0.512 / 0.555	0.431 / 0.494	0.408 / 0.436

Table 5.4: Mean/median hypergroup homogeneities for the three topic models using cross-model similarity measures.

covered by a given model. We present these measurements in Table 5.4. In this table, lower JS distances mean higher homogeneity of the hypergroups. We note that hypergroups based on the BM-LDA model tend to be less homogeneous than hypergroups discovered by the other two models, regardless of the similarity measure used. These differences are statistically significant at $\alpha = 0.01$. This suggests that hypergroups defined based solely on binary-membership links may generally be less consistent. These results are likely explained (at least partially) by the fact that BM-LDA produces less hypergroups, which in turn leads to less homogeneity due to the larger hypergroup size.

Overall, we observe that hypergroups obtained from the multiple-occurrence membership and tag-based models are most homogeneous when the distance JS-MM is used, which suggests that capturing the relations (through membership) and content (through the size of the contributed vocabulary) might indeed be beneficial for hypergroup modeling.

Although evaluation is difficult in practice due to the size of the dataset and lack of ground truth, a subjective evaluation procedure for two of the models (BM-LDA and MM-LDA) has been designed and is presented in the following section.

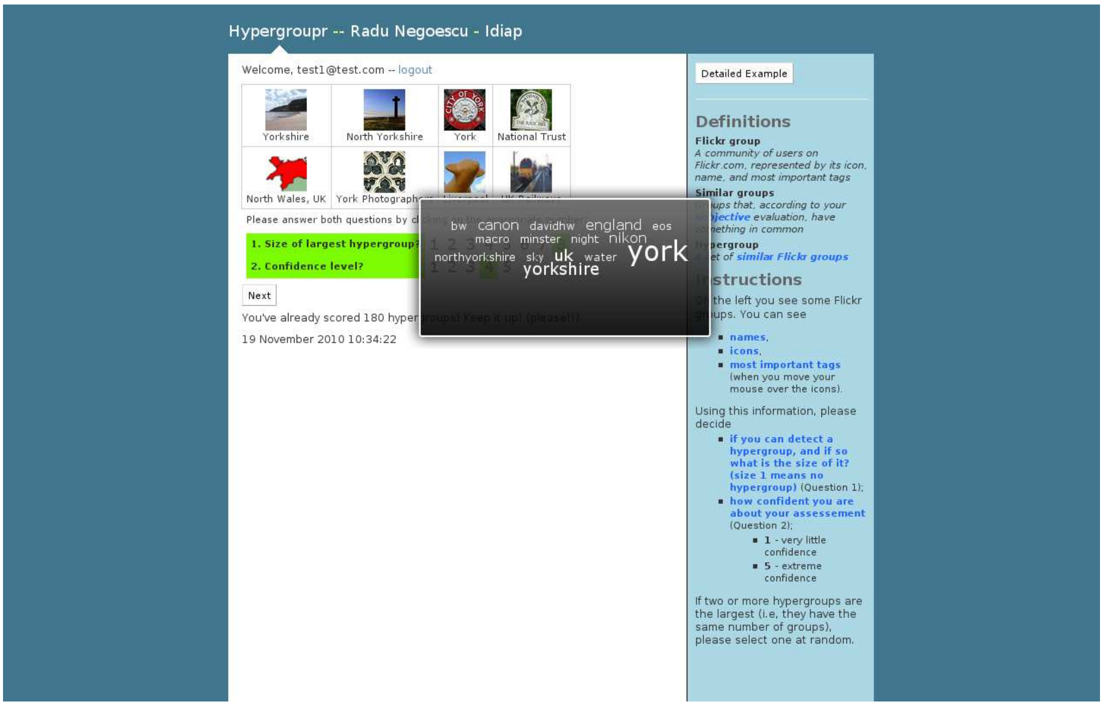
5.4 User Evaluation

As ground truth is very difficult to obtain in the context of our dataset, and measures based on held-out likelihood, such as perplexity, have been shown not to be a necessarily good indicator of the semantic meaningfulness of topics in probabilistic models [14], we have developed a web interface for human evaluation of the hypergroups.

In this section we first describe the annotation tool, then the data used for the annotation, and finally we present the results obtained.

5.4.1 The annotation tool

We present a snapshot of the web interface used for annotations in Figure 5.8. For each cluster, each annotator was shown the set of groups in the cluster, represented by their icons, names, and tag clouds, displayed on mouse-over events.



Hypergroupr -- Radu Negoescu - Idiap

Welcome, test1@test.com -- logout

Please answer both questions by cl

1. Size of largest hypergroup?

2. Confidence level?

1 2 3
1 2 3 4 5
bw canon davidhw england eos
macro minster night nikon
northyorkshire sky uk water york
yorkshire
Next
Detailed Example

You've already scored 180 hyper

19 November 2010 10:34:22

Definitions

Flickr group
A community of users on Flickr.com, represented by its icon, name, and most important tags

Similar groups
Groups that, according to your subjective evaluation, have something in common

Hypergroup
A set of similar Flickr groups

Instructions

On the left you see some Flickr groups. You can see

- **names**,
- **icons**,
- **most important tags** (when you move your mouse over the icons).

Using this information, please decide

- **if you can detect a hypergroup, and if so what is the size of it? (size 1 means no hypergroup)** (Question 1);
- **how confident you are about your assessement** (Question 2);
  - **1** - very little confidence
  - **5** - extreme confidence

If two or more hypergroups are the largest (i.e, they have the same number of groups), please select one at random.

Figure 5.8: Annotation tool - The users are shown a cluster, defined by its composing groups, with icons, names, and tag clouds. The annotators are asked two questions: 1. What is the size of the biggest hypergroup they can detect in this set of groups? 2. How confident are they with their decision?

To the right of the interface, definitions of Flickr groups, similar groups, hypergroups, and instructions for the annotation task were permanently displayed. The annotators had to answer two questions before moving on to the next cluster.

The first question asked them to determine the maximum number of similar groups, in other words, the size of the biggest hypergroup they could form with groups from this cluster. The minimum number of similar groups was 1 (when no two groups were similar), and the maximum number was the number of groups presented in the given cluster. The second question was an assessment on a 5-point rating scale of their own confidence with the decision on the first question.

The clusters were presented in the same order to all annotators.

5.4.2 Data

Our annotation tool allowed a group of annotators to score the perceived homogeneity of the hypergroups, using two samples of hypergroups, from the BM-LDA and MM-LDA models. In order to keep the two sets as comparable as possible we have first randomly selected hypergroups from the BM-LDA model based on the number of groups they contained (between 3 and 14), for a total of 457 hypergroups.

