

# Individual and Social Behavior in Tagging Systems

Elizeu Santos-Neto<sup>§</sup>, David Condon<sup>⊕</sup>, Nazareno Andrade<sup>+§</sup>, Adriana Iamnitchi<sup>⊕</sup>, Matei Ripeanu<sup>§</sup>

Electrical & Computer Engineer<sup>§</sup>  
University of British Columbia  
2332 Mail Mall – KAIS 4075  
Vancouver, BC, Canada  
+1.604.827.4270  
{elizeus,matei}@ece.ubc.ca

Computer Science & Engineering<sup>⊕</sup>  
University of South Florida  
4202 E. Fowler Ave  
Tampa, FL, USA  
+1.813.974.5357  
{dcondon,anda}@cse.usf.edu

Lab. de Sistemas Distribuídos<sup>+</sup>  
Univ. Fed. de Campina Grande  
Av. Aprígio Veloso, 882  
Campina Grande, PB, Brazil  
+55.83.3310.1365  
nazareno@isd.ufcg.edu.br

## ABSTRACT

In tagging systems users can annotate items of interest with free-form terms. A good understanding of the usage characteristics of such systems is necessary to improve the design of current and next generation tagging systems. To this end, this work explores three aspects of user behavior in *CiteULike* and *Connotea*, two systems that include tagging features to support online personalized management of scientific publications. First, this study characterizes the degree to which users re-tag previously published items and reuse tags: 10 to 20% of the daily activity can be characterized as re-tagging and about 75% of the activity as tag reuse. Second, we use the pairwise similarity between users' activity to characterize the interest sharing in these systems. We present the interest sharing distribution across the systems, show that this metric encodes information about existing usage patterns, and attempt to correlate interest sharing levels to indicators of collaboration such as co-membership in discussion groups and semantic similarity of tag vocabularies. Finally, we show that interest sharing leads to an implicit structure that exhibits a natural segmentation. Throughout the paper we discuss the potential impact of our findings on the design of mechanisms that support tagging systems.

## Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software – *information networks*; H.3.5 [Information Storage and Retrieval] Online Information Services – *data sharing*.

## General Terms

Measurement, Design, Experimentation.

## Keywords

Tagging, tag reuse, interest sharing, random null model.

## 1. INTRODUCTION

A *tagging system* allows *users* to associate *tags* to *items*. Such *tagging* feature is commonly found in web-based content-sharing systems (e.g., *Flickr*), social bookmarking systems (e.g., *del.icio.us*, *CiteULike*) and online social networks (e.g.,

*Facebook*). In these systems users generally publish new items, annotate them with tags, and browse and search content via a website. The assignment of a tag to an item is generally referred to as a *tag assignment*.

Tagging is recently recognized for its potential to leverage collaborative production of information that support a wide range of mechanisms such as social search [1], and recommendation [2], although tagging was originally thought as a technique to improve personal content management. A necessary step towards realizing the full potential of tagging for mechanism design is to understand the characteristics of user activity in (collaborative) tagging systems. However, as tagging is a relatively new phenomenon, user behavior characteristics at the individual and social level are relatively unknown.

This paper contributes to addressing this gap and complements previous characterization studies [3-5] by focusing on three novel aspects: *i) item re-tagging*, a measure of the degree to which users re-tag items that are already in the system; *ii) tag reuse*, a measure for the degree to which users reuse a tag to perform new annotations; and, *iii) interest sharing*, a measure for the similarity between users with respect to their tagging activity. In particular, this work answers the following questions:

Q1. *What are the levels of item re-tagging and tag reuse?*

Q2. *What are the characteristics of activity similarity (i.e., interest sharing) among users in these systems?*

Q3. *Does interest sharing relate to other indicators of social behavior (e.g., participation in the same discussion group)?*

Q4. *What are the topology characteristics of the implicit social structure inferred from interest sharing between users?*

The characteristics revealed by our study have practical implications for the design of mechanisms that rely on implicit user interactions such as collaborative search [1, 6]. In particular, our analysis presents evidence of *relatively low item re-tagging* and a much *higher level of tag reuse*. Additionally, the characterization of interest sharing among users shows that both item-based and tag-based interest sharing are concentrated over a small set of user pairs.

Moreover, we show that the observed interest sharing distribution deviates significantly in both concentration and intensity from that of a Randomized Null Model (RNM) that preserves the macro characteristics of the tagging activity of the systems studied but uses random tagging assignments. This deviation underscores the existence of latent usage patterns that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'09, June 29–July 1, 2009, Torino, Italy.

Copyright 2009 978-1-60558-486-7/09/06...\$5.00.

may offer support for optimizing mechanisms such as personalized search, reputation, and spam detection.

Additionally, we seek to understand whether user similarity based on items and tags relates to other indicators of collaboration such as co-participation to discussion groups, and the semantic similarity of their tag vocabularies. We observe that item-based interest sharing does not correlate with explicit social behavior such as co-membership in *CiteULike* groups. We believe that these findings have practical implications on the prediction of social behavior based on implicit similarity.

Finally, we use the interest sharing between users to infer an implicit social structure for the tagging community, which we call *interest-sharing graph*, where users are connected if they have tagged the same item or used the same tags. We find that users self-organize in three distinct regions: *singletons* – users with low activity and unique preferences for items and tags; *small components* – a subset of users with high similarity among them, but isolated from the rest of the system; and a giant component – where most of the activity is concentrated and mixed levels of interest sharing is observed.

This paper is structured as follows. The following section reviews related work. Section 3 describes the traces used in this study. Section 4 presents the characterization of individual user behavior by analyzing the item re-tagging and tag reuse in two tagging systems. Section 5 contains the analysis of implicit user relationships and the characterization of interest sharing across user pairs in the system. Section 6 investigates the correlation between interest sharing and other indicators of collaboration. In Section 7, we present an analysis of the implicit social structure based on the interest sharing relation and Section 8 concludes.

## 2. RELATED WORK

To position our work in the context of related literature, this section is organized along three main topics: First, it presents previous *characterization studies focused on tagging systems*; second, it surveys projects that harness in various forms the concept of *interest sharing*; and, finally, it briefly surveys the *design of system support mechanisms* (e.g., search and recommendation) that could benefit from tagging information.

