COMMUNITY-ORIENTED MODELS AND APPLICATIONS FOR THE SOCIAL WEB

A Dissertation

by

SAID MASOUD ALI KASHOOB

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Computer Engineering

COMMUNITY-ORIENTED MODELS AND APPLICATIONS FOR THE SOCIAL WEB

A Dissertation

by

SAID MASOUD ALI KASHOOB

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	James Caverlee
Committee Members,	Mahmoud El-Halwagi
	Donald Friesen
	Richard Furuta
Head of Department,	Valerie Taylor

May 2011

Major Subject: Computer Engineering

ABSTRACT

Community-Oriented Models and Applications for the Social Web. (May 2011)

Said Masoud Ali Kashoob, B.S., Iowa State University;

M.S., Royal Institute of Technology

Chair of Advisory Committee: Dr. James Caverlee

The past few years have seen the rapid rise of all things "social" on the web from the growth of online social networks like Facebook, to user-contributed content sites like Flickr and YouTube, to social bookmarking services like Delicious, among many others. Whereas traditional approaches to organizing and accessing the web's massive amount of information have focused on content-based and link-based approaches, these social systems offer rich opportunities for user-based and community-based exploration and analysis of the web by building on the unprecedented access to the interests and perspectives of millions of users.

We focus here on the challenge of modeling and mining social bookmarking systems, in which resources are enriched by large-scale socially generated metadata ("tags") and contextualized by the user communities that are associated with the resources. Our hypothesis is that an underlying social collective intelligence is embedded in the uncoordinated actions of users on social bookmarking services, and that this social collective intelligence can be leveraged for enhanced web-based information discovery and knowledge sharing. Concretely, we posit the existence of underlying *implicit communities* in these social bookmarking systems that drive the social bookmarking process which can provide a foundation for community-based organization of web resources.

To that end, we make three contributions:

- First, we propose a pair of novel probabilistic generative models for describing

and modeling community-oriented social bookmarking. We show how these models enable effective extraction of meaningful communities over large real-world social bookmarking services.

- Second, we develop two frameworks for community-based web information browsing and search that are based on these community-oriented social bookmarking models. We show how both achieve improved discovery and exploration of the social web.

- Third, we introduce a community evolution framework for studying and analyzing social bookmarking communities over time. We explore the temporal dimension of social bookmarking and explore the dynamics of community formation, evolution, and dissolution.

By uncovering implicit communities, putting them to use in an application scenario (search and browsing), and analyzing them over time, this dissertation provides a foundation for the study of how social knowledge networks are self-organized, a deeper understanding and appreciation of the factors impacting collective intelligence, and the creation of new information access algorithms for leveraging these communities.

To my family

ACKNOWLEDGMENTS

I am deeply indebted to the many people who, in many different ways, have made the difficulties towards getting this degree bearable, manageable and oftentimes even joyful. My deepest appreciation goes to my advisor Professor James Caverlee. Without his encouragement, guidance and support, this work would not have been completed. During the past few years of my studies, he taught me how to persist in seeking better solutions and solid results, helped me improve my writing and presentation skills, and inspired me by his thoughtful discussions and encouraging comments.

I want to thank Dr. Mahmoud El-Halwagi, Dr. Donald Friesen and Dr. Richard Furuta for sacrificing their valuable time to serve on my graduate committee. I appreciate their valuable and insightful comments about my research and on this dissertation. I am grateful to the many friends and colleagues at Texas A&M University for the numerous discussions on research, graduate courses, and various other issues. I especially want to thank my colleagues at infolab for their help on editing this manuscript. I am grateful for their friendship, help and advice.

I also want to thank the Fulbright program of the U.S. Department of State and the Ministry of Manpower of Oman for financially supporting my doctoral studies.

Last, but not least, I would like to thank all the members of my family, especially my parents, for their continuous love, encouragement, patience and support. Their believe in me is what has brought me this far.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1. Motivation

The past few years have seen the rapid rise of all things "social" on the web from the growth of online social networks like Facebook, to user-contributed content sites like Flickr and YouTube, to social bookmarking services like Delicious, among many others. Unlike top-down hierarchical information architectures that are often brittle and quickly out-dated, this social web promises a flexible bottom-up ("emergent") approach to organizing and managing information centered around *people* and their *social connections* to other people and information resources. This people-centric approach to information management can lead to large-scale user-driven growth in the size and content in the system, bottom-up discovery of "citizen-experts" with specialized knowledge, serendipitous discovery of new resources beyond the scope and intent of the original system designers, and so on. Indeed, this promise is attracting significant strategic investment and support by public health agencies, emergency responders, federal, state and local governments, major companies, and universities, among many others, and has already encouraged new advances in web-based social information sharing [1], online commerce [2], governance [3], citizen journalism [4], and education [5].

As an example, consider the social bookmarking site Delicious. Delicious is a prominent web-based social system that allows users to store bookmarks of web con-

---

The journal model is *IEEE Transactions on Automatic Control.*

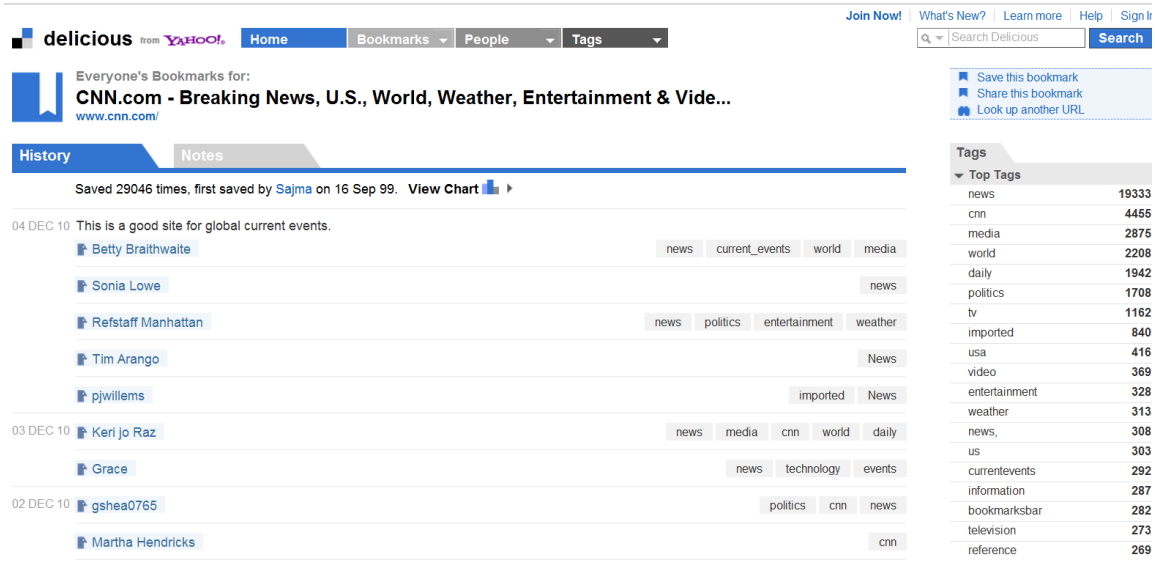


Fig. 1. Social bookmarking example: the CNN webpage's social annotation document

tent, often by associating a simple keyword or phrase with a web page, image, video, or other web media. For example, a user could associate the web resource `www.tamu.edu` with the keywords "university" and "research", while associating `www.cnn.com` with the keywords "news", "weather", and "breaking news". These keyword-based bookmarks (or *tags* or annotations) allow each user to organize a personalized view over web content, for example, by grouping together all "news" webpages of interest to the user. Beyond the purely personalized approach, Delicious users may also share their bookmarks with others, so that collectively the bookmarks of the Delicious community may affect and impact the web experience of other users. While seemingly a simple mechanism without a clear incentive structure for inducing users to bookmark web resources in the first place, much less share bookmarks with strangers, the Delicious bookmarking site alone has grown to over 5 million users who have bookmarked over 180 million unique URLs. In addition to Delicious, similar bookmarking sites have sprung up, including Flickr with more than 5 million images, CiteuLike with

around 5 million scholarly articles, and StumbleUpon with over 9 million users.

Interestingly, while a user's tags can be used in isolation as a form of bookmarking, most social bookmarking systems support the aggregation of keyword-based bookmarks so that web resources themselves may be viewed through the perspective of the millions of users who have collectively organized the web. For example, Figure (2) shows multiple users tagging decisions for the web resource `www.tamu.edu` and Figure (1) shows the top tags applied to the web resource `www.cnn.com`. Along these lines, a number of recent research efforts have studied how social bookmarking can be used for smarter browsing of web content [6], improved search [7], and other forms of information access (e.g., through tag-based clustering [8]).This type of social-powered web organization stands in contrast to traditional approaches for organizing and accessing the web's massive amount of information that have typically focused on *content-based* and *link-based* approaches (e.g., PageRank, HITS). Social bookmarking systems like Delicious, as well as other emerging social systems, offer rich opportunities for new *user-based* exploration and analysis of the web by building on the unprecedented access to the interests and perspectives of millions of users.

2. Research Challenges

Making sense of these social bookmarking systems is challenging, however, and we believe that this situation demands significant research advances to fully realize the new opportunities for large-scale user-based exploration and analysis of the web. In particular, we identify several obstacles to fulfilling the vision of large-scale user-driven organization of the web:

- **User Heterogeneity:** In contrast to "controlled vocabularies" applied by domain experts to organize web resources (e.g., like the Open Directory Project

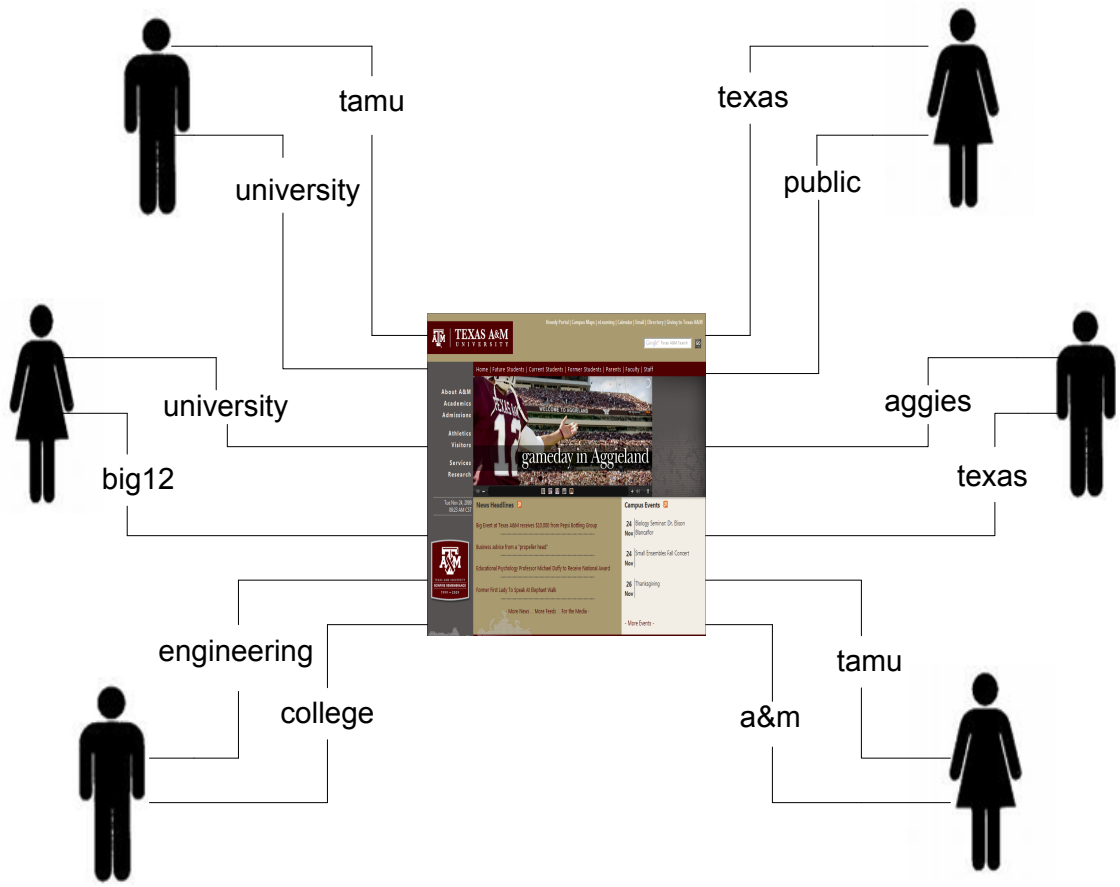


Fig. 2. Social bookmarking example: users applying various tags to a single resource

or Yahoo's web directory), social bookmarking systems rely on a heterogenous bookmarking population that may apply tags that vary greatly in purpose and quality.

- **Lack of Coordination:** Coupled with the user heterogeneity is the lack of coordination in social bookmarking systems. Since bookmarks are typically made in isolation and without explicit coordination with other users (except perhaps implicitly, by viewing the prior tags applied by users on a resource), it is not obvious if the aggregate bookmarks applied by millions of different users should provide any overarching "meaning" to web resources.

- **Scalability:** Social bookmarking systems are already very large and continue to grow and expand. Approaches for leveraging these systems for enhanced organization of the web must be designed with scalability in mind.

- **Temporal Dynamics:** Users may change interests over time and the bookmarks they apply may become stale as the underlying resources change over time. These changes suggest that methods for studying social bookmark oriented web organization should naturally incorporate temporal dynamics.

In this dissertation we are interested in exploring whether we can overcome these challenges toward providing a foundation for user-based exploration and analysis of the web.

3. Overview of Dissertation

This dissertation is concerned with studying the social bookmarking process itself, with designing models that can help us understand this process, and with developing applications that can benefit from the rich data inherent in social bookmarking

systems. The hypothesis of this dissertation research is that an underlying social collective intelligence is embedded in the uncoordinated actions of users on social bookmarking services, and that this social collective intelligence can be leveraged for enhanced web-based information discovery and knowledge sharing. Concretely, we posit the existence of underlying *implicit communities* in these social bookmarking systems that drive the social bookmarking process and can provide a foundation for community-based organization of web resources.

The notion of community is fundamental to the social web – be it friendships on Facebook, groups of similarly-interested users who comment on YouTube videos, collections of Wikipedia contributors who specialize in certain topics, and so on. Social bookmarking systems, as we have seen, aggregate what would appear to be the independent and uncoordinated tagging actions of a large and heterogeneous tagger population, meaning that it is not obvious that communities of users exist or are even detectable. In contrast to explicitly declared group memberships in social systems (which are often stale and fail to reflect the vibrant activity of the system), these implicit communities are necessarily *hidden* from us, but could provide a window into the real-time and dynamic self-organization of these systems.

Toward uncovering and leveraging community on the social web, this dissertation addresses the following research questions:

- How does community manifest itself in social bookmarking systems? And how can we model and effectively extract implicit communities from the large, heterogenous, and uncoordinated actions of millions of users?

- How can we leverage these communities to enhance how users explore the web of socially-tagged resources? And can we improve traditional methods of browsing and retrieval to naturally incorporate the community structure inherent in social

bookmarking systems?

- How does community evolve in social bookmarking systems? Are there some communities which are long-lived while others are short-lived? How do community dynamics over time impact the formation, evolution, and dissolution of communities, their members, and the tags they use?

To address these research questions, this dissertation research makes three unique contributions:

- The first contribution of this dissertation is a pair of probabilistic generative models for describing and modeling community-oriented social bookmarking. These models posit that the observed tagging information in a social bookmarking system is the product of an underlying community structure, in which users belong to implicit groups of interest. We show how these models enable effective extraction of meaningful communities and how they are better suited to modeling social bookmarking data compared to existing content-based models.

- Second, we develop two frameworks for community-based web information browsing and search that are based on our community-oriented social bookmarking models. The first relies on a novel view of web resources that combines traditional content-based modeling with community-oriented bookmarking; our results show that this multi-dimensional view of documents enhances web discovery and browsing. The second framework makes use of the community structure to augment traditional information retrieval ranking methods, and we show how it achieves improved discovery and exploration of the social web.

- Finally, we introduce a community evolution framework for studying and analyzing social bookmarking communities over time by extending the community-

oriented social bookmarking models developed as part of this dissertation. We explore the temporal dimension of social bookmarking and explore the dynamics of community formation, evolution, and dissolution. We show how this approach captures evolution, dynamics, and relationships among the discovered communities, which has important implications for designing future bookmarking systems, anticipating user's future information needs, and so on.

By uncovering these communities, putting them to use in an application scenario (search and browsing), and analyzing them over time, we can enable new avenues of transformative research, including the study of how social knowledge networks are self-organized, a deeper understanding and appreciation of the factors impacting collective intelligence, and the creation of new information access algorithms for leveraging these communities (e.g., a Google for social systems).

The remainder of this dissertation is organized as follows:

- **Chapter II: Background and Related Work.** We begin with the background for this dissertation and survey related work. We provide an overview of existing research and applications in the context of social bookmarking systems pertaining to generating semantics and ontologies, modeling social bookmarking data, and using social bookmarking data for information access and discovery.

- **Chapter III: Community-based Modeling of Social Bookmarking Systems.** This chapter introduces two community-oriented models. The first is the Community-based Categorical Annotation (CCA) Model – which uncovers tag-based communities that represent user interests and interpretations. The second model – the Probabilistic Social Annotation (PSA) Model – captures user activity and the connections between users, tags, and documents leading to an improvement in the categories discovered compared to the CCA model.

Our experimental results on datasets obtained from the Delicious and CiteUlike social bookmarking services show that our models discover more coherent categories of tags and are better suited to handle social bookmarking data compared to existing text-based topic modeling methods. Additionally these models provide a structure of user relationships to other users, to tags, and to resources that can be used to improve information exploration and discovery.

- **Chapter IV: Community-Driven Browsing and Search.** In this chapter we introduce two frameworks that employ the results of our proposed models from the previous chapter to enhance traditional methods of information discovery and retrieval of web resources. Our first framework is based on the Community-based Categorical Annotation (CCA) model. It exploits the apparent differences between the content of the web object and the tags assigned to it (the taggers perspective of the object). We devise similarity measures and explore an approach that utilizes these differences to browse for similar/dissimilar objects in a multi-dimensional space.

  The second framework is based on the probabilistic social annotation model. We develop a novel community-based ranking model for effective community-based exploration of socially tagged web resources. We compare this community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We show that the proposed ranking model results in a significant improvement over these alternatives in the quality of retrieved pages.

- **Chapter V: Temporal Dynamics of Communities in Social Bookmarking Systems.** This chapter considers a time-dependent analysis of social bookmarking services that aims at revealing important characteristics of social book-

marking such as tag evolution, web resource popularity evolution, time-specific interest, user evolution, and community evolution. We begin by observing the social bookmarks on the Delicious social bookmarking service in action for a period of 15 weeks. Our goal is to identify groups of taggers and their interests and how they evolve. To that end we propose a modification to the probabilistic social annotation model that aims to capture community-wide interests over time intervals through modeling user activity and tagging choices. Using the results of this model, we devise community dynamic graph representation that tries to capture users and tags dynamic movement between communities. We show how this representation can enable a closer inspection of communities characteristics as well as that of their constituent users and tags over time. We also show how to use this representation to capture cross-community relationships over time.

- **Chapter VI: Conclusions and Future Work.** We conclude with a summary of the contributions of this dissertation and provide a discussion of future directions that could build on the results presented here.

CHAPTER II

BACKGROUND AND RELATED WORK

1. Introduction

Social bookmarking systems are prime examples of the proliferating and increasingly popular web-based social systems, in which user activity is recorded and traceable on a massive scale. These systems provide new opportunities to explore user perspectives and interests, and to examine the relationship between social interactions and traditional web content. As we have mentioned, social annotations (or tags) are typically simple keywords or phrases that can be attached to an object as informal user-specific metadata. For example, on the Delicious social bookmarking system, a user could tag the web resource `www.espn.com` with tags like "sports", "my-favorites", and "scores". In isolation, a user's annotations can help organize a single user's bookmarks. Since these tags are shared and since many users independently assign tags to the same resource, there is a great opportunity to investigate the presence of latent structures, hidden communities, and the potential impact of these communities on information sharing and knowledge discovery. In the rest of this chapter we survey related work on social bookmarking systems, on web information organization, on topic modeling, and on community discovery.

2. Review of Existing Research on Social Bookmarking Systems

The investigation of social bookmarking and its role in modern computing and information systems has been the topic of many research works over the past few years.

Previous work has addressed various aspects of social bookmarking systems, see [9] for a discussion of the prospects, limitations and value of social bookmarking data for information and knowledge organization. In this section we review research efforts that have dealt with exploiting tagged resources to create semantics and ontologies, model the dynamics of bookmarking systems, and enhance search and retrieval.

2.1. Semantics and Ontologies

Structured data is created by professional curators that have formal education and training. Some examples of structured data systems include the Dewey Decimal System, and the Library of Congress Classification System. These systems consist of taxonomies, ontologies and controlled vocabularies that permit high quality cataloging, categorization and classification of information and resources. However, they are considered to be costly, static, and not scalable.

On the other hand, social bookmarking systems have a very low barrier to entry, and minimal expertise and education requirements as can be seen in Delicious, and Flickr among others. These systems employ free-style tagging with no vocabulary restrictions, no coordination among taggers, and no experts. These systems are inexpensive, dynamic, and scalable.

Now with the emergence of social bookmarking systems, some research works have looked into the effectiveness of social bookmarking systems in producing useful metadata [10], semantics [11, 12], and their usefulness in web classification [13] versus expert classification. Methods for augmenting structured data with free-style user contributed data [14] aim to combine the advantages of both worlds [15] and allow for the creation of emergent knowledge, "knowledge not contained in any one source" [16]. However, social bookmarking systems introduce serious issues such as vocabulary growth and reuse [17], quality selection [18], spam [19, 20, 21] and relevance to

content and query [22].

2.2. Models of Social Bookmarking

In one of the earliest studies of social bookmarking, Golder and Huberman [23] found a number of clear structural patterns in Delicious, including the stabilization of tags over time, even in the presence of large and heterogeneous user communities. This stabilization (which might be counter-intuitive, especially in contrast to the tightly controlled metadata produced by domain experts) suggests a shared knowledge in tagging communities. These results are echoed by Halpin et al. [24], who found a power-law distribution for Delicious tags applied to web pages – meaning that in the aggregate, distinct users independently described a page using a common tagging vocabulary. Similar results can be found elsewhere, including [25], [26], [27], and [28].

The past few years have seen an increased interest in modeling social annotations. Several works that adapt topic-modeling based approaches for modeling social annotations include mapping tags, users, and content to a single underlying conceptual space [12], mapping combined content and tags to an underlying topic space [29], mapping content, tags and additional link information to multiple underlying topic spaces [30]. Additionally, in [31] and [32], the authors assume hidden structures of interests and topics that generate tags for resources. They then are able to discover related resources based on their relevance (distributions) to interests and topics. These results motivate our interest in uncovering hidden communities that could help us understand social bookmarking systems better.

2.3. Social Bookmarking for Search and Recommendation

Tagging's most basic function is to organize resources as a step towards improved browsing and search [33, 7]. Once tagging activities are shared they result in an

impressive source of knowledge that can be used in numerous ways. For example, it can be used to complement link-based search methods [34, 35, 36], to measure resource popularity [6], to build language models for retrieval [29], and to detect trends [37]. Social bookmarking data has also shown potential for improved personalization [38, 39, 40, 41], query expansion [42], and recommender systems [43, 44].

3. Web Information Organization

Information organization and access methods are constantly challenged by the tremendous growth of the web[1]. These methods, which were developed over the past few decades in response to the needs of the newly emerging information (revolution) age, have undergone many reinventions. They went from keyword to text-free indexing, from boolean to algebraic and statistical approaches, and from content to metadata and link-based features. Recently, they have expanded to the social and human aspect of information.

A variety of techniques and algorithms have been developed for information retrieval. They aim to rank a set of document or web pages for a given user query based on a number of features. The most intuitive feature is the document's content where a similarity or probability measure can be computed between the user query and the content. Content based ranking employs similarity measures that take a query and a document and generate a single score which represents the relevance of the query to the document. A type of measure, based on the vector space model [45], calculates cosine similarity between query and document. A second type, based on probabilistic methods [46], calculates the probability of relevance based on query and document. A third type, based on language models [47], calculates the probability

[1] 119 million active domains in the web as of June, 2010 according to http://www.domaintools.com/internet-statistics/

of a document's language model generating the query. The current state-of-the-art content-based retrieval approaches include the BM25 retrieval model [48], the K-means-based retrieval model [49], and the LDA-based retrieval model [50], among a multitude of extensions/hybrids.

For the case of web pages, additional features like hyperlinks between pages, anchor text and web page authority have been successfully used to improve ranking. The web viewed as a graph structure of hyperlinked nodes allows the capture of ties and influences among nodes for ranking and other purposes. This is evident in the PageRank [51] and HITS [52] algorithms. Both algorithms employ citations as votes of confidence for the cited node. The votes are propagated over the graph yielding a global ranking of the nodes. Numerous approaches build on these link based ranking algorithms by incorporating page content using a set of representative topics [53], incorporating external metadata (anchor text, title, and so on) [54], or performing query dependent ranking [55].

A newer source of information on the web is the social footprint of users' interactions with web content and among each others. This can be seen in the form of clicks, query-logs, ratings, comments, tags, and friendships. And with the enormous growth in web users[2], social information will potentially help in devising approaches of information mining, organization and retrieval that can both cope with web growth and suit end user needs.

Recent research work has investigated ways of harnessing social information on the web, for example (using clicks [56], image retrieval [57], collaborative filtering [58, 59], tagging [10] and so on). Since this dissertation focuses mainly on social bookmarking and topic models, we next provide a brief overview of some related

[2] 1,802 million users as of December 2009 with a growth rate of 399.4% in 2000-2009 according to http://www.internetworldstats.com/stats.htm

work on modeling and analyzing social annotations; and on text-based topic modeling, which inspire the models introduced in this dissertation.

4. Topic Modeling

The annotation models presented in this work are inspired by related work in text-based topic modeling. A topic model typically views the words in a text document as belonging to hidden (or "latent") conceptual topics. Prominent examples of latent topic models include Latent Semantic Analysis (LSA) [60], Probabilistic Latent Semantic Analysis (pLSA) [61], and Latent Dirichlet Allocation (LDA) [62]. Topic models are an important component of many information retrieval and language modeling applications. There have been a number of extensions to traditional topic models including applications to hypertext [63] and email networks [64]. Next we present a brief overview of the LDA Model which forms the basis for our models.

4.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei et al., 2003) is a probabilistic generative model that explains the properties of a data corpus by estimating distribution parameters that best fit the observed data. In addition, LDA allows for prediction of a new observation based on the estimated parameters. In LDA documents are seen as a mixture of latent topics, where each topic is a multinomial over words in the vocabulary space. Formally, let Φ be a $K \times V$ matrix with rows representing topics, where each ϕ_k is a distribution over words for topic k, K is the number of topics, and V is the size of corpus vocabulary. Similarly, documents are represented by $M \times K$ matrix Θ, where each row, θ_S, is a distribution over topics for document S. The LDA generative process is as follows:

1. for each topic $z = 1, ..., K$
    - select V dimensional $\phi_z \sim \text{Dirichlet}(\beta)$
2. for each document S_i, $i = 1, ..., M$
    - select K dimensional $\theta_i \sim \text{Dirichlet}(\alpha)$
    - For each word w_j, $j = 1, ..., N_i$
        - Select a topic $z_j \sim \text{multinomial}(\theta_i)$
        - Select a word $w_j \sim \text{multinomial}(\phi_{z_j})$

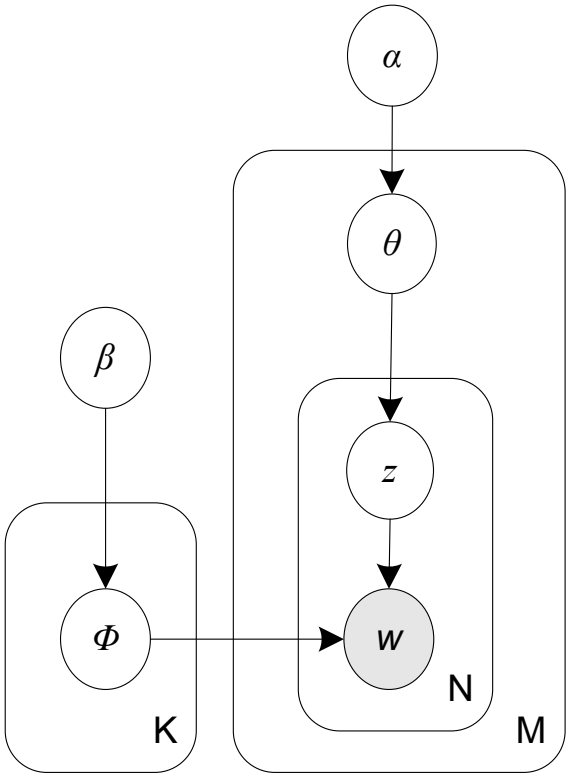


Fig. 3. Graphical representation of the LDA model

Figure (3) depicts the LDA generative process using plate notation. Plates indicate repetition with the variable at the bottom right corner specifying the number of

samples. Gray circles indicate observed variables, empty circles indicate hidden variables, and arrows indicate conditional dependency. The figure shows that parameters α, β and the distributions Φ are sampled once in the process of generating a corpus. The distributions θ_i are sampled once per document. The variables $z_{i,j}$ and $w_{i,j}$ are sampled once for each word position in each document.

Given the parameters α and β, the joint distribution of all variables is given by:

$$\begin{aligned} & p(S_i, z_i, \theta_i, \Phi|\alpha, \beta) \\ =& p(\Phi|\beta) \prod_{j=1}^{N_i} p(w_{i,j}|\phi_{z_{i,j}}) p(z_{i,j}|\theta_i) p(\theta_i|\alpha). \end{aligned}$$

The likelihood of a document S_i is obtained as follows:

$$p(S_i|\alpha, \beta) = \iint p(\theta_i|\alpha) p(\Phi|\beta) \prod_{j=1}^{N_i} p(w_{i,j}|\theta_i, \Phi) d\Phi d\theta_i.$$

The likelihood of the entire corpus $U = \{S_i\}_{i=1}^{M}$ is the product of the likelihoods of all documents:

$$p(U|\alpha, \beta) = \prod_{i=1}^{M} p(S_i|\alpha, \beta).$$

Based on this generative process, a number of standard procedures (e.g., [62], expectation propagation [65], or Gibbs sampling [66]) can be used to infer the distribution over words ϕ_k in each topic k and the distribution over topics θ_i for each document. In our work, we use Gibbs sampling to approximate the underlying distributions because of its simplicity, speed, and accuracy. Next we present Gibbs sampling in the context of the LDA topic modeling approach.

Gibbs sampling [66] is a special case of Markov-chain Monte Carlo methods that estimates a posterior distribution of high dimensional probability distribution. The

sampler draws from a joint distribution $p(x_1, x_2, ..., x_n)$ assuming the conditionals $p(x_i|x_{-i})$ are known, where

$$x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n).$$

Let $\mathbf{S}$, and $\mathbf{z}$ be vectors of length $\sum_i^D N_i$ representing word, and topic assignments, respectively, for the entire corpus. Also let w be a word variable. The joint probability distribution of the LDA model can be factored as:

$$p(\mathbf{S}, \mathbf{z}|\alpha, \beta) = p(\mathbf{S}|\mathbf{z}, \beta)p(\mathbf{z}|\alpha).$$

We derive the Gibbs sampler's update equation as was done in [66] (see Appendix A for details) for the hidden variable of a word token at position i from the joint distribution, noting that $\mathbf{S} = \{\mathbf{S}_i = w, \mathbf{S}_{\neg i}\}$ and $\mathbf{z} = \{\mathbf{z}_i = k, \mathbf{z}_{\neg i}\}$:

$$p(\mathbf{z}_i = k|\mathbf{z}_{\neg i}, \mathbf{S}) \propto \frac{n^w_{k,\neg i} + \beta_u}{\sum_{w=1}^V n^w_{k,\neg i} + \beta_w} \times \frac{n^k_{d,\neg i} + \alpha_k}{\sum_{k=1}^K n^k_{d,\neg i} + \alpha_k} \tag{2.1}$$

where $n^{(\cdot)}_{(\cdot),\neg i}$ is a count excluding the current position assignments of $\mathbf{z}_i$ (e.g., $n^w_{k,\neg i}$ is the count of word w generated by the k-th topic excluding the current position).

Having estimated the topic assignments $\mathbf{z}$, estimates of Φ and Θ are computed as follows:

$$\phi_{z,w} = \frac{n^w_z + \beta_w}{n_z + \sum_w \beta_w}$$

and

$$\theta_{d,z} = \frac{n^z_d + \alpha_z}{n_d + \sum_z \alpha_z}$$

Now for a new unseen web object $\tilde{d}$, the Gibbs sampler can predict its topic distribution and word assignment as follows:

$$\phi_{z,\tilde{w}} = \frac{n_z^{\tilde{w}} + n_z^w + \beta_w}{n_z + \sum_w \beta_w}$$

and

$$\theta_{\tilde{d},z} = \frac{n_{\tilde{d}}^z + \alpha_z}{n_{\tilde{d}} + \sum_z \alpha_z}$$

With this ability to uncover the hidden components that participate in generating the documents, we will see in the coming chapters how it can be extended to uncover the communities of users that constitute the collective intelligence resulting from large scale social bookmarking processes. Next we provide a short introduction to community discovery.

5. Community Detection

Community detection is a rich topic with a number of approaches drawing from sociology and network theory, including [67], [68], [69], [70], [71], and [72]. Communities are groups of cohesive objects (people, documents, nodes, etc) that are closely related based on interactions, similarity, or interests of objects within the group as opposed to objects that lie outside the group. The task of finding these groups and devising measures for their similarity is an open research topic in many areas like social, communication, computer, and biological networks. Example similarity measures that have been developed for community detection include spectral clustering [73] and modularity maximization [74]. These methods among others have been applied to social web data, including bookmarking systems, to detect communities.

Some recent graph-based attempts at social bookmarking analysis include normalized cut methods for finding clusters [75], recursive and k-way partitioning to greedily optimize modularity [76] and recursive spectral bisection [33]. Alternative efforts based on vector space representation include association rules [27], and iterative methods [41]. All these approaches target mainly homogenous systems (e.g. single node-type such as users, or documents). Some works that have adapted these approaches to heterogenous systems (multiple node types) include [77], [78], and [79]. Adapting and integrating some of these approaches with our community-based models is an area we will explore in future research works.

CHAPTER III

COMMUNITY-BASED MODELING OF SOCIAL BOOKMARKING SYSTEMS

1. Introduction

In this chapter we begin our exploration of community in social bookmarking systems by proposing and evaluating two community-oriented models of the social bookmarking process. These models are designed to uncover meaningful community structures implicit in large collections of socially bookmarked web resources. For example, an individual web resource (like an image, video, or web page) in a social bookmarking system is typically associated with a set of tags and a set of users that tagged the resource, but without regard for the common community-driven motivations of the users who applied the tags in the first place. Can we uncover these hidden communities (see Figure 4) purely through an analysis of the observed tags? Modeling the fundamental bookmarking processes that ultimately lead to the observed tags in a social bookmarking system is essential to better understand this new social media, to improve the design of future bookmarking systems, and to drive our insights into leveraging community for enhanced web organization.