The annotation effort thus started with eleven annotators and 457 clusters, all but one computer scientists from the same research institute. The clusters were presented in the same order, randomized once for all annotators, in order to maximize the number of annotators per cluster. Of the eleven annotators, eight annotators (3 females and 5 males, all except one computer scientists, all except two with no Flickr experience as users) scored at least 103 BM-LDA hypergroups, while the other 3 only roughly 60. These three annotators were subsequently removed from the annotation process based on their low completion rate.

We then collected the 771 groups composing this subset of 103 hypergroups, and extracted from the MM-LDA model the hypergroups containing them. We then also filtered these hypergroups by size with the same constraints as before, and kept the 98 of them which had the highest number of overlapping groups with the previous subset. The 98 hypergroups contained a total of 742 groups. The same eight annotators then assessed these 98 hypergroups. This second annotation task was subject to a one time payment of 30 Swiss francs.

We show in Figure 5.9 the histograms of the cluster sizes in both models. A double-tailed t-test fails to reject the null hypothesis that the two distributions come from populations with equal means at 5% significance level. We can thus assume that although the sampling process for hypergroups of the MM-LDA model follows a different path from the one of the BM-LDA model, the two samples are comparable.

5.4.3 Data analysis

With the above procedure, we obtained two distinct annotation datasets: the first, on 103 clusters from the BM-LDA model, and the second on 98 clusters from the MM-LDA one.

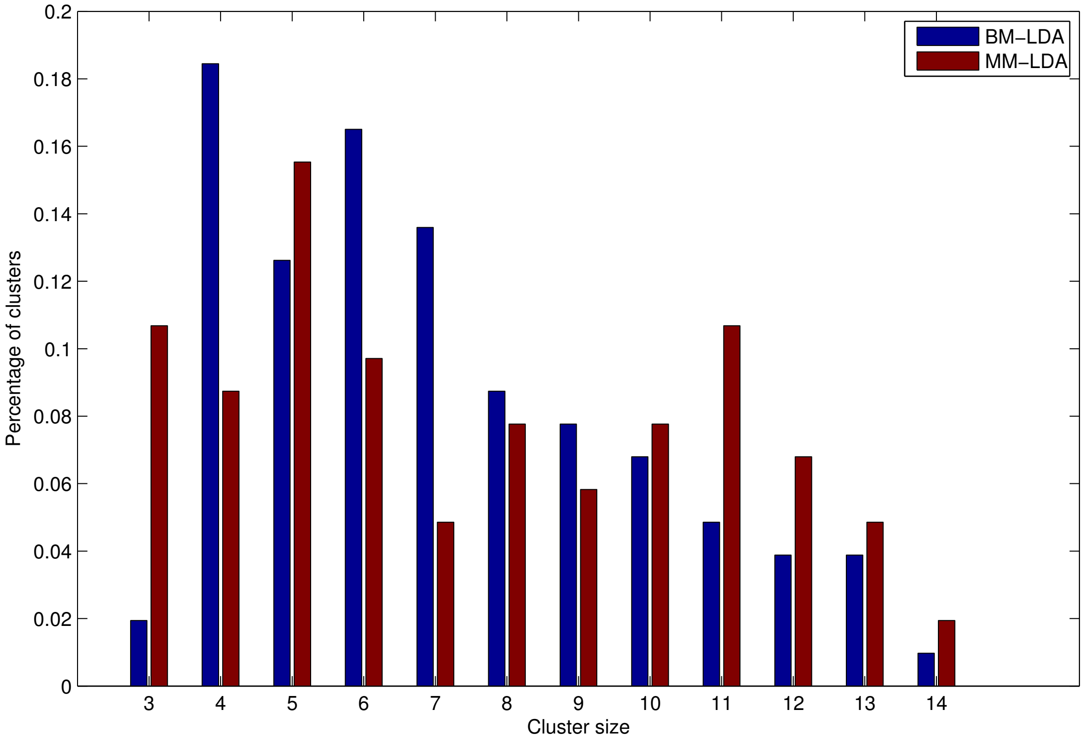


Figure 5.9: Cluster sizes histograms for the two models - On the left BM-LDA and on the right MM-LDA. The two distributions show no statistically significant difference in means at $\alpha = 0.05$, with $p = 0.273$ and confidence interval [-1.296 0.369].

	BM-LDA	MM-LDA
Under "2nd worst-case" line	15%	12%
Above "50-50 case" line	45%	56%
Above "2nd best-case" line	15%	23%

Table 5.5: Purity stats over all hypergroups, for the binary-membership (BM-LDA) and multiple-membership (MM-LDA) models.

For each annotated cluster we have eight annotations composed of two values: 1) the number of similar groups; 2) the confidence of the annotator. We defined a measure called *cluster purity* for a comparable measure of quality across hypergroups. Formally this measure is $\rho = \frac{N_{sg}-1}{N_g-1}$, where N_{sg} is the number of similar groups detected by the annotator, and N_g is the size of the cluster. The intuition behind subtracting 1 from both nominator and denominator comes from a group recommendation scenario: for each of the groups in the similar groups subset, ρ represents the ratio of good recommendations from the remaining cluster members. This measure penalizes clusters where no hypergroups were detected ($N_{sg} = 1$) by reducing purity to 0.

5.4.3.1 Cluster purities

We show in Figures 5.10 and 5.11 the plots of mean cluster purities for BM-LDA and MM-LDA against each cluster size, as well as three thresholds: the red continuous line with cross markers represents the "2$^{\text{nd}}$ worst-case scenario" purity, obtained when only two similar groups are found; the green dash-dotted line with cross markers represents the "2$^{\text{nd}}$ best-case scenario", obtained when only one group is dissimilar from all the others; and finally, the violet dashed line with x markers represents the "50-50 case scenario", when half the groups in a cluster are similar. Obviously, the worst case scenario corresponds to zero purity, and the best one to unity. Some overlapping points have been drawn slightly off to the left and right of the corresponding x value (the size of the cluster), in order to give a visual indication of the real number of clusters in each region.