**Characterization studies:** Given the increasing popularity of tagging systems, the research community started to study their usage patterns and to propose models that explain and predict user behavior in these systems. Golder and Huberman [3] analyze *del.icio.us* – a social bookmarking tool that allows users to share and tag URLs. They show that the relative proportion of each tag occurrence in a given item converges over time and this pattern can be modeled by the Eggenberger-Polya’s urn model.

Along a similar line, Cattuto et al. [7] study tag co-occurrence patterns to understand properties of tag association. They observe that the distribution of tag co-occurrence is similar to a power law and propose a process that models the observed tag co-occurrence patterns based on the Yule-Simon’s stochastic process [8].

Other related work concentrates on the characterization of the utility of tagging. Chi and Mytkowicz [4], for example, focus on understanding the impact of user population and content growth on navigability in *del.icio.us*. They use an information theoretical framework to quantify the navigability of a tagging

system. They conclude that, as the user population grows, it becomes harder to find content – thus, navigability is impaired, because tags lose their informational value (i.e., the ability to narrow down a list of relevant items). Similarly, but using a different approach, Suchanek et al. [9] study how meaningful tags are. They quantify the semantic properties of tags and provide a model that enables an evaluation of how much tag suggestions influence tagging behavior.

Our study complements these previous studies and provides a new take on the characterization of tagging systems. This study focuses on new questions that reveal unexplored characteristics of user behavior at individual level (item re-tagging and tag reuse) and at social level (interest sharing between users) in tagging systems.

**Interest sharing analysis.** A graph-centric approach is an alternative way of characterizing tagging systems, where users are connected by edges based on the similarity of their activity. This approach has been used to characterize scientific collaborations, the web, and peer-to-peer networks in [10]. Li et al. [11] target the problem of finding users with similar interests in online social networking sites. In particular, they use a *del.icio.us* data set to define implicit links between users based on the similarity of their tags. First, they support the intuition that tags accurately represent the content by showing that tags assigned to a URL match to a great extent the keywords that summarize that URL. Next, they design and evaluate a system that clusters users based on similar interests and identify topics of interests in the community. Cattuto et al. [12] study *Bibsonomy* [13] and show the existence of small-world patterns in a tri-partite network formulation of tagging systems. The network connects users, items and tags in a hypergraph. Krause et al. [14] also explores the topology of a tagging system, but the one formed by items similarity, to compare the *folksonomy* inferred from search logs and tagging systems. Their results show similarities which suggest that keywords can be considered as tags to URLs.

Our study differs from these previous investigations in two aspects: First, the interest sharing characterization focuses on the system-wide concentration and intensity of user interest sharing, as opposed to solely topological characteristics. Second, we attempt to validate the shared user interest observed from tagging activities by using external information such as membership to discussion groups and semantic similarity of tag vocabularies (Section 6).

**Mechanism design:** system characterization work was primarily motivated by its potential impact on system design. Thus, several studies propose to exploit characteristics of tagging systems to improve particular mechanisms. For example, Yanbe et al. [15] suggest using the tagging assignments generated in tagging systems to improve the quality of web searches. In particular, they use *del.icio.us* to improve PageRank [16] search rankings and show that this approach improves item freshness at the top of the ranking without sacrificing relevance.

Yahia et al. [1] propose and evaluate novel top-k querying techniques that explore similarity among user vocabularies (i.e., tags used) in tagging systems. They show that the traditional top-k query techniques lead to efficient processing at the expense of prohibitive space overhead and propose novel heuristics to regulate the space-time tradeoff. Central to their

approach is the construction of a network of users based on their activity similarity. Using a one-month long trace from *del.icio.us*, they show that clustering users based on tag similarity leads to improvements in both space and execution time.

Sigurbjörnsson et al. [2] study tag recommendation strategies and tagging behavior in *Flickr*. Their analysis of tag categories based on WordNet<sup>1</sup> data shows that tags form a wide spectrum of categories that include location, subject description, and time. The proposed recommendation strategies achieved up to 97% of probability of finding a good tag among the top-5 recommended tags.

The present work complements these studies by providing evidence that tagging activity can be useful to support information retrieval mechanisms. For instance, the results we present on item re-tagging and tag reuse (Section 4) confirm that tagging systems are a good source of fresh content to compose search results, as users are constantly discovering new items and tagging them. Moreover, we present novel insights about the individual and social behavior of users in tagging systems via the characterization of interest sharing and its relation to other indicators of collaborative behavior.

### 3. DATA SETS

This section describes the tagging systems and the data sets considered in this study. Both systems we analyze, *CiteULike* and *Connotea*, are meant to help users organize references to scientific publications. We chose to characterize *CiteULike* and *Connotea* in order to contribute to a broader understanding of tagging systems, as most of previous studies concentrate on the popular yet generic system *del.icio.us*. Our preliminary intuition is that a study of more specialized tagging systems—in this case, for managing academic publications—may reveal social structures that may be harder to identify in generic systems.

In these systems each user maintains a *library*: a collection of citation records linked to on-line articles or webpages maintained on the publisher website. User may assign *tags* to *items* in their own library, or in other user’s library, if the latter is public. Tags may serve to group items, as a form of categorization, or to help finding items in the future [3]. The tagging activity can be private (i.e., only the user who generated the activity can access it) or public. The analysis presented in the next sections concentrates on the public portion of the activity. A user can see what (public) tags other users assigned to an item, thus the user is able to reinforce the choice of tags as appropriate by repeating the tags previously assigned to that item.

An item can be added to a user’s library (an action often referred to as item *posting*) in three ways: *i*) browse popular scientific literature portals (e.g., ACM Portal, IEEE Explorer, arXiv.org) and use their features that automate item posting; *ii*) search for items already present in other users’ libraries and add them to her own library; and, *iii*) post a new item manually.

Table 1 presents a summary of the data sets used in this investigation. The data sets consist of all tagging activity since the creation of each community until January’09, more than two

years of user activity for each. The *CiteULike* dataset is available directly from its website. For *Connotea*, we built a crawler that leverages *Connotea*’s API to collect tagging activity since December 2004. Note that we do not have access to browsing records: that is, the trace contains information about when items are posted and when an item was annotated with a given tag, but we do not know whether the tag is subsequently used by a user to navigate through the system, for example. The data sets are ‘cleaned’ to exclude sources of noise (such as the default ‘no-tag’ annotations) that may affect our analysis.