Our approach for studying and modeling social web objects is inspired by related work in text-based topic modeling. A topic model typically views the words in a text document as belonging to hidden (or "latent") conceptual topics. Prominent examples of latent topic models include Latent Semantic Analysis (LSA) [60], Probabilistic Latent Semantic Analysis (pLSA) [61], and Latent Dirichlet Allocation (LDA) [62]. Topic models are an important component of many information retrieval and language modeling applications. There have been a number of extensions to traditional

topic models including applications to hypertext [63], email networks [64], and social bookmarking data [12, 29, 31, 30].

To uncover the underlying structure of social bookmarking systems, we propose two models – the Community-based Categorical Annotation (CCA) [80] and the Probabilistic Social Annotation(PSA) [81], that take into account the additional features that come with social web objects (tags, and users). These models differ from previous efforts on modeling social bookmarking data in the following ways: (i) groups of users with a common understanding or similar interpretations of resources are represented as a community; (ii) each community has a world view encoded by a set of categories; and (iii) we recognize that the annotation process involves a user choice to participate in annotating an object and a subsequent decision on the type of tags to be used and we place that at the core of our models.

In the rest of this chapter, we present each model in turn and compare the proposed models to two prominent probabilistic topic models (Latent Dirichlet Allocation and Pachinko Allocation) via an experimental study of the popular Delicious, Flickr, and CiteULike bookmarking services. We find that the proposed community-based annotation models identify more coherent implicit structures than the alternatives and are better suited to handle unseen social annotation data.

2. Preliminaries and Reference Model

We consider a universe of discourse $\mathcal{U}$ consisting of D socially bookmarked resources: $\mathcal{U} = \{O_1, O_2, ..., O_D\}$. We view each resource O_i by both its intrinsic content C_i and the social annotations (or tags) S_i attached to it by the community of users. Hence, each resource is a tuple $O_i = \langle C_i, S_i \rangle$ where the content and the social annotations are modeled separately. We call the social annotations S_i applied to a resource its

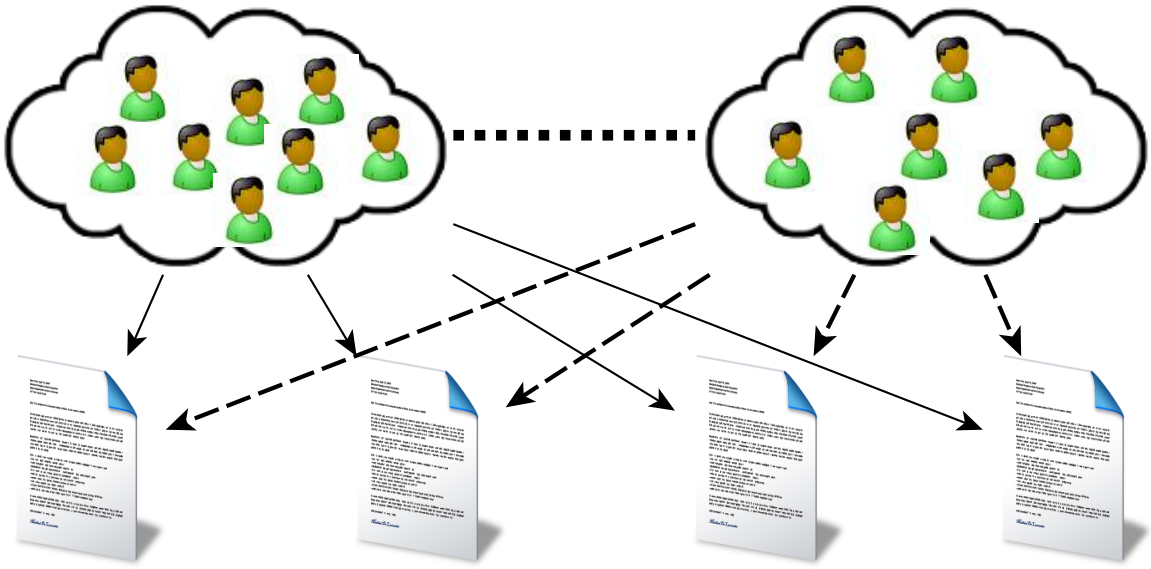


Fig. 4. Latent communities of users engaging in social bookmarking

social annotation document. For example, the resource corresponding to a web page tagged on the Delicious bookmarking service would consist of the HTML contents of the web page as well as the *social annotation document* generated by the members of the Delicious service. A social annotation document can be modeled by the set of tags and their frequencies:$S_i = \{\langle tag_j, freq(tag_j)\rangle\}$. We will additionally model the particular user applying the tag, as well as the time at which each tag was applied. Note that the CCA model introduced in the following section implicitly models users via tag co-occurrence.

[Definition] Social Annotation Document: *For a resource $O \in \mathcal{U}$, we refer to the collection of tags assigned to the resource as the resource's social annotation document S, where S is modeled by the set of users and the tags they assigned to the resource: $S = \{\langle user_j, tag_j\rangle\}$.*

Our proposed models are derived from the LDA topic modeling approach proposed by Blei et al. [62]. An LDA-based model can be easily adapted to social

annotations by considering the document unit to be a social annotation document and the underlying topics to be social annotation categories. Since LDA is typically used in document-based modeling and not tag-based modeling, we shall refer to the adapted version as TagLDA for clarity.

TagLDA: TagLDA views a tag document as a mixture of latent categories (or topics), where each category is a multinomial over tags in the vocabulary space. Formally, let Φ be a $K \times V$ matrix representing categories, where each ϕ_k is a distribution over tags for category k, K is the number of categories, and V is the size of tag vocabulary. Similarly, object are represented by $D \times K$ matrix Θ, where each θ_S is a distribution over categories for object S.

The TagLDA generative process is as follows:

1. for each category $z = 1, ..., K$
    - select V dimensional $\phi_z \sim \text{Dirichlet}(\beta)$
2. for each object S_i, $i = 1, ..., D$
    - select K dimensional $\theta_i \sim \text{Dirichlet}(\alpha)$
    - For each tag t_j, $j = 1, ..., N_i$
        - Select a category $z_j \sim \text{multinomial}(\theta_i)$
        - Select a tag $t_j \sim \text{multinomial}(\phi_{z_j})$

At this point, the estimation procedures introduced in Chapter II can be applied to uncover the latent structure that underlies the described generative process. When using the Gibbs sampling method, the update equation is the same as equation (2.1), presented in the related work chapter, with tags substituted for words and categories substituted for topics.

TagLDA provides a foundation for discovering communities in social bookmarking systems. Fundamentally, however, a social annotation document is a collaborative effort among many taggers, whereas TagLDA is a topic model with no notion of authorship or community. In essence TagLDA can be used to discover social annotation categories over tags, but not social annotation communities over users, since users are not explicitly modeled in the generation process. Recent work on author-topic models [64] has added the concept of "author" to the LDA model, but fundamentally these models are designed to model text documents that have a single (or a few) authors. In contrast, a social annotation document is the product of (potentially) hundreds of authors. These observations suggest a new approach. Our first model is an LDA extension that can uncover communities based solely on tags. We further extend the model to include users and time.

3. Community-based Categorical Annotation (CCA) Model

In this section we propose a probabilistic generative model that aims to model the social annotation process. By modeling the communities of interest that engage in social bookmarking and the implicit categories that each community considers, we develop the Community-based Categorical Annotation (CCA) Model. The CCA model views a *category* as a mixture of tags and a *community* as a mixture of categories. Hence, a community of interest is inherently composed of the tags that it uses.

[Definition] Social Annotation Community: *A social annotation community c is a composite of tag categories that are related through the underlying process that generated social bookmarking data, and presumably represents the outlook of a certain group of users and their interest when they participate in the tagging process.*

[Definition] Social Annotation Category: *A social annotation category z is*

a probability distribution over tags in the vocabulary V such that $\sum_{t \in V} p(t|z) = 1$, where $p(t|z)$ indicates membership strength for each tag t in category z.

3.1. Generating Social Annotations with CCA

We begin with an example. Suppose we have an image of a Tyrannosaurus rex. The collaborative tagging environment allows this object to be tagged by users with various interests, expertise, and in various human languages. Hence, the social annotation document associated with this image may include tags that were applied by a scientist e.g., tags like `cretaceous` and `theropod`), by an elementary school student (e.g., tags like `meat-eater` and `t-rex`) and by a French-speaking tagger (e.g., tags like `carnivore` and `lézard-tyran`).

We view the underlying groups that form around these interests, expertise, and languages as distinct *communities*. For each community, there may be some number of underlying *categories* that inform how each community views the world. Continuing our example, the scientist community may have underlying categories centered around Astronomy, Biology, Paleontology, and so on. For each object, the community selects tags from the appropriate underlying category or mixture of categories (e.g., for tagging the dinosaur, the tags may be drawn from both Biology and Paleontology).

In practice, these communities and categories are *hidden* from us; all we may observe is the social annotation document that is a result of these communities and the categories they have selected.

Formally, CCA assumes a corpus of D social annotation documents drawn from a vocabulary V of tags, where each social annotation document S_i is of variable length N_i. The model assumes that the tags in a social annotation document are generated from a mixture of L distinct communities, where each community is a mixture of hidden categories K_l, and where each category is a mixture of tags. Therefore, the

tagging process involves two steps: 1) the selection of a community from which to draw tags and 2) the selection of the categories that influence the preference over tags based on the object's content, and the tagger's perception/understanding of the content. The CCA tag generation process is illustrated in Figure (5) and described here:

1. for each community $c = 1, ..., L$
   - for each category $z = 1, ..., K_c$
     - select V_c dimensional $\phi_z \sim$ Dirichlet(γ)
2. for each object S_i, $i = 1, ..., D$
   - Select L dimensional $\kappa \sim$ Dirichlet(α)
   - for each community $c = 1, ..., L$
     - select K_c dimensional $\theta_c \sim$ Dirichlet(β)
   - For each tag position $S_{i,j}$, $j = 1, ..., N_i$
     - Select a community $c_{i,j} \sim$ multinomial(κ_i)
     - Select a category $z_{i,j} \sim$ multinomial($\theta_{c_{i,j}}$)
     - Select a tag $t_{i,j} \sim$ multinomial($\phi_{z_{i,j}}^{c_{i,j}}$)

A social annotation document's community distribution $\kappa_i = \{\kappa_{i,j}\}_{j=1}^{L}$ is sampled from a Dirichlet distribution with parameter $\alpha = \{\alpha_i\}_{i=1}^{L}$. A community's category distribution $\theta_i = \{\theta_{i,j}\}_{j=1}^{K}$ is sampled from a Dirichlet distribution with parameter $\beta = \{\beta_i\}_{i=1}^{K}$. A category's tag distribution $\phi_z = \{\phi_{z,i}\}_{i=1}^{|V|}$ is sampled from a Dirichlet distribution with parameter $\gamma = \{\gamma_i\}_{i=1}^{|V|}$. The generative process creates a social annotation document by sampling for each tag position $S_{i,j}$ a community $c_{i,j}$ from a multinomial distribution with parameter κ_i, a category $z_{i,j}$ from a multinomial

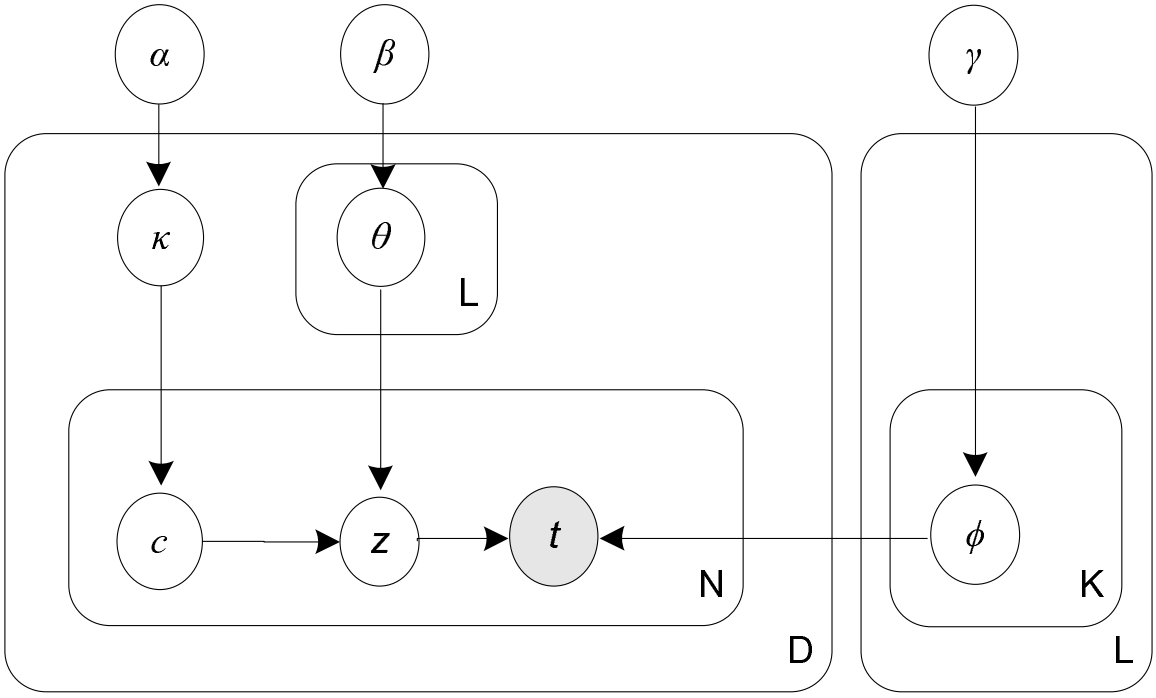


Fig. 5. Graphical representation of the CCA model.

distribution with parameter $\theta_{c_{i,j}}$. A tag is then sampled for that position from a multinomial distribution with parameter $\phi_{z_{i,j}}^{c_{i,j}}$.

Based on the model, we can write the likelihood that a tag position $S_{i,j}$ in a social annotation document is assigned a specific tag t as:

$$p(S_{i,j} = t|\kappa_i, \Theta, \Phi) =$$
$$\sum_{l=1}^{L}\sum_{k=1}^{K_l} p(S_{i,j} = t|\phi_k^l)p(z_{i,j} = k|\theta_l)p(c_{i,j} = l|\kappa_i).$$

Furthermore, the likelihood of the complete social annotation document S_i is the joint distribution of all its variables (observed and hidden):

$$p(S_i, z_i, c_i, \kappa_i, \Theta, \Phi|\alpha, \beta, \gamma) =$$
$$\prod_{j=1}^{N_i} p(S_{i,j}|\phi_{z_{i,j}}^{c_{i,j}})p(z_{i,j}|\theta_{c_{i,j}})p(c_{i,j}|\kappa_i).$$

Integrating out the distributions κ_i, Θ, and Φ and summing over c_i and z_i gives the marginal distribution of S_i given the priors:

$$p(S_i|\alpha,\beta,\gamma) = \iiint p(\kappa_i|\alpha)p(\Theta|\beta)p(\Phi|\gamma) \times \prod_{j=1}^{N} p(S_{i,j}|\kappa_i,\Theta,\Phi)d\Phi d\Theta d\kappa_i$$

Finally our universe of discourse $\mathcal{U}$ consisting of all D social annotation documents occurs with likelihood:

$$p(\mathcal{U}|\alpha,\beta,\gamma) = \prod_{i=1}^{D} p(S_i|\alpha,\beta,\gamma)$$

3.2. Parameter Estimation and Inference

The CCA model provides a generative approach for describing how social annotation documents are constructed. But our challenge is to work in the reverse direction – taking a set of social annotation documents and inferring the underlying model (including the hidden community and category distributions). This entails learning model parameters κ, Θ, and Φ (the distributions over communities, categories, and tags, respectively).

Previous work that aimed at recovering similar hidden structure from joint posterior distributions has shown that exact computation of these parameters is intractable. There exists, however, several approximation methods in the literature for solving similar parameter estimation problems (like in LDA), including expectation maximization [62], expectation propagation [65], and Gibbs sampling. In this dissertation, we adopt Gibbs Sampling (see [66] for a thorough treatment) which is a special case of Markov-chain Monte Carlo methods that estimates a posterior distribution of a high-dimensional probability distribution. The sampler draws from a joint distribution $p(x_1, x_2, ..., x_n)$ assuming the conditionals $p(x_i|x_{-i})$ are known,

where $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$.

For community assignment c, category assignment z, tag assignment t of tag positions in a corpus, and given the parameters α, β and γ, the joint probability function can be factored into:

$$p(S, z, c|\alpha, \beta, \gamma) \quad = \quad p(S|c, z, \gamma)p(z|c, \beta)p(c|\alpha).$$

We derive the Gibbs sampler's update equation for the hidden variables (community, and category) from the joint distribution (in a similar fashion to the approach in Appendix A) and arrive at:

$$p(c_i, z_i|c_{-i}, z_{-i}, t) \propto \frac{n_S^{c_i} - 1 + \alpha_{c_i}}{n_S - 1 + \sum_c \alpha_c} \times \frac{n_{z_i}^{t_i} - 1 + \gamma_{t_i}}{n_{z_i} - 1 + \sum_t \gamma_t} \times \frac{n_S^{z_i} - 1 + \beta_{z_i}}{n_S^{c_i} - 1 + \sum_z \beta_z}$$

where t_i is the tag at position i, z_i is the category, c_i is the community, S is the object, $n_S^{c_i}$ is the count of positions in the object assigned to community c_i, n_S is the length of the object, $n_{z_i}^{t_i}$ is the count of positions with category z_i and tag t_i in the corpus, n_{z_i} is the count of positions with category z_i in the corpus, and $n_S^{z_i}$ is the count of positions with category z_i in the object.

The first factor represents the weight of community c_i in the object, the second represents the contribution of the tag at position i to category z_i in the entire corpus, while the third factor represents the weight of category z_i in the object.

Having estimated the community assignment c and category assignment z, esti-

mates of Φ,Θ and κ are computed as follows:

$$\phi_{z,t} = \frac{n_z^t + \gamma_t}{n_z + \sum_t \gamma_t}$$

$$\theta_{i,z} = \frac{n_S^z + \beta_z}{n_S^c + \sum_z \beta_z}$$

$$\kappa_{i,c} = \frac{n_S^c + \alpha_c}{n_S + \sum_c \alpha_c}$$

Now for a new unseen social annotation document $\tilde{S}$, the Gibbs sampler can predict its tag assignment as follows:

$$\phi_{z,\tilde{t}} = \frac{n_z^{\tilde{t}} + n_z^t + \gamma_t}{n_z + \sum_t \gamma_t}$$

where $n_z^{\tilde{t}}$ is the count of positions with category z and tag t in the unseen object. Its category distribution is:

$$\theta_{\tilde{S},z} = \frac{n_{\tilde{S}}^z + \beta_z}{n_{\tilde{S}}^c + \sum_z \beta_z}$$

where $n_{\tilde{S}}^z$ is the count of positions with category z in the unseen object, and its community distribution is:

$$\kappa_{\tilde{S},c} = \frac{n_{\tilde{S}}^c + \alpha_c}{n_{\tilde{S}} + \sum_c \alpha_c}$$

where $n_{\tilde{S}}^c$ is the count of positions with community c in the unseen object.

3.3. Applying CCA to Flickr and Delicious

Given the categorical annotation model, we next apply the model to two prominent social bookmarking services – Flickr (for images) and Delicious (for web pages).

Flickr dataset: For Flickr, we began a crawl from the tag cloud at `http://flickr.com/photos/tags`. We have identified 1,578,437 images that have been an-

notated by 42,156 unique users who have used 156,127 unique tags. For the experiments in this paper, we considered a sample of 92,000 images that have been tagged by 44,980 unique tags. We normalize the data and train the categorical annotation model with 90,000 objects and use the rest for testing.

Delicious dataset: Like Flickr, the Delicious crawler starts with a set of popular tags. Our crawler has discovered 607,904 unique tags, 266,585 unique web pages annotated by Delicious, and 1,068,198 unique users. Of the 266,585 total web pages, we have retrieved the full HTML for a randomly chosen set of 47,852 pages. We filter this set to keep only pages in English (We identify English pages using the TextCat implementation based on [82]) with a minimum length of 20 words, leaving us with 27,572 web pages with 16,216 unique annotations. Since many of the pages annotated by Delicious are primarily text documents, we also parsed the text of each document for an analysis discussed later in the paper. We use $20,000$ of the objects to train our model and the remaining $7,572$ are used for testing.

3.4. Revealing Hidden Categories

One challenge to discovering latent structure in social annotations is to identify the appropriate number of hidden categories and hidden communities of interest that generated the observed data. Since the hidden categories and communities are not directly observed, we must use some unsupervised method.

In this section, we begin by considering the simplified case of a single community, but with an unknown number of hidden categories. We revisit this assumption in the following section. To identify the number of categories, we rely on a standard measure from information theory – perplexity. Perplexity measures how well a model (here the categorical annotation model built over a training set) predicts a test sample, and it is has been widely used in text-based topic modeling (e.g., [62, 29]). We measure

perplexity on a held-out set $\tilde{D}$ using the parameters of an estimated model $\mathcal{M}$ for a given dimension (or category) K for the hidden variable:

$$Perp(\tilde{D}) = \exp - \frac{\sum_{d=1}^{\tilde{D}} \log P(S_d|\mathcal{M})}{\sum_{d=1}^{\tilde{D}} N_d}$$

where

$$\log P(S_d|\mathcal{M}) = \sum_{t=1}^{V} n_d^{(t)} \log \left(\sum_{k=1}^{K} \phi_{k,t}\theta_{d,k} \right)$$

and $n_d^{(t)}$ is the number of times terms t was observed in document S_d and N_d is the length of S_d. The variable ϕ is a model parameter while the variable θ is computed for the held-out set. Low perplexity values indicate a good selection of the number of categories for the hidden variable given a corpus.

We experimented with different category dimensions for both Flickr and Delicious. The perplexity as a function of the number of categories for Flickr is shown in Figure (6). The horizontal axes show the number of categories and the vertical axes show the perplexity values. Notice the decrease in perplexity as the number of categories increase, as well as the different rates of decrease. For Delicious, we observe a similar curve as shown in Figure (7) , but with a "knee" at 40 categories. Based on these results, we selected 70 categories for Flickr and 40 categories for Delicious.

Given the choice of the number of categories for both Flickr and Delicious, what are the discovered categories? And are they semantically coherent? In Tables (I) and II, we report the most significant annotations for a sample of 10 of the discovered categories in each dataset ranked by probability of tag given a category $\phi_{z,t}$. We find that overall the discovered categories appear to be semantically meaningful. As future work, it will be interesting to evaluate these discovered categories in a concrete application setting (e.g., tag-based information retrieval).

To further illustrate the revealed categories, we report in Table (III) the most relevant documents per category for 10 of the Delicious categories. We rank the

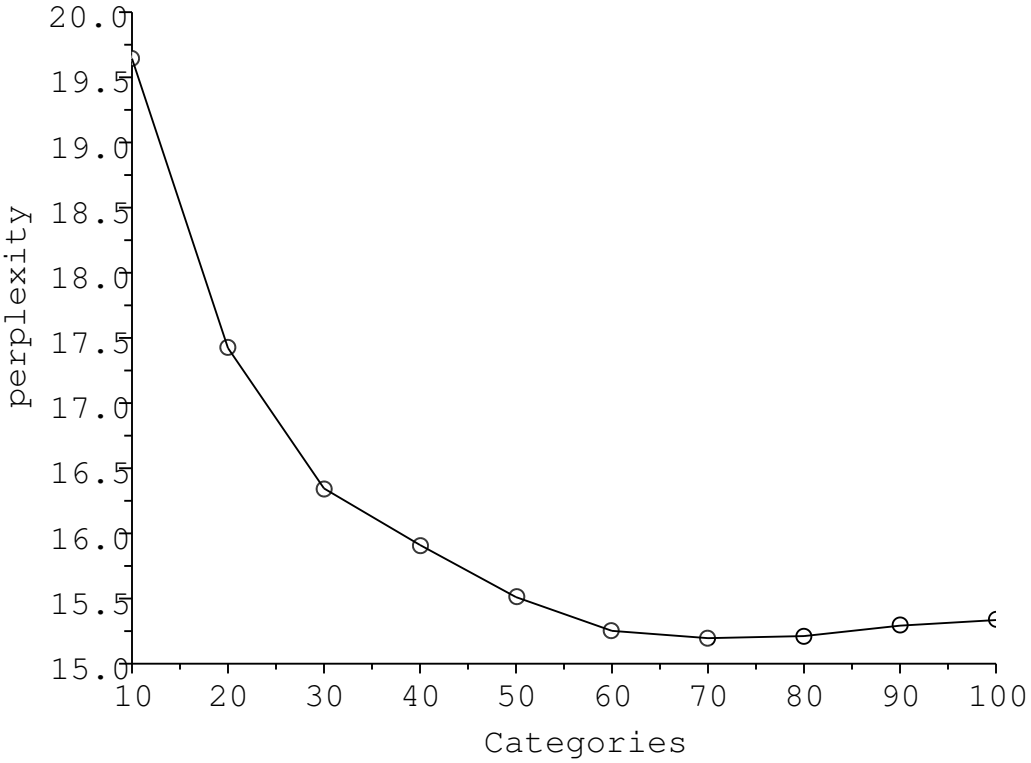


Fig. 6. CCA-based category perplexity for Flickr.

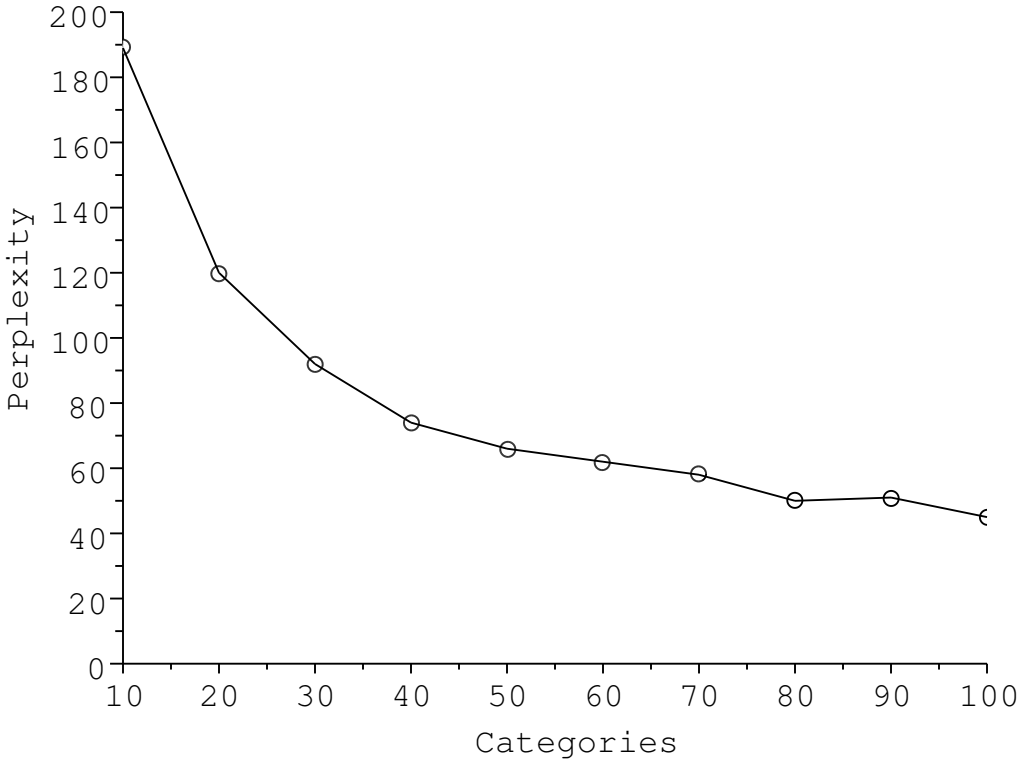


Fig. 7. CCA-based category perplexity for Delicious.

Table I. Flickr: 10 of the 70 discovered categories and the most likely tags per category (in order of $\phi_{z,t}$).

Cat 0:	boat, sport, itali, water, torino, athlet, ship, turin, sundai, sail, oar, rower, competit, ...
Cat 1:	canada, veteran, vancouv, memori, war, remembrancedai, dai, ontario, remembr, ...
Cat 2:	portrait, face, hand, woman, photoshop, hair, girl, color, lip, photograph, self, retrato, ...
Cat 3:	build, citi, architectur, old, urban, tower, histor, skyscrap, skylin, stone, center, librari, ...
Cat 4:	water, river, blue, reflect, bridg, fish, sky, boat, canon, artist, washington, mountain, ...
Cat 5:	mountain, winter, snow, landscap, lake, switzerland, cold, montagna, alp, trek, ...
Cat 6:	art, graffiti, paint, urban, streetart, street, tag, draw, sticker, illustr, abstract, artist, ...
Cat 7:	cat, anim, love, kitten, cute, kitti, pet, gato, felin, chat, gatto, bunni, rabbit, heart, ...
Cat 8:	train, railwai, tourist, tourism, station, laura, railroad, unitedkingdom, ride, york, locomot,...
Cat 9:	food, cook, cake, restaur, chocol, dinner, sweet, eat, minnesota, yummi, wine, bake, ...

Table II. Delicious: 10 of the 40 discovered categories and the most likely tags per category (in order of $\phi_{z,t}$).

Cat 0:	webdesign, design, inspir, web, resource, templat, galleri, award, web2.0, websit, ...
Cat 1:	secur, financ, monei, .net, storag, invest, backup, asp.net, c#, busi, econom, bank, ...
Cat 2:	googl, mobil, calendar, phone, sync, api, voip, cellphon, comparison, nokia, sm, ...
Cat 3:	mac, osx, appl, wiki, softwar, ipod, macosx, app, applic, tool, ssh, wikipedia, quicksilv, ...
Cat 4:	educ, math, learn, resourc, teach, kid, technolog, mathemat, school, interact, elearn, ...
Cat 5:	tutori, howto, photoshop, tip, refer, guid, adob, articl, resourc, effect, trick, text, ...
Cat 6:	photographi, photo, imag, galleri, flickr, camera, slideshow, mindmap, stock, space, ...
Cat 7:	rubi, rail, rubyonrail, host, nyc, amazon, web, http, authent, s3, webhost, develop, ...
Cat 8:	fun, humor, funni, comic, cool, geek, interest, entertain, humour, del.icio.us, cartoon, ...
Cat 9:	video, visual, anim, movi, tv, film, youtub, motiongraph, motion, stream, media, ...

documents using the probability of a category given a document $\theta_{i,z}$. We find that the quality of these results is consistent across categories.

3.5. Discovering Communities

Given the results of uncovering hidden categories, we next turn to the nature of results when we estimate both the communities of interest (which recall are composed of underlying categories) and the categories within each community (which recall are composed of tags). Experimentally, we have run the CCA model with several community/category combinations, and in Table (IV) we report a representative result for 5 communities and 5 categories for Delicious.

Note how the communities of interest are centered around categories that share some thematic relationship. For example, Comm2 is a "Lifestyle" community of interest with categories related to shopping, travel, food, and books. In the flat single community analysis of the previous section, these types of categories would either be combined into a single category of interest, blurring the distinct interests of each category, or the categories may be separated but not linked by community. Here, we see how the CCA model provides a hierarchical layer for grouping related categories by their common community of interest. Further, note that the two more technically minded communities are indeed quite distinct – Comm3 is centered around "web 2.0" from a consumer point-of-view (with categories related to YouTube, blogs, and social networking), whereas Comm4 is centered around "web 2.0" from a development point-of-view (with categories related to different web development tools and languages). These results are encouraging and in the following sections, we explore techniques to further refine the quality of community formation.

Table III. Top 4 most relevant documents per category ranked by $\theta_{i,z}$ (showing 10 of the 40 categories)

Category 0 (Web design)	**Category 5 (Photoshop)**
http://www.Webbyawards.com/Webbys/current.php?season=12	http://psdtuts.com/photo-effects-tutorials/applying-a-realistic-tattoo/
http://www.coolhomepages.com/	http://abduzeedo.com/creating-smoke
http://vandelaydesign.com/blog/galleries/minimal-Websites-designs/	http://psdtuts.com/text-effects-tutorials/create-a-spectacular.../
http://www.designlicks.com/flash/index.php	http://psdtuts.com/tutorials-effects/seriously-cool-photoshop.../
Category 1 (Banking and money)	**Category 6 (Photography)**
https://www.fidelity.com/	http://hirise.lpl.arizona.edu/earthmoon.php
http://home.ingdirect.com/	http://www.boston.com/bigpicture/2008/05/cassini_nears_four.../
http://www.chase.com/	http://wildphoto.smugmug.com/
http://www.wamu.com/personal/default.asp	http://www.boston.com/bigpicture/2008/06/martian_skies.html
Category 2 (Calendar syncing and messaging)	**Category 7 (Ruby)**
http://www.gcalsync.com/	http://ec2onrails.rubyforge.org/
http://oggsync.com/	http://code.macournoyer.com/thin/
http://www.clickatell.com/pricing/message_cost.php	http://www.hostingrails.com/
http://www.davesWebsite.com/software/gsync/	http://mongrel.rubyforge.org/
Category 3 (Apple/Mac)	**Category 8 (Fun and humor)**
http://www.magnetk.com/expandrive	http://www.dilbert.com/
http://macntfs-3g.blogspot.com/	http://www.achewood.com/
http://code.google.com/p/macfuse/	http://xkcd.com/162/
http://www.sccs.swarthmore.edu/users/08/mgorbach/MacFusionWeb/	http://www.sarcasmsociety.com/
Category 4 (Education)	**Category 9 (Video and movies)**
http://school.discoveryeducation.com/schrockguide/assess.html	http://www.netflix.com/MemberHome
http://www.learningpage.com/	http://www.netflix.com/
http://edhelper.com/	http://joox.net/
http://www.teach-nology.com/	http://www3.alluc.org/alluc/

Table IV. Communities, their categories, and the most likely tags per category (in order of $\phi_{z,t}$).

Comm 0	Cat 0	art, design, paper, drawing, diy, fun, cool, animation, toys, crafts...
	Cat 1	humor, funny, fun, comics, geek, blog, comic, humour, cool, webcomic...
	Cat 2	dictionary, reference, language, english, writing, tools, thesaurus, slang ...
	Cat 3	games, game, fun, flash, gaming, free, online, puzzle, secondlife, charity...
	Cat 4	imported, misc, firefoxbookmarks, bookmarks, firefoxtoolbar, myspace, computer...
Comm 1	Cat 0	science, math, kids, mathematics, reference, education, astronomy, games, physics,...
	Cat 1	howto, productivity, lifehacks, tips, gtd, reference, diy, tutorial, blog, organization...
	Cat 2	education, learning, resources, teaching, elearning, technology, free, video, language, online...
	Cat 3	photography, photo, photos, images, flickr, art, tools, graphics, free, image...
	Cat 4	web20, tools, wiki, collaboration, presentation, mindmap, online, software, powerpoint, free...
Comm 2	Cat 0	shopping, environment, design, green, tshirts, home, sustainability, clothing, activism, shop...
	Cat 1	travel, reference, maps, airline, world, guide, flights, seating, airlines, geography...
	Cat 2	books, reference, library, research, history, literature, ebooks, free, archive, writing...
	Cat 3	news, blog, technology, magazine, politics, blogs, tech, culture, daily, media...
	Cat 4	food, cooking, recipes, blog, recipe, music, reviews, reference, blogs, howto...
Comm 3	Cat 0	tools, web20, mobile, productivity, software, collaboration, calendar, phone, widgets, web...
	Cat 1	video, tv, streaming, videos, movies, media, youtube, free, television, online...
	Cat 2	business, marketing, advertising, startup, internet, technology, trends, entrepreneurship, ideas, media...
	Cat 3	web20, social, community, socialnetworking, twitter, collaboration, tools, socialsoftware, networking, web...
	Cat 4	blog, web20, blogs, blogging, news, rss, aggregator, web, tools, technology...
Comm 4	Cat 0	javascript, ajax, programming, framework, development, web20, web, library, webdev, yahoo...
	Cat 1	wordpress, blog, themes, theme, plugin, blogging, plugins, blogs, templates, design...
	Cat 2	php, opensource, cms, software, web, email, development, drupal, ecommerce, programming...
	Cat 3	css, webdesign, web, design, html, reference, tutorial, webdev, standards, development...
	Cat 4	programming, java, development, reference, c, net, database, tutorial, sql, tools...