We summarize in Table 5.5 the statistics of the purity measures over all hypergroups for the two models. The first statistic, the percentage of hypergroups performing under the 2$^{\text{nd}}$ worst case line, shows a slightly smaller number for MM-LDA, with 12% opposed

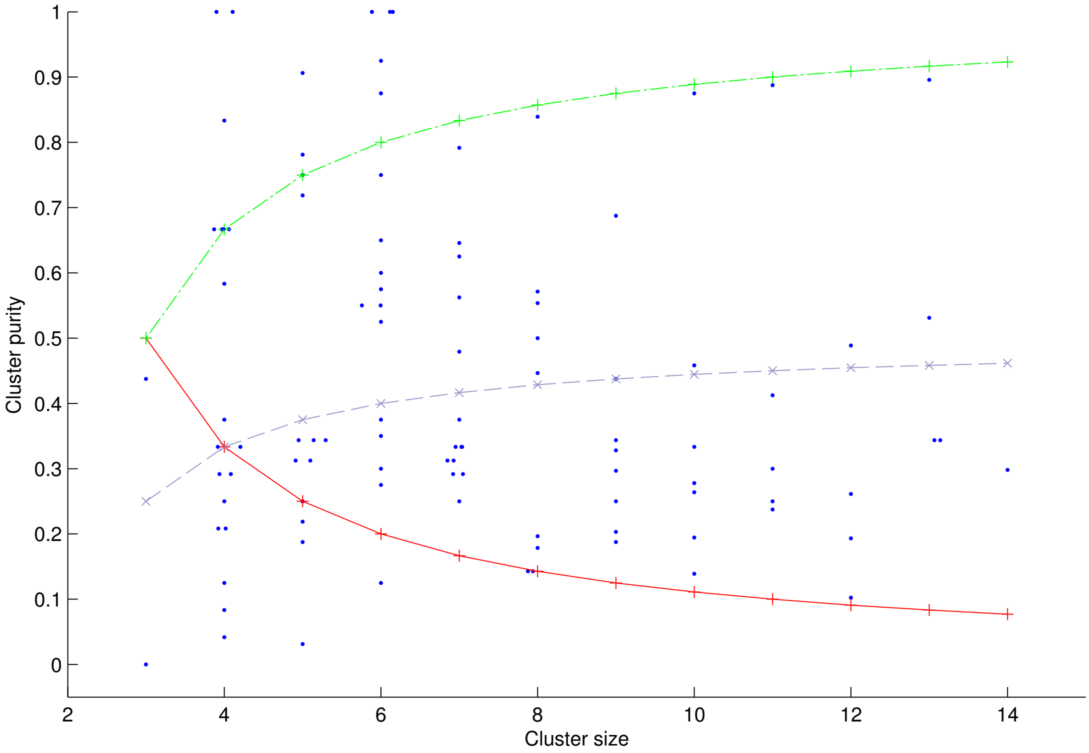


Figure 5.10: Cluster purities for the BM-LDA model - The x-axis shows the size of the clusters, where some of the points are drawn at positions slightly off to the right and left, for completeness of visualization when several clusters of the same size have the same purity.

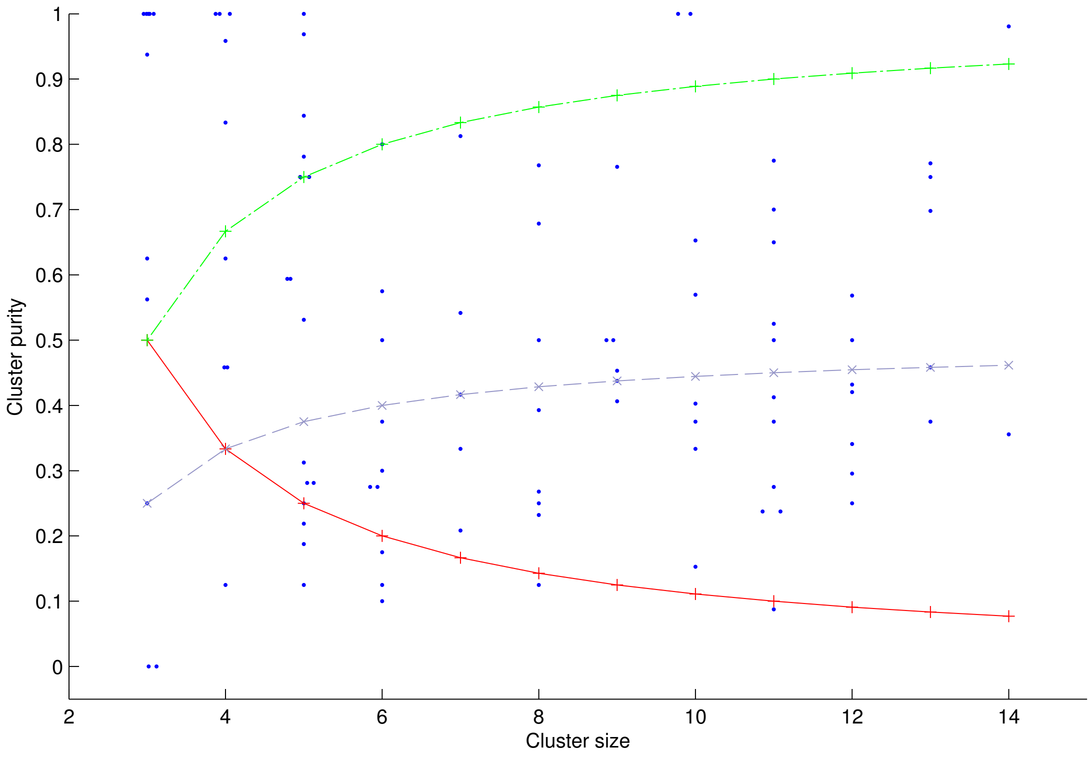


Figure 5.11: Cluster purities for the MM-LDA model - The x-axis shows the size of the clusters, where some of the points are drawn at positions slightly off to the right and left, for completeness of visualization when several clusters of the same size have the same purity.

to 15% of the hypergroups in the case of BM-LDA. The second line of the table shows the percentage of hypergroups with at least half the groups being considered similar by the 8 annotators (on average), with 45% for BM-LDA, and 56% for MM-LDA. In more than half of the hypergroups based on the MM-LDA model the annotators considered at least half of the groups to be similar. Finally, we show the percentage of hypergroups performing better than the "2nd best-case" scenario, with 15% for BM-LDA, and 23% for MM-LDA. Thus, for nearly a quarter of the hypergroups found by the MM-LDA model, only at most one group was judged to be an outlier for the cluster.