**Table 1: Summary of the data sets used.**

	<b>CiteULike</b>	<b>Connotea</b>
<b>Activity period</b>	11/2004 – 01/2009	12/2004 – 10/2008
<b># Users</b>	40,327	34,742
<b># Items</b>	1,325,565	509,311
<b># Tags (distinct)</b>	274,982	209,759
<b># Tag Assignments</b>	4,835,488	1,671,194

### 4. TAG REUSE AND ITEM RE-TAGGING

In a tagging system where users mostly introduce new items and tags, efficiently harnessing information based on collective action is difficult, if not impossible, as information about new items and tags are hard to predict. Thus, understanding this dimension of user behavior can help estimating the potential efficiency of techniques that rely on similarity of past user activity (e.g., recommender systems). To this end, this section analyzes the degree to which items are repeatedly tagged and tags reused. In particular, the goal of this section is to address the following questions:

*Q1.1: What are the levels of item re-tagging and tag reuse?*

*Q1.2: Is most of the activity generated by returning users or by an influx of new users?*

In the rest of this section, we first formalize the metrics *item re-tagging* and *tag reuse*. Second, it characterizes the levels of item re-tagging and tag reuse as well as the level of activity generated by returning users. Finally, it discusses the implications of the usage characteristics discovered.

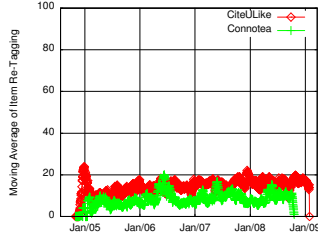
#### 4.1 Levels of Item Re-tagging and Tag Reuse

An item is re-tagged if one or more users tag it again (with the same or different terms) after it was tagged once. Similarly, a tag is reused if it appears in the trace more than once (for the same or different items) with different timestamps. We aim to determine which portion of the activity falls in these categories.

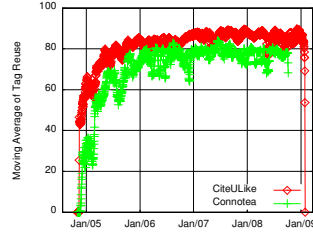
A tagging system is composed of a set of users, items and tags, respectively denoted by  $U, I, T$ . The tagging activity is a set of tuples  $(u, i, t, \tau)$ , where  $u \in U$  is a user who tagged item  $i \in I$  with tag  $t \in T$  at time  $\tau$ .

Let  $I^d$  be the set of items which are tagged in a given day  $d$ , that is the set of all items tagged at time  $\tau$  such that  $\tau \in d$ . Thus, the level of item re-tagging during a given day  $d$  is determined as follows:

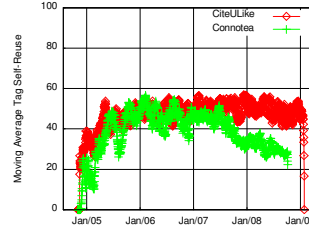
<sup>1</sup> <http://wordnet.princeton.edu/>



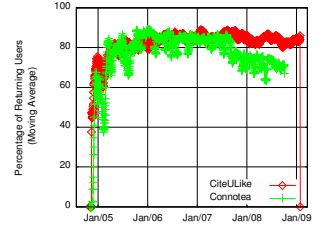
**Figure 1: Daily item re-tagging (20-day moving average)**



**Figure 2: Daily tag reuse (20-day moving average)**



**Figure 3: Tag self-reuse (20-day moving average)**



**Figure 4: Percentage of daily activity generated by returning users.**

$$ir(d) = \frac{\left| \left( \bigcup_{k=0}^{d-1} I^k \right) \cap I^d \right|}{|I^d|} \quad (1)$$

Tag reuse is defined in a similar way based on  $T^d$  as the set of items tagged in a given day  $d$ .

We determine the aggregate level of item re-tagging and tag reuse in *CiteULike* and *Connotea* as expressed by the median of daily item re-tagging and tag reuse over the entire traces.

Table 2 shows that both *CiteULike* and *Connotea* have *relatively low levels of item re-tagging*, yet *high levels of tag reuse*. To test whether these aggregate levels are a result of stable behavior over time, Figure 1 and Figure 2 present the moving average (with a window of 20 days) of daily item re-tagging and tag reuse. Qualitatively, these results show that after a brief period of bootstrapping, item re-tagging and tag reuse levels remain stable. Quantitatively, daily tag reuse is approximately five times larger than item re-tagging in *CiteULike* and eight times in *Connotea*.

**Table 2: A summary of average daily item re-tagging, tag reuse and activity generated by existing users.**

		CiteULike	Connotea
<b>Re-Tagged Items</b>	Median	16.2%	6.9%
	Std. Dev.	16.1%	6.8%
<b>Reused Tags</b>	Median	92.4%	76.2%
	Std. Dev.	8.9%	17.5%

On one hand, from the perspective of personal content management, the levels of item re-tagging and tag reuse, together with the much larger number of items than tags in the systems, suggest that users exploit tags as an instrument to categorize items according to topics of interest. On the other hand, the relatively high level of tag reuse suggests that users may have common interest over some topics, but not necessarily over specific items, as the low item re-tagging hints.

A question that arises from the above observations is whether the levels of item re-tagging and tag reuse are generated by the same user or by different users. We observe that virtually none of the item re-tagging events are produced by the user who originally introduced the item to the system: generally, users *do not* add new tags to describe the items they collected and annotated once.

On the other hand, as illustrated by Figure 3, about 50% of tag reuse is *self-reuse* (i.e., the reuse of a tag by a user who already used it first). This level of self-reuse indicates that users will often tag multiple items with the same tag, a behavior consistent with the use of tagging for item categorization and personal content

management, as discussed above. In Section 5 we investigate further the social aspect of tag reuse by defining and measuring interest sharing among users.

## 4.2 Is Most of the Activity Generated by New Users?

To understand whether the observed low level of item re-tagging is generated by a high rate of new users joining the community, we estimate the levels of activity generated by returning users (as opposed to new users that join the community). Figure 4 shows that, after a short bootstrap period, the level of tagging activity generated by returning users remain stable at about 80% over the rest of the trace for both *CiteULike* and *Connotea* (up to Jan/2008, when it seems to have an increase in the influx of new users). Thus, the low levels of item re-tagging are not the result of a constant stream of new users joining the community and introducing new items, but the outcome of expanding interests of existing users.