3.6. Summary

The collaborative tagging process is driven by the users' various interests, expertise, background, and language. We view the underlying groups that form around these interests, expertise, and languages as distinct *communities*. For each community, there may be some number of underlying *categories* that inform how each community views the world, i.e., selects tags. The CCA generative model aims to model the social annotation process by modeling the communities of interest that engage in social bookmarking and the implicit categories that each community considers when making tagging decisions. It assumes that a *community* is a mixture of categories and a *category* is a mixture of tags.

Formally, CCA assumes a corpus of D social annotation documents drawn from a vocabulary V of tags, where each social annotation document S_i is of variable length N_i. The model assumes that the tags in a social annotation document are generated from a mixture of L distinct communities, where each community is a mixture of hidden categories K_l, and where each category is a mixture of tags. Therefore, the tagging process involves the selection of a community from which to select categories and the selection of the category that specifies the preference over tags based on the object's content, and the tagger's perception/understanding of the content.

In practice, these communities and categories are *hidden* from us; all we may observe is the social annotation document that is a result of these communities and the categories they have selected. We use Markov-chain Monte Carlo (MCMC) methods that simulate the posterior distribution to estimate the underlying structure of the social bookmarking process. Given a collection of social annotation documents we are able to infer this underlying structure, i.e., community and category distributions.

Experimentally, we have run the CCA model with several community/category

combinations and our results show how the communities of interest are centered around categories that share thematic relationships. We also experimented with different category dimensions for our Flickr and Delicious datasets and found that overall the discovered categories are semantically meaningful.

The CCA model enables us to verify the presence of semantically meaningful categories that are also grouped together to form communities or themes. In addition it allows us to test our basic assumptions about the existence of differences between the process generating text documents and of that generating tags for existing documents, as we will see in the next chapter. The model, however, does not explicitly model the user which is an important component in the tag generation process. Next, we expand this simple model to include user activity explicitly.

4. Probabilistic Social Annotation (PSA) Model

In this section we propose a probabilistic generative model that extends the CCA model from the preceding section to include the individual users that participate in tagging an object along with the tags they used. By including the users in the modeling process, the resulting communities now have distributions over users that explicitly specify a user membership probabilistically in each community. Hence, we need to redefine the *social annotation community* from the previous section.

[Redefinition] Social Annotation Community: *A social annotation community c is composed of a probability distribution over users in U such that $\sum_{u \in U} p(u|c) = 1$, where $p(u|c)$ indicates membership strength for each user u in community c*

4.1. Generating Social Annotation with PSA

Rather than modeling the tag generation process as if tags are generated regardless of user, the Probabilistic Social Annotation (PSA) model combines the $\langle user, tag \rangle$ generation process; a natural consequence of such an approach is the discovery of user communities in addition to tag categories.

Formally, the PSA model assumes a corpus of D social annotation documents drawn from a vocabulary of V tags and U users, where each social annotation document S_i is of variable length N_i. The model assumes that the $\langle user, tag \rangle$ pairs in a social annotation document are generated from a mixture of L distinct communities, where each community is a mixture of users that view the world based on a set of K_l hidden categories, and where each category is a mixture of tags. Therefore, the tagging process involves two steps: 1) the selection of a community from which to draw users and 2) the selection of the categories that influence the user's view or preference over tags based on the object's content, and the tagger's perception of the content.

Let $\mathbf{S}_i, \mathbf{z}$, and $\mathbf{c}$ be vectors of length N_i representing $\langle user, tag \rangle$ pair, category, and community assignments, respectively, in a social annotation document. The PSA model generation process is illustrated in Figure (8) and described here:

1. for each community $c = 1, ..., L$
    - Select U dimensional $\tau_c \sim$ Dirichlet(δ)
    - for each category $z = 1, ..., K_c$
        - select V_c dimensional $\phi_z \sim$ Dirichlet(γ)
2. for each object $\mathbf{S}_i$, $i = 1, ..., D$
    - Select L dimensional $\kappa \sim$ Dirichlet(α)

- for each community $c = 1, ..., L$
  - select K_c dimensional $\theta_c \sim$ Dirichlet(β)
- For each position $S_{i,j}$, $j = 1, ..., N_i$
  - Select a community $\mathbf{c}_{i,j} \sim$ multinomial(κ_i)
  - Select a user $\mathbf{S}^u_{i,j} \sim$ multinomial($\tau_{\mathbf{c}_{i,j}}$)
  - Select a category $\mathbf{z}_{i,j} \sim$ multinomial($\theta_i^{\mathbf{c}_{i,j}}$)
  - Select a tag $\mathbf{S}^t_{i,j} \sim$ multinomial($\phi_{\mathbf{c}_{i,j}}^{\mathbf{z}_{i,j}}$)

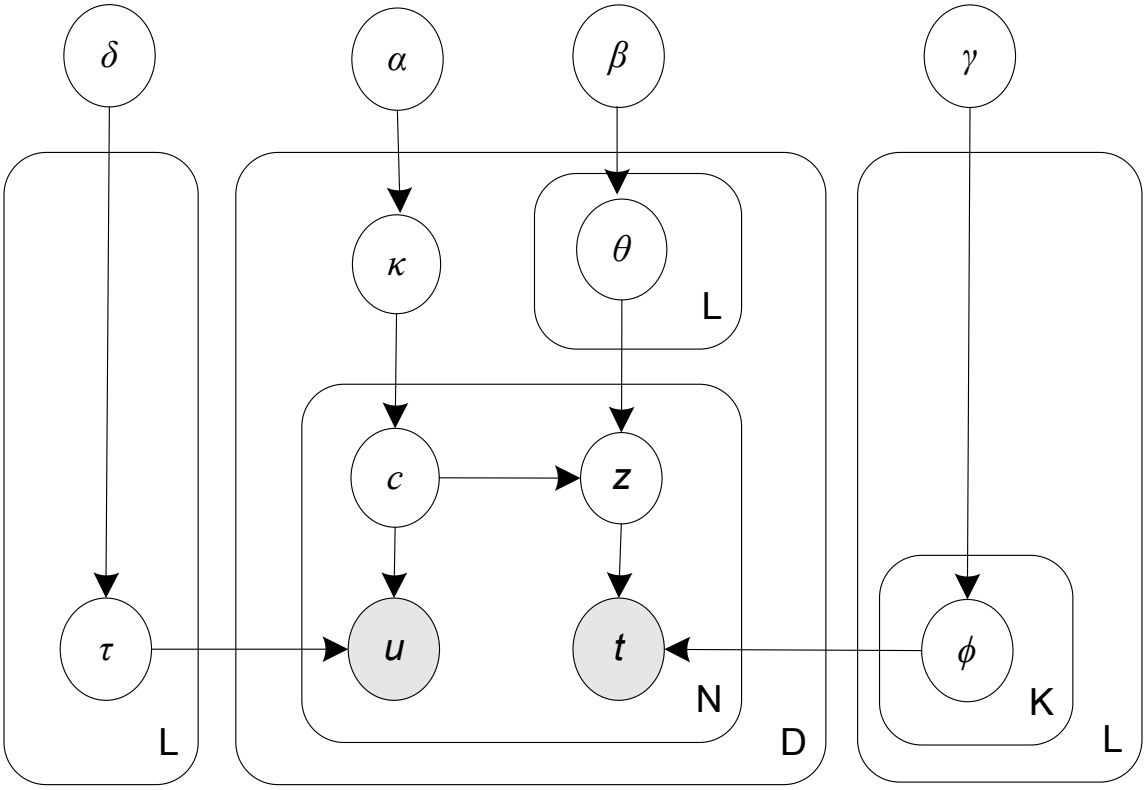


Fig. 8. Probabilistic Social Annotation model (PSA)

A social annotation document's community distribution $\kappa_i = \{\kappa_{i,j}\}_{j=1}^{L}$ (representing the communities interest in the object) is sampled from a Dirichlet distribution with parameter $\alpha = \{\alpha_i\}_{i=1}^{L}$. A per object community's category distribution

$\theta_i^c = \{\theta_{i,j}^c\}_{j=1}^K$ (representing the community interpretation of the object) is sampled from a Dirichlet distribution with parameter $\beta = \{\beta_i\}_{i=1}^K$. A category's tag distribution $\phi_z = \{\phi_{z,i}\}_{i=1}^{|V|}$ (representing a topic of interest) is sampled from a Dirichlet distribution with parameter $\gamma = \{\gamma_i\}_{i=1}^{|V|}$. Finally, A community's user distribution $\tau_c = \{\tau_{c,i}\}_{i=1}^{|U|}$ (representing a group of users with common interests) is sampled from a Dirichlet distribution with parameter $\delta = \{\delta_i\}_{i=1}^{|U|}$. The generative process creates a social annotation document by sampling for each position $\mathbf{S}_{i,j}$ a community $\mathbf{c}_{i,j}$ from a multinomial distribution with parameter κ_i, a category $\mathbf{z}_{i,j}$ from a multinomial distribution with parameter $\theta_i^{\mathbf{c}_{i,j}}$. A user is then sampled for that position form a multinomial distribution with parameter $\tau_{\mathbf{c}_{i,j}}$. Similarly a tag is sampled for that position from a multinomial distribution with parameter $\phi_{\mathbf{c}_{i,j}}^{\mathbf{z}_{i,j}}$.

Based on the model we can write the likelihood that a position $\mathbf{S}_{i,j}$ is assigned a specific $\langle user, tag \rangle$ pair $\{u, t\}$ as:

$$p(\mathbf{S}_{i,j} = \{u,t\}|\kappa_i, \Theta, \Phi, \tau) = \sum_{l=1}^{L} p(\mathbf{S}_{i,j}^u = u|\tau_l) p(\mathbf{c}_{i,j} = l|\kappa_i) \left(\sum_{k=1}^{K_l} p(\mathbf{S}_{i,j}^t = t|\phi_l^k) p(\mathbf{z}_{i,j} = k|\theta_i^l) \right)$$

As before, the likelihood of the complete social annotation document $\mathbf{S}_i$ is the joint distribution of all its variables (observed and hidden):

$$p(\mathbf{S}_i, \mathbf{z}_i, \mathbf{c}_i, \kappa_i, \Theta, \Phi, \tau|\alpha, \beta, \gamma, \delta) = \prod_{j=1}^{N_i} p(\mathbf{S}_{i,j}^t|\phi_{\mathbf{c}_{i,j}}^{\mathbf{z}_{i,j}}) p(\mathbf{S}_{i,j}^u|\tau_{\mathbf{c}_{i,j}}) p(\mathbf{z}_{i,j}|\theta_i^{\mathbf{c}_{i,j}}) p(\mathbf{c}_{i,j}|\kappa_i)$$

Integrating out the distributions κ_i, Θ, τ and Φ and summing over $\mathbf{c}_i$ and $\mathbf{z}_i$ gives the marginal distribution of $\mathbf{S}_i$ given the priors:

$$p(\mathbf{S}_i|\alpha,\beta,\gamma,\delta) = \int\int\int\int p(\kappa_i|\alpha)p(\Theta|\beta)p(\Phi|\gamma)p(\tau|\delta)$$
$$\prod_{j=1}^{N_i}\sum_{\mathbf{c}_{i,j}} p(\mathbf{S}^u_{i,j}|\tau_{\mathbf{c}_{i,j}})p(\mathbf{c}_{i,j}|\kappa_i)\left(\sum_{\mathbf{z}_{i,j}} p(\mathbf{S}^t_{i,j}|\phi^{\mathbf{z}_{i,j}}_{\mathbf{c}_{i,j}})p(\mathbf{z}_{i,j}|\theta_i^{\mathbf{c}_{i,j}})\right)$$
$$d\Phi d\Theta d\tau d\kappa_i$$

Finally the universe of discourse $\mathcal{U}$ consisting of all D social annotation documents occurs with likelihood:

$$p(\mathcal{U}|\alpha,\beta,\gamma,\delta) = \prod_{i=1}^{D} p(\mathbf{S}_i|\alpha,\beta,\gamma,\delta)$$

4.2. Parameter Estimation and Inference

Similar to CCA, the PSA model provides the generative approach for producing social annotation documents. The goal is to recover the structures that produced these social annotation document – taking a set of social annotation documents and inferring the underlying model (including the hidden community and category distributions). This entails learning model parameters κ, τ, Θ, and Φ (the distributions over communities, users, categories, and tags, respectively).

To learn model parameters we follow the same approach used in LDA to approximate the posterior distributions by Gibbs sampling. Let $\mathbf{S}, \mathbf{z}$, and $\mathbf{c}$ be vectors of length $\sum_i^D N_i$ representing $\langle user, tag\rangle$ pair, category, and community assignments, respectively, for the entire corpus. Also let u and t be user and tag variables. Following the approach used in [66] the joint probability distribution of the PSA model can be factored as:

$$p(\mathbf{S}^u, \mathbf{S}^t, \mathbf{z}, \mathbf{c}|\alpha, \beta, \gamma, \delta) = p(\mathbf{S}^u|\mathbf{c}, \delta)p(\mathbf{c}|\alpha)p(\mathbf{S}^t|\mathbf{z}, \mathbf{c}, \gamma) \\ p(\mathbf{z}|\mathbf{c}, \beta).$$

We derive the Gibbs sampler's update equation (See Appendix B for details) for the hidden variables (community, and category) from the joint distribution and arrive at:

$$p(\mathbf{z}_i = k, \mathbf{c}_i = l|\mathbf{z}_{\neg i}, \mathbf{c}_{\neg i}, \mathbf{S}^t, \mathbf{S}^u) \propto \\ \frac{n^u_{l,\neg i} + \delta_u}{\sum_{u=1}^{U} n^u_{l,\neg i} + \delta_u} \times \frac{n^t_{lk,\neg i} + \gamma_t}{\sum_{t=1}^{V} n^t_{lk,\neg i} + \gamma_t} \\ \times \frac{n^{lk}_{S,\neg i} + \beta_{lk}}{\left(\sum_{k=1}^{K_l} n^{lk}_{S} + \beta_{lk}\right) - 1} \times \frac{n^{l}_{S,\neg i} + \alpha_{l}}{\left(\sum_{l=1}^{L} n^{l}_{S} + \alpha_{l}\right) - 1} \tag{3.1}$$

where $n^{(\cdot)}_{(\cdot),\neg i}$ is a count excluding the current position assignments of $\mathbf{z}_i$ and $\mathbf{c}_i$ (e.g., $n^t_{lk,\neg i}$ is the count of tag t generated by the k-th category of the l-th community excluding the current position).

For the purpose of inference of new unseen web objects based on a model $\mathcal{M}$, the update equation for the Gibbs sampler is the following:

$$p(\tilde{\mathbf{z}}_i = k, \tilde{\mathbf{c}}_i = l|\tilde{\mathbf{c}}_{\neg i}, \tilde{\mathbf{z}}_{\neg i}, \tilde{S}^t, \tilde{S}^u, \mathcal{M}) \propto \\ \frac{\tilde{n}^u_{l,\neg i} + n^u_l + \delta_u}{\sum_{u=1}^{U} \tilde{n}^u_{l,\neg i} + n^u_l + \delta_u} \times \frac{\tilde{n}^t_{lk,\neg i} + n^t_{lk} + \gamma_t}{\sum_{t=1}^{V} \tilde{n}^t_{lk,\neg i} + n^t_{lk} + \gamma_t} \\ \times \frac{n^{lk}_{\tilde{S},\neg i} + \beta_{lk}}{\left(\sum_{k=1}^{K_l} n^{lk}_{\tilde{S}} + \beta_{lk}\right) - 1} \times \frac{n^{l}_{\tilde{S},\neg i} + \alpha_{l}}{\left(\sum_{l=1}^{L} n^{l}_{\tilde{S}} + \alpha_{l}\right) - 1} \tag{3.2}$$

where where $n^{(\cdot)}_{(\cdot),\neg i}$ are counts from the given model $\mathcal{M}$, $\tilde{n}^{(\cdot)}_{(\cdot),\neg i}$ are counts from the new objects, and $\tilde{S}$ is a new unseen object.

4.3. PSA: Simplified Version

Alternatively, we can simplify the Probabilistic Social Annotation model by assuming that each community of users agrees on a single category view of the world. We can then combine the *community* variable c and *category* variable z into a single hidden variable. As a result, the model finds a distribution over tags as well as a distribution over users, capturing both the users' similarities/interests and their world view simultaneously. The generation process is illustrated in Figure (9) and works as follows:

1. for each community $c = 1, ..., L$
    - Select U dimensional $\tau_c \sim \text{Dirichlet}(\alpha)$
    - select V dimensional $\phi_c \sim \text{Dirichlet}(\gamma)$
2. for each object $\mathbf{S}_i$, $i = 1, ..., D$
    - Select L dimensional $\theta_i \sim \text{Dirichlet}(\beta)$
    - For each position $\mathbf{S}_{i,j}$, $j = 1, ..., N_i$
        - Select a community $\mathbf{c}_{i,j} \sim \text{multinomial}(\theta_i)$
        - Select a user $\mathbf{S}^u_{i,j} \sim \text{multinomial}(\tau_{\mathbf{c}_{i,j}})$
        - Select a tag $\mathbf{S}^t_{i,j} \sim \text{multinomial}(\phi_{\mathbf{c}_{i,j}})$

An advantage of simplifying the model is a lower computational complexity. A drawback, however, is you restrict members of a community to a single world view; which slightly reduces model generalization to unseen data as seen in our results.

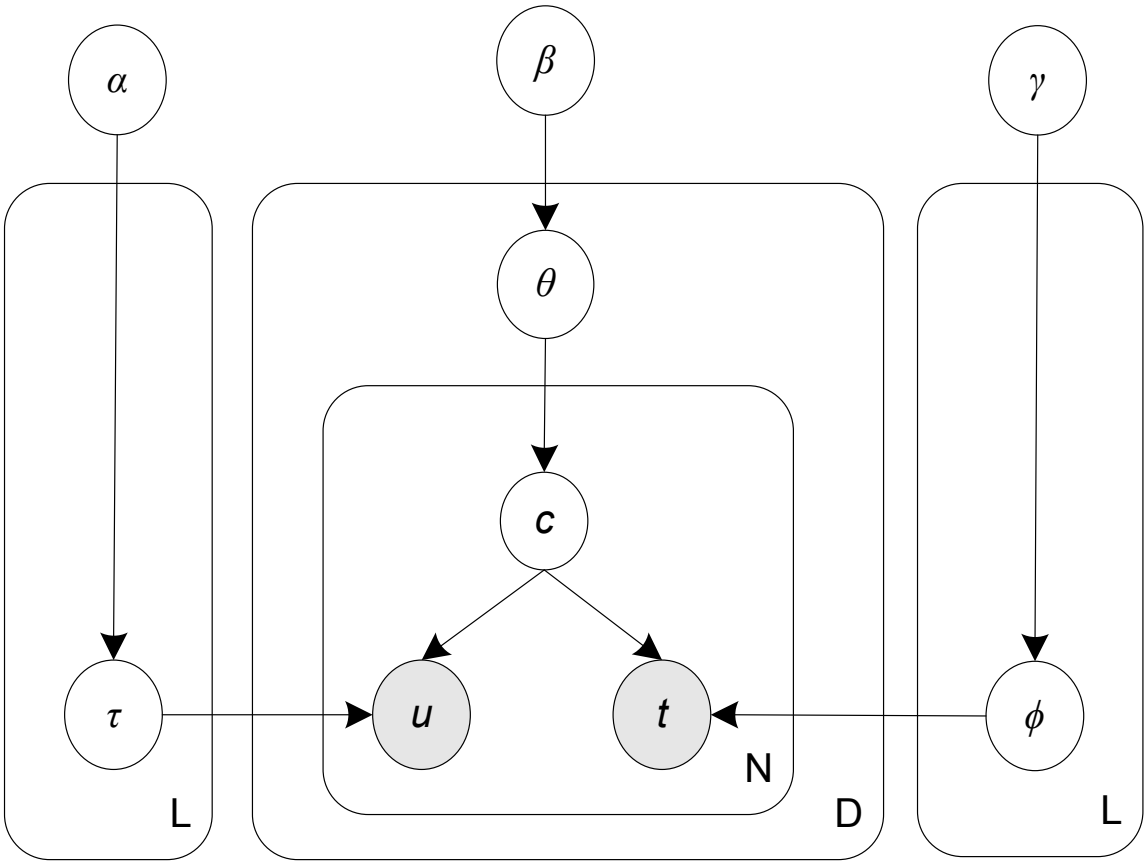


Fig. 9. Simplified Probabilistic Social Annotation model (simplePSA)

The update equation for the Gibbs sampler (3.1) reduces to:

$$p(\mathbf{c}_i = k|\mathbf{c}_{\neg i}, \mathbf{S}^t, \mathbf{S}^t) \propto \tag{3.3}$$
$$\frac{n^u_{k,\neg i} + \alpha_u}{\sum_{u=1}^{U} n^u_{k,\neg i} + \alpha_u} \times \frac{n^t_{k,\neg i} + \gamma_t}{\sum_{t=1}^{V} n^t_{k,\neg i} + \gamma_t}$$
$$\times \frac{n^k_{S,\neg i} + \beta_k}{\left(\sum_{k=1}^{K} n^k_S + \beta_k\right) - 1}$$

and the Gibbs sampler predictive update equation (5.1) becomes:

$$p(\tilde{\mathbf{c}}_i = k|\tilde{\mathbf{c}}_{\neg i}, \tilde{S}^t, \tilde{S}^u, \mathcal{M}) \propto \frac{\tilde{n}^u_{k,\neg i} + n^u_k + \alpha_u}{\sum_{u=1}^{U} \tilde{n}^u_{k,\neg i} + n^u_k + \alpha_u} \times \frac{\tilde{n}^t_{k,\neg i} + n^t_k + \gamma_t}{\sum_{t=1}^{V} \tilde{n}^t_{k,\neg i} + n^t_k + \gamma_t} \times \frac{n^k_{\tilde{S},\neg i} + \beta_k}{\left(\sum_{k=1}^{K} n^k_{\tilde{S}} + \beta_k\right) - 1} \tag{3.4}$$

4.4. Applying PSA to CiteULike and Delicious

In this section we evaluate the quality of the PSA models over two prominent social bookmarking services: Delicious and CiteULike. We use the Delicious dataset previously described and the following CiteULike dataset:

CiteULike dataset: We collected from CiteULike a dataset of 80,833 resources, with 50,491 unique tags and 12,496 unique users. We normalize the data and train the different models with 79,000 objects and use the rest for testing.

Our goal is to evaluate how well the model predicts previously unseen data and the quality of the discovered latent structures.

We compared PSA and the simplified version of PSA (simplePSA) against TagLDA and a tag-based version of Pachinko Allocation (PAM) [83]. PAM is an LDA extension that aims to capture correlations among latent topics using a directed acyclic graph. We focus here on the four level PAM where the internal nodes in the tree represent supertopic/subtopic distributions and leaf nodes are distributions over the vocabulary space. For our purposes, we refer to the super-topics as communities, subtopics as categories, and our documents are collections of tags assigned by various users.

We use the public implementations of LDA and PAM distributed in the Mallet toolkit [84]. Hyperparameters for both models are set to toolkit standard: for

LDA ($\alpha = 50/K, \beta = 0.01$) and for PAM ($\alpha = 50/K, \beta = 0.001$) with optimization enabled for both models. For the PSA models we experimented with several combinations of hyperparameters. We also estimate hyperparameters using the fixed-point iteration method in [85]. The results we compare with other models are run with hyperparameters ($\alpha = 0.1, \beta = 1, \gamma = 0.1, \delta = 0.1$) and optimization enabled.

For all models, a Gibbs sampler starts with randomly assigned communities/categories, runs for 2000 iterations with optimization every 50 iterations and an initial burnin-period of 250 iterations. For TagPAM we experiment with three communities sizes $(8, 12, 16)$ each with 20 to 100 categories. For PSA we experiment with community category combinations that result in total number of categories from 20 to 100. simplePSA and TagLDA – which have no community/category hierarchy – are run from 20 to 100 categories (although recall that simplePSA models user communities and tag categories simultaneously)

4.5. Evaluation

We compare all the models using two metrics: 1) ability to predict previously unseen data and 2) quality of discovered latent structures.

We evaluate each model's generalization to unseen data using the empirical likelihood method [83]. To compute empirical likelihood we generate 1000 documents based on the models generative process. We then build a multinomial over the vocabulary space from these samples. Finally, we compute the empirical likelihood of a held out testing set using the obtained multinomial over the vocabulary space.

For quality evaluation we solicit judgment on the coherence of discovered categories. Categories from each model were anonymized and put in random order. Each evaluator is asked to judge the category coherence by trying to detect a theme from the category's top 10 terms. Coherence is graded on a $0 - 3$ scale with 0 being poor

coherence and 3 excellent coherence. The evaluators are also asked to report the number of terms that deviate form the theme they thought the category represented.

We use testing sets of size 1%, and 10% of the size of the training set for testing all models. The empirical likelihood results are consistent for the two sets, therefore we report results from the smaller set.

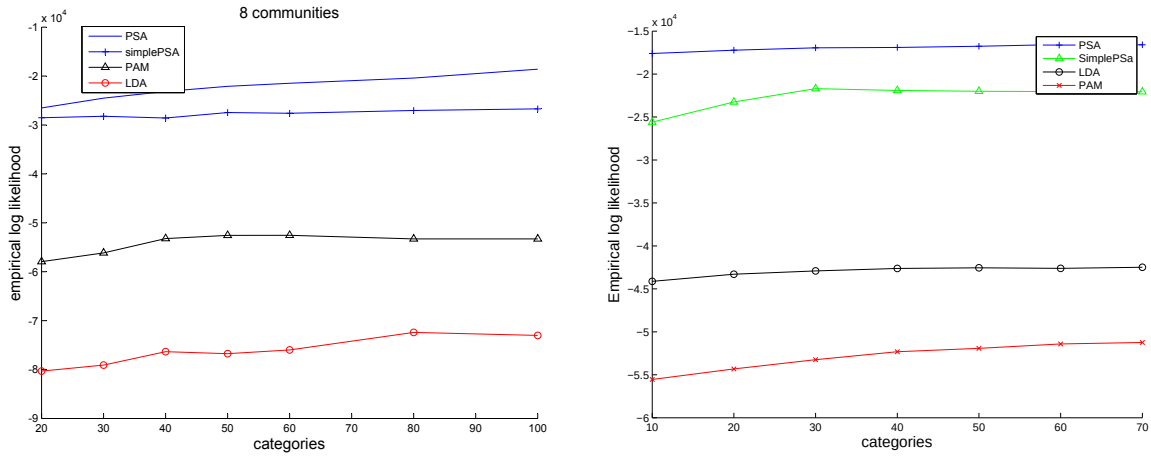
We plot the empirical likelihood results in Figures (10,11, and 12). The y-axes show the empirical log likelihood and the x-axes show the number of categories. Focusing on Figure (10), the PSA model performs the best, followed by simplePSA, TagPAM and TagLDA respectively. PSA and simplePSA performance improves with increasing number of categories, with PSA spiking at 35 categories then slowly continuing to improve. simplePSA behaves similarly with its initial spike at 50 categories. TagPAM's performance improves initially, peaks around $40 - 50$ categories, then decreases slightly and stabilizes. Likewise, the performance of TagLDA peaks around 40 categories, decreases slightly, then peaks again at 80 categories. Figures (11, and 12) show similar results with improved performance for TagPAM when number of communities is increased. Still the PSA model performs better than TagPAM.

Based on the above results, 40 categories lead to good performance in all models. For our user study we present the discovered 40 categories from each model for coherence evaluations. A sample of these categories is shown in Table (V).

A group of four evaluators judged the categories' coherence and noted the deviating terms. The evaluation results are as shown in Table (VI). Evaluating coherence, we can see from the table that on average, PSA and simplePSA perform the best followed by TagPAM and TagLDA respectively. PSA shows on average a 6% improvement over TagLDA and a 7% improvement over TagPAM. We also look at the number of deviating terms from the perceived theme and observe similar improvements.

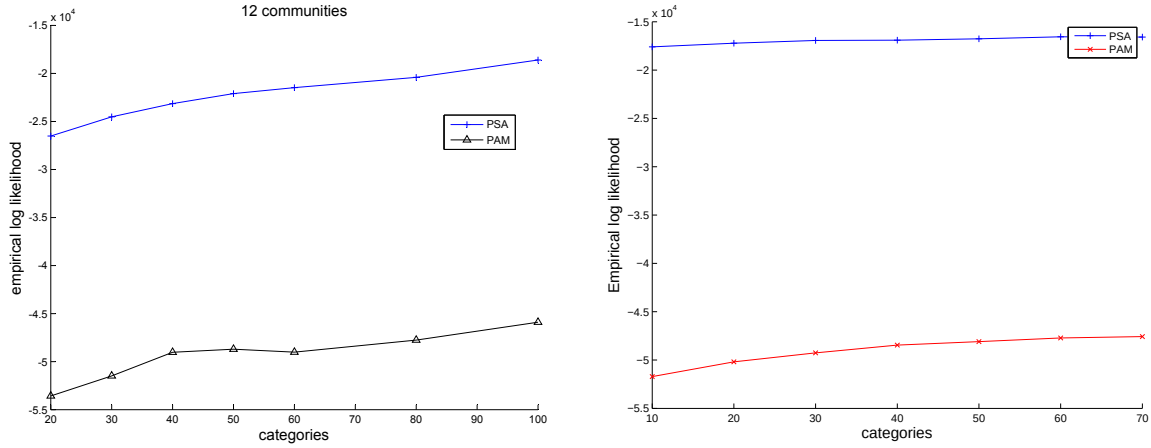
Table V. Sample categories uncovered by the different models

LDA	airlin flight seat hotel businesscard airfar airplan card vacat ticket firefox scalabl cluster amazon scale jobsearch ec blueprint tumblr mapreduc tumblelog screensav fileshar religion evolut bibl opensoci restaur christian foodblog lego steampunk vista
PAM	airlin flight seat hotel webcom airfar airplan vacat ticket xkcd cheap scalabl cluster speed scale deploy number shoe tune concurr mapreduc bandwidth religion cloud evolut bibl tagcloud oreilli christian comedi flex flashcard ibm
PSA	airlin flight seat hotel airfar airplan vacat ticket cheap trip holidai django cach scalabl cluster eclips amazon scale j2ee memcach trac s3 religion recycl bibl eco evolut christian lego consumer knot ecolog garden
simplePSA	airlin flight seat airfar airplan deal ticket coupon bargain cheap hotel authent scalabl cluster openid rest amazon webhost scale opensoci s3 ec2 religion evolut bibl christian sga lego mckaysheppard genet atheism dna church



(a) Delicious (b) CiteULike

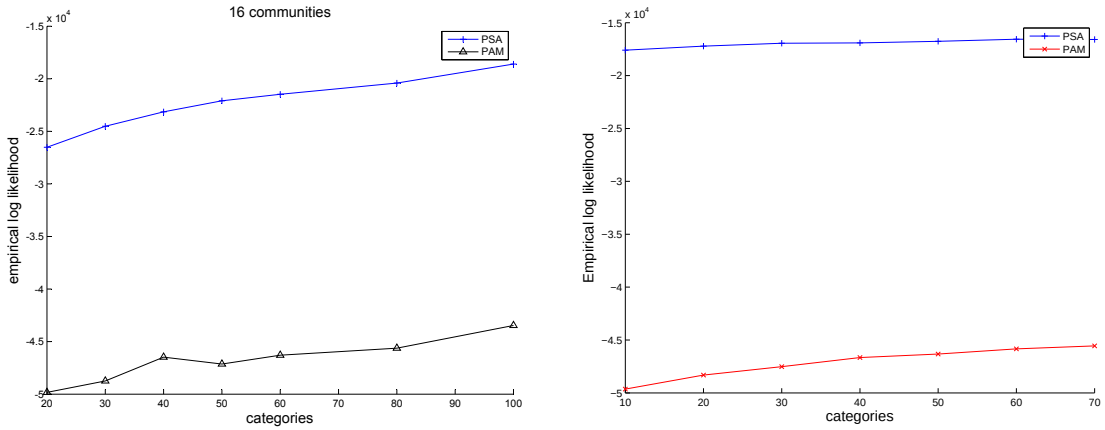
Fig. 10. Empirical likelihood results for 8 communities



(a) Delicious (b) CiteULike

Fig. 11. Empirical likelihood results for 12 communities

In Figures (13, and 14) we report the detailed evaluation scores and deviating terms, respectively, for all categories from all four models. Figure (13) shows the number of categories from each model and the coherence scores they received. Notice that PSA and simplePSA have higher number of categories receiving a score of 2 or higher compared to TagPAM and TagLDA . We can also see that PSA and simplePSA



(a) Delicious (b) CiteULike

Fig. 12. Empirical likelihood results for 16 communities

have lower number of categories receiving a score of 1 or lower compared to TagPAM and TagLDA. Figure (14) shows the number of categories from each model versus the number of deviating terms. Again PSA and simplePSA have a higher number of categories containing small number of deviating terms compared to TagPAM and TagLDA and they have a lower number of categories containing large number of deviating terms compared to TagPAM and TagLDA. We present example communities, their top tags, and their top users discovered by our PSA model for both Delicious and CiteULike in Tables (VII) and (VIII), respectively.