We also look in Figure 5.12 at the average cluster purities over the eight annotators for the two models, shown here in histogram form. The y-axis shows the percentage of clusters in each model falling in 10 bins spread over the [0..1] interval of cluster purities. The mean cluster purity for the BM-LDA model is 0.44 (median 0.34), and the mean for the MM-LDA hypergroups is 0.52 (median 0.48). The null hypothesis that the two distributions have the same means cannot be rejected by a two-tailed t-test at the 5% significance level. In a simulation with synthetic data based on the empirical values of the means and variances obtained for the two models, statistical significance is observed when the mean purity for the MM-LDA model is higher than 0.57, for the same number of samples $N = 100$.

5.4.3.2 Quality of annotations

The eight annotators first saw the BM-LDA clusters, and one month later, the MM-LDA clusters. One interesting observation is that, in both annotation sets, there is a high correlation between the cluster purity and the annotator self-reported confidence score, with $r_{BM-LDA} = 0.768$ for the BM-LDA model, and $r_{MM-LDA} = 0.812$ for the MM-LDA model, and the p-values equal to $p_{BM-LDA} = 2.585e - 21$ and $p_{MM-LDA} = 3.333e - 24$ respectively. One might expect human confidence to go down with the increase in the number of groups in a cluster, however, the correlation coefficient between the cluster size and annotator confidence is $r_{BM-LDA} = -0.112$ for the BM-LDA model, and $r_{MM-LDA} = -0.153$ for the MM-LDA one, showing little correlation. These two observations are an encouraging indication that the annotators felt confident to assess large clusters (10 to 14 groups) as well as they did small and medium clusters.

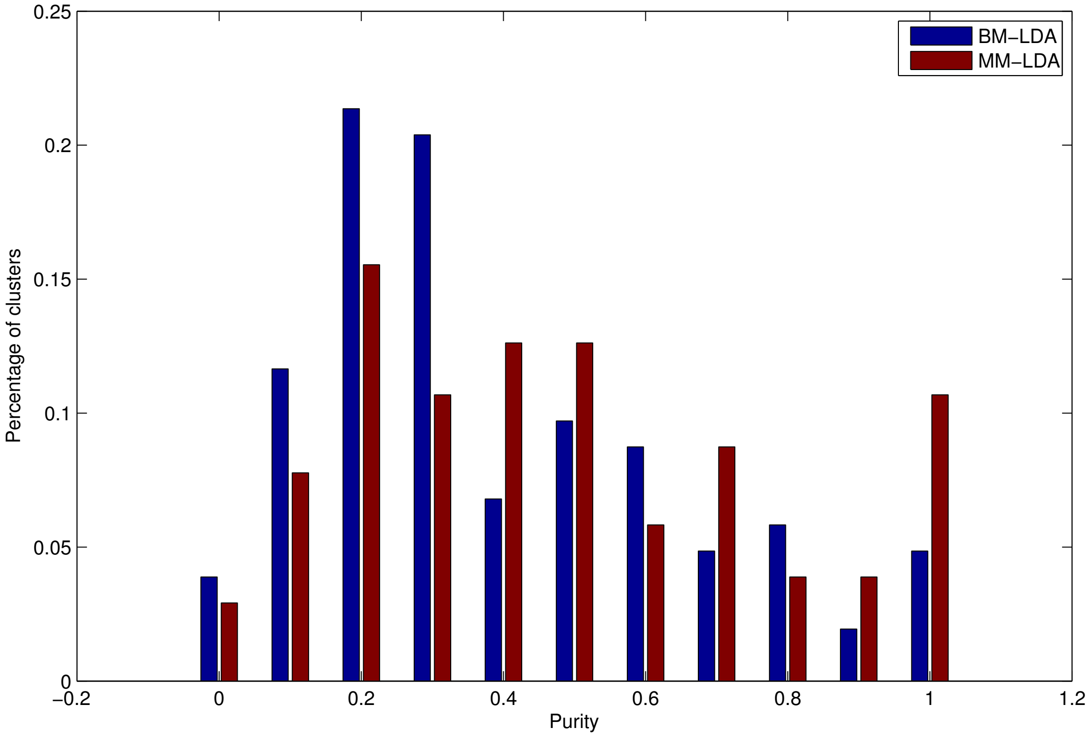


Figure 5.12: Cluster purity histograms for the two models - On the left BM-LDA and on the right MM-LDA. The two distributions show no statistically significant difference in means at $\alpha = 0.01$ with $p = 0.039$ and confidence interval [-0.177 0.020].

We show in Figure 5.13 the mean cluster purity of each annotator on the two models. We plotted the average purity over the 103 clusters of the BM-LDA model with blue star markers, and with red cross markers the average purity over the 98 clusters of the MM-LDA model. We observe a clear trend for all annotators except one, in which the average purities over the MM-LDA model are higher than those over the BM-LDA model.