## 4.3 Summary and Implications

The observed user behavior impacts the efficiency of systems that rely on the inferred similarity among items such as recommender systems. On one hand, the relatively low level of item re-tagging suggests a highly sparse data set (i.e., if we attempt to connect users based on common items tagged). A sparse data set poses challenges to the design of recommender systems because it makes difficult to predict future user preferences over items, as it is common to recommender systems to rely on the similarity of users, for instance, to select recommendation candidates.

On the other hand, the higher level of tag reuse confirms that tagging has the potential to circumvent, or at least alleviate, the sparsity problem, as described above. The tags and users associated to each item could not only serve to link items and build the item-to-item structure, but could also potentially provide semantic information about items. This information may help, for instance, design better citation management tools for the research community.

Despite the sparse data set problem, the fact that users tend to permanently add fresh content, as indicated by the low level of item re-tagging, highlights that an approach similar to that proposed by Yanabe et al. [15] would be useful in a search portal for academic content. They suggest considering activity from social bookmarking systems such as *del.iciou.us* to improve the freshness and relevance of search results produced by a search engine. Portals for academic publications, such as Google Scholar, could exploit this fact to improve the freshness and relevance of their search results by using a combination of the now traditional PageRank search algorithm and annotations from systems like *CiteULike*, *Connotea* and *Bibsonomy*.

## 5. INTEREST SHARING

The analysis of item re-tagging and tag reuse in the previous section shows that virtually no item is re-tagged by the user who first published it. This implies that the observed level of re-tagging is the result of *different* users annotating the same item, which we call *interest sharing*.

This section defines and characterizes the *interest sharing* in *CiteULike* and *Connotea* with the goal to capture the similarity between users as implied by their tagging activity. Analyzing the system-wide interest sharing is relevant for information retrieval mechanisms such as search engines tailored for tagging systems [1, 17], as they exploit the similarity among users to determine the relevance of query results.

In particular, this section focuses on two questions. The first aims to characterize interest sharing in the real systems we study. To complement this analysis and highlight possible usage patterns, the second question demands an independent comparison basis to contrast the observed interest-sharing distributions.

*Q2.1: How is interest sharing distributed across the system?*

*Q2.2: Is the observed concentration of interest sharing high?*

Let us consider  $U$  the set of users in a tagging system. The activity of a user  $u_k \in U$  is expressed by  $I_k$  and  $T_k$ , which are respectively the set of items annotated and the set of tags used by  $u_k$ . The interest sharing between two users, as implied by the similarity of their activity, is denoted by the function  $w: U \times U \mapsto \mathfrak{R}$ .

We explore two interest sharing functions based on item and tag activity, and we refer to them as *item-based* and *tag-based* interest sharing. To estimate similarity between sets we use the *Asymmetric Jaccard Similarity Index* [18]. We note that our use of the Jaccard Index is not new: Stoyanovich et al. [5] used the index to model shared user interest in *del.icio.us* and evaluate its efficiency in predicting future user behavior. Chi, Pirolli and Lam [19] applied the symmetric index to determine the diversity of users and its impact in a social search setting. However, we go one step further, as we explore the system wide characteristics of interest sharing and the implicit social structure inferred from it (Section 6).

The *item-based* interest-sharing metrics are defined as follows (the tag-based version is defined similarly and denoted by  $w_T$ ):

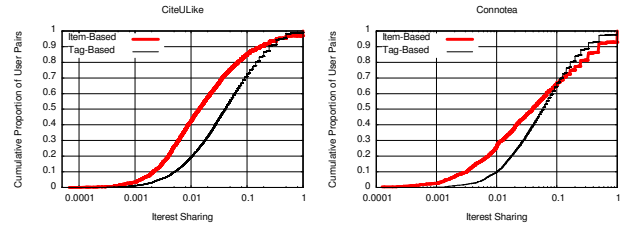
**Definition:** *The level of item-based interest sharing between two users,  $k$  and  $j$ , as perceived by  $k$ , is the ratio between the size of the intersection of the two item sets and the size of the item set of that user.*

$$w_I(k, j) = \frac{|I_k \cap I_j|}{|I_k|} \quad (2)$$

Equation 2 captures how much the interests of a user  $u_k$  match those of another user  $u_j$ , from the perspective of  $u_k$ . We use the asymmetric similarity index rather than the symmetric version (which uses the size of the union of the two sets as the denominator in Equation 2) because the asymmetric definition models personal views of similarity, which are more relevant for personalized content delivery.

### 5.1 How is Interest Sharing Distributed Across the System?

This section presents the distribution of interest sharing between user pairs in *CiteULike* and *Connotea*. We first find that for the item-based interest sharing, approximately 99.9% of user pairs in *CiteULike* tag no items in common (i.e., share no interest and consequently have  $w_I(k, j) = 0$ ). In *Connotea*, the percentage is virtually the same: 99.8%. For the tag-based interest sharing, the percentage of user pairs sharing no tags ( $w_T(k, j) = 0$ ) is slightly lower: 83.8% and 95.8%, for *CiteULike* and *Connotea*, respectively. Such sparsity in the pairwise user similarity supports the conjecture that users are drawn to tagging systems primarily by their personal content management needs [3], as opposed to the desire of collaborating with others.



**Figure 5: Distributions for item-based and tag-based interest sharing in *CiteULike* and *Connotea* (for pairs of users with non-zero sharing).**

The rest of this section focuses on the remaining user pairs that either tag items in common or use at least one tag in common. We start by determining the cumulative probability distribution (CDF) of item- and tag-based interest-sharing for these sets of user pairs in *CiteULike* and *Connotea*.

Figure 5 shows that, in *CiteULike*, the typical intensity of tag-based interest sharing is higher than its item-based counterpart. More precisely, the item-based interest sharing is lower than 0.1 ( $w_I(k, j) \leq 0.1$ ) for more than 80% of the user pairs, while the tag-based interest sharing is lower than 0.1 ( $w_T(k, j) \leq 0.1$ ) for about 75% of the user pairs. Similarly, in *Connotea*, for more than 65% user pairs item- or tag-based interest sharing are lower than 0.1. This is not surprising: after all, both systems include two to three times more items than tags.