Average	LDA	PAM	PSA	simplePSA
score	1.51	1.49	1.60	1.59
number of deviating terms	5.72	5.7	5.34	5.35

Table VI. Coherence evaluation results

4.6. The Role of Users

The improvements achieved by or models are due primarily to the inclusion of the user as a generated variable. Smaller improvement comes from the hierarchical structure of

Table VII. Delicious: communities, their top tags and their top users

community	top tags	top users and their top tags
Comm 0	dictionari translat encyclopedia slang thesauru grammar spanish french acronym linguist etymolog answer diccionario vocabulari ital	user 11985: web,freestylelanguag, video, onlin, italian, german, dictionari, translat... user 73941: word, cultur, resourc, languag, internet, vocabulari, buzzword, lexicon, speech, idiom... user 24880: photographi, fileshar, folksonomi, english, copyright,resourc, document, thesauru, data... user 8846: gener, usabl, english, audio, articl, publish, write, literaci, dictionari, student, resourc...
Comm 1	mindmap diagram brainstorm visio oreilli shoe whiteboard anatomi flowchart mind uml bodi brain conceptmap wirefram meet graphicorga	user 8226: podcast,visual,audio,music,draw, brainstorm,folksonomi,talotool, creativ, todo, share, social... user 36411: document, new, innov, photo, creativ, idea, visual, health, mindmap, graph, list, usabl, draw... user 30704:mindmap, grid, video, imag, art, technolog, anim, elearn, portfolio, framework, chart, design... user 12696: photo, wiki, audio, mindmap, collabor, brainstorm, sound, imag, graphic, art, free, elearn...
Comm 2	dn bank financi credit bill budget auction ebai date domain loan lend invoic palett weather calcul trade currenc clock frugal	user 7587: innov,energi, infrastructur, usa, blog, nielsen, communicat, media, strategi, mackai, democraci... user 18375:program, articl, cool, bank, howto, product, onlin, daili,blog, bill ,colour, financ, monei... user 65615: develop, softwar, program, job, search, callcent, book, servic, commun, secur, monei, career ... user 2165: interest, program, list, websit, review, financ, altern, lifestream, admin, access, job, custom...
Comm 3	distro vmware debian tweak emul recoveri driver vm vista laptop kernel wine registri livecd gentoo boot sandbox usb thunderbird uninstal	user 55702:refer, freeservic, search, orpdw, installed, educ, applic, secur, sourceforg, operatingsystem... user 4959: tip, develop, virtual, architectur, hack, share, technolog, host, widget, system, cluster, monitor ... user 24559:lifehack, todo, develop, technolog, applic, tutori, util, network, audio, ubuntu, maco, resourc... user 23732: secur, extens, tutori, howto, educ, develop, linux, ubuntu, hardwar, backup, design, new...
Comm 4	airlin flight seat airfar airplan deal ticket coupon bargain cheap hotel question vacat trip aviat discount sport fly metafilt price	user 2147: busi, network, televis, book, guid, directori, mashup, rate, locat, mobil, flight, answer... user 59837: onlin, web, homeschool, recip, game, teach, travel, food, hotel, photographi, blog, internet ... user 65681:travel, airlin, blog, transport, flight, ticket, airfar, vacat, refer, map, deal, airplan, airport... user 5187: map, airlin, blog, flight, refer, web20, aggreg, tool, resourc, internet, social, daili, commun...

Table VIII. CiteULike: communities, their top tags and their top users

community	top tags	top users and their top tags
Comm 0	mt human model dna cognit represent	user 3261: represent, concept, imageri, english, model, memori, languag, comprehens...
	circuit semant state chart protein	use 1323: mt,autosom, africa,african,admixtur,european, asia...
	concept power forens knowledg low	user 5876: circuit, power,literatur,low, electron,integr,cmo,digit,logic,chip...
	imageri memori literatur system	user 5944: protein,us,non,support,govt,research, sequenc,molecular,model,structur...
Comm 1	semant web ontolog inform	user 109:visual,inform,collabor,semant,tag,data,retriev, web,ontolog,mine...
	visual semanticweb servic	user 1122: toread,ui,web,thesi,distribut,sa,self,relat, architectur...
	mine data knowledg retriev rdf	user 272:digit,librari,metadata,collabor,semant,web,user, name,cluster,ontolog...
	webservic search model manag	user 3061:im,abc,diplomarbeit,learn,open,sourc,bildung, postmodern,commun...
	extract xml databas autonom	user 262: semant,commun,social,ontolog,web,cop,km,folksonomi, network,blog...
Comm 2	ag review signal	user 146: ag,stress,oxid,signal,gene,redox,ro,mitochondria, regul,antioxid...
	stress oxid cancer	user 1740:dynam,energi,molecular,simul, basi,calcul,state, function,structur,potenti...
	cell protein fret dna	user 2693: chromatin,spermatozoa,dna,review,nucleu,nuclear,rat, structur,replic,lipid...
	gene 5 dynam camp arabidopsi	user 3628: fret,camp,signal,second,messeng,fluoresc,sensor, analysi,diffus...
	apoptosi chromatin flower ro regul	user: 4152: caveolin,shear,endotheli,apoptosi,integrin, cholesterol,akt,rac,mechan...
Comm 3	robot numer method	user: 5723: us,non,ph,support,genet,govt,anim,physiolog,gene...
	surfac null	user 3518: network,local,robot,sensor,system,mobil,wireless,estim, algorithm,distribut...
	system network us non local	user: 5931:numer,method,surfac,null,problem,initi,valu, black,hole,gravit...
	cognit problem genet ph adapt	user 4810: cognit,embodi,robot,agent,artifici,and,intellig,situat,system,life...
	valu model sensor and initi	user 574: poverti,sozial,inequ,und,wirtschaft,incom,measur,data,imput,panel ...
Comm 4	program type haskel	user 778: model,simul,lisp,dsl,program,oop,gi,dss,logic,water ...
	theori function languag	user 760: type,extens,pars,syntax,continu,linear,dynam,control,macro,program ...
	monad compil logic concurr	user 869: program,function,type,haskel,semant,monad,lazi,call,by,evalu...
	analysi model calculu scheme semant	user 5187: theori,program,type,theorem,prove,logic,synthesi,calculu,proof,lambda...
	lambda dsl system virtual categori	user 808: sla,call,asr,pedagogi,model,pronunci,languag,linguist,good,nlp...

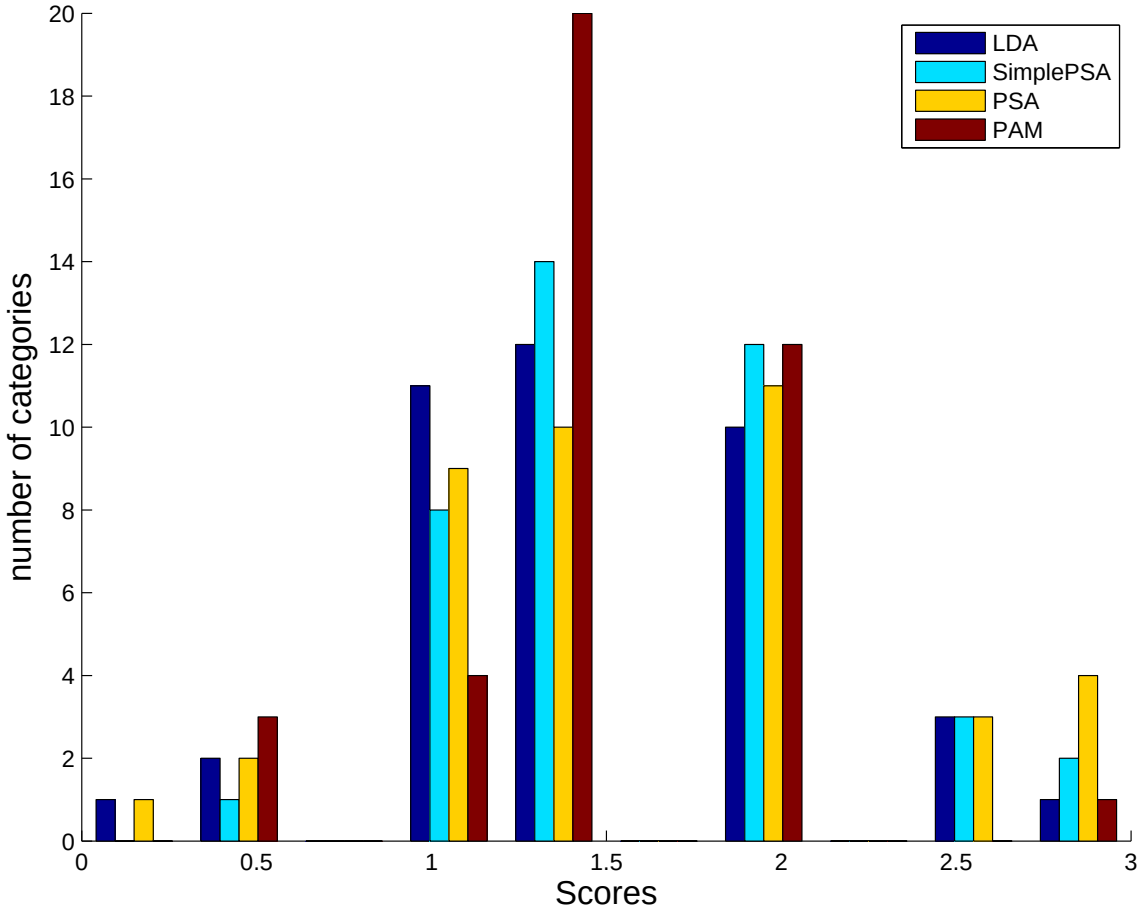


Fig. 13. User study results: shows the count of categories from each model and their respective score (0 to 3), with 0 representing no coherence and 3 representing excellent coherence

communities and categories that we introduce. This is clearly evident in Figure (10). Notice the performance of PSA compared to that of simplePSA. In this section, we show how the introduction of the user as a generated variable in the social annotation process impacts the Gibbs sampler.

As an example, suppose we have a corpus with a tag vocabulary of length 3, $V = \{w_1, w_2, w_3\}$, two users, $U = \{u_1, u_2\}$, and two communities $L = \{C_1, C_2\}$. Also

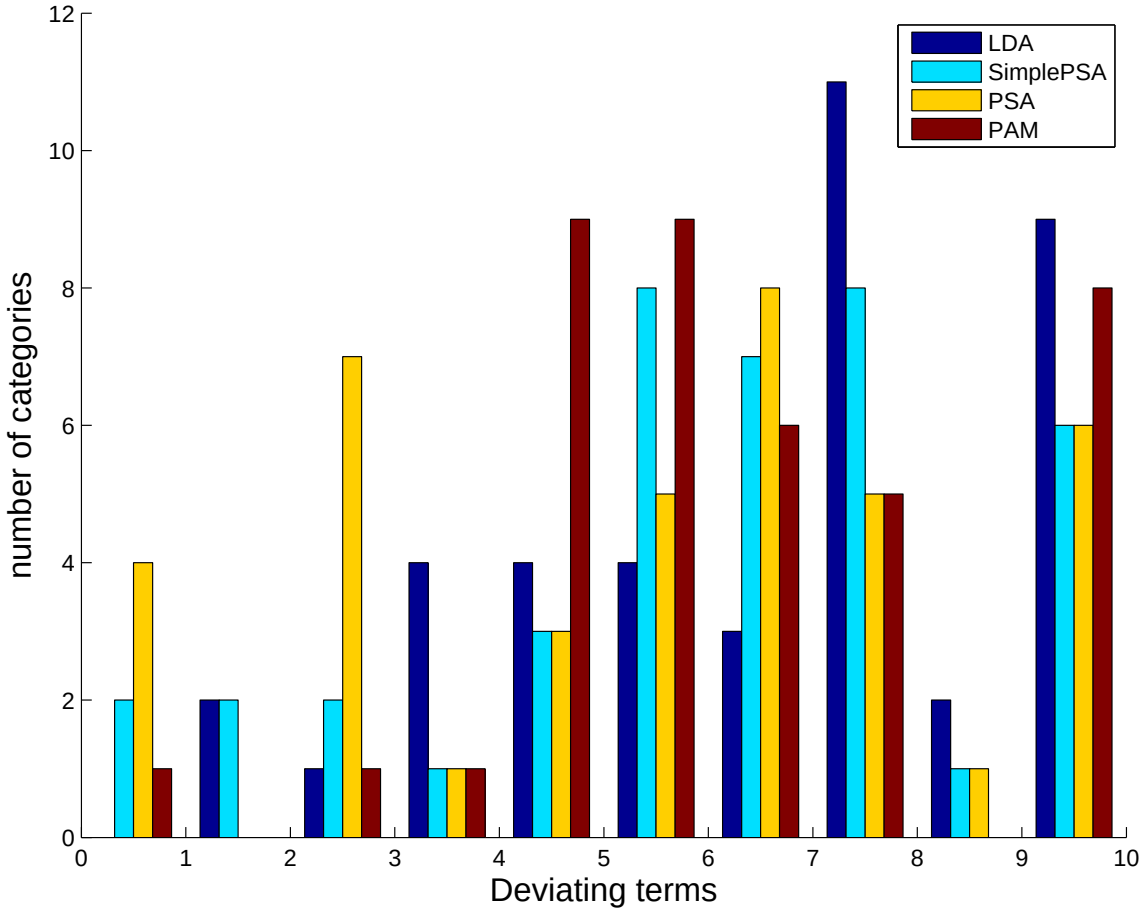


Fig. 14. User study results: shows the count of categories from each model and their respective number of deviating terms

suppose the corpus contains a single document of length 6:

$$\mathbf{S} =< u_1, w_1; u_1, w_2; u_1, w_2; u_1, w_1; u_2, w_2; u_2, w_3 > .$$

Assume the priors to be uniform over tags, users, and communities. Table (IX) compares the impact of excluding/including users on the Gibbs sampler. Suppose the Gibbs sampler has completed $n-1$ iterations and the resulting community assignments are as shown in the 5_{th} row of the table. To illustrate this difference, we use Equation (3.3) to compute the probability of community assignment for the word at

position 1 of the corpus. Remember that our Gibbs sampler excludes the current position, so at the beginning of iteration n two words of the corpus belong to community C_1 and the other three words belong to community C_2.

When calculating the update probabilities in the case that excludes users we ignore the first factor of Equation (3.3). Notice that w_1 which occurs at position 4 is assigned to community C_1. The Gibbs sampler gives almost equal chance for this word to belong to either community. This is because of two competing factors: (i) the majority of the words in this document already belong to community C_2, so one factor favors C_2; (ii) at the same time the word w_1 has already been assigned once to community C_1 which balances both outcomes.

Now let us consider the case in which users are included. Here the Gibbs sampler clearly prefers community C_1 over C_2 for this position. The reason being that the user associated with the word at position 1 which also happen to be the user most interested in this document had already been assigned twice to C_1. We can point to at least three advantages of including users in social annotation modeling 1) faster convergence; users co-occurrences and associations with tags resolve ties leading to faster consensus 2) better results in terms of quality of categories as shown in our user study 3) additional clustering of users that can be useful in numerous application.

4.7. Summary

Rather than modeling the tag generation process as if tags are generated regardless of user, the Probabilistic Social Annotation (PSA) model treats both tags and users as generated variables that are drawn in pairs; a natural consequence of such an approach is the discovery of user communities in addition to tag categories. Similar to the CCA model, the PSA model assumes a corpus of D social annotation documents drawn from a vocabulary of V tags with the addition of a vocabulary of U users. The model

assumes that the $\langle user, tag \rangle$ pairs in a social annotation document are generated from a mixture of L distinct communities, where a community now is a mixture of users that view the world based on a set of K_l hidden categories, and where each category is still a mixture of tags. Thus, the tagging process involves the selection of a community from which to draw users as well as their tagging categories and a second selection step among these categories for the one that will influence the user's view or preference over tags based on the object's content, and the tagger's perception of the content.

Alternatively, we can simplify the Probabilistic Social Annotation model by assuming that each community of users agrees on a single category view of the world. We can then combine the *community* variable and *category* variable into a single hidden variable. As a result, the model finds a distribution over tags as well as a distribution over users, capturing both the users' similarities/interests and their world view simultaneously.

Again, we employ MCMC methods to uncover the underlying structure (community and category distributions). We experiment with several community category combinations on our CiteULike and Delicious datasets. We also compare the results of our models to results obtained using existing traditional topic models that cannot model the user role in social bookmarking (LDA and PAM [62, 83]).

Our experiments compare the above models using two metrics: (i) ability to predict previously unseen data; and (ii) quality of discovered latent structures. Our experimental results show that our models do generalize better to unseen social annotation documents as well as improve the quality of latent structures discovered. The improvements achieved by or models are due primarily to the inclusion of the user as a generated variable.

5. Computational Complexity

As we have seen above, learning model parameters via Gibbs sampling involves iterations over the entire corpus that sample conditional probabilities for communities and categories at each position. let N be number of iterations, L be number of communities, K be number of categories, D be number of documents, and S be average document length. The computational complexity of CCA and PSA using Gibbs sampling is $\mathcal{O}(NLKDS)$. That is a factor of L higher than LDA and simplePSA. The complexity of PAM is $\mathcal{O}(N(L+K+1)DS)$.

6. Summary

Understanding the social annotation process is essential to modeling the collective semantics centered around large-scale social annotations; which is the first step towards potential improvements in information discovery and knowledge sharing. In this chapter, we have introduced two novel probabilistic generative models of the social annotation process, emphasizing the user/community role as a major actor in this domain. We have compared our models to two prominent topic models (PAM and LDA). Our experimental results show improvements in models generalizing to unseen social annotation documents as well as improvements in the quality of latent structures they discovered. Next we consider applications based on the results of the social annotation models we have introduced here.

Table IX. Gibbs sampling example with and without users

	Without Users							With Users					
corpus	w_1	w_2	w_2	w_1	w_2	w_3		u_1, w_1	u_1, w_2	u_1, w_2	u_1, w_1	u_2, w_2	u_2, w_3
Sampling	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Iter: $n-1$	C_1	C_1	C_2	C_1	C_2	C_2		C_1	C_1	C_2	C_1	C_2	C_2
Iter: n	$P(c_1 = C_1) = 0.48$ $P(c_1 = C_2) = 0.52$	C_1	C_2	C_1	C_2	C_2		$P(c_1 = C_1) = 0.83$ $P(c_1 = C_2) = 0.17$	C_1	C_2	C_1	C_2	C_2

CHAPTER IV

COMMUNITY-DRIVEN BROWSING AND SEARCH

1. Introduction

With the proliferation of web based social systems has come a commensurate interest in leveraging this wealth of collaborative community-based information for improving information access and discovery. For example, there has been growing excitement at augmenting traditional web search and browsing through the incorporation of tag information from social bookmarking services, e.g., [6], [8], [86],[7], [12], [41],[87].

In this chapter we present two frameworks that employ the results of our proposed models from the previous chapter to enhance traditional methods of information discovery and retrieval of socially tagged web objects. The proposed models uncover a structure underlying the social bookmarking process that probabilistically relates objects, tags and users to a community and category hierarchy. In this structure we can determine related objects based on their community and category distributions, related users based on their community memberships, and related tags based on their category or community memberships. We can also determine related communities and categories based on their user and tag distributions as well as object memberships.

Our first framework is based on the Community-based Categorical Annotation (CCA) model [80]. It exploits the apparent differences between the content of the web object and the tags assigned to it (the taggers perspective of the object). We devise similarity measures and explore an approach that utilizes these differences to browse for similar/dissimilar objects in a multi-dimensional space. We term this the Topics-Category Browsing Framework.

The second framework is based on the Probabilistic Social Annotation (PSA) model [81]. We develop a novel community-based ranking model for effective community-based exploration of socially-tagged web resources. We compare this community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We term this the Community-based Exploration Framework [88].

2. Topics-Category Browsing Framework

One important aspect of social bookmarking is the relationship between an object's content and its social annotation. Previous efforts have unified these two views to generate both the content and the annotations through a single process. Their intuition is that the author of a document and the social annotators of a document are driven by the same motivations. Indeed, there is evidence that many tags applied to a web page can also be found in the text of the page [86]. This leads one to believe that documents with similar content will have similar tags and vice versa. Our first task is to examine this hypothesis, i.e., do web objects with similar content have similar tags and vice versa?

Once we have shown that objects with similar content do not necessarily have similar tags and vice versa, we explore how this disparity can be used to better browse socially tagged web objects. When there are differences between a web document content and its tags, it is possible to improve information discovery by considering multidimensional similarity measures. That is instead of computing similarity between web objects using content alone or a combined space of content and tags, two separate spaces can be used. We illustrate this approach for integrating the annotation-based model with content-based approach for web object browsing.

The idea is to represent the objects content and annotations in reduced dimensional spaces of topics and categories. We then explore objects based on both their categorical and topical similarity (and dissimilarity) to a candidate query object. Given this query object, We measure its similarity to all other objects which are then classified into three views: (i) objects that are similar in topic space and similar in category space; (ii) objects that are similar in topic space, but dissimilar in category space and; (iii) objects that are dissimilar in topic space, but similar in category space. The user is then presented with the top ranked documents in each view and can select an object of interest to view or to further use it to browse for similar and dissimilar objects.

2.1. Categories vs. Content-based Topics

Now that we have seen in the preceding chapter how the CCA model can identify hidden categories and communities of interest that are used to drive the social annotation process, we revisit the relationship between an object's content and its social annotation document (recall $O_i = \langle C_i, S_i \rangle$). Previous efforts have unified these two views to generate both the content and the annotations through a single process (e.g., [12, 29]). Such a unified view, however, would seem to be meaningful for annotated objects that are primarily text (like web pages). It is less clear how to unify the content and annotation generation process for non-textual objects like images and videos. To examine whether this unified document content and social annotation model is even reasonable for primarily text-based web pages, we assume the processes generating both are separate.

We assume web object content to be generated from a latent structure of *topics* and its annotations are assumed to be, separately, generated from a latent structure of *categories*. Now we use our CCA model to uncover these underlying categories, and

use the LDA model to uncover the underlying topics. We are interested to understand if the underlying topic modeling approach for generating a document is the same as the categorical modeling approach for generating a social annotation document.

To measure the similarity of the content and annotation generation processes, we compare pairs of topics and categories. If the two processes are similar, we would expect to see many similar topic/category pairs. The similarity of a category topic pair can be measured using the Jensen-Shannon distance [89] for comparing two probability distributions p and q over an event space X:

$$JS(p,q) = \frac{KL(p,m) + KL(q,m)}{2},$$

where $m = \frac{p+q}{2}$, and $KL(p,q)$ is the Kullback-leibler divergence defined as:

$$KL(p,q) = \sum_{x \in X} p(x) \cdot \log \left(\frac{p(x)}{q(x)} \right).$$

Computing the JS-distance between a topic and category requires that we represent each topic or category z by a probability vector ϕ_z over the union of the tag vocabulary space and the content vocabulary space.

2.2. Categories and Topics on Delicious

We evaluate the content and annotation disparity on the Delicious dataset, (Flickr could not be used since it annotates images). We considered the 40 categories discovered using the CCA model and additionally ran LDA [62] on the document content of the collected web pages and identified 40 latent topics (again using perplexity). In Figure (15) we present JS-distance computed for all (topic,category) pairs. The x-axis shows the categories, the y-axis shows the topics, and the z-axis shows (1−JS-distance). We use (1−JS-distance) for visibility where similar pairs will show as large

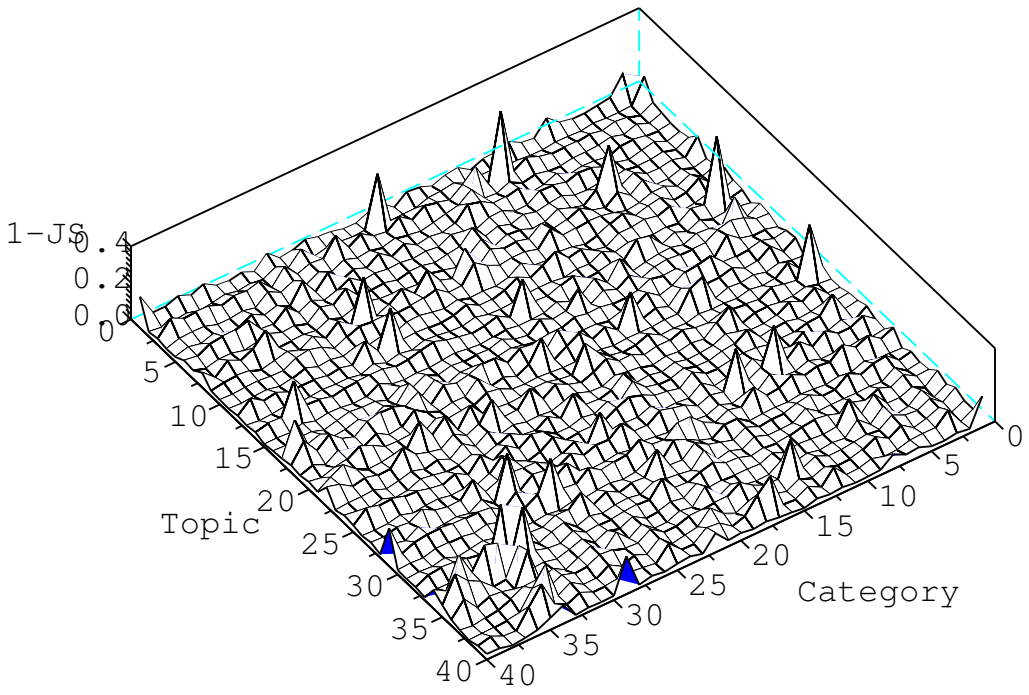


Fig. 15. Topic versus category similarity

spikes on the plot.

While there are some clear spikes, for the majority of topics there is no clear mapping to related categories, and vice versa. Hence we believe that the categorical annotation model identifies semantically coherent hidden categories that are not the same as the topics discovered through the application of a traditional content-based topic model – which further validates the need to separately model and study the collective intelligence annotation process from the content-generation process.

To further understand this separation, we also examined the set of social annotation document pairs that are categorically similar, where we considered pairs with

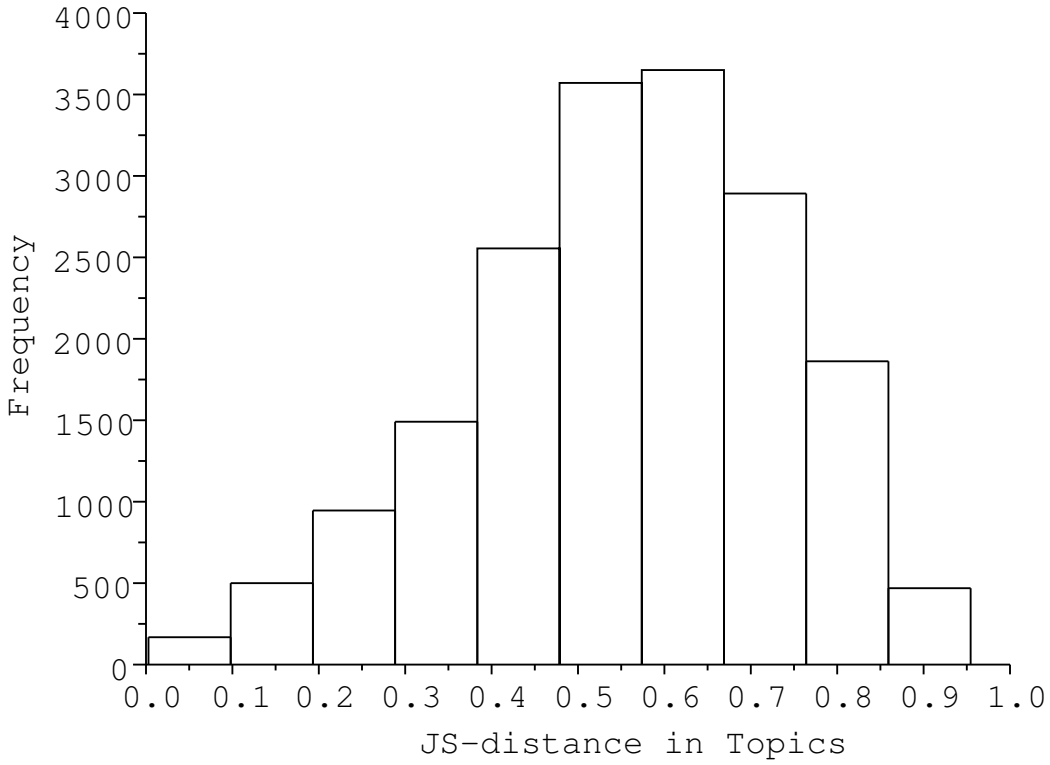


Fig. 16. Jensen-Shannon distance distribution in categories: objects with < 0.1 JS-distance in category space

JS-divergence less than 0.1 in the categorical space. How topically similar are these documents? Do documents that share similar tags also share similar content? In Figure (16), we report the JS-distance between these categorically-similar objects over their *content-based topic similarity*. Note how many of these categorically-similar web pages are quite dissimilar in topic space. In other words, objects tagged with similar tags do not necessarily have similar content.

Conversely, we also considered the set of web page pairs in our Delicious dataset that had a JS-divergence less than 0.1, where we measured the JS-divergence over the topics associated with each document. We find that many of these topically-

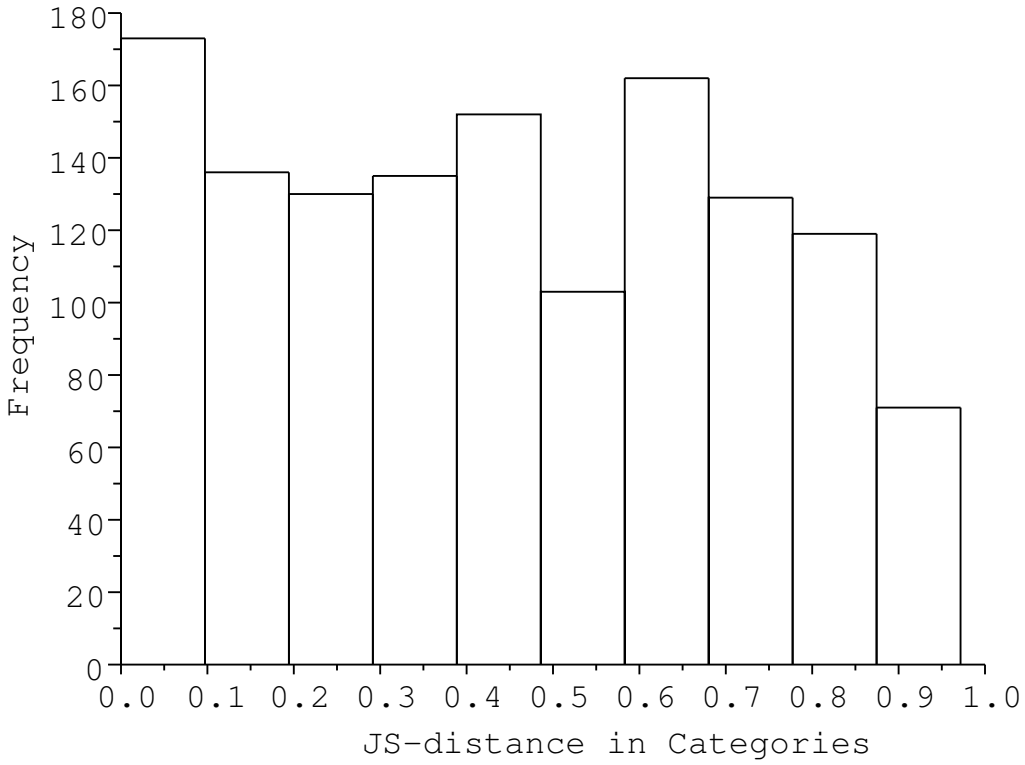


Fig. 17. Jensen-Shannon distance distribution in copics: objects with < 0.1 JS-distance in topics space

similar web pages are quite dissimilar in categorical space as shown in Figure (17). These results echo what we saw in Figure (16), that two documents may share many keywords in common (i.e., are topically similar), but their view from the community of social annotations is quite different.

2.3. Browsing in Topic and Category Spaces

We briefly illustrate one way to use both the annotation-based categorical model and the content-based topic approach for discovery and exploration of web objects. The main idea is to explore objects based both on their categorical and topical similarity

Category Space	Topic Space: JS < 0.1	Topic Space: JS > 0.9
JS < 0.1	http://www.ryanheise.com/cube/	http://www.anniston.lib.al.us/readalikes.htm
	http://www.alchemistmatt.com/cube/5by5cube.h	http://www.nintendo8.com/toplist/more/
	http://www.chessandpoker.com/rubiks-cube-sol	http://hca.gilead.org.il/
	http://williambader.com/museum/cubes/cubes.h	http://www.mcpl.lib.mo.us/readers/
	http://peter.stillhq.com/jasmine/rubikscubesoluti	http://www.viceteam.org/
	http://www.scaredcat.demon.co.uk/rubikscube/t	http://www.netlibrary.net/Collections.htm
	http://www.ryanheise.com/cube/beginner.html	http://www.forgottenbooks.org/
	http://www.ryanheise.com/cube/beginner.html#	http://www.gamelib.com.br/
	http://www.rubikssolver.com/	http://www.vistaicons.com/
	http://www.howtodothings.com/hobbies/how-to-	http://www.earlyword.com/
	http://thearufam.brinkster.net/cube/yy/	http://www.wyrdysm.com/games.php
JS > 0.9	http://tutorial.math.lamar.edu/	
	http://www.purplemath.com/modules/quadform.htm	
	http://edweb.tusd.k12.az.us/ibeneli/flash.html	
	http://www.mathgoodies.com/lessons/vol5/intro_integers.html	
	http://www.purplemath.com/modules/index.htm	
	http://davis.wpi.edu/~matt/courses/soms/	
	http://www.ee.ic.ac.uk/hp/staff/www/matrix/property.html	
	http://www.edhelper.com/math_grade1.htm	
	http://www.mathleague.com/help/integers/integers.htm	
	http://www.incompetech.com/graphpaper/	
	http://incompetech.com/graphpaper/	
	http://www.degraeve.com/reference/specialcharacters.php	

Fig. 18. Browsing in category and topic spaces

(and dissimilarity) to a candidate query object.

Here, we consider an example web page in the Delicious dataset concerned with the popular 1980s-era Rubik's Cube and several methods for solving the puzzle. The vocabulary for this page is overwhelmingly mathematical and based solely on the content this document is classified under the mathematics topic with high probability. However, the document also clearly belongs to the games and puzzles category (and this is reflected in the tags assigned to it). Given this query document, in Figure(18), we show the most relevant documents to our query document based on three views: (i) similar in topic space and similar in category space – these documents are primarily

mathematical approaches to Rubik's cube and similar puzzles; (ii) similar in topic space, but dissimilar in category space – these documents are primarily mathematical documents; and (iii) dissimilar in topic space, but similar in category space – these documents are primarily about games and puzzles.

2.4. Summary

Our browsing framework exploits the relationship between web objects content (text) and their social annotations (tags applied by various users). We show that when there are differences between a web document content and its tags, it is possible to improve information discovery by considering multidimensional similarity measures: computing similarity measures on separate content and annotation spaces instead of content alone or a combined content and tags space.

A prerequisite to computing similarity is to represent objects in reduced dimensionality spaces using CCA and LDA models. In this step, object content is assumed to be generated from a latent structure of *topics* and its annotations are assumed to be generated from a latent structure of *categories*. With these structures in hand, we compute objects similarity in the *topics* and *categories* spaces using the Jensen-Shannon distance defined above. We find that differences do exist between a web object content and its annotations and illustrate how it can be used improve browsing of socially tagged web objects.

The idea is to explore objects based on both their categorical and topical similarity (and dissimilarity) to a candidate query object. Given this query object, we measure its similarity to all other objects which are then classified into three views: (i) objects that are similar in topic space and similar in category space; (ii) objects that are similar in topic space, but dissimilar in category space and; (iii) objects that are dissimilar in topic space, but similar in category space. The user is then presented

with the top ranked documents in each view and can select an object of interest to view or to further use it to browse for similar and dissimilar objects.

A question remains, why are social tags of a web object different from its content? We attribute the cause to taggers having interpretations, perceptions, or interests on the object content that is different from the content creator. We see it because the social bookmarking medium allows for these differences to materialize and further complement the object's content. This however would be worth further detailed investigations.

3. Community-based Exploration Framework

Another important aspect of social bookmarking is the way in which users are handled and how they are modeled as part of tag generation process. Recent work on author-topic models [64] has added the concept of "author" to the topic models, but fundamentally these models are designed to model text documents that have a single (or a few) authors. In contrast, our probabilistic social annotation models in the preceding chapter account for the fact that many authors independently participate in the creation of social bookmarking. This leads to the discovery of user communities in addition to tag categories.

Here we leverage the underlying structure of the social bookmarking process uncovered by our models towards improving community-based and user-based search and exploration in the hopes of improving information access over socially tagged documents. Our goal is to enhance a baseline ranking function based on the BM25 ranking framework by considering community information derived from the generative annotation models. Concretely, we consider two approaches for leveraging this community information: (1) ranking by query-community relevance; and (2) ranking

by user-community relevance.