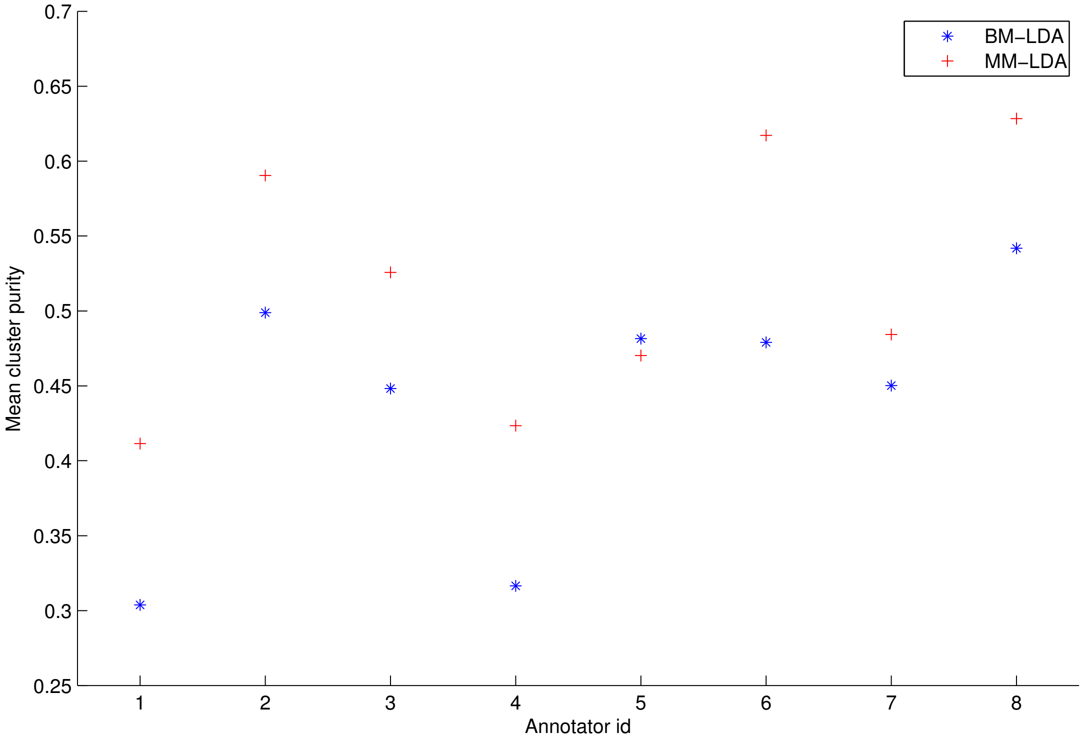


Figure 5.13: Cluster purity averages for the eight annotators across the two models - In blue and with $*$ markers the BM-LDA annotations, and in red and $+$ markers, the MM-LDA ones. For all but one of the annotators the average cluster purity over the MM-LDA model is higher.

We show some more examples of clusters and annotator answers in Figures 5.14 through 5.16. Figure 5.14 shows a 10-groups cluster from the MM-LDA model that was annotated quite differently by the annotators. Three annotators found only 2 similar groups, three other annotators found 5 similar groups, while the remaining two annotators found 7 and 9 respectively. The final purity for this cluster is $\rho = 0.40$. Although the annotators had access to the tag clouds of each group in an overlayed tooltip, it appears that lack of domain knowledge may have influenced the assessements of certain clusters. For example, the groups *EOS Squad* and *350D Digital Rebel XT Group* are both groups related to Canon EOS digital cameras, *Canon EF-S 10-22* is the name of a wideangle Canon lens, and so on. In Figure 5.15 we show a 5-groups cluster, with purity $\rho = 0.84$. Five annotators found all 5 groups to be similar, one annotator found 4 similar groups, and the other two annotators found only 3 groups to be similar. A big, 12-groups cluster is shown in Figure 5.16, where annotations were again slightly divergent, with two annotators who found 4 similar groups, one who found 5 similar groups, two others who found 6 similar groups, and the remaining three annotators finding 7 similar groups. The final purity is therefore $\rho = 0.43$. As a contrasting, and rare example of a large cluster with great annotator agreement, we show in Figure 5.17 a 14-groups cluster with purity $\rho = 0.98$. All annotators but one found all 14 groups to be similar, while the last annotator thought only 12 groups were similar. These examples give a better idea of the challenges faced by the annotators in assessing the homogeneity of the clusters.

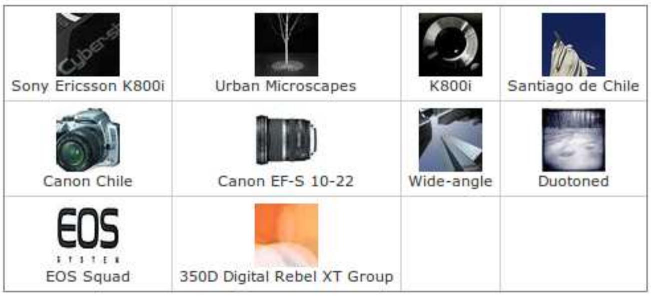


Figure 5.14: A 10-groups cluster from the MM-LDA model - Three annotators found only 2 similar groups, three other annotators found 5 similar groups, while the remaining two annotators found 7 and 9 respectively. The final purity for this cluster is $\rho = 0.40$.

5.5 Conclusions

In this chapter, we have proposed a method to discover hyper-communities in Flickr. By finding groups of similar groups we enable users to find somewhat unpopular groups that do not show up at the top of traditional search results. We have shown that the affinity propagation clustering algorithm yields homogeneous hypergroups, regardless

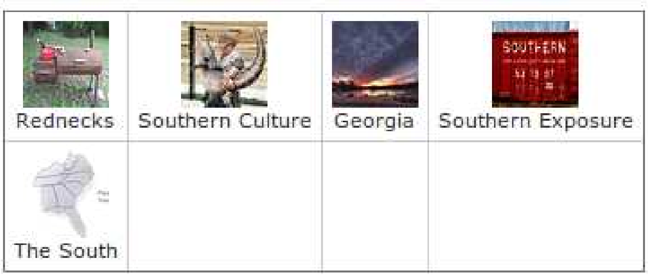


Figure 5.15: A 5-groups cluster from the MM-LDA model - In all five groups' tag clouds, the tags ***georgia*** and ***south*** are amongst the top 20 tags. Two annotators judged that only 3 groups were similar, one annotator found 4 similar groups, and the rest of 5 annotators judged all 5 groups to be similar. The final purity, averaged over all 8 annotators, is $\rho = 0.84$.

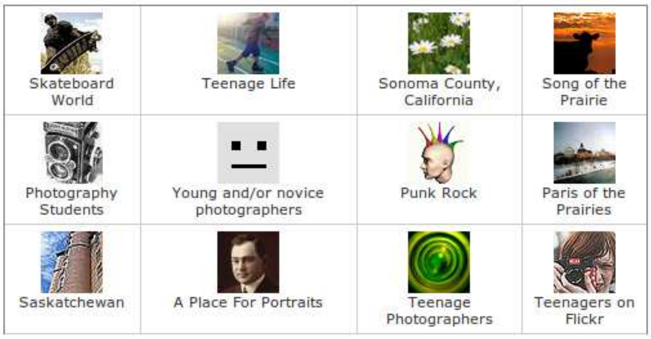


Figure 5.16: A 12-groups cluster from the MM-LDA model - Two annotators found 4 similar groups, one annotator found 5 similar groups, two annotators found 6 similar groups and the other three annotators found 7 similar groups, with the final average purity $\rho = 0.43$.

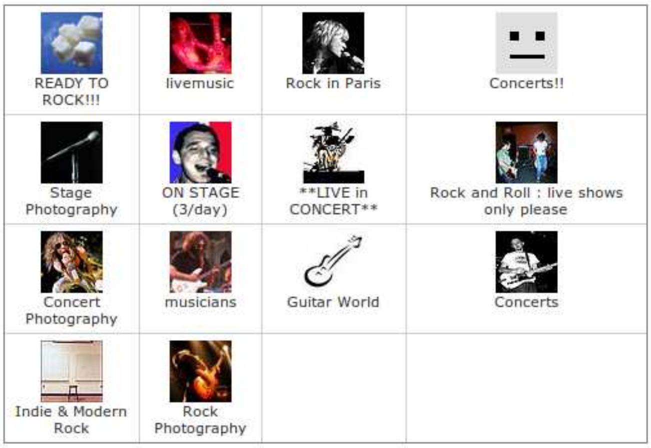


Figure 5.17: A 14-groups cluster from the MM-LDA model - All annotators but one found all groups to belong together, with the last annotator finding only 12 of the groups to be similar. The final purity is $\rho = 0.98$.

of the underlying model used, with better results for the model based on both content and membership links. Hypergroups found this way tend to be of relatively small sizes.