The results in Figure 5 also show a relative difference between the distributions of item- and tag-based interest sharing. This difference suggests the existence of latent organization among users as reflected by their fields of interest. Considering that *CiteULike* and *Connotea* are ultimately citation management systems, the observed relative difference between the levels of interest sharing may be due to a large number of user pairs which are interested in the same high-level domain (e.g., *computer networks*), but diverge regarding the interests over specific sub-domains (e.g., *internet routing* versus *firewall traversal techniques*). Thus, users are more likely to use similar tags drawn from a vocabulary specific to the general domain for annotating items from different and more specific sub-domains. In the next section, we discuss an experiment to validate this intuition.

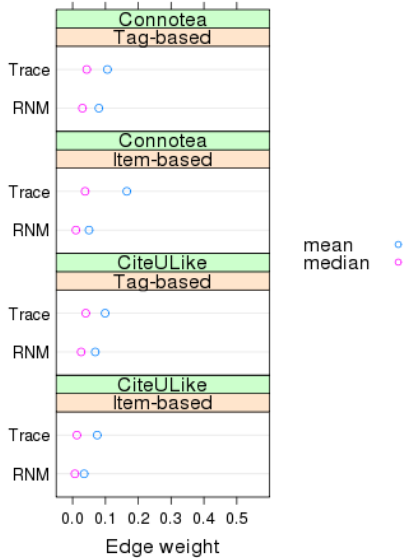
### 5.2 Comparing with a Baseline

The goal of this section is to better understand the interest sharing levels we observe: Are they high or low? Is interest sharing evenly

distributed over all users or significantly concentrated among a small number of them?

For this investigation, we need a comparison basis of a high or low interest sharing distribution. We opt for a referential of low interest sharing, which we build as a randomized null model (RNM) [20] by simulating a tagging system that preserves the main macro-characteristics of the tagging activity (i.e., the number of items, tags, and users, as well as item and tag popularity and user activity distributions) of the systems we study, but where users make random tag assignments. The interest sharing levels in the RNM are therefore those that would happen by chance, if two users tagging the same item is a result of random individual behavior, rather a product of their common interests. If the interest-sharing levels or its concentration in the data we observe are significantly higher or more intense than the values found in the RNM, then the underlying process from which the observed interest sharing emerges carries more information than a random one.

To test this hypothesis we compare the two sets of data (real and RNM-generated) in terms of the numbers of user pairs with non-zero interest sharing and the interest-sharing intensity distribution. We have used the RNM to generate five synthetic traces corresponding to the real systems we analyze. For the rest of this section, the RNM results represent averages over the five RNM traces. We confirmed that the five synthetic traces represent a large enough sample to guarantee a 95% confidence interval for the average interest sharing observed from the RNM simulations.



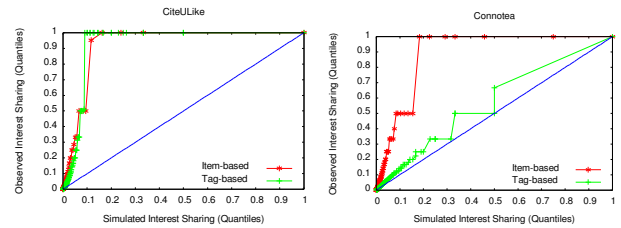
**Figure 6: A comparison between the observed interest sharing (Trace) and the RNM. It is clear that the RNM deviates from the actual trace. It is worth noting that the 95% confidence interval for the mean is narrow on the order of  $\pm 0.001$ .**

The number of user pairs that share some item-based interest (i.e.,  $w_i(k, j) > 0$ ) is approximately three times smaller in the real systems than in the RNM simulated ones. Tag-based interest sharing follows a similar trend. *This implies that interest sharing (and, consequently similarity between users) is more concentrated in the real systems we study than in our simulated RNM.*

Next, we compare the intensity of interest sharing for the user-pairs that have non-zero interest. Figure 7 presents the Q-Q plots that directly compare the distributions of interest-sharing levels derived from the actual trace and those derived from the simulated RNM. A deviation from the diagonal indicates a difference between these distributions: The higher the points are above the diagonal, the higher the observed interest-sharing levels are compared to the interest sharing originated from the tagging activity generated by the RNM.

These results indicate that, although the absolute values for the observed interest sharing are low, user’s tag choices are far from random. For further insight, Figure 6 directly compares the average and the median of the RNM simulated interest sharing to those observed in *CiteULike* and *Connotea*. In *CiteULike*, the average and median for both item- and tag-based interest-sharing intensity are over three times larger than those for the RNM-generated traces. In *Connotea* the same observation is valid for the item-based interest sharing.

We note that the only interest-sharing distribution that is close to the one produced by the random model is for *Connotea*’s tag-based interest sharing (Figure 7). However, there is a significant deviation from randomness: real activity trace leads to three times fewer user-pairs that share interest than the corresponding RNM.



**Figure 7: Q-Q plots that compare the interest sharing distributions for the observed vs. simulated (i.e., the RNM model) for *CiteULike* and *Connotea*.**

### 5.3 Summary and Implications

This section provides a characterization of interest-sharing distributions in *CiteULike* and *Connotea* and an analysis of the observed distributions. The latter is performed by comparing the interest-sharing distributions in the real systems and those produced by a randomized null model (RNM). The RNM preserves the macro characteristics of the systems we investigate, yet uses random tagging assignments.

The comparison highlights two main characteristics of the interest sharing: first, interest sharing is significantly more concentrated in the real traces than in the RNM-generated activity: in quantitative terms, three times fewer user pairs share interests in the real traces. Second, most of the time, the observed interest-sharing intensity is significantly higher than that generated by the RNM equivalent.

The following abstract model offers a possible explanation for these observations. Let us consider that the set of tags that can be assigned to an item is largely limited by the set of topics that item is related to. In this case, intuitively, the probability of choosing a tag is conditional to the set of topics the item is related to. At one extreme, the maximum diversity of topics occurs when there is a one-to-one mapping between topics and tags in the system. That is, each tag introduces a different topic. The RNM simulates the

other extreme, a single topic that encompasses all tags in the system.