In the first approach, we aim to boost each document ranking by two factors: (i) the query term importance in each community; and (ii) the document importance in each community. The first factor boosts the ranking of documents that contain query terms considered important by the community regardless of document's community preference. The second factor boosts the ranking of documents that are preferred by the community regardless of them containing the query term. The net effect boosts the ranking of documents that are both preferred by the community and contain query terms that are considered important by the community. These factors can either be averages over all communities or maximum likelihood estimates obtained from the most representative community for the given query.

In the second approach, we aim to leverage user community information to enhance search and exploration. To that end, we consider the community membership for each user as determined by our model. Knowing a user's community strength, we can favor documents that are most preferred from the user's community, even if the user has never tagged the document. This approach constitutes to factors: one that accounts for community preference for a document and the second accounts for user membership in that community. Similarly, the factors can either be averages over all communities or maximum likelihood estimates obtained from the most representative community for the given user.

We use rank aggregation methods [90] to combine the results of the baseline (BM25), the query-community, and the user-community ranking functions. Our experimental results show that the inclusion of query-community and user-community rankings improves over the baseline retrieval method. In particular, We compare community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We

find that the proposed ranking model results in a significant improvement over these alternatives (from 7% to 22%) in the quality of retrieved pages.

3.1. Community-based Ranking

In the preceding chapter we presented the PSA models for discovering implicit social bookmarking communities. We now turn our attention to leveraging this information for community-based exploration of socially tagged documents. Since implicit user communities can be extracted from large-scale social bookmarking services, it may be advantageous to leverage this community structure to enhance user-based exploration over the web of socially tagged resources. Instead of guiding users to resources that a user already knows about (via his own tags) or that are globally well-known (e.g., a web page that many other users have already tagged), we seek to develop new community-based exploration approaches that emphasize the community's implicit view (e.g., to identify resources that are relevant to the implicit *sports* community).

Our goal is to leverage the discovered community structure to implicitly connect users, tags, and resources for more effective information exploration and discovery. Concretely, we propose a novel community-based ranking model that is designed with this intuition in mind. We illustrate our approach using results from the simplified PSA (SimplePSA) model but the approach is general enough to apply to the PSA and other models.

Applying the SimplePSA model to a collection of user, tag, and resource tuples produces the following distributions:

- For each community, we have a probability distribution over all users $\tau_c = \{\tau_{c,i}\}_{i=1}^{|U|}$
- For each community, we have a probability distribution over all tags $\phi_c =$

$\{\phi_{c,i}\}_{i=1}^{|T|}$

- For each resource, we have a probability distribution over communities $\theta_i = \{\theta_{i,j}\}_{j=1}^{L}$,

Based on these discovered distributions, we can, for example, identify implicitly related users and implicitly related tags based on their common community membership. These relationships can be used to automatically suggest related tags, to recommend unknown users (and their collection of bookmarks) to interested users, and so forth. Similarly, we can identify implicitly related resources based on their community distribution (e.g., to identify similar resources based on the communities that are interested in them), and support other forms of social exploration.

While the possibilities are quite large for applying the discovered community-based information from the SimplePSA model, we examine in the rest of this section two approaches for ranking resources based on the community's perspective. Concretely, we consider: (i) a query-community ranking approach that maps a user's topical interest (expressed as a query) to resources preferred by communities with a similar topical interest; and (ii) a user-community ranking approach that re-ranks all resources based on the user's implicit community, regardless of the query. In both cases, we are interested to examine if the discovered implicit community structure can enable more effective ranking than traditional (non-community) based approaches.

3.1.1. Query-Community Ranking

In the first approach, we aim to boost a baseline resource ranking by two factors: (i) the query term importance in each community; and (ii) the resource importance in each community. The first factor boosts the ranking of resources that contain query terms considered important by the community regardless of document's community

preference. The second factor boosts the ranking of resources that are preferred by the community regardless of them containing the query term. The net effect boosts the ranking of resources that are both preferred by the community and contain query terms that are considered important by the community.

This query-community ranking is defined as product of likelihoods of query terms relevance to resources as follows:

$$Score(S, Q) = \prod_{t \in Q} p(S|t)$$

Now, $p(S|t)$, the resource relevance to a query is computed over all communities using the two factors mentioned above, the community preference for the document and the community preference for the tag as follows:

$$p(S|t) = \sum_{c=1}^{L} p(S|c)p(c|t) \tag{4.1}$$

Using Bayes' rule we further expand each factor to be:

$$p(S|c) = \frac{p(c|S)p(S)}{p(c)} \quad \text{and} \quad p(c|t) = \frac{p(t|c)p(c)}{p(t)}$$

and by substituting into (4.1) we finally get:

$$p(S|t) = \sum_{c=1}^{L} \frac{p(c|S)p(t|c)p(S)}{p(t)}$$

The quantities $p(c|S)$ and $p(t|c)$ are readily available from the SimplePSA model results:

$$p(c|S) = \theta_S^c \quad \text{and} \quad p(t|c) = \phi_c^t$$

The document prior probability $P(S)$ and the tag prior probability $P(t)$ are collection dependent and we compute them as follows:

$$P(S) = \frac{|S|}{\sum_{i \in R} |S_i|} \quad \text{and} \quad P(t) = \frac{tf(t)}{\sum_{i \in T} tf(i)}$$

where, $|S|$, is the length of document S, and $tf(t)$ is the count of tag t.

3.1.2. User-Community Ranking

In the second approach, we aim to boost resource ranking by user community information. To that end, we consider the community membership for each user as determined by our model. Knowing a user's community strength, we can favor resources that are most preferred from the user's community, even if the user has never tagged the resource. This approach constitutes two factors: one that accounts for community preference for a resource and the second accounts for user membership in that community. This user-community ranking is defined as follows:

$$Score(S, u) = \sum_{c=1}^{L} p(S|c)p(c|u)$$

Using Bayes' rule again we expand $p(S|c)$ and $p(c|u)$ into:

$$p(c|u) = \frac{p(u|c)p(c)}{p(u)} \quad \text{and} \quad p(S|c) = \frac{p(c|S)p(S)}{p(c)}$$

Then substituting into the previous equation we get:

$$Score(S,u) = \sum_{c=1}^{L} \frac{p(u|c)p(c|S)p(S)}{p(u)}$$

These quantities are, again, readily available from the SimplePSA model results:

$$p(u|c) = \tau_c^u \quad \text{and} \quad p(c|S) = \theta_S^c$$

The document prior probability is computed is in the previous section and the user prior probability is computed as:

$$P(u) = \frac{uf(u)}{\sum_{i \in U} uf(i)}$$

where $uf(u)$ is the count of user u in the collection.

3.2. Rank Aggregation

In both cases: the query-community score and the user-community score, we can combine each individual score with a baseline query-resource score to arrive at a final score for each socially tagged document with respect to a query. In this work, as a baseline ranking approach, we adapt the popular BM25 retrieval model to the context of retrieval on social bookmarking systems [48]. For a user who is interested in searching the web of socially tagged resources, we can adapt the BM25 ranking over $\mathcal{U}$ for a query Q by scoring each social tagging document S:

$$Score_{BM25}(S,Q) = \sum_{t \in Q} IDF(t) \frac{f(t,S)(k_1+1)}{f(t,S) + k_1(1 - b + b\frac{|S|}{avgL})}$$
$$IDF(t) = \log \frac{D - n(t) + 0.5}{n(t) + 0.5}$$

where, $f(t,S)$ is the frequency of tag t in a social tagging document S, $|S|$ is total tags in S, $avgL$ is the average length of documents in $\mathcal{U}$, D is the total number of documents, $n(t)$ is the number of documents containing tag t. $k1$ and b are free parameters which we take to be their typical settings: $k1 = 2$, $b = 0.75$. BM25 provides a baseline for ranking resources by considering the presence of query terms in the socially tagged document, but makes no attempt to incorporate community or latent topic information.

To combine this baseline with the query-community score and the user-community score, we rely on rank aggregation, which is the task of combining voters' rankings of a set of candidates to obtain a single ranking for the set. It is a well known problem encountered in many contexts; especially in social choice theory. In our case the candidates are social web documents and the voters are the BM25 and the proposed community-based ranking functions. To combine the rankings produced by the ranking functions, we adopt a simple positional method known as Borda's Rule [90]. In Borda's rule each candidate is awarded a point for each competitor candidate ranked below it. Candidates are finally ranked by their accumulated points.

3.3. Ranking Over Socially Tagged Resources

Now we examine how community-based ranking models introduced above perform over socially tagged web objects. In addition to basing them on the BM25 mode results, we compare their results against two alternative state-of-the-art retrieval

models: (i) Cluster-based retrieval using K-means clustering; and (ii) LDA-based retrieval. While these retrieval models have been developed in the context of text-based retrieval, we adapt each to the context of retrieval on social bookmarking systems as described in the following brief sections.

Cluster-based Retrieval (K-means): The first approach is a tag-based implementation of cluster-based retrieval introduced by Liu and Croft [49]. Cluster-based retrieval hypothesizes that by grouping text documents (in our case, social tagging documents), the quality of ranking can be improved by smoothing each document with the rest of the documents in the cluster (in essence, asserting that similar documents will satisfy the same information need). In practice, we use K-means clustering to cluster all social tagging documents; we set $k = 40$ based on the results of the previous experiment. Documents in each cluster are then combined to build a unigram language model, i.e., a multinomial distribution over its vocabulary space. Ranking in this case is based on clusters instead of documents.

$$Score(cluster, Q) = \prod_{t \in Q} p(t|cluster).$$

The quantity $p(t|cluster)$ is computed from a cluster language model smoothed by a background model as follows:

$$p(t|cluster) = \lambda \frac{tf(t, cluster)}{\sum_t tf(t, cluster)} + (1 - \lambda) \frac{tf(t, Coll)}{\sum_t tf(t, Coll)}$$

where $tf(t, cluster)$ is the count of tag t in the cluster and $tf(t, Coll)$ is the count of tag t in the entire collection. The free parameter ($\lambda = 0.5$) controls the smoothing proportion. The cluster-based ranking can then be combined with the per-document

BM25 score using rank aggregation as in the case of community-based ranking.

LDA-based Retrieval: The second approach we consider follows Wei and Croft [50] to incorporate the LDA based document representation for retrieval. Given the inferred distributions from TagLDA, we can define an LDA-based ranking function as follows:

$$Score(S, Q) = \prod_{t \in Q} p(t|S)$$

Now, $p(t|S)$, the query likelihood given the document is computed over all tag topics using two factors: the document preference for the topic and the topic preference for the tag as follows:

$$p(t|S) = \sum_{z=1}^{K} p(z|S)p(t|z) \tag{4.2}$$

The quantities $p(z|S)$ and $p(t|z)$ are available from the TagLDA model results, $p(z|S) = \theta_S^z$ and $p(t|z) = \phi_z^t$. The LDA-based scores can then be combined with the per-document BM25 scores using rank aggregation.

3.3.1. Tag-based Retrieval

To evaluate the quality of community-based ranking and to be fair across all models, we first consider retrieval using only tags (since BM25, LDA, and K-means do not model the user as SimplePSA does). We select three sets of tags with the following criteria:

Rare tags: Six rare tags, that is tags that occur on at most 5 resources.

Unambiguous tags: Eight pairs of unambiguous tags, where we pair the tag *"tool"*

Table X. Tag-based retrieval results

	NDCG@10				%change by BM25+SimplePSA over		
	BM25	BM25+Kmeans	BM25+LDA	BM25+SimplePSA	BM25	BM25+Kmeans	BM25+LDA
Unambiguous tags	0.42	0.46	0.46	0.52	0.22	0.12	0.12
Rare tags	0.67	0.78	0.76	0.81	0.2	0.04	0.07
Popular tags	0.65	0.67	0.69	0.58	-0.09	-0.12	-0.15

with a number of tags such as *"finance", "music", "health", "social", "game"*, making the pair, *"music tool"*, very specific in what is expected to be retrieved.

Popular tags: Twelve popular tags, tags like *"new", "program", "resource", "howto", "blog"*.

For each of these tag sets we retrieve the top ten relevant documents per query using each ranking model – BM25, BM25+Kmeans, BM25+LDA, and BM25+SimplePSA. The results for each query are presented to four judges to determine their relevance to the query on a scale of 1 to 5 with 1 being least relevant and 5 most relevant. The judgements scores are analyzed using Normalized Discounted Cumulative Gain (NDCG)[91]. For a list of graded resources, NDCG computes the gain of each resource in the list based on its grade and rank and accumulates the gains over the list up to a specified position. Table (X) presents the NDCG@10 for each ranking model across the three types of queries, as well as the percent change by the proposed community-based ranking model BM25+SimplePSA versus the other three approaches.

First, consider the set of rare tags. Suppose our collection has 3 documents that carry the tag x. When a user searches for this tag using a traditional retrieval methods, e.g. BM25, those 3 documents will be returned as relevant and all other documents are given a score of zero, or a corpus wide smoothing score. However, there might be documents in the collection that do not carry the query tag but are relevant to the query, e.g., documents tagged with synonyms of the query term, or misspellings, or topically relevant tags. Community-based tag grouping (as in SimplePSA) could help improve results for this kind of query and our results support this conjecture. As Table (X) shows, the SimplePSA model results in the best ranking quality for rare tag queries, improving on BM25 by 20%, improving on K-means by 4%, and on LDA by 7%. We attribute this improvement to the ability of the SimplePSA model to

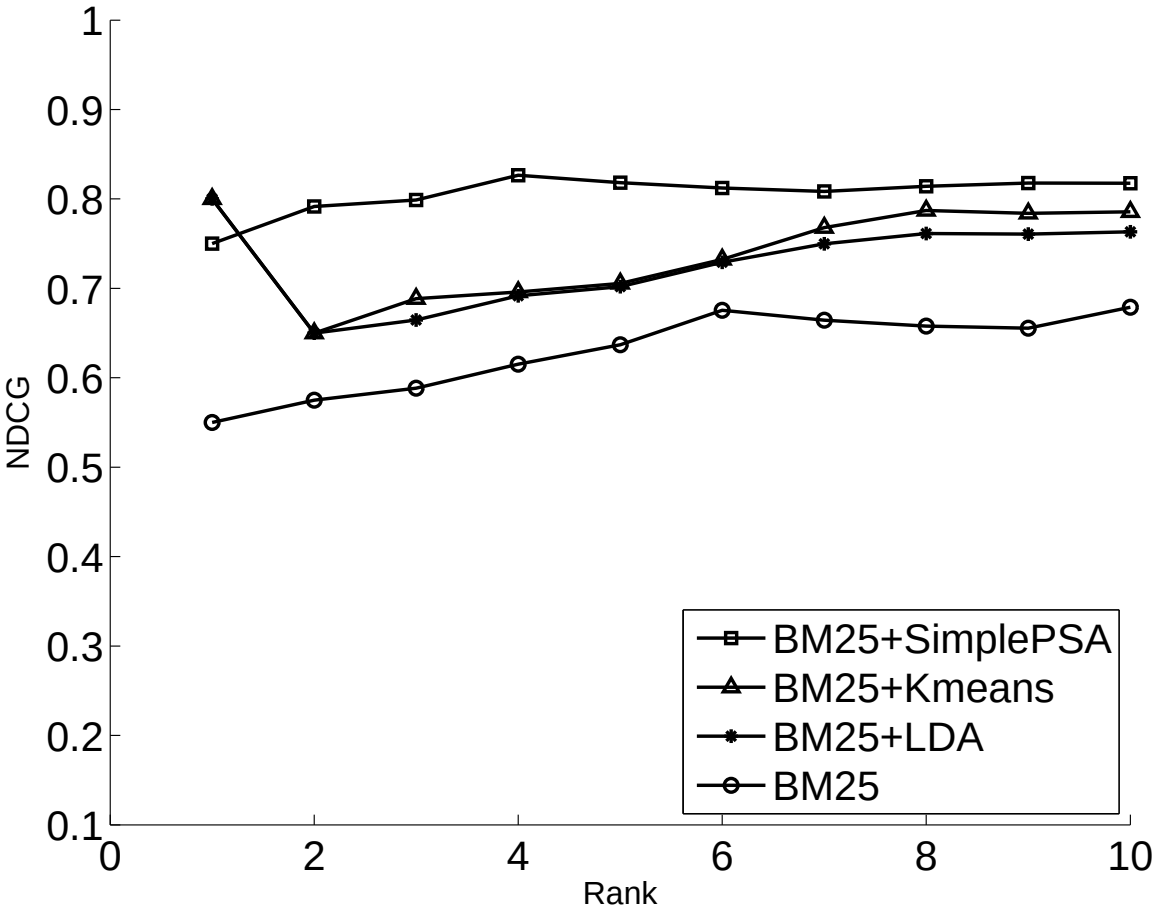


Fig. 19. Ranking quality for rare tag queries

better fit social bookmarking data than LDA or K-means. See Figure (19) for more detailed NDCG results for rank positions up to 10.

Second, for the set of unambiguous tags, the intuition is that a tag such as "tool" is popular, general and belongs uniformly to many communities while the the other tags are specific and could be prominent in small number of communities. For a traditional retrieval model the results are dominated by documents relevant to the general term due to high term frequency in document that might not be relevant to the second term. At the same time, the scores of the more specific term are not prevalent enough to make it to top positions in the retrieved list. Community-based

scores can help bring those documents that are considered valuable to the second terms's community up in the list. As Table (X) shows, the SimplePSA model results in the best ranking quality for unambiguous tag queries, improving on BM25 by 22%, improving on K-means by 12%, and on LDA by 12%. See Figure (20) for more detailed NDCG results for rank positions up to 10.

Finally, for the set of popular tags, their popularity makes them belong uniformly to many communities making the community structure, in this case, of little benefit. The community structure might actually degrade the retrieval performance by promoting documents that are too general and perceived uniformly across communities, which we suspect to be the case in our results. Another issue with this kind of queries, is the difficulty to evaluate relevance when query terms are vague. This was evident in the disagreements among judges scores for popular tag results.

To test judge biases in scoring the results for the different queries we use the Wilcoxon signed rank test [92], which given two paired samples of measurements, tests if the differences come from a symmetric distribution with zero median against the alternative that differences do not have a zero median. In our case if the differences have a zero median, we can conclude that the judges biases are not significant and that there is significant agreement on how the results are ranked. The Wilcoxon signed rank test results are shown in Table (XI). When the P-value < 0.05, we reject the hypothesis of zero median and conclude that there is significant disagreement among judges scores. This is seen only in the case of popular tags.

3.3.2. User-based Retrieval

Now that we have seen how community-based ranking can improve tag-based retrieval, we next consider how user-based retrieval can be improved. The goal of this section is to show the benefit of user modeling in social bookmarking as is done in the

Table XI. Wilcoxon signed rank test of agreement between judges scores with 95% confidence

Query group	P-value(two-tailed)
Unambiguous tags	0.10
Rare tags	0.39
Popular tags	< 0.0001
User-based retrieval tags	0.39

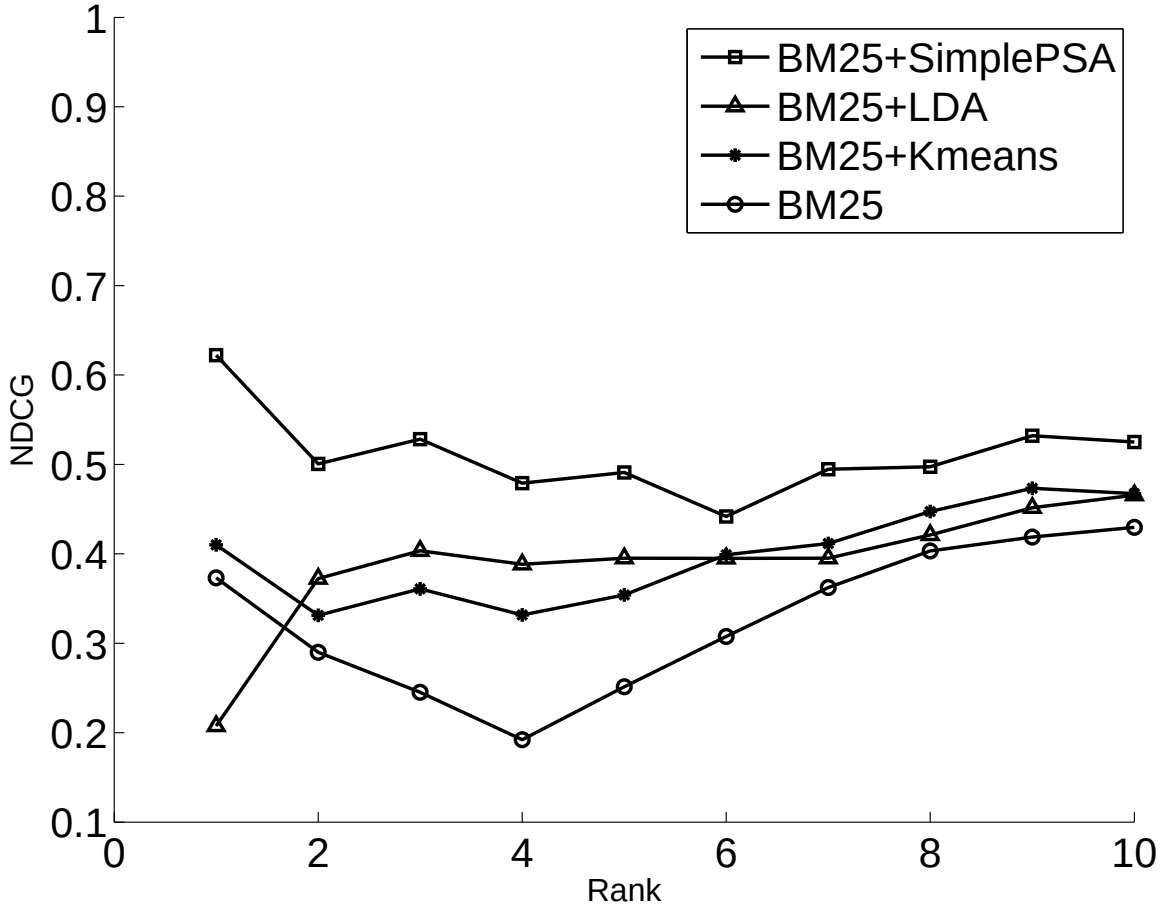


Fig. 20. Ranking quality for unambiguous queries

SimplePSA model.

To that end, we select five users that exhibit interest for some of the tags we used in the previous experiment, i.e., top users from the most representative community for the tag. For each user and tag combination, we retrieve the top ten relevant documents. These results are judged for relevance as was previously done on a scale of 1 to 5. The NDCG of judges results are as shown in Table (XII). The results for SimplePSA model include both the query-community (SimplePSA) ranking and the user-community (SimplePSA(U)) ranking. Notice that SimplePSA performs best with 47%, 43% and 60% improvement over BM25, K-means, and LDA respectively. These improvements show that user-based community structure uncovered by the SimplePSA model helps improve ranking of tagged resources.

3.4. Summary

Our second application framework utilizes the user-community and tag-community structures uncovered by modeling the social bookmarking process. With these structures in hand, we are able to augment traditional ranking methods with new approaches that make use of relationships exhibited by social web documents. In particular, having users as in integral component of the social web allow for the grouping of users based on interests, interpretation and background knowledge. These groups can be used to bias document rankings based on the user and the query term relevance to the group. We illustrate the benefits provided by results of our models in a framework that augments traditional ranking methods with two new approaches that are based on query-community relevance and user-community relevance.

We compare the results of these two approaches to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We find that the inclusion of query-community and user-

Table XII. User-based retrieval results

NDCG@10				%change by BM25+SimplePSA(U) over		
BM25	BM25+Kmeans	BM25+LDA	BM25+SimplePSA(U)	BM25	BM25+Kmeans	BM25+LDA
0.39	0.40	0.36	0.58	0.47	0.43	0.60

community rankings improves over the baseline retrieval method (BM25) as well as the other two clustering models in retrieving quality pages.

4. Conclusions

In this chapter we presented two frameworks that employ the results of our proposed models from the previous chapter to enhance traditional methods of information discovery and retrieval of socially tagged web objects. Our first framework is based on the CCA model and exploits the disparity between web object content and its social annotation to browse for similar/dissimilar objects in a multidimensional space. The second framework based on the PSA model utilizes the user-community and tag-community structures to improve ranking.

Although our approaches suggest an important role for socially contributed data in advancing information discovery, there are a number of limitations to its application and generalization to social bookmarking systems at large. First, LDA based approaches, in general, including our SimplePSA model require global knowledge and perform many iterations to uncover latent variables. Hence, applying thses models in an on-line system is computationally expensive. Second, our SimplePSA model, as does LDA, assumes a fixed number of latent variables and does not consider the temporal aspect of tagging. Therefore, it cannot capture growth and evolution. Third, our assumption of global user communities does not capture individual user behavior. In addition, the lack of standard corpora for social bookmarking data makes evaluating and comparing results of different research methods difficult. Furthermore, results based on individually collected corpora need to be verified for generalization to different social bookmarking systems.

However, some of these limitations can be overcome. A combination of a both

on-line and off-line approach can solve the processing requirements of LDA-based models. Also, there are methods for dynamically discovering the number of latent variables (see for example [93]). Next we will incorporate time into the SimplePSA model to capture community evolution over time.

CHAPTER V

TEMPORAL DYNAMICS OF COMMUNITIES IN SOCIAL BOOKMARKING SYSTEMS

1. Introduction

The preceding chapters of this dissertation have approached the social bookmarking process from a time-independent perspective. But in practice, social bookmarks are generated over time. Bookmarks for a certain web resource grow and evolve as the resource is discovered by more users. Also, new resources are added to the social web over time. User interests and bookmarking behavior possibly vary over time too. In this chapter, we conduct time-dependent analysis of the social bookmarking services that aims at revealing important characteristics of social bookmarking such as tag evolution, web resource popularity evolution, time-specific interest, user evolution, and community evolution.

Potential applications that can benefit from time-based analysis for social bookmarking services are: (i) social bookmarking systems: tagger behavior and its evolution can help in the design of better bookmarking systems; (ii) marketing and advertisement: tagger interests and evolution can serve as marketing indicators; (iii) information availability: tagger interests and evolution can help anticipate user's future information needs; and (iv) information access and organization: user interests evolution can be used enhance tag cloud design which in turn improves browsing of the social web.

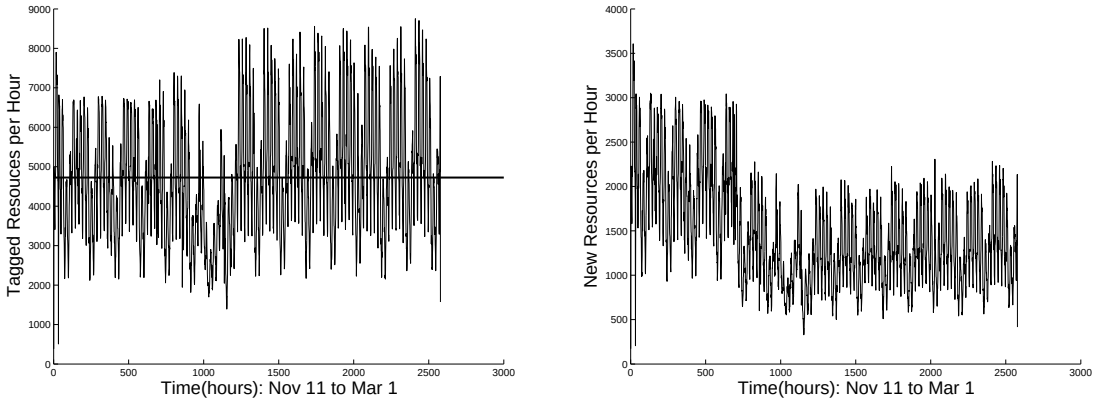
We begin by observing the social bookmarks on the Delicious social bookmarking service in action for a period of 15 weeks. The data collected allows us to determine

general social bookmarking behavioral statistics such as posts per day, unique tags per day, taggers per resource and so on. We then present our approach to model social bookmarking systems over time. Our goal is to identify groups of taggers and their interests and how they evolve. To that end we propose a modification to the probabilistic social annotation model (SimplePSA) in [81] that aims to capture community-wide interests over time intervals through modeling user activity and tagging choices. Using the results of this temporal modeling approach to social bookmarking services, we devise community dynamic graph representation that tries to capture users and tags dynamic movement between communities. We show how this representation can enable a closer inspection of communities characteristics as well as that of the constitute user and tags over time. We also show how to use this representation to capture cross-community relationships over time.

2. Temporal Social Bookmarking Data and Features

Our dataset was collected from Delicious' recent feed over a period of 15 weeks (November 11th, 2009 to March 1st, 2010). It consists of 13,405,322 unique postings over 3,778,338 unique URLs, performed by 641,021 users using 1,504,147 unique tags.

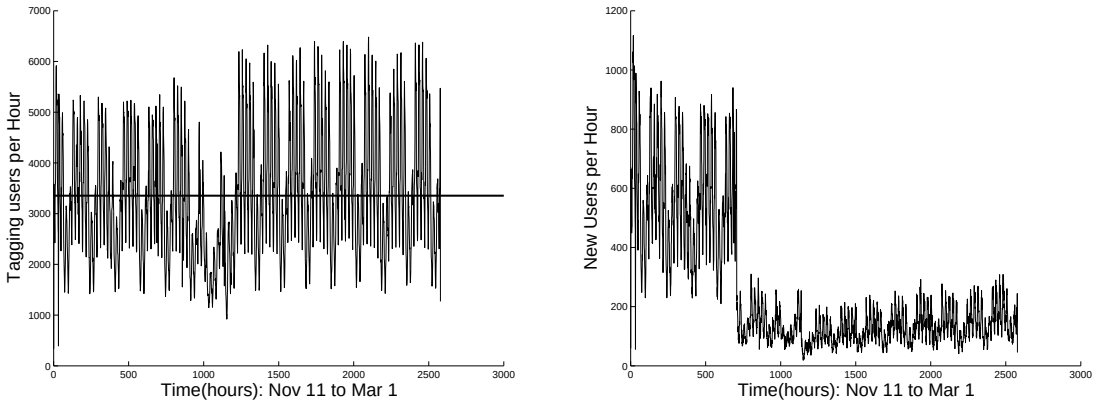
Some interesting observations about the dataset are the number of tagged resources over time, the number of taggers over time, and the number of tags used. In Figure (21) we show resource activity over the observed period. The x-axes show the time interval in hours and the y-axes show the number of resources. Figure (21) (a) presents the number of unique resources tagged per hour. On average, there are $4,726$ unique resources tagged every hour. Figure (21) (b) shows that on average about $1,000$ of these resources have not been observed in any previous time interval.



(a) Unique resources tagged per hour

(b) New unique resources tagged per hour

Fig. 21. Resource activity per hour over 15 week period: (a) shows the number of resources tagged per hour and (b) shows the number of new resources never seen before



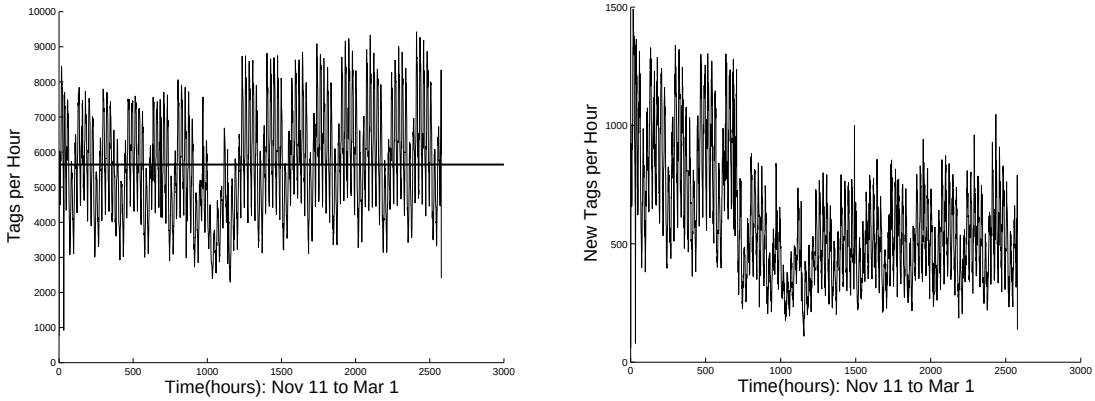
(a) Unique active taggers per hour

(b) New unique active taggers per hour

Fig. 22. Tagger activity per hour over 15 week period: (a) shows the number of taggers active per hour and (b) shows the number of new taggers never seen before

In Figure (22) we present the taggers activity over time. The x-axes show the time interval in hours and the y-axes show the number of taggers. Figure (22) (a) presents the number of unique active taggers per hour. On average, there are $3,353$

unique taggers every hour. Figure (22) (b) shows that on average about 100 taggers in each hour have not been observed in any previous time interval. Finally, in Figure (23) we present tags' usage over time. The x-axes show the time interval in hours and the y-axes show the number of tags. Figure (23) (a) presents the number of unique tags used per hour. On average, taggers use $5,644$ unique tags every hour. Figure (23) (b) shows that on average about 500 of these tags have not been observed in any previous time interval.



(a) Unique applied tags per hour (b) New unique applied tags per hour

Fig. 23. Tags per hour over 15 week period: (a) shows the number of tags used per hour and (b) shows the number of new tags never seen before

Generally, we observe that the overall Delicious social bookmarking community gets around 110K resources, 80K taggers, and 130K tags on a daily basis. We also can conclude that this community maintains growth in the number of resources, users and tags; with resources having the highest growth rate followed by tags then by taggers. The drop in activity around the $1,000^{th}$ hour is due to Christmas time and the New Year. Notice that a drop in the number of new taggers arriving into the system occurs around the 700^{th} hour. We take this as an indicator that most taggers are active at least once a month. A similar drop but less pronounced occurs with new tags around

the same time. A much slighter drop is seen in the case of resources. A one month duration is sufficient to capture the majority of taggers and tags but not of resources.