Human annotation of a sample of clusters shows that the discovered hypergroups are indeed meaningful, and confirms our hypothesis that similar groups share content and/or members. We have also shown that using information derived from topic models, such as the number of relevant topics, can give insights into the structure and quality of the hypergroups.

We have also proposed a method to assess the homogeneity of discovered hypergroups based on similarity measures employed by the clustering process. Our results seem to encourage the use of fused information coming from content and relations, such as is the case for the MM-LDA model.

The annotation process raised an interesting question regarding the underlying annotator motivation for completing the task, and the influence that a monetary reward may have not only on the completion rate, but also on the perceived quality of the results. While recent studies seem to indicate monetary rewards on Amazon's Mechanical Turk [11] do not influence the quality of the data produced by annotators, this question should probably be investigated as part of the task itself, taking into consideration the different underlying motivations of the annotators. The possibility of conducting large-scale evaluation experiments with this type of infrastructure remains as an issue for future work.

In our work we examined one type of fusing content and relations, but clearly other fusing methods are also possible, and they could be investigated in the future. Also, although we have designed three distinct models, other representations of groups may be envisaged, in which more importance is given to who tagged what, and possibly when. Finally, a prototype of group search by using hypergroups, which contains a number of challenges for effective visualization and discovery, should also be subject of future work.

6

Conclusions

We review in this chapter the four major contributions of our work, and equally important, we review some of its limitations. For some of the issues raised here, solutions are not straightforward, but they may become so in the near future.

6.1 Contributions

As mentioned in the introduction of this thesis, the overall goal of our work was to achieve a deep understanding of online social media communities (photo communities in particular) using large-scale data. Furthermore, making use of this understanding, we aimed to build viable unsupervised models for online user and community modeling.

We started by analyzing Flickr, one of the most popular online photo sharing communities. Our analysis pointed out the different modalities in which users share photos with the world and with specific groups of interest. Using a large-scale dataset with roughly 7 million photos and more than 22,000 users, we showed that users contribute a substantial amount of their photo collections in online communities called Flickr Groups. Regardless of whether they pay or not for the membership, users share on average 30% of their collections in groups, allowing these communities to emerge as content-rich entities.

In an attempt to analyze the extent to which other social media repositories are similar to Flickr, we performed a large-scale comparative analysis of Flickr and Kodak Gallery using a dataset of over 5 million images and 10,000 users, and we found certain differences at the level of the raw vocabularies, induced by the users' motivations and system design choices, as well as some similarities.

The analysis carried out in Chapter 2 laid the foundation for our modeling task. Thus we proposed a probabilistic topic model for Flickr groups and users alike, and we jointly modeled these two types of entities starting from a bag-of-tags representation. We found that the topic model can indeed learn meaningful topics. We then showed the model to be useful in comparing and finding similar entities at a more abstract level, regardless of their type. We also proposed several direct and indirect applications of the topic-based representation of entities, such as entity discovery, and search-through-topic extensions to traditional search paradigms.

After having jointly modeled entities from the same system with promising results, we turned our attention to the task of jointly modeling entities from different media systems. We proposed a probabilistic topic model for the joint modeling of Flickr and Kodak Gallery users, and we showed that the effects of the users' motivations and needs can be strongly observed also at the topic level, in what we called the Kodak Moments and Flickr Diamonds, two sets of sharing behaviors corresponding to family-oriented sharing and exposure-seeking users respectively. We also believe this work shows the potential that social media modeling has as a complement to small-scale ethnographic studies, which are very time and effort-intensive.

Finally, we proposed a method to for hyper-community detection in Flickr starting from existing communities (Flickr groups). By building probabilistic topic models on top of three different bag-of-tags representations of groups, we applied a deterministic clustering algorithm and partitioned a dataset of roughly 11,000 entities in hypergroups. Apart from objective measures of homogeneity of the discovered hyper-communities, we also proposed a user evaluation of two of the models, by building an annotation interface and gathering data from human observers. This study showed that hypergroups discovered by our models are generally homogeneous.

6.2 Limitations and future work

Pragmatic reasons such as data availability partially restrict the choice of social media systems that researchers are able to analyze. Flickr has been since its very early life a system that exposed its data through an API, and this encouraged us to make use of its rich social and multimedia content data for most of our analysis and models. Whenever possible, we tried to extend our analysis and models to other online photo media systems, and this led to our comparative study of Flickr and Kodak Gallery. The biggest limitations of this comparative study stem from the inherent differences of the two systems, such as the lack of a tagging system in Kodak Gallery, or the system design choice for default sharing options (public versus private sharing of photos). Although some pre-processing was applied to the Kodak Gallery image captions, it is not clear to what point the large-scale differences in vocabulary might be explained by the different ways tags and free text are used to describe images. Also, while we strongly believe that design decisions impact the data created by users of such systems, we have not addressed the problem of providing guidelines for system designers, as such an interpretation of our results is not straightforward. It is also important to investigate the measure in which a system that implements a successful model for any kind of task (tag expansion, search, or recommendation) may influence the tagging behavior of its users. This could be the subject of cross-field research in the future.

During the first stages of the thesis, one of the goals to explore was the use of image-content as input for our models, and to examine to what extent image features can help improve results with respect to metadata alone. However, large-scale datasets such as ours also imply non-negligible time costs, and computational complexity remained one of the other major hindering factors in achieving this goal. Adding image features into the models is clearly a direction to explore in the future.

Other kinds of features may also prove very useful in increasing the accuracy of social media community models, such as time-related information, which could lead to time-dynamic models for the study of the evolution of users, groups, and hypergroups. Depending on the modeling approach, computational complexity could increase significantly, and efficient algorithms might be needed for this kind of analysis to be feasible. We did not address this objective in this thesis.