However, in real systems, the interests for each individual user are limited to a finite set of topics that is likely to determine the tag vocabulary they use. This leads to a concentration of interest-sharing, as implied by the tag similarity, on few user pairs, yet at higher intensity than that produced by the RNM – a behavior observed in the real system we study.

Using the intuition of this model, we conjecture that the sharp difference between the tag-based in *CiteULike* and *Connotea* is explained by a more uniform set of interests in *Connotea* than in *CiteULike* (an observation supported by the smaller size of the user population and relatively smaller tag vocabulary). An approach to test this conjecture is: *i*) to model tag choices for each item as a nested Chinese Restaurant Process [21] that takes into account the conditional probabilities mentioned above; *ii*) to estimate the topic structure in *CiteULike* and *Connotea* and verify by how much they differ. We plan to explore this model in the future.

Finally, we believe that the observed divergence between the observed and the RNM interest sharing distributions shows that interest sharing embeds information about user self-organization according to their preferences. This information, in turn, could be exploited by mechanisms that rely on implicit relation between users.

## 6. SHARED INTEREST AND INDICATORS OF COLLABORATION

The previous section demonstrates that interest sharing in the systems we study is distributed differently than in a community modeled by an RNM that preserves the same macro characteristics. We attempt to explain this deviation from randomness and turn our attention to possible correlations between interest sharing and other elements of user behavior observable in these systems. This section summarizes our experience in this direction by addressing the following question:

*Q3. Does interest sharing relate to higher-level social behavior?*

It is important to mention that the number of externally observable elements of user behavior to which we have access is limited by the design of the systems themselves (e.g., the tagging systems collect limited information on user attributes) and by our limited access to data (e.g., we do not have access to browsing traces). One *CiteULike* feature, however, is useful: *CiteULike* allows users to form explicit discussion groups to share items among a selected subset of users – an indicator for user collaboration in the system. We thus explore to what degree interest sharing is correlated with group co-membership (and thus with collaboration) in *CiteULike*.

Along the same lines, we use a second external signal: *semantic* similarity between tag vocabularies. More specifically, we test the hypothesis that item-based interest sharing relates to *semantic similarity* between user vocabularies. The underlying assumption here is that users that (have the potential to) collaborate employ similar vocabularies.

The rest of this section presents in details the methodology and the results of these two experiments. In brief, our conclusions are:

- The lack of interest sharing between two users is , then they are

unlikely to collaborate (i.e., they do not participate in the same groups and the similarity between their vocabularies is low). We find, however, evidence that positive item-based interest sharing is related to higher vocabulary similarity between users than zero interest sharing; an indication that high usage similarity makes collaboration possible.

- On the other side, we find no statistical correlation between the *intensity* of interest sharing and the collaboration levels as implied by group co-membership or vocabulary similarity.

### 6.1 Group Membership

We first observe that only 14% of users declare membership to one or more groups. For this section we limit our analysis to user pairs for which both users are members of at least one group. Also, the analysis focuses on groups that have two or more users (about 50% of all groups) as groups with only one user are not representative of potential collaboration. The goal is to explore the possible relationship between item-based interest sharing and co-membership in one or more groups.

Let  $H_u$  be the set of groups the user  $u$  participates in. Thus, we determine the group-based similarity  $w_H(u,v)$  between two users  $u$  and  $v$  using the asymmetric Jaccard index, defined as in Equation 2. Based on this similarity definition, we study whether the intensity of item-based interest sharing between two users with non-zero interest sharing (i.e.,  $w_I(u,v) > 0$ ) correlates with group membership similarity.

We find no correlation between  $w_I(u,v)$  – the item-based similarity – and  $w_H(u,v)$  – the group-based similarity. More precisely, Pearson’s correlation coefficient is approximately 0.12, and Kendall’s  $\tau$  is about 0.05. To put these correlation results in perspective, we look at group similarity for two distinct groups of user pairs: those with zero and positive interest sharing. We observe that, although the group information is relatively sparse, pairs of users with no interest sharing are much less likely to have a group in common than those with positive interest sharing. (*The figures are only 0.6% of users with  $w_I(u,v) = 0$  have  $w_H(u,v) > 0.012$ , while 4% of the users with  $w_I(u,v) \geq 0$  have  $w_H(u,v) > 0.2$* )

These observations suggest that, although users share interest over items, and may implicitly benefit from each other tagging activity (e.g., finding content of interest on each other library), this may not imply that they actively engage into explicit collaborative behavior, such as participation to the same discussion groups. Conversely, the lack of interest sharing is strongly related to the lack of collaborative behavior.

### 6.2 Semantic Similarity of Tag Vocabularies

This section substitutes the group-based similarity, used in the previous section, by the *semantic* similarity between the tag vocabularies of the corresponding users (i.e. the set of tags a user has applied to items posted to *CiteULike*). It presents the metric used to estimate the semantic similarity of tag vocabularies, discusses methodological issues and, finally, presents the results.

**Estimating semantic similarity:** In order to estimate the semantic similarity between individual tags, we use WordNet, a lexical database, which includes semantic relations between word senses such as synonymy (the same or similar meaning) and hypernymy/hyponymy (one term is a more general sense of the other). Different methods have been implemented for quantifying semantic similarity by using WordNet. In particular,

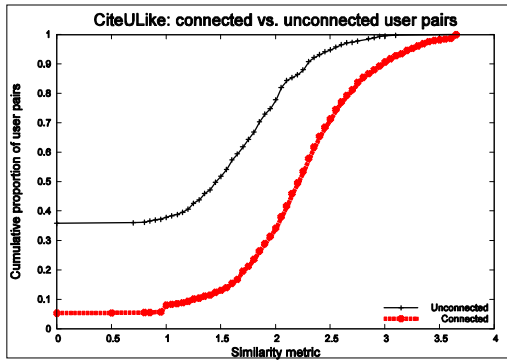
WordNet::Similarity – a Perl module – provides a set of semantic similarity measures [22].

For our experiments, we use the Leacock-Chodorow similarity metric [23], as previous experiments [24], based on human judgments, suggest that it performs best in capturing semantic similarity. The metric is derived from the negative log of the path length between two word senses in the WordNet “is-a” hierarchy, and is only usable between pairs of words where both have one or more noun senses.