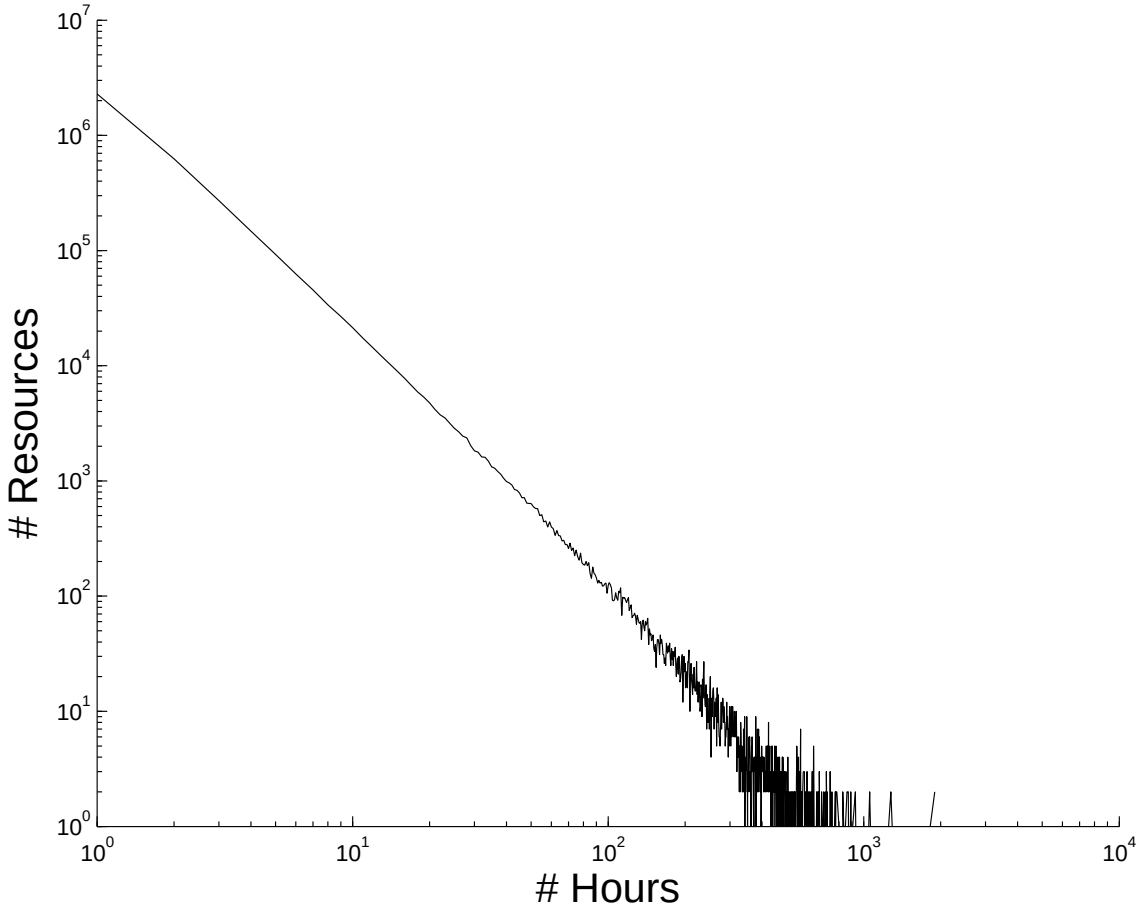


Fig. 24. The number of resources and the corresponding number of hours in which they were observed at least once

Next, we turn to examine the reoccurrence of resources, taggers and tags over time during the observed period. For each resource, tagger, and tag, we want to observe the corresponding number of hours in which they were active at least once.

In Figure (24), we plot the count of resources and the corresponding time intervals in which they appeared at least once. It shows us that there are more than 1M resources that were observed in only one time interval each (one hour), while there

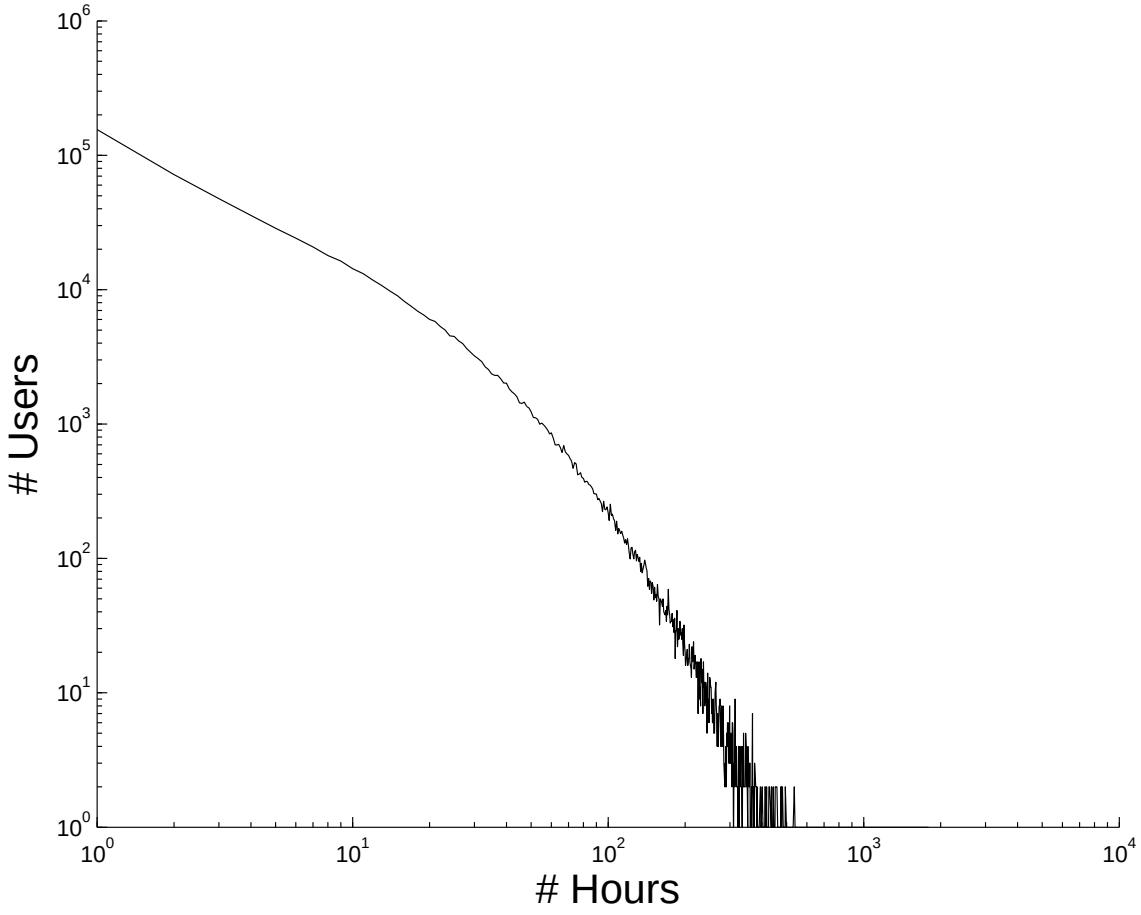


Fig. 25. The number of taggers and the corresponding number of hours in which they were active

are 10 resources that were observed in at least 300 time intervals each. These features can be of interest for identifying popular resources or spam.

Similarly, for taggers we plot in Figure (25) the number of taggers and the corresponding number of hours in which they were active. We see that there are more than 100K taggers that were active only in one time interval. On the other hand, 10 taggers were seen in between 200 to 300 time intervals each. These features can be of interest for identifying prolific taggers, spammers, trend makers, as well as bots.

Finally, Figure (26) shows the number of tags and the count of hours in which

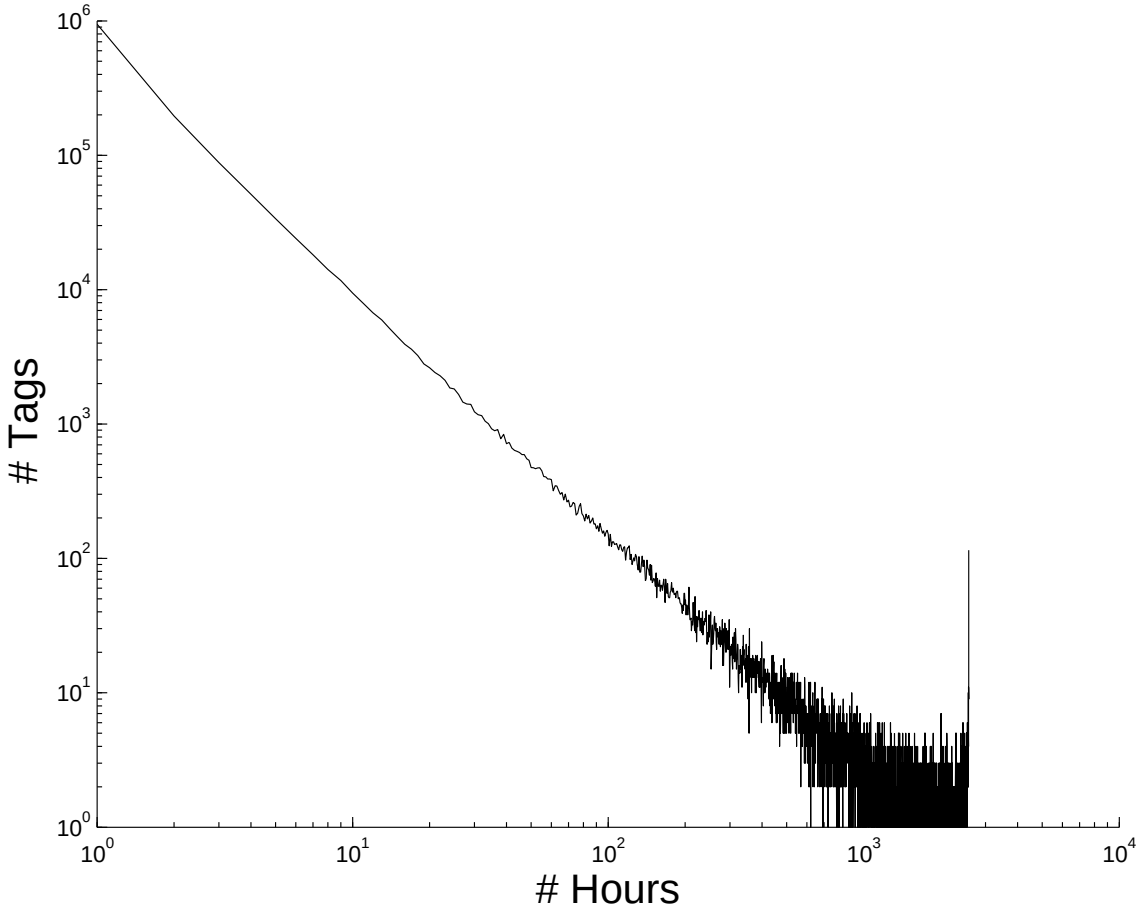


Fig. 26. The number of tags and the corresponding number of hours in which they were observed at least once

they were used at least once. The figure indicates that there are about 1M tags that were seen in only one time interval each, while about 10 tags appeared in $1,000$ time intervals each. These features can be useful to determine trends, classification, and spam terms.

We can see that resources, taggers, and tags reoccurrence follow the common power law distribution where a few elements are very active and the majority have very low reoccurrence. Notice that the resources and the tags reoccurrence form a straight line on the log-log scale while the taggers reoccurrence forms a curved line

(slower decline in the number of taggers for increasing time interval counts) which indicates that a large number of taggers are active unlike the case for active tags and active resources.

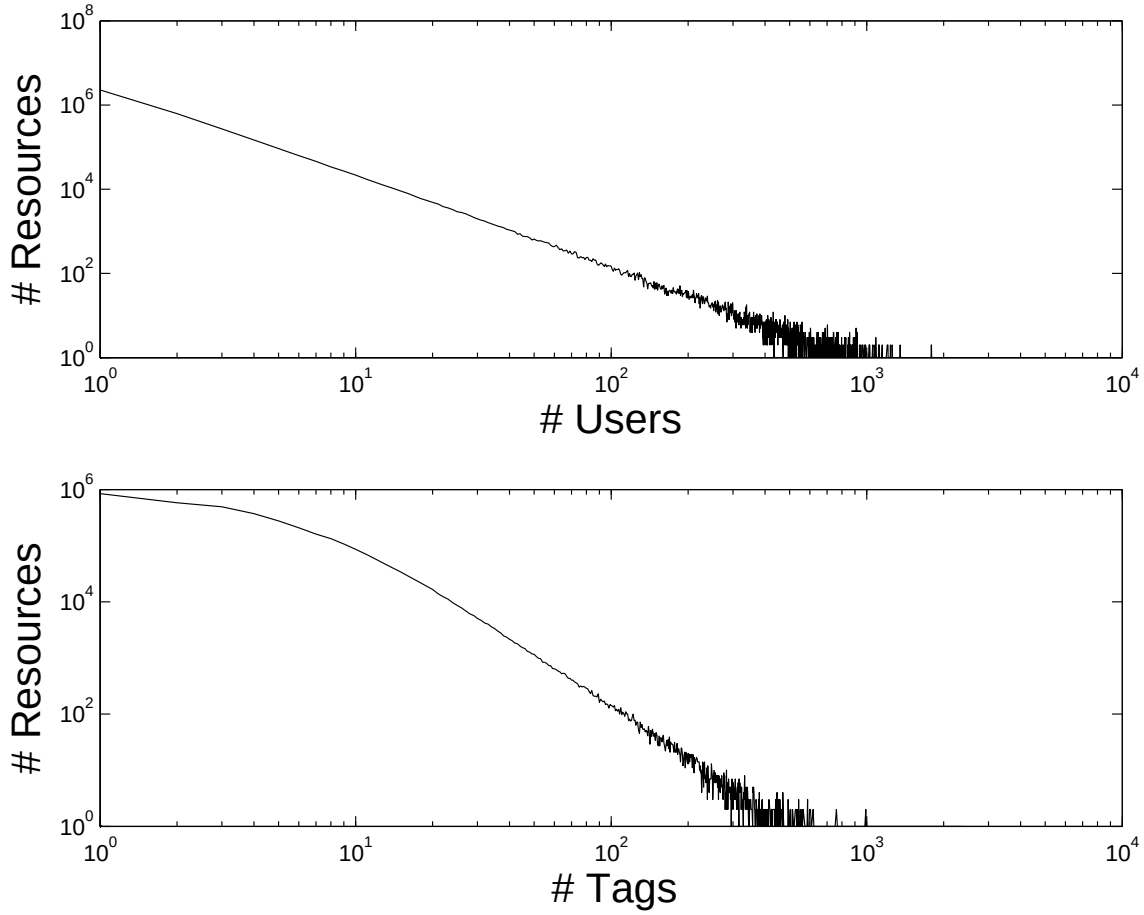


Fig. 27. Resources interactions with: taggers (top) and tags (bottom)

To further illustrate the characteristics of social bookmarking systems, we examine the co-occurrences among resources, taggers, and tags. We start by looking at how resources interact with taggers and tags. Figure (27) plots the number of resources and the corresponding number of taggers and tags that they co-occurred with. Notice in the top subfigure that there are more than 1M resources that are tagged by just

one tagger each, while there are around 100 resources that are tagged by about 200 taggers each. In the bottom subfigure we see that there are about 1M resource with only one tag each and about about 100 resources with about 150 tags each. Once again we observe the phenomenon that a few resources are popular, attracting many users and tags, while the majority get minimal exposure to users and very few tags.

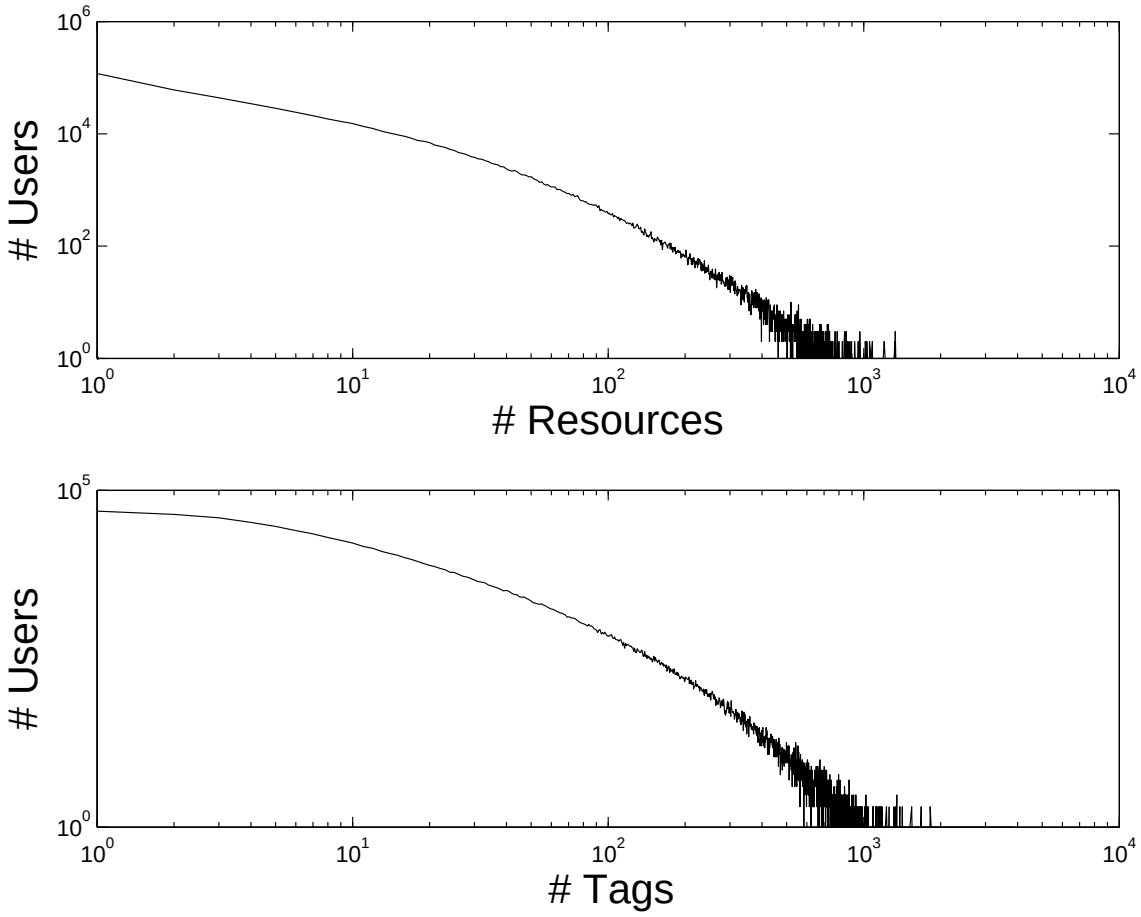


Fig. 28. Taggers interactions with: resources (top) and tags (bottom)

Now we look at how taggers interact with resources and tags. In Figure (28) we show similar observation for taggers. In the top subfigure, we observe that there are about 1M taggers that tag just one resource each, while there are 10 taggers that tag

more than 1,000 resources each. In the bottom subfigure, we see that there are about 100K taggers that use just one tag each, and there are 10 taggers that use about 1,000 tags each.

Finally, we look at how tags interact with taggers and resources. Figure (29) presents similar counts for tags. In the top subfigure we observe that there are more than 100K tags that are used by just on tagger each, while there are 10 tags that are used by about 1,000 taggers each. The bottom subfigure shows that there are about 1M tags that appear on only one resource each, while there are 100 tags that appear on about 1,000 resources each.

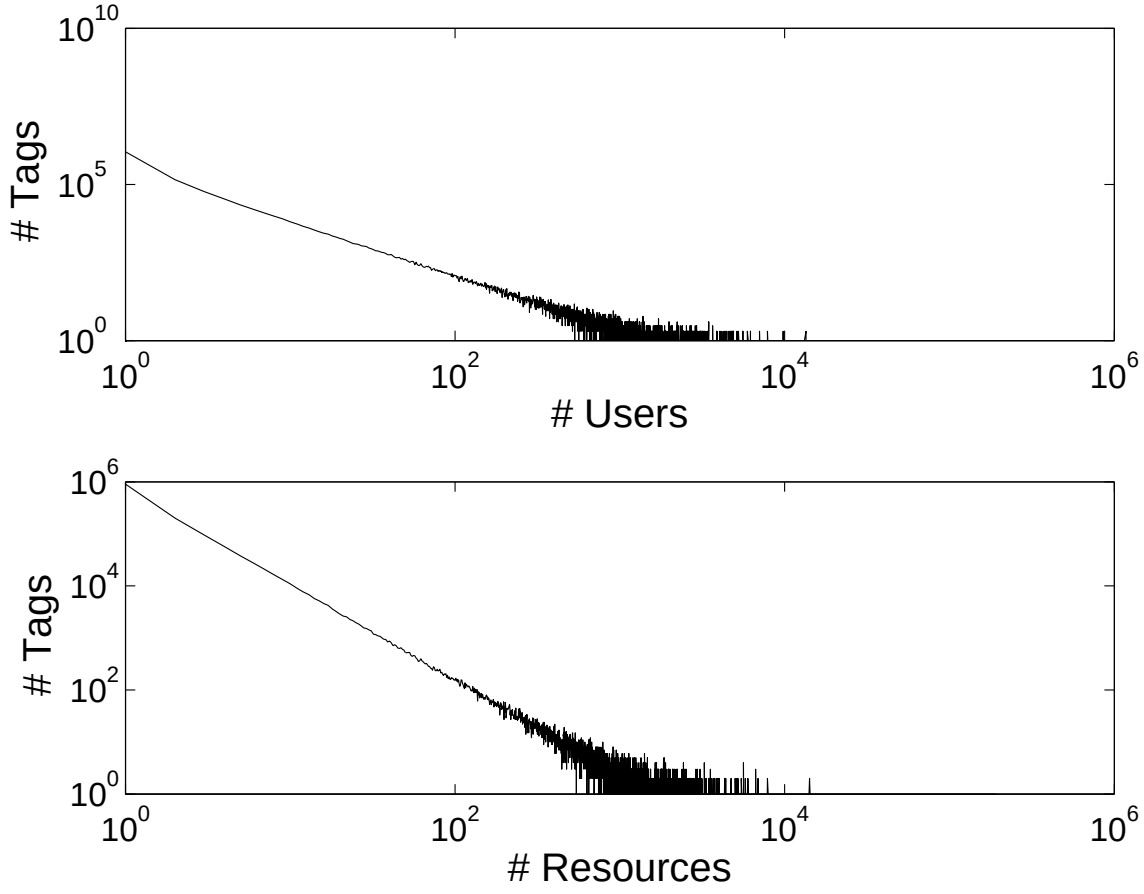


Fig. 29. Tags interactions with: taggers (top) and resources (bottom)

Although the Delicious feed could be throttled, filtered or delayed internally the Figures above suggest a minimum activity rate for all three types (resource, user, tag) as well a minimum growth estimate.

3. The Segmented Community-based Tagging Model

In previous chapters, we have adapted a text-based topic modeling approach to handle social bookmarking data, where we consider the document unit to be the collection of all tags and users applied to a particular resource. We call this collection of tags and users applied to a resource its *social tagging document*.

[Definition] Social Tagging Document: *For a resource $r \in \mathcal{U}$, we refer to the collection of tags assigned to the resource as the resource's social tagging document S, where S is modeled by the set of users and the tags they assigned to the resource:* $S = \{\langle user_j, tag_j \rangle\}$.

In the simplified PSA (simplePSA) approach , we posit the existence of L communities that are implicit in the universe of discourse $\mathcal{U}$, where each community is composed of users and tags that are representative of the community's perspective. Since community membership is not fixed, we model membership as a probability distribution, where each user and tag has some probability of belonging to any community.

[Definition] Social Tagging Community: *A social tagging community c is composed of (i) a probability distribution over users in U such that $\sum_{u \in U} p(u|c) = 1$, where $p(u|c)$ indicates membership strength for each user u in community c; and (ii) a probability distribution over tags in the vocabulary T such that $\sum_{t \in T} p(t|c) = 1$, where $p(t|c)$ indicates membership strength for each tag t in community c.*

Extending the simplePSA modeling approach to capture the temporal nature of

social bookmarking services can be accomplished in several ways. A straightforward extension is to still view the basic unit to be the collection of tags and users applied to a particular resource but limited to those occurring in a specific time period (e.g., by hour, day, week, etc.). Now all tagging activity that occurs in a given time interval is assumed (as previously done) to be drawn from global latent structures of communities of users and their associated tag vocabulary. In addition these global structures can now change from one time interval to the next. We call this approach, the Segmented Community-based Tagging (SCTAG) Model.

Our intuition is that reasonably long time intervals (for example a week) will contain a mixture of tagged resources that can potentially reveal the current global interests as well as a classification of the different taggers and the tags they use. Additionally observing the system over consecutive time intervals will reveal the evolution of interests, user groups, as well as tag groups.

The SCTAG model partitions the annotations applied to a single resource into K segments based on the time the annotation was applied. For example, Figure (30) shows a sample resource (the CNN page, `www.cnn.com`) being split into June and July 2010 segments. This processing step is performed for all resources. It results in collections of social tagging documents ordered by time. Each collection can be modeled separately using the simplePSA model presented in the previous chapters. But to capture the changes from one time segment to the next we require the simplePSA model to use latent structures learned from earlier time segments as a prior for learning the structures of later time segments (see Figure (31)).

3.1. Generative Process

The generative process for the SCTAG model in Figure (31) works as follows:

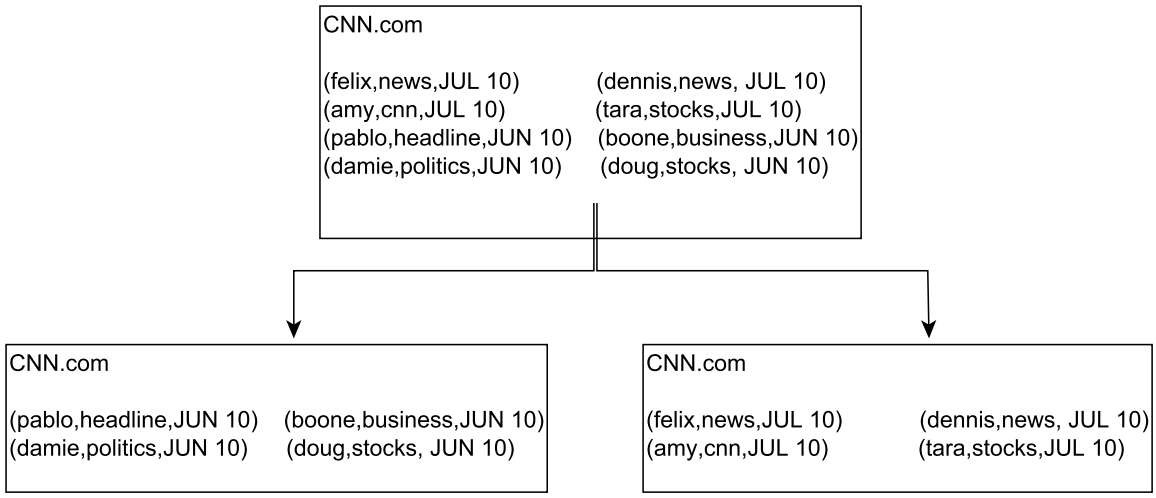


Fig. 30. Splitting annotations to time segments (month)

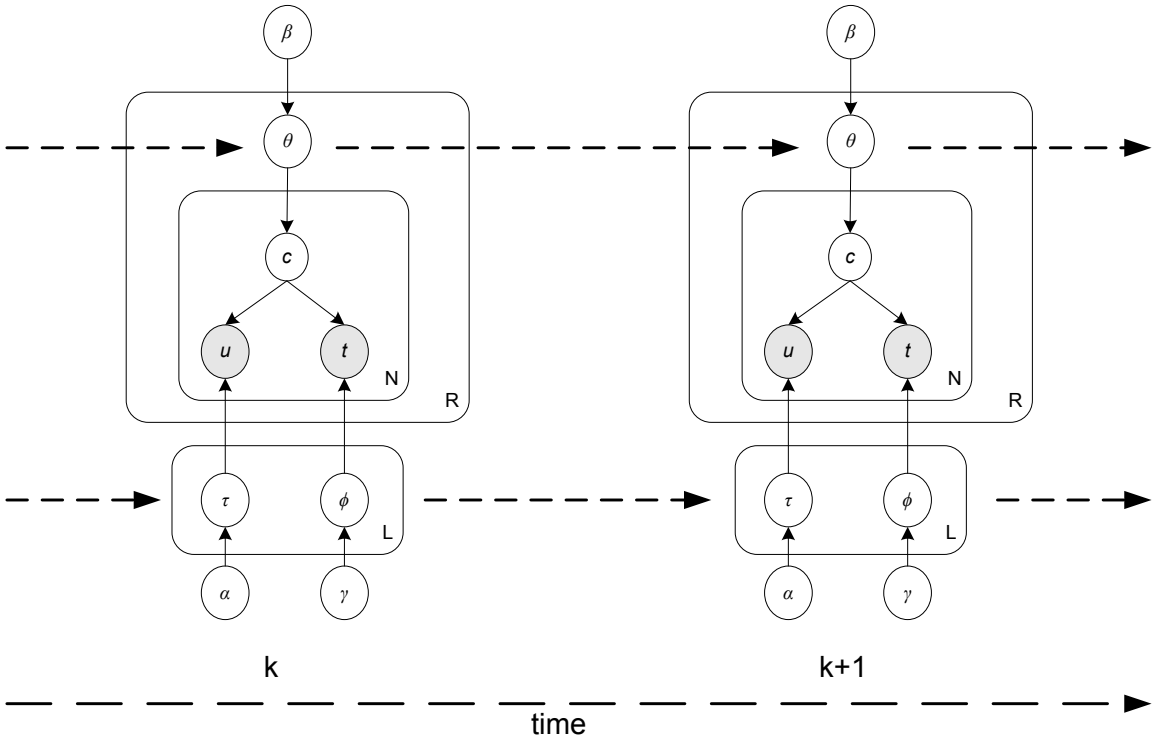


Fig. 31. Segmented Community-based Tagging model (SCTAG)

- for time segment $k = 2, ..., K$

  1. for each community $c = 1, ..., L$
     - Select U dimensional $\tau_c^k \sim \text{Dirichlet}(f(\tau_c^{k-1}, \alpha))$
     - select V dimensional $\phi_c^k \sim \text{Dirichlet}(f(\phi_c^{k-1}, \gamma))$
  2. for each object $\mathbf{S}_i$, $i = 1, ..., D$
     - Select L dimensional $\theta_i^k \sim \text{Dirichlet}(f(\theta_i^{k-1}, \beta))$
     - For each position $\mathbf{S}_{i,j}$, $j = 1, ..., N_i$
       * Select a community $\mathbf{c}_{i,j} \sim \text{multinomial}(\theta_i^k)$
       * Select a user $\mathbf{S}_{i,j}^u \sim \text{multinomial}(\tau_{\mathbf{c}_{i,j}}^k)$
       * Select a tag $\mathbf{S}_{i,j}^t \sim \text{multinomial}(\phi_{\mathbf{c}_{i,j}}^k)$

The first time segment, $k = 1$, has no prior latent structure and therefor reduces to SimplePSA model. Each consecutive time segment augments the prior structure with current observation points allowing for evolutionary behaviors to be observed. Next we use Gibbs sampling to estimate the latent structures of the SCTAG generative process.

3.2. Parameters Estimation with Gibbs Sampling

The generative process shown above describes how temporal social tagging documents are created. Our goal here is to take collections of social tagging documents that we assume are the product of such a generative process and recover the underlying hidden structures of communities, their users and their tags. More specifically, we learn model parameters τ, θ, and ϕ (the distributions over communities, users, and tags, respectively) for each time segment.

Let $\mathbf{S}$ and $\mathbf{c}$ be vectors of length $\sum_i^R N_i$ representing $\langle user, tag \rangle$ pair, and community assignments, respectively, for the collection in a single time segment, k. Also let u and t be user and tag variables. Suppose the latent structures learned in the previous time segment $k-1$ is known, $\mathcal{M}_{k-1}$. Following the approach used in [66] we derive the following Gibbs sampler's update equation for assigning communities to user, tag pairs in time segment k:

$$p(\mathbf{c}_i = l | \mathbf{c}_{\neg i}, S^t, S^u, \mathcal{M}_{k-1}) \propto \tag{5.1}$$
$$\frac{n^u_{l,\neg i} + \tilde{n}^u_l + \alpha_u}{\sum_{u=1}^U n^u_{l,\neg i} + \tilde{n}^u_l + \alpha_u} \times \frac{n^t_{l,\neg i} + \tilde{n}^t_l + \gamma_t}{\sum_{t=1}^V n^t_{l,\neg i} + \tilde{n}^t_l + \gamma_t}$$
$$\times \frac{n^l_{S,\neg i} + \beta_l}{\left(\sum_{l=1}^L n^l_S + \beta_l\right) - 1}$$

where $n^{(\cdot)}_{(\cdot),\neg i}$ is a count excluding the current position assignments of $\mathbf{c}_i$ (e.g., $n^t_{l,\neg i}$ is the count of tag t generated by the l-th community excluding the current position). $\tilde{n}^{(\cdot)}_{(\cdot),\neg i}$ are prior counts from the preceding time segment. The prior counts are set to zero for the first time segment as well as for new users and tags that appear for the first time in later segments.

4. Community Discovery

In our experiments, we segment the dataset into weekly time segments as was discussed in previous sections. This results in 15 sub-collections each with $90,000$ social tagging documents on average. We run the SCTAG model on the first sub-collection to determine the number of potential communities. We set the model hyperparameters to $(\alpha = 0.9, \beta = 0.1, \gamma = 0.01)$ and vary the number of communities from 20 to 160.

Our goal here is to first confirm that the model works and to also determine an

Table XIII. A sample top tags in communities

	top tags
Comm 0	video movi film stream youtub media entertain cinema televis onlin free documentari subtitl towatch anim multimedia clip download review watch live show list filmmak recommend vimeo seri flv...
Comm 1	learn educ elearn web2 train teach research technolog onlin moodl resourc knowledg open pedagogi eportfolio opensourc virtual blog commun ict instruct secondlif lm assess virtualworld dog theori...
Comm 2	map api googl geographi googlemap gp gi mashup geo locat data local geoloc visual googleearth geocod earth cartographi world develop refer foursquar geologi geotag tool mapa webservic...
Comm 3	socialmedia facebook market social media socialnetwork trend brand web2 busi strategi advertis roi mashabl casestudi 2010 research measur socialmedia twitter digit internet polici...
Comm 4	fashion shop blog cloth magazin design style inspir shirt vintag cultur beauti trend tshirt shoe moda men accessori store retro art bag hipster lifestyl jewelri cool...

appropriate number of communities. Since the focus is not on optimizing the number of communities discovered we elect to fix the community parameter to 100 since it resulted in the most cohesive top-20 tags with the least overlap across discovered communities. A sample of the discovered communities top tags are shown in Table (XIII).

To illustrate the benefit of using this model, we compare these discovered communities to the top frequency tags observed during one hour intervals in our dataset. Table (XIV) shows a sample of most frequently used tags per hour. Notice the overwhelming presence of “web design” and “programming”. Contrary to the impression one gets of a lack of community structure based on the most frequent tags, a topic modeling based approach reveals some interesting communities.

After the initial step of determining the number of communities, the SCTAG model is run on the remaining time segments where a preceding time segment result serves as a prior for the following time segment.

Table XIV. Top frequency tags per hour

Hour								
1	2	...	1001	1002	...	2001	2002	...
tools	design	...	design	design	...	tools	design	...
design	tools	...	tools	blog	...	webdesign	webdesign	...
google	programming	...	blog	webdesign	...	tools	tools	...
software	webdesign	...	webdesign	inspiration	...	blog	inspiration	...
programming	inspiration	...	programming	software	...	video	blog	...
reference	blog	...	software	video	...	inspiration	web	...
inspiration	software	...	video	programming	...	web	video	...
blog	google	...	inspiration	art	...	programming	resources	...
webdesign	web	...	free	tutorial	...	free	programming	...
web2.0	reference	...	tutorial	reference	...	css	reference	...
science	tutorial	...	reference	web	...	software	software	...
tutorial	free	...	art	development	...	reference	free	...
resources	development	...	development	photography	...	tutorial	css	...
web	art	...	music	music	...	jquery	development	...
research	video	...	howto	howto	...	web2.0	art	...
opensource	education	...	web2.0	web2.0	...	resources	javascript	...

5. Community Evolution

Previously we have used user, tag, and resource interactions to capture communities of users and their tag vocabulary. Now, we observe these communities over time and try to capture how they change. We define change based on two aspects of the community: i) the users forming the community and ii) the tags representing the community perspective.