A set of open questions remain at the level of model and vocabulary parameters: what is an appropriate number of words to keep in a million user system, which may also be multi-lingual? What is the appropriate number of topics of interest to learn in a probabilistic topic model, considering the size of the vocabulary? While we made a compromise between the size of the vocabulary, the number of topics, and model complexity in order to analyze our methods, empirical or theoretical answers to these questions are not yet readily available, although this is an active research field in machine learning and data mining. Some other types of models, such as hierarchical topic models, or models that explicitly model the hierarchical structure that may exist within Flickr groups could also be investigated. As the amount of public data constantly increases, questions pertaining to the scalability of such probabilistic models to very-large scale datasets are becoming more and more important. Although studies such as ours with tens of thousands of users are several orders of magnitude larger than traditional ethnographic studies, models that work with hundreds of millions of users (or content items) need to be designed and analyzed. This might be difficult to accomplish in an academic environment without easy access to data of such magnitude.

One of the areas that proved most challenging was the evaluation of the models we proposed. Although experiments can be designed for retrieval or recommendation scenarios, evaluation is often subjective. User studies are an alternative evaluation method, but they suffer from lack of scalability, both in terms of size of the annotated data, as well as in terms of annotator population. A possible solution is given maybe surprisingly by social media itself, crowd-sourcing annotations becoming a simpler task via, for example, Amazon's Mechanical Turk service. However, such solutions bring about their own problems, such as the reliability of annotators and the additional administrative overhead. Our own annotation process brought up an interesting question regarding the underlying annotator motivations (monetary or otherwise) for completing the task. We believe that this question should be investigated in the future as part of the annotation task itself.

Finally, it is likely that the future will bring about more openness from existing and new social media systems, thus allowing researchers to more easily study and compare model performance across different social media communities.

References

[1] William M. Vander Weyde. at George Easteman House Museum, Sept. 2008. http://flickr.com/photos/george_eastman_house/sets/ 72157607377134096/. 2

[2] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *CHI '07: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, San Jose, CA, USA, 2007. 12

[3] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proc. of the 2007 Conf. on Digital Libraries*, Vancouver, BC, Canada, 2007. 12

[4] G. Amato and U. Straccia. User profile modeling and applications to digital libraries. pages 184–197, 1999. 80

[5] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *CHI '07: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, San Jose, CA, USA, 2007. 4, 12, 31, 42

[6] T. L. Berg and D. Forsyth. Automatic Ranking of Iconic Images. Technical report, U.C.Berkeley, 2007. 12

[7] S. Berkovsky, D. Heckmann, and T. Kuflik. Addressing challenges of ubiquitous user modeling: Between mediation and semantic integration. *Advances in Ubiquitous User Modelling*, page 1–19, 2009. 80

[8] S. Berkovsky, T. Kuflik, and F. Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008. 80

[9] 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*, page 2003. MIT Press, 2004. 77

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal Machine Learning Research*, 3, 2003. 44, 47, 81, 96, 100

[11] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk. *Perspectives on Psychological Science*, 6(1):3, 2011. 126

[12] F. Carmagnola and F. Cena. User identification for cross-system personalisation. *Information Sciences*, 179(1-2):16–32, 2009. 80

[13] F. Carmagnola, F. Cena, O. Cortassa, C. Gena, and A. Toso. A preliminary step toward user model interoperability in the adaptive social web. *UBIDEUM 2007*, page 29, 2007. 80

[14] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading Tea Leaves: how Humans Interpret Topic Models. In *NIPS'09: Proc. of the 23rd Int'l Conf. on Neural Information Processing Systems*, Vancouver, CA, Dec. 2009. 47, 114

[15] H. M. Chen, M. H. Chang, P. C. Chang, M. C. Tien, W. H. Hsu, and J. L. Wu. Sheepdog: Group and Tag Recommendation for Flickr Photos by Automatic Search-based Learning. In *MM '08: Proc. of the 16th ACM Int'l Conf. on Multimedia*, Vancouver, Canada, 2008. 42

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object R etrieval. In *Proc. 11th Int'l. Conf. on Computer Vision*, Rio de Janeiro, Brazil, 2007. 32

[17] D. Comaniciu, V. Ramesh, and P. Meer. Real-time Tracking of Non-Rigid Objects using Mean Shift. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2000. 59

[18] M. De Choudhury. Modeling and Predicting Group Activity over Time in Online Social Media. In *HT '09: Proc of the 20th ACM Conf on Hypertext and Hypermedia*, Torino, Italy, 2009. 42, 43

[19] M. De Choudhury, H. Sundaram, Y.-R. Lin, A. John, and D. Duncan Seligmann. Connecting Content to Community in Social Media via Image Content, User Tags and User Communication. In *Intl. Conf. on Multimedia and Expo (ICME)*, New York, NY, USA, 2009. 42, 43

[20] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990. 44

[21] P. Dolog and M. Schäfer. A framework for browsing, manipulating and maintaining interoperable learner profiles. *User modeling 2005*, page 397–401, 2005. 80

[22] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, page 16–25, 2007. 101

[23] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proc. of the 15th Intl. Conf. on World Wide Web*, Edinburgh, Scotland, 2006. 12, 76

[24] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2), 2005. 101

[25] M. Egger, K. Fischbach, P. Gloor, A. Lang, and M. Sprenger. Deriving Taxonomies from Automatic Analysis of Group Membership Structure in Large Social Networks. In *Lecture Notes in Informatics, vol 154, Proc. of Informatik 2009*, Lubeck, 2009. 42, 102, 103

[26] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007. 99

[27] G. Gonzalez, B. Lopez, and J. L. d. l. Rosa. A multi-agent smart user model for cross-domain recommender systems. *Beyond Personalization*, 2005. 80

[28] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004. 82, 85, 97

[29] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001. 44, 45, 46, 47, 100

[30] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries for large collections of geo-referenced photographs. In *WWW '06: Proc. of the 15th Intl. Conf. on World Wide Web*, Edinburgh, Scotland, 2006. 12

[31] T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *Communications, IEEE Transactions on*, 1967. 59

[32] W. Kammergruber, M. Viermetz, and C. Ziegler. Discovering communities of interest in a tagged On-Line environment. In *2009 International Conference on Computational Aspects of Social Networks*, pages 143–148, Fontainebleau, France, 2009. 101, 102

[33] 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 *MULTIMEDIA '07: Proc. of the 15th ACM Intl. Conf. on Multimedia*, Augsburg, Germany, 2007. 12

[34] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data-mining*, Philadelphia, PA, USA, 2006. 12, 42

[35] K. Lerman and L. Jones. Social Browsing on Flickr. In *Proc. of Intl. Conf. on Weblogs and Social Media (ICWSM)*, Boulder, CO, U.S.A., March 2007. 12, 13, 42, 43, 98