We define the similarity  $sim(t_1, t_2)$  between two tags  $(t_1, t_2)$  as the maximum Leacock-Chodorow similarity between every available noun sense of  $t_1$  and  $t_2$ . Thus, the semantic similarity between the tag vocabularies  $T_u$  and  $T_v$  of two users,  $u$  and  $v$ , as perceived by  $u$ , is denoted by  $w_V(u, v)$ , and determined by the ratio between the sum over the pairwise tag similarities and the size of  $u$ ’s vocabulary, as expressed by Equation 3 below. We note that this metric is based on the Modified Hausdorff Distance (MHD) used by [25].

$$w_V(u, v) = \frac{\sum_{t_1 \in T_u, t_2 \in T_v} sim(t_1, t_2)}{|T_u|} \quad (3)$$

**Methodological issues and limitations:** There are two practical issues with the metric above. First, to avoid biasing our estimates, if two users assigned the same tags to the same item, we omit them from their vocabularies, before determining the aggregate similarity. Second, a limitation of this metric is that it uses only tags that have noun senses in WordNet. Tags applied by *CiteULike* users may be words or phrases from any language, abbreviations, or even arbitrary strings invented by the user. WordNet consists mainly of common English words. However, while only 17.3% of unique tags in our data had noun senses in WordNet, these tags accounted for 61.1% of all tagging activity. In future work, we plan to explore an approach similar to Suchanek et al. [9] to extend the coverage of the dictionary to encompass proper names.



**Figure 8: CDF of tag vocabulary similarity for user pairs with positive (top) and zero (bottom) interest sharing.**

**Results:** We test whether there was a significant difference in tag vocabulary similarity between two random samples of user pairs: one where all user pairs ( $n = 500$  pairs) have zero item-based interest sharing and one with positive item-based interest sharing ( $n = 2000$  pairs). This analysis shows that the vocabularies of user pairs with positive interest sharing are *significantly more similar* than those of user pairs with no interest sharing (Figure 8). In

particular, the average semantic vocabulary similarity for unconnected users,  $\mu_u = 1.20$  ( $\pm 0.12$  – 90% c.i.), is almost two times smaller than that for connected pairs  $\mu_c = 2.18$  ( $\pm 0.03$  – 95% c.i.). This salient difference in the vocabulary similarity suggests that the item-based similarity embeds information about the “language” shared by the users to describe the items they are interested in. Finally, we determine the correlation between the between the item-based interest sharing and the tag vocabulary similarity. We find that neither Pearson’s nor Kendall’s correlation is significant. More precisely, for user pairs where  $w_I(u, v) > 0$ , the Pearson’s  $r$  is  $-0.09$ , and Kendall’s  $\tau$  is  $-0.11$ .

### 6.3 Summary and Implications

This section takes a first step towards understanding the relationship between the interest sharing and observable activity that captures collaborative behavior. First, we look at correlations between the item-based interest sharing and the group-based similarity. The observations indicate that although the intensity of item-based interest sharing does not correlate with explicit collaborative behavior, as implied by co-membership on groups, user pairs that have interest sharing are more likely to participate to similar groups.

Second, we evaluate the relationship between item-based similarity and the *semantic* similarity of tag vocabularies. We discover that, although the two do not yield a Pearson’s correlation, item-based similarity does embed information about the expected semantic similarity between user vocabularies.

These results have main implications on mechanisms that aim at predicting collaborative behavior, as they could exploit item-based similarity to set expectations about group-based and vocabulary-based similarity. Moreover, one could easily use deviations from observed relationship between item-based similarity and the two indicators of collaborative behavior presented here to detect malicious user behavior.

## 7. ANALYZING THE IMPLICIT SOCIAL STRUCTURE

This section uses the interest-sharing relations between users to construct an implicit social network and to study its characteristics. The implicit social structure is informally defined as the *interest-sharing graph*, where two users are connected if they tagged the same item or used the same tag. In particular, this section addresses the following question:

*Q4: What are the topology characteristics of the item- and tag-based interest-sharing graphs?*

The *interest-sharing graph* is a data structure inspired by the data sharing-graph introduced by Iamnitchi et al. [10] to study how people share interest in data on the web and in peer-to-peer systems. We define the interest-sharing graph as a directed graph whose nodes represent users. Users are connected in this graph if their interests are similar. The *intensity* of shared interest between two users, as defined in Equations 3 and 4, determines the edge weights. Naturally, we investigate two types of interest-sharing graphs: item-based and tag-based.

The item-based interest-sharing graph is a directed graph defined as  $G = (U, E, w_I)$ , where  $U$  is the set of nodes that corresponds to users,  $w_I$  is the edge weight function and  $E$ , the set of directed

edges, is formally denoted by  $E = \{(u_k, u_j) | w_l(u_k, u_j) > 0\}$ . The tag-based interest-sharing graph is defined similarly.

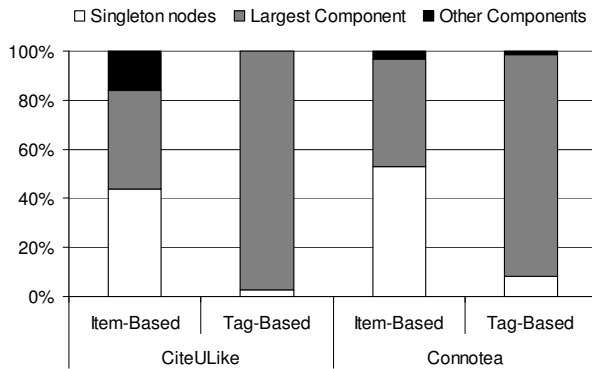
Note that the *existence* of an edge is symmetric:  $(u_k, u_j) \in E \Leftrightarrow (u_j, u_k) \in E$ . However, the edge weight (Equations 3 and 4) is determined by the source node and can be different for each direction.

**Segmentation:** Figure 9 presents the percentages of nodes in each region of the graph. The first observation is that the *item-based* interest-sharing graphs for both *CiteULike* and *Connotea* are highly segmented, with a massive number of singleton nodes and a set of small components and one giant component. This reflects a set of users organized in several clusters with disjoint interests. More importantly, it shows that there is a large set of users with *unique* interests over the set of items they consume.

The segmentation of the topology for the *tag-based* interest-sharing graphs is less pronounced with a smaller number of singleton nodes and connected components. This indicates that even though user interests are diverse in terms of items, users use similar tags to annotate their items of interest.