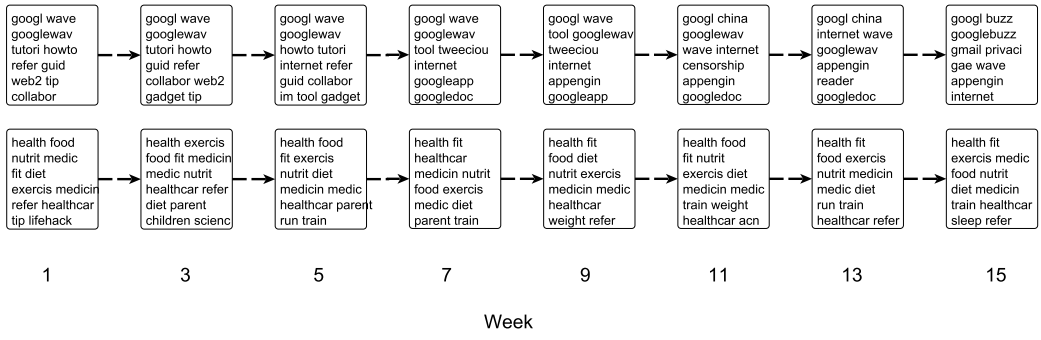


Fig. 32. A community's topic evolution over time

Let us now inspect how community interests change over time. An example of community interest evolution over time segments is shown in Figure (32). We present two sample communities about "Google tools" and "Health" along with their top-10 tags over the 15 week period. Notice how the "Google tools community" is initially concerned with "Google wave" applications and "collaboration" in weeks $1-9$ (November to December 2009, during which Google Wave was released) then switched to "Google china", "Google buzz", and "privacy" in weeks $11-15$ (January to February 2010, during which the Google China hack scandal and Google Buzz privacy issues occurred). On the other hand the "health community" can be seen as more stable with interests continually represented by tags such as "food", "exercise", "health care", and "medicine".

The SCTAG model gives us a per community distribution over users as well as a distribution over tags. Focusing on just one community over a sequence of time segments, we can measure changes on both distributions over users and over tags using a measure like the Jensen-Shannon (JS) distance (defined in the preceding chapter).

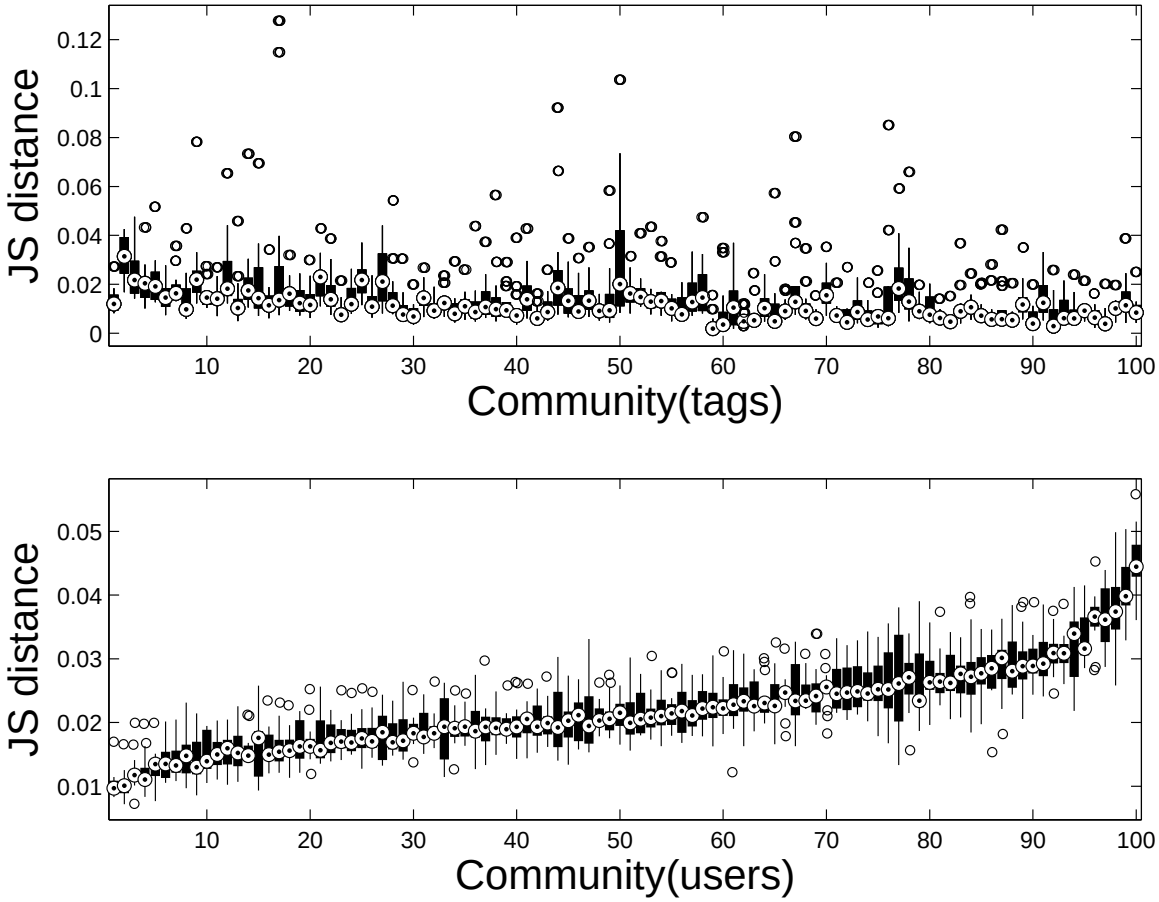


Fig. 33. JS-distance per community over time (sorted by users)

We present the results of the JS-distance per community over time in both user and tag spaces in Figure (33). The results are sorted by the distance over the user space. Notice that communities vary in their distance over the user space as shown in the bottom subfigure. This indicates evolutionary dynamics of user membership in the communities. This however can not be said about the tag space as is shown in

the top subfigure. Notice that all communities have relatively similar distances in the tag space despite their distances in the user space, meaning that communities with high user churn and those with low user churn all have relatively stable interests.

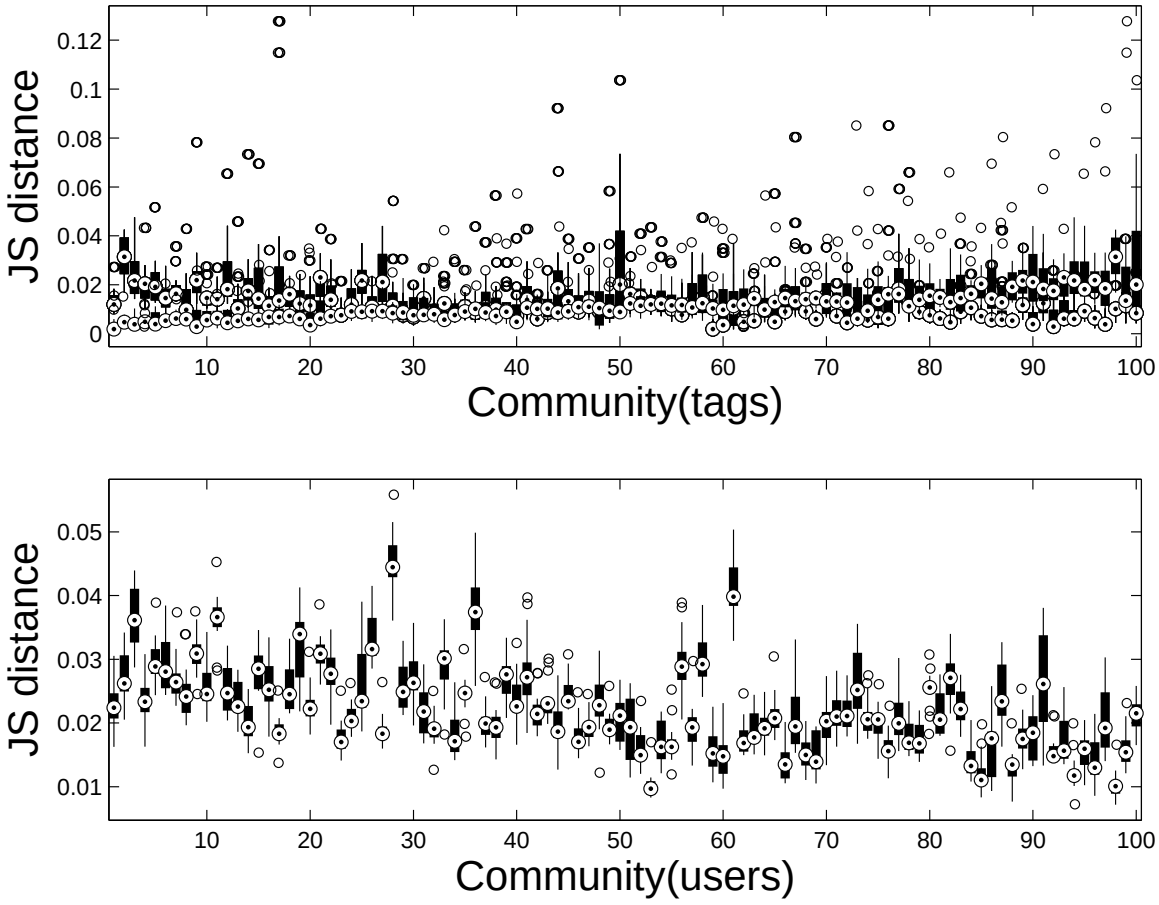


Fig. 34. JS-distance per community over time (sorted by tags)

The same trend can be observed when the results are sorted by the distance in the tag space, as is shown in Figure (34). Based on this we can conclude that communities tend to evolve more on their user space than on their tag space. That is, users tend to change their community membership over time more often than do tags. This is an expected result as user membership in communities represents the user transient interest while tag membership in communities represents a thematic

classification of the tags.

6. Community Dynamics

To determine inter-community relationships we can apply the same Jensen-Shannon distance to measure the overlap over users and tags between two communities. This overlap can also be measured as a Jaccard or cosine distance. However, these methods are too expensive as they require pairwise comparisons over the community space. Alternatively, we develop a simpler method for comparing communities focusing on the user and the tag spaces and their community assignment.

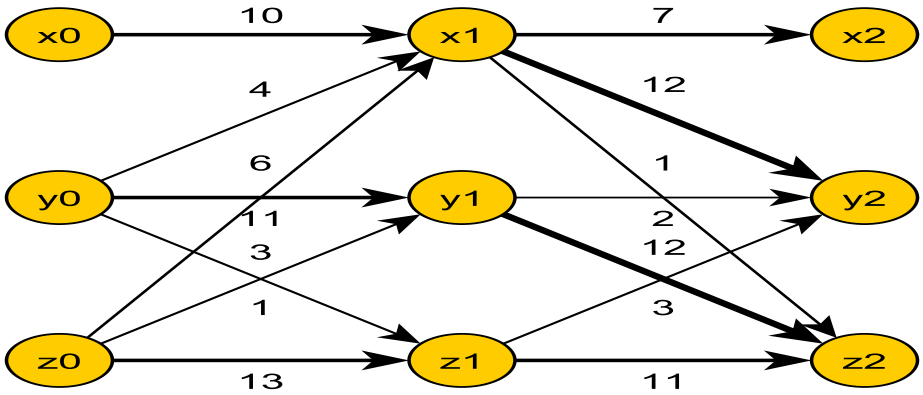


Fig. 35. Example transitions across communities over time

Our method takes the results of the SCTAG model and tracks each user and tag, and their top community assignment in each time segment. From this step, we can determine the paths that users or tags take over time in relation to communities. For example, in Figure (35), we have three communities spread over three time segments. An example path is $(x0, x1, y2)$, representing users or tags assigned to community x at times 0 and 1; and to community y at time 2. By simply counting the number of users or tags transitions between communities over time segments, we can capture

community evolution, similar communities, communities that evolve together and others that deviate over time.

Formally, let there be a graph $G = (V, E)$ where the set of nodes V represents the community space spanning all time segments, and the set of edges E represents users' community assignment transition from one time segment to the next. An edge $e = (x_0, y_1)$ indicates that a user with community assignment x at time 0 got assigned to community y at time 1. The weight of edge e, $w(e)$, represents the number of users or proportion of users that had made the same community assignment transition. For example, in Figure (35), 13 users assigned to community z at time 0 were assigned to community z at time 1, while 3 of these users got assigned to community y at time 2.

Using this graph setup we ask questions of community relationships and stability. Are there stable communities? Do stable communities have core members? And what is the size of this stable core? We consider stability to be influenced by the following factors: the community core, the community parters, and the community joiners.

[**Definition**] **Community Core:** A social tagging community c has a core consisting of users or tags that are successively assigned to community c in two consecutive time segments, $CC(c) = \{u \in \cap_k^{k+1} U_k \texttt{ s.t. } c = argmax_{c=1,..,L}\ \tau_{c,u}^k\}$. For tags, $CC(c) = \{t \in \cap_k^{k+1} T_k \texttt{ s.t. } c = argmax_{c=1,..,L}\ \phi_{c,t}^k\}$.

[**Definition**] **Community Parters:** A social tagging community c parters are a set of users or tags that are assigned to community c at time segment $k-1$ but not at time segment k:

$PA(c) = \{u \in \cap_k^{k+1} U_k \texttt{ s.t. } c = argmax_{c=1,..,L}\ \tau_{c,u}^{k-1}\ and\ c \neq argmax_{c=1,..,L}\ \tau_{c,u}^k\}$.
And for tags:

$PA(c) = \{t \in \cap_k^{k+1} T_k \texttt{ s.t. } c = argmax_{c=1,..,L}\ \phi_{c,t}^{k-1}\ and\ c \neq argmax_{c=1,..,L}\ \phi_{c,t}^k\}$.

[**Definition**] **Community Joiners:** A social tagging community c joiners are a set of users or tags that are not assigned to community c at time segment $k-1$ but

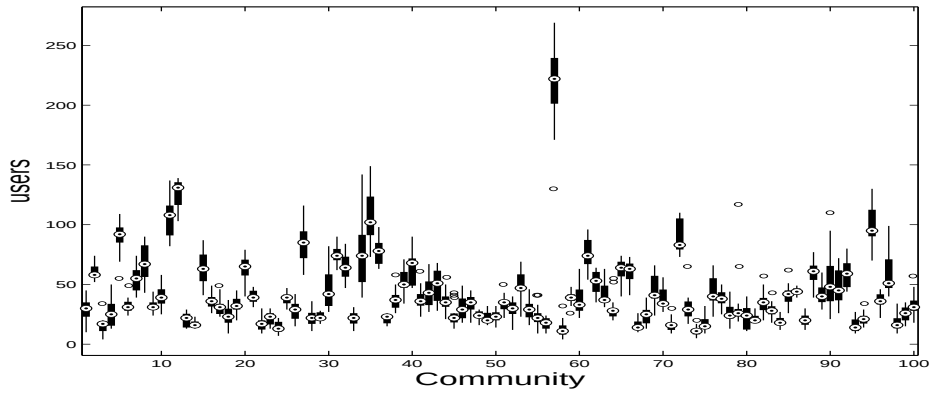


Fig. 36. Average number of users assigned to community

are in time segment k:

$$JO(c) = \{u \in \cap_k^{k+1} U_k \ \texttt{s.t.}\ c \neq argmax_{c=1,..,L}\ \tau_{c,u}^{k-1} \ and \ c = argmax_{c=1,..,L}\ \tau_{c,u}^{k}\ \}.$$

And for tags:

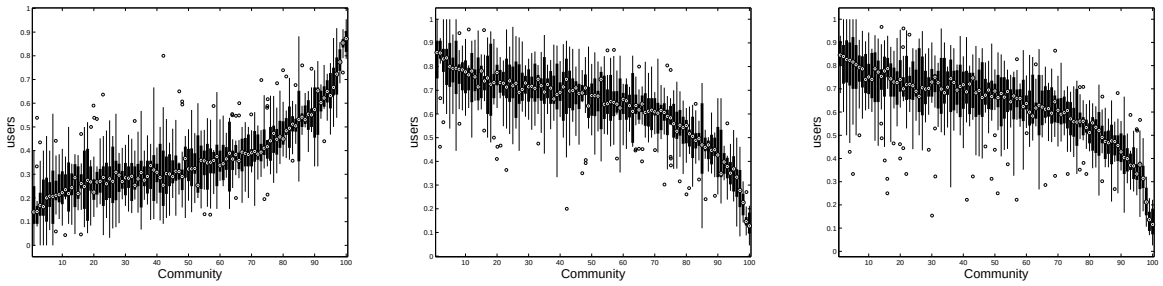
$$JO(c) = \{t \in \cap_k^{k+1} T_k \ \texttt{s.t.}\ c \neq argmax_{c=1,..,L}\ \phi_{c,t}^{k-1} \ and \ c = argmax_{c=1,..,L}\ \phi_{c,t}^{k}\}$$

For example, consider the communities x, y, z shown in Figure (35) with 3 time segments. The core for community x is at most 7, its parters are 13 and its joiners are 10. Respectively for communities y and z, their cores are at most 2 and 11, their parters are 19 and 6, and their joiners are 6 and 14.

6.1. Users and Tags in Communities

We start by showing how the user space is assigned to the discovered communities. Figure (36) shows the average number of users assigned to each community over the observed time period. The average number of users assigned to a community fall in the range $[12, 300]$.

Transforming the SCTAG discovered communities into the community dynamics graph introduced earlier allows us to observe the community core, community parters, and community joiners. We present the results for users and tags (core, parters,



(a) Core (b) Parters (c) Joiners

Fig. 37. Proportion of users per community

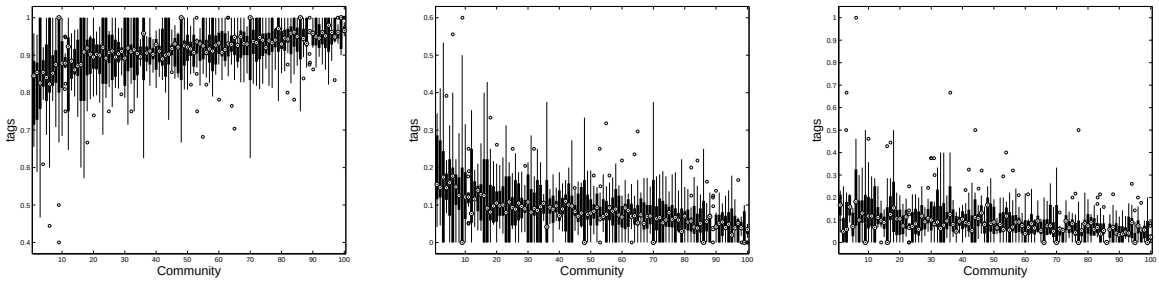
joiners) per community in Figures (37) and (38), respectively. For users , the figure shows the mean proportion of users and its variance for all three types. Notice the communities with small core user proportion less than 0.25 are the ones that have high mean proportion of joiners and parters, as is expected. We call these communities *high user churn communities*. These communities form about one fourth of the total number of communities. Similarly communities with high core user proportion greater than 0.6 are the ones that have low mean proportion of joiners and parters. We call these *low user churn communities*. These communities form a smaller fraction of the discovered communities, about one tenth. The majority of communities have a mean core user proportion in the range $[0.25, 0.60]$, meaning the majority of communities maintain between one quarter to one half of their user memberships over time. Next we take a closer look at high and low user churn communities.

6.2. Low and High User Churn Communities

In Table (XV), we present 5 low churn communities and 5 high churn communities. The first column shows the mean core user proportion for the community over the observed time period, the second column shows the standard deviation, and the third shows the average core user size. The fourth column, shows the top tags representing

Table XV. Community user churn (core tags are shown in red)

low churn			
mean	std. dev.	avg. size	tags
0.868	0.051	44.13	bandom brendon spencer fic ryan patd author slash gerard frank pair fandom torchwood jack pete
0.643	0.087	59.46	recip food cook dessert bake vegetarian chocol blog cake chicken pasta bread cooki breakfast
0.642	0.059	60.8	net drupal asp develop program tutori microsoft modul mvc cm howto facebook silverlight pattern
0.623	0.074	100.2	rubi rail rubyonrail program develop tutori test gem plugin web xmpp deploy howto github api
0.607	0.046	125.8	socialmedia market facebook social media twitter socialnetwork web2 advertis blog busi brand trend
High churn			
mean	std. dev.	size	tags
0.168	0.1212	12.6	document write refer program tutori howto wiki tip latex vim django develop python manual editor
0.173	0.1229	16.8	shop ecommerc commerc import busi url onlin websit web paypal magento auction internet info
0.193	0.1178	16.2	search googl present searchengin tool powerpoint web2 engin internet web visual research bing
0.201	0.1330	14.3	confer event forum commun collabor video calendar access present web2 tool web webconferenc
0.218	0.0856	28.8	video movi stream film onlin free youtub televis media entertain download search cinema



(a) Core (b) Parters (c) Joiners

Fig. 38. Proportion of tags per community

the community interest. By our definition the low churn communities have high mean core user proportion and high churn communities have low mean core user proportion. Notice the low churn communities have lower variance around the mean compared to high churn communities. Also low churn communities cores are much larger than high churn communities. Now by inspecting the top tags for the low churn communities we notice that they have specialized, and narrow interests. For example, the top one is about "Fan Fiction Stories", the following one is about "Cooking and Recipes", and the third is about "Web Programming". On the other hand, high churn communities have more generic interests like "Shopping", "Search tools", and "Videos". Also in Table (XV), we highlight core tags (in red) among the top tags for each community. In general, low churn communities have higher counts of core tags compared to high churn communities.

Next we look at how low and high churn communities top tags evolve over time. We present in Table (XVI) a sample low churn community interested in "Politics" and another high churn community interested in "Audio and Sound". For each community we list the top tags in time segments $\{1, 5, 9, 13\}$. We also highlight the core tags in red. Generally, we see no dramatic differences between how top tags of low and high churn communities evolve over time. But we can see that over time more core tags are

Table XVI. Community evolution (core tags are shown in red)

Community evolution	
week	Low Churn
1	polit govern histori new econom law usa cultur blog refer research war activ obama media statist militari crime women polici
5	polit govern econom usa histori china new economi cultur obama activ war polici women law militari intern world gender femin
9	polit govern econom usa histori activ terror cultur war new law women economi obama societi femin polici gender islam crime
13	polit govern usa econom activ cultur obama new histori law women femin war polici gender militari race india economi corrupt
week	High Churn
1	audio podcast sound music book free read video mp3 literatur resourc librari blog audiobook web2 tool educ record speech
5	audio podcast sound free video speech mp3 music multimedia audiobook record bbc radio resourc text award listen teen
9	audio podcast video music sound free multimedia bbc mp3 record audiobook speech radio sampl award resourc media text
13	audio podcast sound music video text mp3 audiobook speech multimedia record free award radio voic ipod media sampl nois

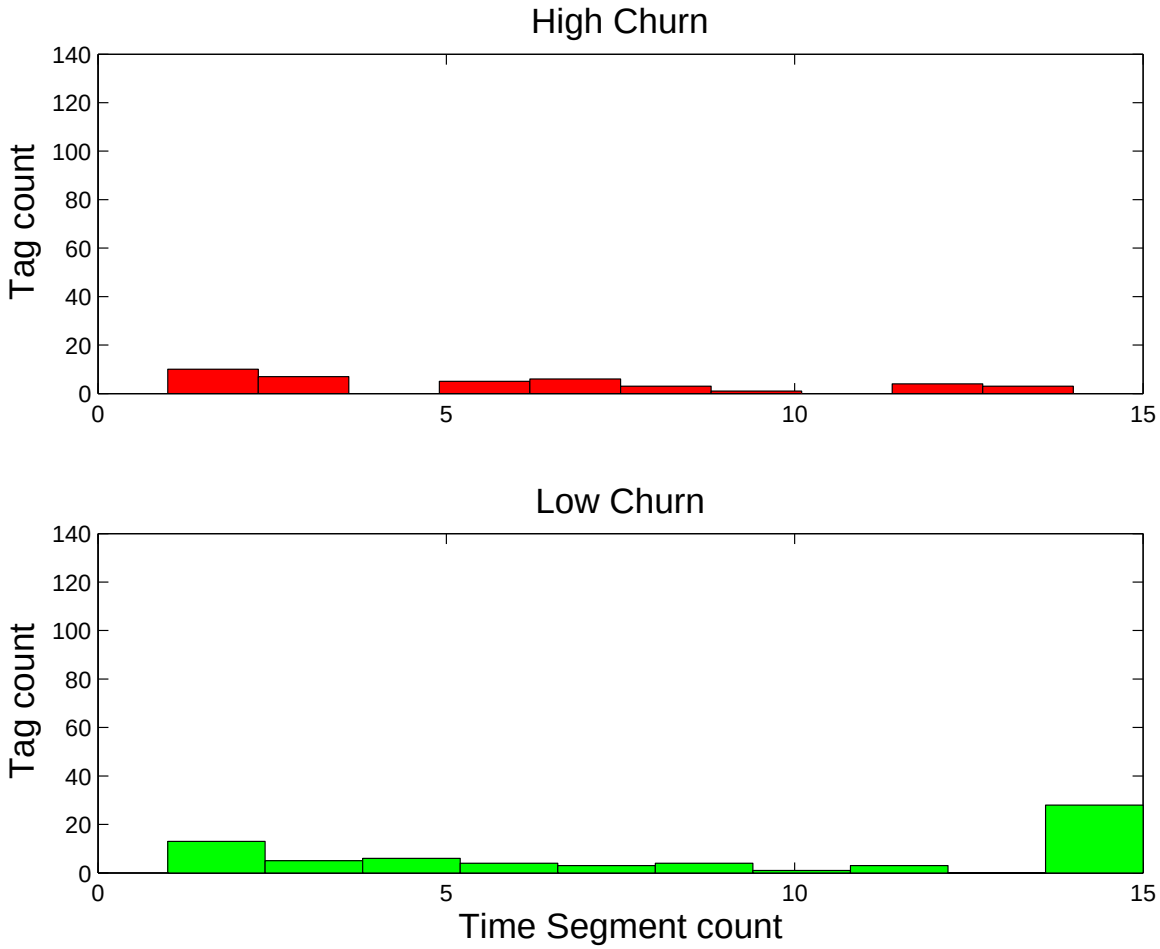


Fig. 39. Core tags behavior in low and high churn communities

represented in the communities top tags for both low and high churn communities.

6.3. Core Users and Tags in Communities

This leads us to the question of how core tags and core users behave in low and high churn communities. To observe this behavior we compute the number of core tags and core users and the corresponding number of time segments in which they occur for both low and high churn communities.

In Figure (39) we show the results for core tags. Notice that in low churn

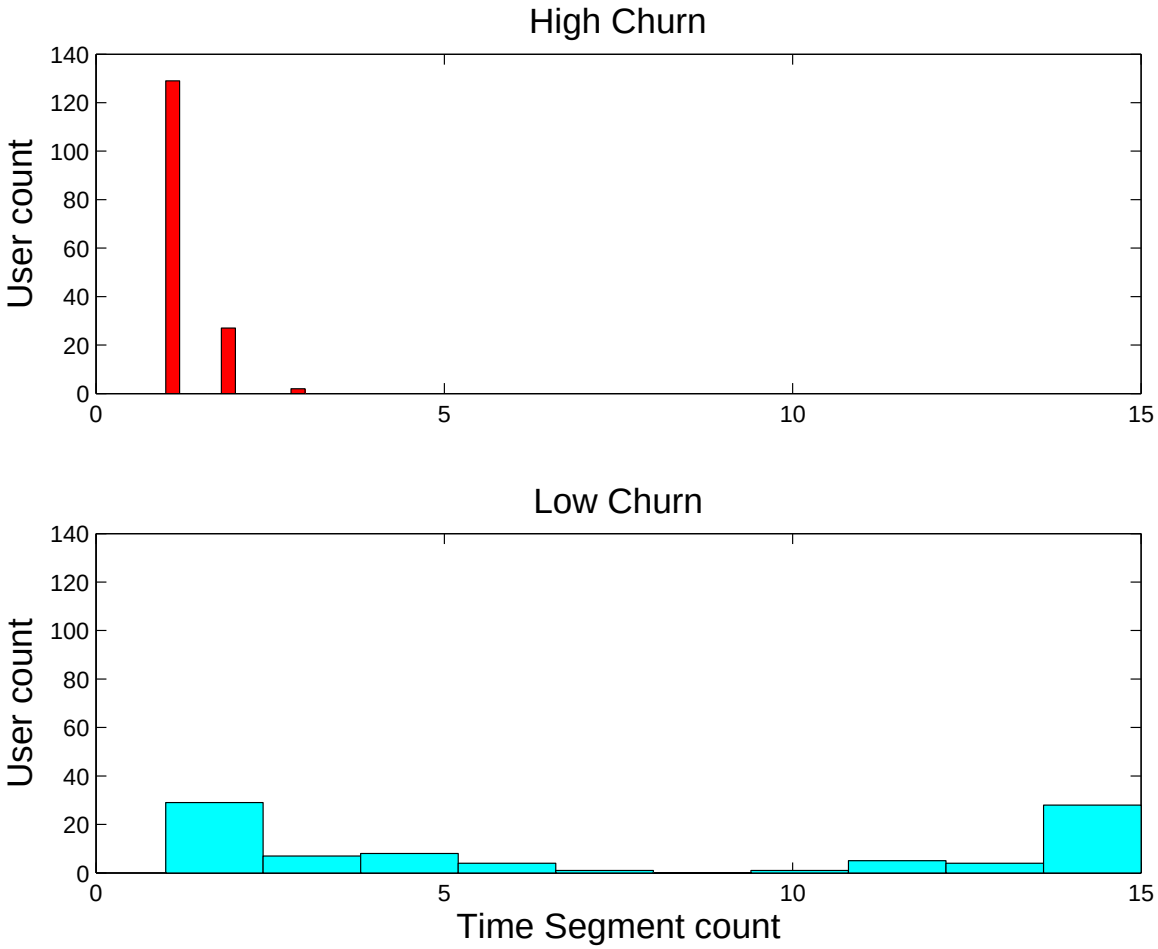


Fig. 40. Core users behavior in low and high churn communities

communities about 20 core tags occur in all 15 time segments (the entire observed period). On the other hand, for high churn communities, no single core tag occurred in all 15 time segments. Generally, we observe, in low churn communities, that more core tags occur over many time segments. In high churn communities we see that core tags occur over fewer time segments.

We present the results for core users in Figure (40). For low churn communities, more than 20 core users occur in 15 time segments while no single core user occurred in more than 3 time segments in high churn communities. Again, we observe that core

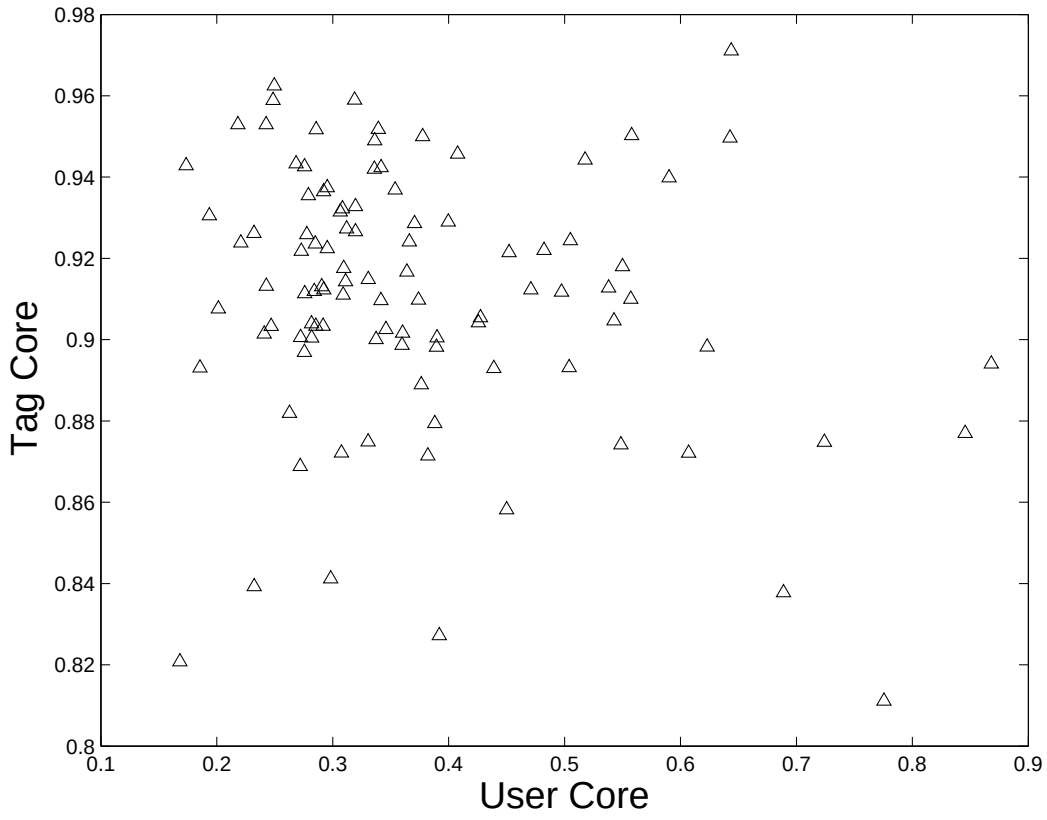


Fig. 41. Core users and core tags' relation to community evolution

users occur over many time segments in low churn communities while majority of core users occur in only one single time segment for the case of high churn communities. Notice that the behavior of core users in high versus low churn communities is more stark than that of core tags.

To illustrate how a community's core tags and core users correlate, we view the communities on the xy-coordinates, with the x-axis representing a community's core user proportion and the y-axis representing its core tag proportion. The results are shown in Figure (41). Notice that all communities have core tag proportions greater that 0.8 while the core user proportion vary widely.

6.4. Community Relationships

In the above sections, we considered individual communities and how they change over time based on their users and tags. Now we look at how communities relate and interact with each other. For example, suppose we are given a community that is interested in the topic "Health". Here we ask, how does this community relate to other communities, for example "Cooking", or "Politics"? What are the closest communities to it, and how does their relationship behave over time?

Our approach focuses on the Joiner and the Parters per community. In Figure (42), we show the "Politics" community and how it relates to other communities over time. For this purpose we consider the users that at some time segment have been members of the "Politics" community. For these users we look at what transitions they make over time, which communities do they transition to when leaving the "Politics" community and which communities do they come from when joining it? The figure shows us other communities that users with interests in "Politics" have shown interest in. Some examples of these communities are "Jobs", "Climate", "Social Media", "Culture", "Finance" and "Health".

We show similar results for the "Health" community in Figure (43). The figure shows examples of communities that users who were members of the "Health" community were also interested in, such as "Politics", "Finance", "Cooking", "Science", "Travel", and "Iphone-dev". In both examples, we see that user transitions across communities is an indicator of some implicit bond between the communities.

7. Summary

Observing the social bookmarking process over time offers many interesting insights. A one month duration can capture the majority of taggers and tags but not resources.