[36] K. Lerman, A. Plangrasopchok, and C. Wong. Personalizing Results of Image Search on Flickr. In *AAAI workshop on Intelligent Techniques for Web Personlization*, Vancouver, Canada, 2007. 12, 13, 42, 43

[37] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic Evolution of Social Networks. In *KDD '08: Proc. 14th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, 2008. 31

[38] R. Lienhart and M. Slaney. PLSA on Large Scale Image Databases. In *ICASSP '07: Proc. of the 2007 Intl. Conf. on Acoustics, Speech and Signal Processing, Honolulu, Hawaii*, 2007. 12

[39] Y. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data*, 3(2):1–31, 2009. 101

[40] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Blog community discovery and evolution based on mutual awareness expansion. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 48–56, Fremont, CA, USA, 2007. 101

[41] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proc. of the 17th Conf. on Hypertext and Hypermedia*, 2006. 12, 42

[42] A. D. Miller and W. K. Edwards. Give and Take: a Study of Consumer Photo-sharing Culture and Practice. In *CHI'07: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, San Jose, CA, USA, 2007. 12, 31, 35, 42, 79, 93

[43] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC '07: Proc. 7th ACM SIGCOMM Conf. on Internet Measurement*, pages 29–42, New York, NY, USA, 2007. ACM. 31

[44] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. 46

[45] R. A. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez. Flickr Hypergroups. In *MM '09: Proc. of the 17th ACM Intl. Conf. on Multimedia*, Beijing, China, Oct. 2009. 96

[46] R. A. Negoescu and D. Gatica-Perez. Analyzing Flickr Groups. In *CIVR '08: Proc. of the Intl. Conf. on Image and Video Retrieval*, Niagara Falls, Canada, July 2008. 12

[47] R. A. Negoescu and D. Gatica-Perez. Topickr: Flickr Groups and Users Reloaded. In *MM '08: Proc. of the 16th ACM Intl. Conf. on Multimedia*, Vancouver, Canada, Oct. 2008. 31, 98

[48] R. A. Negoescu and D. Gatica-Perez. Modeling Flickr Communities through Probabilistic Topic-based Analysis. *IEEE Transactions on Multimedia*, 2010. 41

[49] R. A. Negoescu and D. Gatica-Perez. Flickr groups: Multimedia communities for multimedia analysis. In X.-S. Hua, M. Worring, and T.-S. Chua, editors, *Internet Multimedia Search and Mining*. Bentham Science Publishers, in press.

[50] R. A. Negoescu, A. C. Loui, and D. Gatica-Perez. Kodak moments and flickr diamonds: how users shape large-scale media. In *Proceedings of the international conference on Multimedia*, page 1027–1030, 2010. 12, 79

[51] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed Inference for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, volume 20, 2007. 76

[52] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006. 101

[53] T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh. Hyper-Community detection in the blogosphere. 2010. 101, 102, 103

[54] O. Nov, M. Naaman, and C. Ye. What drives content tagging: the case of photos on flickr. In *CHI '08: Proc of the 26th SIGCHI Conf. on Human Factors in Computing Systems*, Florence, Italy, 2008. 4, 42

[55] J. Park and S. Ram. Information systems interoperability: What lies beneath? *ACM Transactions on Information Systems (TOIS)*, 22(4):595–632, 2004. 80

[56] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR'07: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2007. 12, 32

[57] J. P. Pickett, editor. *The American Heritage Dictionary of the English Language*. Houghton Mifflin, January 2000. 13

[58] A. Plangprasopchok and K. Lerman. Constructing Folksonomies from User-specified Relations on Flickr. In *WWW '09: Proc. of the 18th Intl. Conf. on World Wide Web*, Madrid, Spain, 2009. 42

[59] C. Prieur, D. Cardon, J.-S. Beuscart, N. Pissard, and P. Pons. The Strength of Weak Cooperation: a Case Study on Flickr. Retrieved on Jan 21, 2010 from http://arxiv.org/abs/0802.2317, Feb 2008. 42

[60] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR'07: Proc. of the 30th Intl. Conf. on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007. 12

[61] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Auai '04: Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States, 2004. 47, 100

[62] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in Folksonomies: Social Link Prediction from Shared Metadata. In *Wsdm '10: Proc. of the 3rd ACM Int'l. Conf. on Web Search and Data Mining*, New York, NY, USA, 2010. 32

[63] P. Schmitz. Inducing Ontology from Flickr Tags. In *WWW '06: Proc. of the Workshop on Collaborative Tagging*, Edinburgh, Scotland, 2006. IW3C2. 12

[64] P. Schmitz. Leveraging community annotations for image adaptation to small presentation formats. In *MULTIMEDIA '06: Proc. of the 14th ACM Intl. Conf. on Multimedia*, Santa Barbara, CA, USA, 2006. 12

[65] A. Singla and I. Weber. Camera Brand Congruence in the Flickr Social Graph. In *WSDM '09: Proc. of the 2nd ACM Intl. Conf. on Web Search and Data Mining*, Barcelona, Spain, 2009. 42

[66] N. A. Van House. Flickr and Public Image-sharing: Distant Closeness and Photo Exhibition. In *CHI'07: Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, USA, 2007. 4, 12, 13, 16, 17, 31

[67] R. van Zwol. Flickr: Who is Looking. In *WI '07: Proc. of the Intl. Conf. on Web Intelligence*, San Jose, CA, USA, 2007. 4, 12, 13, 42

[68] M. Viviani, N. Bennani, and E. Egyed-Zsigmond. A survey on user modeling in multi-application environments. In *2010 Third International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services*, pages 111–116, Nice, France, 2010. 81

[69] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and their attributes. In *NIPS'05: Proc. 19th Conf. on Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2005. 100

[70] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang. PLDA: Parallel Latent Dirichlet Allocation for Large-scale Applications. In *Proc. of 5th Int. Conf. on Algorithmic Aspects in Information and Management*, 2009. 76

[71] K. Q. Weinberger, M. Slaney, and R. V. Zwol. Resolving tag ambiguity. In *Proceeding of the 16th ACM international conference on Multimedia*, page 111–120, 2008. 12

[72] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 927–936, 2009. 101, 102

[73] YouTube. Youtube Factsheet. Retrieved on Jan 18, 2011, http://www.youtube.com/t/fact_sheet, Jan. 2011. 2

[74] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An LDA-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics, 2007 IEEE*, page 200–207, 2007. 101