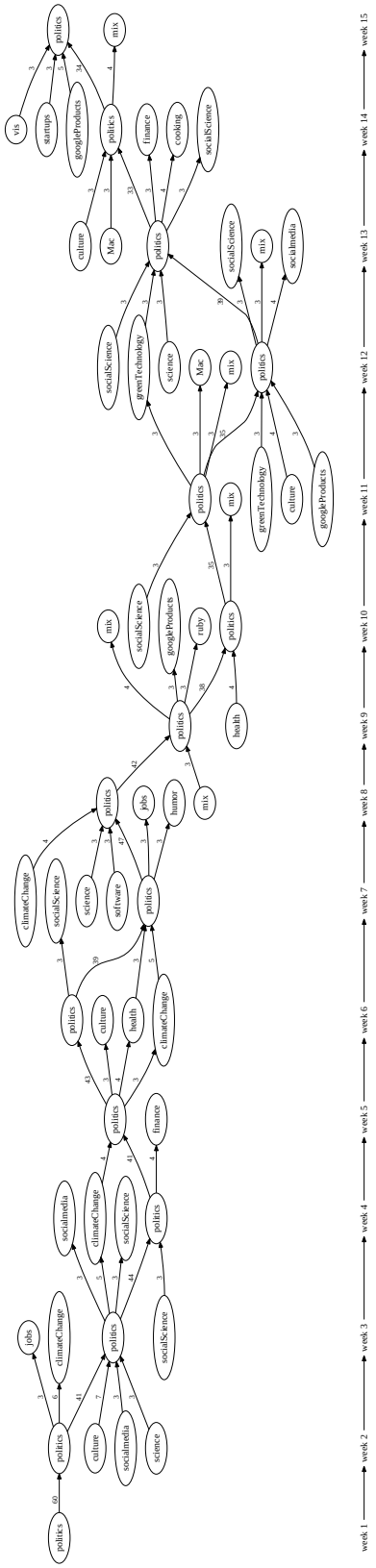


Fig. 42. "Politics" community users transition to/from other communities over time

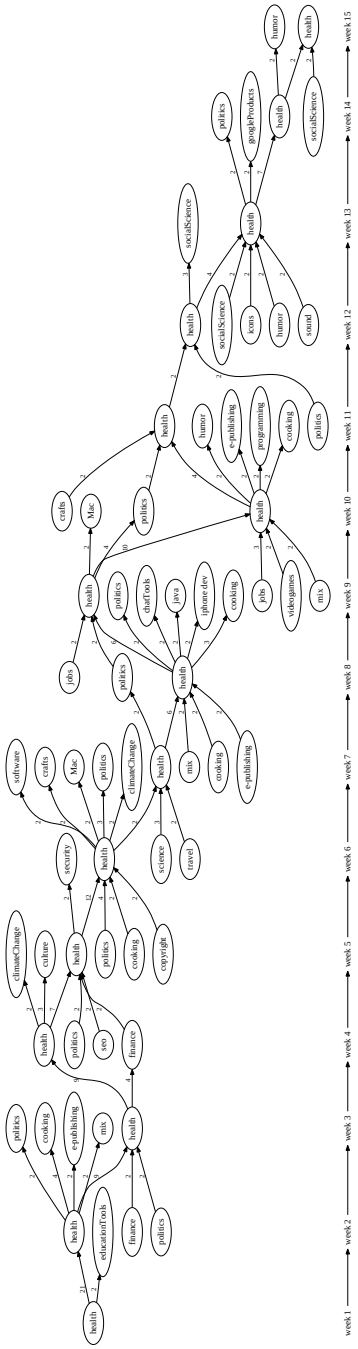


Fig. 43. “Health” community users transitions to/from other communities over time

Resources, taggers, and tag re-occurrences follow the common power law distribution where a few elements are very active and the majority have very low re-occurrences. Co-occurrences among the three types show that a few resources, users and tags are popular while the the majority have minimal exposure. The plots however suggest a sustained daily tagging activity indicating growth in tags, users, and resources.

Extending the SimplePSA approach to model social bookmarking services over time involves segmenting the social tagging document for each resource. The model is then used to uncover structures in each time segment and uses a preceding time segment structure's as a prior to determining the structures of the following time segment. This modeling process allows for the detection of communities and their evolution. A further inspection shows that communities are more dynamic along their user distributions than along their tag distributions.

We have introduced a community dynamics representation of communities and their users and tags transitions that allows us to further inspect the behavior of users and tags in relation to communities. Based on this representation we observe low and high user churn communities. We see that low user churn communities have specialized and narrow interests versus the more generic interests of high user churn communities. We also observe that over time core tags gain more prominence in the community's top tags.

Our examination of how core users and tags behave over time in both low and high user churn communities reveals that: i) core tags and core users in low user churn communities are present in many time segments, ii) core tags and core users in high churn communities are present in only a few time segments. We also examine the correlation between core user proportions and core tag proportions over communities and conclude that no correlation is present. Our illustration of how different communities are related based on user interest transitions over time using the community

dynamics graphs shows that we can identify related communities that are meaningful.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

1. Conclusions

Understanding and modeling the collective semantics centered around large-scale social annotations is a promising research avenue with potential implications for information discovery and knowledge sharing. As a step in this direction, in this dissertation, we study the social bookmarking process itself from a community-oriented perspective, design models that help understand it, and propose applications that can benefit from its rich data. Concretely, we make three unique contributions:

- The first contribution is a pair of probabilistic generative models for describing and modeling the social annotation process. These models posit that the observed tagging information in a social bookmarking system is the product of an underlying community structure, in which users belong to implicit groups of interest. The first model – the Community-based Categorical Annotation (CCA) Model – uncovers this latent structure by identifying tag categories that represent user interests, interpretations, etc. The second model – the Probabilistic Social Annotation (PSA) Model – captures user activity and the connections between users, tags, and documents leading to an improvement in the categories discovered compared to the CCA model. Our experimental results on datasets obtained from the Delicious and CiteUlike social tagging communities show that our models discover more coherent categories of tags and are better suited to handle social bookmarking data compared to existing text-based topic

modeling methods. Additionally these models provide a structure for describing user relationships to other users, to tags, and to resources that can be used to improve information exploration and discovery.

- The second contribution includes two frameworks for web-based information exploration and discovery that are based on our models of the social bookmarking process. In the first framework, we propose a document similarity measure that utilizes the apparent differences between document content and document tags and the underlying structures that produced them. This enables a novel view of documents based on their distributions over their generative hidden structures in the content space and the tag space. In this multi-dimensional view, documents can be classified as similar/dissimilar in both tag and content spaces. Users are then able to find documents with similar content and similar tags, documents with similar content and dissimilar tags, and so on. We illustrate our proposed exploration framework on datasets obtained from the Delicious community and our results show that this multi-dimensional view of documents would enhance data discovery and browsing.

  The second framework we propose makes use of the community structure of users and the categories they use to augment traditional ranking methods for achieving improved discovery and exploration of social web objects. Including the user in the social bookmarking generation process results in groupings of users based on interactions with tags and resources. This allows for the design of similarity measures that make use of user associations within communities and categories. With that in mind, we introduce two approaches for leveraging this community information: (1) ranking by query-community relevance; and (2) ranking by user-community relevance. We compare the results of these

approaches with the state-of-the-art BM25 ranking model as a baseline, as well as with results obtained from existing generative topics models. Our initial experimental results show our models achieve improvement in ranking results over existing topic models which in turn perform better than the standard BM25 model.

- Finally, we turn to time-based analysis of social bookmarking data. By spreading social bookmarking data over a time-line we can compute time-based statistics of tags, resources, and users and their evolution. We can also apply our models to this data and observe the evolution of the system's collective intelligence.

  Observing the social bookmarking process over time offers many interesting insights. A month long observation can capture the majority of taggers and tags but not resources. Resources, taggers, and tag re-occurrences follow the common power law distribution where a few elements are very active and the majority have very low re-occurrences. Co-occurrences among resources, users and tags show that a few resources, users and tags are popular while the the majority have minimal exposure. The sustained daily tagging activity indicates growth in tags, users, and resources.

  Extending the SimplePSA approach to model social bookmarking over time involves segmenting the social tagging document for each resource. The model is able to uncover structure in each time segment and uses a preceding time segment's structure as prior to determining the structure of the following time segment. This modeling process allows the detection of communities and their evolution. A further inspection of the communities and their constituent users and tags dynamics shows that communities are more dynamic along their

user distributions than along their tag distributions.

We introduce a community dynamics representation of communities and their users and tags transitions that allows us to further inspect the behavior of users and tags in relation to communities. Based on this representation we observe low and high user churn communities. We see that low user churn communities have specialized and narrow interests versus the more generic interests of high user churn communities. We also observe that over time core tags gain more prominence in the community's top tags.

Our examination of how core users and tags behave over time in both low and high user churn communities reveals that i) core tags and core users in low user churn communities are present in many time segments, ii) core tags and core users in high churn communities are present in only a few time segments. We also examine the correlation between core user proportions and core tag proportions over communities and conclude that no correlation is present. Our illustration of how different communities are related based on user interest transitions over time using the community dynamics graphs shows that we can identify related communities that are meaningful.

2. Future Work

Although our approach suggest an important role for socially contributed data in advancing information discovery, there are a number of limitations to its application and generalization to social bookmarking systems at large. First, LDA based approaches, in general, including our models require global knowledge and perform many iterations to uncover latent variables. Hence, using them on-line is difficult. Second, our models, as does LDA, assume a fixed number of latent variables. Third, our

assumption of global user communities does not capture individual user behavior. In addition, the lack of standard corpora for social bookmarking data makes evaluating and comparing results of different research methods difficult. And, results based on human judges of individually collected corpora need to be verified for generalization to different social bookmarking systems.

However, some of these limitations can be overcome. A combination of a both on-line and off-line approach can solve the processing requirements of LDA-based models. Also, there are methods for dynamically discovering the number of latent variables (see for example [93]). Finally, these are some tasks that we would like to further investigate in future research work:

- Consider more fine-grained hierarchical models of the social annotation process.
- Construct individual user models in addition to the global user communities.
- Expand the integrated browsing model and verify it over standard datasets.
- Extend the scope of experimental validation to other social bookmarking communities.
- Investigate alternative temporal modeling approaches of social bookmarking data with different granularity, segmenting approaches, localization and personalization.
- Infuse topic modeling approaches with graph-based and other approaches that include metadata such as ratings, and anchor text.

REFERENCES

[1] "Twitter," December 2010. [Online]. Available: http://twitter.com/

[2] P. Leitner and T. Grechenig, "Collaborative shopping networks: Sharing the wisdom of crowds in e-commerce environments," in *21st Bled eConference*, 2008, pp. 321 – 335.

[3] T. Willard, *Social Networking and Governance for Sustainable Development*, 2009. [Online]. Available: http://www.iisd.org/pdf/2009/social_net_gov.pdf

[4] "Digital journal," December 2010. [Online]. Available: http://www.digitaljournal.com/

[5] B. Bercovitz, F. Kaliszan, G. Koutrika, H. Liou, Z. Mohammadi Zadeh, and H. Garcia-Molina, "Courserank: A social system for course planning," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2009, pp. 1107–1110.

[6] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the International Conference on World Wide Web*, 2007, pp. 501–510.

[7] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su, "Towards effective browsing of large scale social annotations," in *Proceedings of the International Conference on World Wide Web*, 2007, pp. 943–952.

[8] C. H. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," in *Proceedings of the International Conference on World Wide Web*, 2006, pp. 625–632.

[9] G. Macgregor and E. McCulloch, "Collaborative tagging as a knowledge organisation and resource discovery tool," *Library Review*, vol. 55, no. 5, pp. 291–300, 2006.

[10] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Ht06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, 2006, pp. 31–40.

[11] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd, "Evaluating similarity measures for emergent semantics of social tagging," in *Proceedings of the International Conference on World Wide Web*, 2009, pp. 641–650.

[12] X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," in *Proceedings of the International Conference on World Wide Web*, 2006, pp. 417–426.

[13] M. G. Noll and C. Meinel, "Exploring social annotations for web document classification," in *Proceedings of the ACM Symposium on Applied Computing*, 2008, pp. 2315–2320.

[14] P. Mika, "Ontologies are us: A unified model of social networks and semantics," in *International Semantic Web Conference*, ser. Lecture Notes in Computer Science, vol. 3729. International Semantic Web Conference 2005, November 2005, pp. 522–536.

[15] S. Christiaens, "Metadata mechanisms: From ontology to folksonomy ... and back," in *OTM Workshops*, ser. Lecture Notes in Computer Science, 2006, pp. 199–207.

[16] T. Gruber, "Collective knowledge systems: Where the social web meets the

semantic web," in *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, February 2008, pp. 4–13.

[17] U. Farooq, Y. Song, J. M. Carroll, and C. L. Giles, "Social bookmarking for scholarly digital libraries," *IEEE Internet Computing*, vol. 11, no. 6, pp. 29–35, 2007.

[18] S. Sen, J. Vig, and J. Riedl, "Learning to recognize valuable tags," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, ser. IUI '09, 2009, pp. 87–96.

[19] B. Liu, E. Zhai, H. Sun, Y. Chen, and Z. Chen, "Filtering spam in social tagging system with dynamic behavior analysis," in *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, 2009, pp. 95–100.

[20] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, ser. AIRWeb '09, 2009, pp. 41–48.

[21] N. Neubauer, R. Wetzker, and K. Obermayer, "Tag spam creates large non-giant connected components," in *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, ser. AIRWeb '09, 2009, pp. 49–52.

[22] M. J. Carman, M. Baillie, R. Gwadera, and F. Crestani, "A statistical comparison of tag and query logs," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 123–130.

[23] S. Golder and B. A. Huberman, "The structure of collaborative tagging systems," August 2005. [Online]. Available: http://arxiv.org/abs/cs/0508082

[24] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proceedings of the International Conference on World Wide Web*, 2007, pp. 211–220.

[25] C. Cattuto, V. Loreto, and L. Pietronero, "Collaborative tagging and semiotic dynamics," May 2006. [Online]. Available: http://arxiv.org/abs/cs.CY/0605015

[26] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto, "Vocabulary growth in collaborative tagging systems," April 2007. [Online]. Available: http://arxiv.org/abs/0704.3316

[27] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in *Proceedings of the International Conference on World Wide Web*, 2008, pp. 675–684.

[28] C. Veres, "The language of folksonomies: What tags reveal about user classification," *Natural Language Processing and Information Systems*, vol. 3999/2006, pp. 58–69, 2006.

[29] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles, "Exploring social annotations for information retrieval," in *Proceedings of the International Conference on World Wide Web*, 2008, pp. 715–724.

[30] D. Ramage, P. Heymann, C. D. Manning, and H. G. Molina, "Clustering the tagged web," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2009, pp. 54–63.

[31] A. Plangrasopchok and K. Lerman, "Exploiting social annotation for automatic resource discovery," April 2007. [Online]. Available: http:

//arxiv.org/abs/0704.1675

[32] A. Plangprasopchok and K. Lerman, "Modeling social annotation: A bayesian approach," November 2008. [Online]. Available: http://arxiv.org/abs/0811.1319

[33] G. Begelman, P. Keller, and F. Smadja, "Automated tag clustering: Improving search and exploration in the tag space," in *Proceedings of the International Conference on World Wide Web*, 2006.

[34] S. Kolay and A. Dasdan, "The value of socially tagged urls for a search engine," in *Proceedings of the International Conference on World Wide Web*, 2009, pp. 1203–1204.

[35] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Can social bookmarking enhance search in the web?" in *Proceedings of the Joint Conference on Digital Libraries*, 2007, pp. 107–116.

[36] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2008, pp. 195–206.

[37] R. Wetzker, T. Plumbaum, A. Korth, C. Bauckhage, T. Alpcan, and F. Metze, "Detecting trends in social bookmarking systems using a probabilistic generative model and smoothing," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[38] M. J. Carman, M. Baillie, and F. Crestani, "Tag data and personalized information retrieval," in *Proceedings of the 2008 ACM Workshop on Search in Social Media*, ser. SSM '08, 2008, pp. 27–34.

[39] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang, "Personalized tag recommendation using graph-based ranking on multi-type interrelated objects," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 540–547.

[40] S. Bateman, M. J. Muller, and J. Freyne, "Personalized retrieval in social bookmarking," in *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, ser. GROUP '09, 2009, pp. 91–94.

[41] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 155–162.

[42] J. Wang and B. D. Davison, "Explorations in tag suggestion and query expansion," in *Proceedings of the 2008 ACM Workshop on Search in Social Media*, ser. SSM '08, 2008, pp. 43–50.

[43] J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining recommendations using tags," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, ser. IUI '09, 2009, pp. 47–56.

[44] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles, "Real-time automatic tag recommendation," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.

[45] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[46] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," *Information Pro-*

cessing & Management, vol. 36, no. 6, pp. 779–808, 2000.

[47] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 275–281.

[48] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, pp. 333–389, April 2009.

[49] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 186–193.

[50] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 178–185.

[51] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Computer Networks and ISDN Systems*, 1998, pp. 107–117.

[52] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[53] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the International Conference on World Wide Web*, May 2002, pp. 517–526.

[54] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 459–460.

[55] M. Richardson and P. Domingos, "The intelligent surfer: Probabilistic combination of link and content information in pagerank," in *Advances in Neural Information Processing Systems*, 2001, pp. 1441–1448.

[56] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.

[57] S. Santini, A. Gupta, and R. Jain, "Emergent semantics through interaction in image databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, pp. 337–351, May 2001.

[58] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*, 2007, vol. 4321, ch. 9, pp. 291–324.

[59] J. Golbeck and J. Hendler, "Filmtrust: Movie recommendations using trust in web-based social networks," in *IEEE Consumer Communications and Networking Conference*, vol. 1, 2006, pp. 282–286.

[60] 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*, vol. 41, no. 6, pp. 391–407, 1990.

[61] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[62] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Journal of Machine Learning Research*, vol. 3, April 2003, pp. 993–1022.

[63] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Latent topic models for hypertext," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2008, pp. 230–239.

[64] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "The author-recipient-topic model for topic and role discovery in social networks, with application to Enron and academic email," in *Workshop on Link Analysis, Counterterrorism and Security*, 2005, pp. 33–44.

[65] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2003, pp. 352–359.

[66] G. Heinrich, "Parameter estimation for text analysis," Tech. Rep., 2004. [Online]. Available: http://www.arbylon.net/publications/text-est.pdf

[67] R. Albert and A. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[68] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 7821–7826, 2002.

[69] R. Albert, H. Jeong, and A.-L. Barabasi, "The diameter of the world wide web," *Nature*, vol. 401, pp. 130–131, 1999.

[70] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, no. 393, pp. 440–442, 1998.

[71] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *ACM SIGCOMM Computer Communication Review*,

1999, pp. 251–262.

[72] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 066111, 2004.

[73] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2001, pp. 849–856.

[74] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, June 2006.

[75] A. Java, A. Joshi, and T. Finin, "Detecting commmunities via simultaneous clustering of graphs and folksonomies," in *Proceedings of the Tenth Workshop on Web Mining and Web Usage Analysis*, August 2008, pp. 1–+.

[76] J. Ruan and W. Zhang, "An efficient spectral algorithm for network community discovery and its applications to biological and social networks," in *ICDM '07*, 2007, pp. 643–648.

[77] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. E. J. Newman, "Random hypergraphs and their applications," *Physical Review E*, vol. 79, no. 6, pp. 066 118–+, 2009.

[78] T. Murata, "Modularity for heterogeneous networks," in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, 2010, pp. 129–134.

[79] V. Zlatić, G. Ghoshal, and G. Caldarelli, "Hypergraph topological quantities for tagged social networks," *Physical Review E*, vol. 80, no. 3, p. 036118, September 2009.

[80] S. Kashoob, J. Caverlee, and Y. Ding, "A categorical model for discovering latent structure in social annotations," in *International Conference on Weblogs and Social Media*, 2009, pp. 1–+.

[81] S. Kashoob, J. Caverlee, and E. Khabiri, "Probabilistic generative models of the social annotation process," in *Proceedings of the IEEE International Conference on Computational Science and Engineering*, 2009, pp. 42–49.

[82] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.

[83] W. Li and A. Mccallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the International Conference on Machine Learning*, 2006, pp. 577–584.

[84] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[85] T. P. Minka, "Estimating a dirichlet distribution," 2003. [Online]. Available: http://research.microsoft.com/~minka

[86] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2008.

[87] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Can social bookmarking enhance search in the web?" in *Proceedings of the Joint Conference on Digital Libraries*, 2007, pp. 107–116.

[88] S. Kashoob, J. Caverlee, and K. Y. Kamath, "Community-based ranking of the social web," in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, 2010, pp. 141–150.

[89] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[90] D. Coppersmith, L. Fleischer, and A. Rudra, "Ordering by weighted number of wins gives a good ranking for weighted tournaments," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2006, pp. 776–782.

[91] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[92] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[93] 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*, 2004, pp. 17–24.

APPENDIX A

DERIVING THE GIBBS SAMPLING EQUATION FOR LDA MODEL

Using the LDA model we factor the joint probability distribution as follows:

$$p(\mathbf{S}, \mathbf{z}|\alpha, \beta) = p(\mathbf{S}|\mathbf{z}, \beta)p(\mathbf{z}|\alpha).$$

We can now derive the terms on the right one at a time. The first term $p(\mathbf{S}|\mathbf{z}, \beta)$ can be derived from a multinomial on the observed word counts given the topics:

$$p(\mathbf{S}|\mathbf{z}, \Phi) = \prod_{i=1}^{N} p(w_i|z_i) = \prod_{i=1}^{N} \phi_{z_i,w_i}$$

Splitting the product into a product over word vocabulary and another over topics we get:

$$p(\mathbf{S}|\mathbf{z}, \Phi) = \prod_{k=1}^{K} \prod_{w=1}^{V} (\phi_{k,w})^{(n_k^w)}$$

Where $(n_k^w)$is a count of the times word w is seen in topic k. We then obtain the $p(\mathbf{S}|\mathbf{z}, \beta)$ by integrating out Φ

$$\begin{aligned} p(\mathbf{S}|\mathbf{z}, \beta) &= \int p(\mathbf{S}|\mathbf{z}, \Phi)p(\Phi|\beta)d\Phi \\ &= \int \prod_{z=1}^{K} \frac{1}{\Delta(\beta)} \prod_{w=1}^{V} (\phi_{z,w})^{(n_k^w)+\beta_w-1} d\phi_z \\ &= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z + \beta)}{\Delta(\beta)} \end{aligned}$$

where

$$\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^{dim\vec{\alpha}} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{dim\vec{\alpha}} \alpha_k)},$$

is the Dirichlet delta function, Γ is the Gamma function, and

$$\vec{n}_z = \{n_z^w\}_{w=1}^{V}$$

is vector counting occurrences of words w with topic z.

Similarly, We derive $p(\mathbf{z}|\alpha)$ from

$$\begin{aligned} p(\mathbf{z}|\Theta) &= \prod_{i=1}^{N} p(z_i|d_i) = \prod_{m=1}^{M}\prod_{k=1}^{K} p(z_i = k|d_i = m) \\ &= \prod_{m=1}^{M}\prod_{k=1}^{K} (\theta_{m,k})^{n_m^k} \end{aligned}$$

and by integrating over Θ we get:

$$\begin{aligned} p(\mathbf{S}|\alpha) &= \int p(\mathbf{S}|\Theta)p(\Theta|\alpha)d\Theta \\ &= \int \prod_{m=1}^{M} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} (\theta_{m,k})^{n_m^k+\alpha_k-1} d\theta_m \\ &= \prod_{m=1}^{D} \frac{\Delta(\vec{n}_m + \alpha)}{\Delta(\alpha)}. \text{ Where } \vec{n}_m = \{n_m^k\}_{k=1}^{K}. \end{aligned}$$

The joint probability distribution is then:

$$p(\mathbf{S}, \mathbf{z}|\alpha, \beta) = \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z + \beta)}{\Delta(\beta)} \times \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \alpha)}{\Delta(\alpha)}$$

Using the joint probability distribution above, we can derive the update equation for the Gibbs sampler as follows:

$$\begin{aligned} p(z_i = k|\mathbf{z}_{\neg i}, \mathbf{S}) &= \frac{p(\mathbf{z}, \mathbf{S})}{p(\mathbf{S}, \mathbf{z}_{\neg i})} = \frac{p(\mathbf{S}|\mathbf{z})}{p(\mathbf{S}_{\neg i}|\mathbf{z}_{\neg i})p(\mathbf{S}_i)} \times \frac{p(\mathbf{z})}{p(\mathbf{z}_{\neg i})} \\ \propto & \frac{\Delta(\vec{n}_z + \beta)}{\Delta(\vec{n}_{z,\neg i} + \beta)} \frac{\Delta(\vec{n}_m + \alpha)}{\Delta(\vec{n}_{m,\neg i} + \alpha)} \\ \propto & \frac{\Gamma(n_k^w + \beta_w)\Gamma(\sum_{w=1}^{V} n_{k,\neg i}^w + \beta_t)}{\Gamma(n_{k,\neg i}^w + \beta_t)\Gamma(\sum_{w=1}^{V} n_k^w + \beta_t)} \frac{\Gamma(n_m^k + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^k + \alpha_k)}{\Gamma(n_{k,\neg i}^w + \alpha_k)\Gamma(\sum_{k=1}^{K} n_m^k + \alpha_k)} \\ \propto & \frac{n_{k,\neg i}^w + \beta_u}{\sum_{w=1}^{V} n_{k,\neg i}^w + \beta_w} \times \frac{n_{d,\neg i}^k + \alpha_k}{\sum_{k=1}^{K} n_{d,\neg i}^k + \alpha_k} \end{aligned}$$

where $n_{(\cdot),\neg i}^{(\cdot)}$ is a count excluding the current position assignments of $\mathbf{z}_i$ (e.g., $n_{k,\neg i}^w$ is

the count of word w generated by the k-th topic excluding the current position).

APPENDIX B

DERIVING THE GIBBS SAMPLING EQUATION FOR PSA MODEL

Using the PSA model we factor the joint probability distribution as follows:

$$p(S^u, S^t, z, c|\alpha, \beta, \gamma, \delta) = p(S^u|c, \delta)p(c|\alpha)p(S^t|z, c, \gamma)p(z|c, \beta).$$

We can now derive the terms on the right one at a time. The first term $p(S^u|c, \delta)$ can be derived from a multinomial on the observed user counts given the communities:

$$p(S^u|c, \tau) = \prod_{i=1}^{N} p(u_i|c_i) = \prod_{i=1}^{N} \tau_{c_i,u_i}$$

Splitting the product into a product over communities and another over users we get:

$$p(S^u|c, \tau) = \prod_{l=1}^{L}\prod_{u=1}^{U}(\tau_{l,u})^{(n_l^u)}$$

Where $(n_l^u)$is a count of the times user u is seen in community l. We then obtain the $p(S^u|c, \delta)$ by integrating out τ

$$\begin{aligned} p(S^u|c, \delta) &= \int p(S^u|c, \tau)p(\tau|\delta)d\tau \\ &= \int \prod_{l=1}^{L} \frac{1}{\Delta(\delta)} \prod_{u=1}^{U}(\tau_{l,u})^{(n_l^u+\delta_u-1)}d\tau \\ &= \prod_{l=1}^{L} \frac{\Delta(\vec{n}_l + \delta)}{\Delta(\delta)} \end{aligned}$$

where

$$\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^{dim\vec{\alpha}} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{dim\vec{\alpha}} \alpha_k)},$$

is the Dirichlet delta function, Γ is the Gamma function, and

$$\vec{n}_l = \{n_l^u\}_{u=1}^{U}$$

is vector counting occurrences of user u with community l.

Similarly, We derive $p(c|\alpha)$ from

$$\begin{aligned} p(c|\kappa) &= \prod_{i=1}^{N} p(c_i|d_i) = \prod_{m=1}^{D}\prod_{l=1}^{L} p(c_i = l|d_i = m) \\ &= \prod_{m=1}^{D}\prod_{l=1}^{L} (\kappa_{m,l})^{n_m^l} \end{aligned}$$

and by integrating over κ we get:

$$\begin{aligned} p(c|\alpha) &= \int p(c|\kappa)p(\kappa|\alpha)d\kappa \\ &= \int \prod_{m=1}^{D} \frac{1}{\Delta(\alpha)} \prod_{l=1}^{L} (\kappa_{m,l})^{n_m^l+\alpha_l-1} d\kappa_m \\ &= \prod_{m=1}^{D} \frac{\Delta(\vec{n}_m + \alpha)}{\Delta(\alpha)}. \text{ Where } \vec{n}_m = \{n_m^l\}_{l=1}^{L}. \end{aligned}$$

And the term $p(S^t|z, c, \gamma)$ from

$$\begin{aligned} p(S^t|c, z, \Phi) &= \prod_{i=1}^{N} \Phi_{c_i,z_i,t_i} \\ &= \prod_{l=1}^{l}\prod_{k=1}^{K_l} \prod_{i:c_i=l,z_i=k} p(t_i = t|c_i = l, z_i = k) \\ &= \prod_{l=1}^{L}\prod_{k=1}^{K_l}\prod_{t=1}^{V} (\Phi_{l,k,t})^{n_{lk}^t} \end{aligned}$$

and by integrating over Φ we get:

$$\begin{aligned} p(S^t|c, z, \gamma) &= \int p(S^t|c, z\Phi)p(\Phi|\gamma)d\Phi \\ &= \int \prod_{l=1}^{L}\prod_{k=1}^{K_l} \frac{1}{\Delta(\gamma)} \prod_{t=1}^{V} (\Phi_{l,k,t})^{n_{lk}^t+\gamma_t-1} d\Phi_{lk} \\ &= \prod_{l=1}^{L}\prod_{k=1}^{K_l} \frac{\Delta(\vec{n}_{lk} + \gamma)}{\Delta(\gamma)}. \text{ Where } \vec{n}_{lk} = \{n_{lk}^t\}_{t=1}^{V}. \end{aligned}$$

Finally, the term $p(z|c,\beta)$ from

$$\begin{aligned} p(z|c,\Theta) &= \prod_{i=1}^{N} p(z_i|c_i,d_i) \\ &= \prod_{m=1}^{D}\prod_{l=1}^{l}\prod_{k=1}^{K_l} p(z_i=k|c_i=l,d_i=m) \\ &= \prod_{m=1}^{D}\prod_{l=1}^{l}\prod_{k=1}^{K_l} (\Theta_{m,l,k})^{n_m^{lk}} \end{aligned}$$

and by integrating over Θ we get:

$$\begin{aligned} p(z|c,\beta) &= \int \prod_{m=1}^{D}\prod_{l=1}^{L}\prod_{k=1}^{K_l} \frac{1}{\Delta(\beta)} (\Theta_{m,l,k})^{n_m^{lk}+\beta_{lk}-1} d\Theta_m \\ &= \prod_{m=1}^{D}\prod_{l=1}^{L} \frac{\Delta(\vec{n}_m+\beta)}{\Delta(\beta)}. \text{ Where } \vec{n}_m = \{n_m^{lk}\}_{k=1}^{K_l}. \end{aligned}$$

The joint probability distribution is then:

$$\begin{aligned} p(S^u,S^t,c,z|\alpha,\beta,\gamma,\delta) &= \prod_{l=1}^{L} \frac{\Delta(\vec{n}_l+\delta)}{\Delta(\delta)} \prod_{l=1}^{L}\prod_{k=1}^{K_l} \frac{\Delta(\vec{n}_{lk}+\gamma)}{\Delta(\gamma)} \\ &\quad \prod_{m=1}^{D}\prod_{l=1}^{L} \frac{\Delta(\vec{n}_m+\beta)}{\Delta(\beta)} \prod_{m=1}^{D} \frac{\Delta(\vec{n}_m+\alpha)}{\Delta(\alpha)}. \end{aligned}$$

Using the joint probability distributed derived above, we can derive the update

equation for the Gibbs sampler as follows:

$$
\begin{aligned}
& p(z_i = k, c_i = l | \mathbf{z}_{\neg i}, \mathbf{c}_{\neg i}, \mathbf{t}, \mathbf{u}) = \frac{p(\mathbf{z}, \mathbf{c}, \mathbf{t}, \mathbf{u})}{p(\mathbf{z}_{\neg i}, \mathbf{c}_{\neg i}, \mathbf{t}, \mathbf{u})} \\
= & \frac{p(\mathbf{u}|\mathbf{c})}{p(\mathbf{u}_{\neg i}|\mathbf{c}_{\neg i})p(\mathbf{u}_i)} \frac{p(\mathbf{t}|\mathbf{z}, \mathbf{c})}{p(\mathbf{t}_{\neg i}|\mathbf{z}_{\neg i}, \mathbf{c}_{\neg i})p(\mathbf{t}_i)} \frac{p(\mathbf{z}|\mathbf{c})}{p(\mathbf{z}_{\neg i}|\mathbf{c}_{\neg i})} \frac{p(\mathbf{c})}{p(\mathbf{c}_{\neg i})} \\
\propto & \frac{\Delta(\vec{n}_l + \delta)}{\Delta(\vec{n}_{l,\neg i} + \delta)} \frac{\Delta(\vec{n}_{lk} + \gamma)}{\Delta(\vec{n}_{lk,\neg i} + \gamma)} \frac{\Delta(\vec{n}_S + \beta)}{\Delta(\vec{n}_{S,\neg i} + \beta)} \frac{\Delta(\vec{n}_S + \alpha)}{\Delta(\vec{n}_{S,\neg i} + \alpha)} \\
\propto & \frac{\Gamma(n^u_l + \delta_u)\Gamma(\sum_{u=1}^{U} n^u_{l,\neg i} + \delta_u)}{\Gamma(n^u_{l,\neg i} + \delta_u)\Gamma(\sum_{u=1}^{U} n^u_l + \delta_u)} \frac{\Gamma(n^t_{lk} + \gamma_t)\Gamma(\sum_{t=1}^{V} n^t_{lk,\neg i} + \gamma_t)}{\Gamma(n^t_{lk,\neg i} + \gamma_t)\Gamma(\sum_{t=1}^{v} n^t_{lk} + \gamma_t)} \\
& \frac{\Gamma(n^{lk}_S + \beta_{lk})\Gamma(\sum_{k=1}^{K_l} n^{lk}_{S,\neg i} + \beta_{lk})}{\Gamma(n^{lk}_{S,\neg i} + \beta_{lk})\Gamma(\sum_{k=1}^{K_l} n^{lk}_S + \beta_{lk})} \frac{\Gamma(n^l_S + \alpha_l)\Gamma(\sum_{l=1}^{L} n^l_{S,\neg i} + \alpha_l)}{\Gamma(n^l_{S,\neg i} + \alpha_l)\Gamma(\sum_{l=1}^{L} n^l_S + \alpha_l)} \\
\propto & \frac{n^u_{l,\neg i} + \delta_u}{\sum_{u=1}^{U} n^u_{l,\neg i} + \delta_u} \times \frac{n^t_{lk,\neg i} + \gamma_t}{\sum_{t=1}^{V} n^t_{lk,\neg i} + \gamma_t} \times \frac{n^{lk}_{S,\neg i} + \beta_{lk}}{\left(\sum_{k=1}^{K_l} n^{lk}_S + \beta_{lk}\right) - 1} \times \frac{n^l_{S,\neg i} + \alpha_l}{\left(\sum_{l=1}^{L} n^l_S + \alpha_l\right) - 1}
\end{aligned}
$$

where $n^{(\cdot)}_{(\cdot),\neg i}$ is a count excluding the current position assignments of z_i and c_i (e.g., $n^t_{lk,\neg i}$ is the count of tag t generated by the k-th category of the l-th community excluding the current position).

VITA

Said Masoud Ali Kashoob received his Bachelor of Science degree in computer engineering from Iowa State University in 2000. He later received his Master of Science degree in computer science from the Royal Institute of Technology(KTH), Sweden in 2004. In 2011, he completed his doctoral studies in the Computer Science and Engineering Department at Texas A&M University in College Station, Texas under the supervision of Dr. James Caverlee.

Said's research interests have included information retrieval and discovery, statistical analysis, topic models, and social networks. Currently, he is a faculty member at the College of Technology in Salalah, Oman and can be reached by email at skashoob@gmail.com.