To better understand the relationship between user activity and the topology of the interest-sharing graph, we determine the percentage of activity produced by singleton users compared to the rest of the network. The activity produced by the singleton nodes accounts for approximately 20% of item activity in *CiteULike* and 11.3% in *Connotea*. This indicates that by narrowing the study to consider only the nodes with degree one or higher (as we do in the rest of this section), we are also considering the largest portion of user activity in both systems.

**Analysis:** The observations discussed above corroborate with previous results on the evolution of the structure of online friendship networks. In particular, Kumar et al. [26] show that the friendship network in *Flickr* and *Yahoo! 360* are formed by a giant component, which is orbited by smaller components and isolated nodes. Kumar et al. argue that a possible explanation for the emergence of such structure in *Flickr*, for example, can be obtained by modeling three types of users that differ in their *linking* strategies: first, there are users who actively reach out to other users by creating links; second, there are users who are more conservative on linking to others; and, third, there are users who do not create links at all.



**Figure 9: The percentage of nodes that are found in each of three structural regions of the graph: singletons, largest connected component, and other components.**

In the context of tagging systems such as *CiteULike* and *Connotea*, a similar interpretation is possible, where three classes of users co-exist: first, the users who are interested in a large set of topics, which, given the definition of the interest sharing graph, implies an increased chance to connect to more users; second, the users focused on a specific topic which are less likely to join a larger community; and, finally, users who are interested in rather unique items or using personalized tags, hence unlikely to have activity similar to that of other users.

**Summary and Implications:** The main result of this section shows that the item-based interest structure is much more segmented than its tag-based counterpart. While most of the current tagging-based tools offer social browsing (i.e., discovering users with similar interests) via shared items, we conjecture that the lower segmentation of tag-based interest sharing allows for content discovery and, at least in citation management systems, could benefit from a *tag-based social browsing* interface. The idea is to allow users to navigate through each others libraries via *tag-user* association, as opposed to *item-user* association. The graph structure discussed above implies that users would be able to reach many more users if they navigate based on tag similarity than on item-similarity.

## 8. CONCLUSIONS

In this work we analyzed user behavior characteristics at the individual and aggregate level in tagging systems. To this end, we studied two tagging systems: *CiteULike* and *Connotea*, and focused on three aspects of these systems: *i) item re-tagging*, a measure for the degree to which users re-tag the items already existing in the system; *ii) tag reuse*, a measure for the degree to which users reuse a tag perform new annotations; and, *iii) interest sharing*, a measure for the similarity between users with respect to their tagging activity. Our main observations are summarized as follows:

1. The characterization of item re-tagging and tag reuse suggests that the individual need for organizing content is a stronger motivation for tagging than collaborating with others to categorize it.
2. The distribution of interest sharing across the system (and consequently similarity between user's activities) is significantly more concentrated than that of a system with the same macro characteristics yet where random tagging assignments are used. This highlights the existence of information in the tagging activity and the fact that this information is not lost by our proposed interest-sharing metric. Mechanisms to support tagging systems such as personalized search and malicious content detectors may be able to exploit this information.
3. Our attempts to uncover possible correlations between item-based similarity and external indicators of collaboration between users (e.g., group co-membership) have mixed results. While we find evidence that *lack* of item-based similarity is related to the *lack* of collaboration, we do not find a statistical correlation between the *intensity* of interest sharing and the collaboration levels.
4. Finally, we find that the structure of interest sharing is highly segmented, which suggests that user population self organize into clusters of interests.

The implications of these results relate to: *i) recommender*

systems – as the sparsity of user similarity demand more sophisticated techniques to achieve better precision and recall results; ii) *malicious user detection* – as spam detection mechanisms, specially tailored for tagging systems, could use deviations from the interest sharing characteristics of a non-malicious user population to detect malicious users; iii) *support for collaborative behavior* – as the information embedded in the interest sharing metric can still be harnessed, if not to predict collaborative behavior, at least to prune out the unlikely one.

**Future work:** i) *Activity traces:* Traces limit us to a particular analysis – similarity-based on explicit activity. In possession of click traces we could make a more comprehensive analysis of interest sharing; ii) *Static graph:* While we use static interest-sharing graph, it would be interesting to generalize the interest-sharing graph definition into a temporal graph, where edges are considered as a function of time. This would be useful to understand a potential “decay factor” of interest sharing between users, which has obvious implications to the graph structure, and therefore to mechanisms that rely on it; iii) *Semantic similarity only considers noun senses present in WordNet:* In a future work, however, we could exploit a combination of WordNet and YAGO – a dictionary that contains commonly used proper names.

Apart from solidifying the results we presented here by addressing the limitations above, there are two main directions we plan to explore: first, to investigate whether existing generative stochastic models could explain the emergence of the observed interest sharing patterns; and, second, to assess the impact of the observed behavior on the efficiency of existing mechanisms that support tagging systems, such as malicious user detection.

**Acknowledgments:** The authors thank Chris Hall, who provided group membership traces from CiteULike; and, Roger Yin for his help on the Connotea crawler. Elizeu Santos-Neto was partially supported by a British Columbia Innovation Council Fellowship. Nazareno Andrade was supported by a Graduate Student Exchange Program grant from the Government of Canada. Adriana Iamnitchi was partially supported under NSF grant CNS-0831785.

## 9. REFERENCES

- [1] S. Yahia et al. "Efficient network aware search in collaborative tagging sites", In *VLDB'08*, pp. 710-721, 2008.
- [2] B. Sigurbjörnsson and R. v. Zwol, "Flickr tag recommendation based on collective knowledge," in *WWW'08*.
- [3] S. Golder and B. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, pp. 198-208, 2006.
- [4] E. H. Chi and T. Mytkowicz, "Understanding the efficiency of social tagging systems using information theory," in *HT'08*.
- [5] J. Stoyanovich et al., "Leveraging tagging to model user interests in del.icio.us," in *Proceeding of the AAAI Spring Symposium on Social Information Processing*, 2008,
- [6] B. Evans and E. Chi, "Towards a model of understanding social search," in *CSCW '08*, pp. 485-494, 2008.
- [7] C. Cattuto et al. "Semiotic dynamics and collaborative tagging", *PNAS*, vol. 104, pp. 1461-1464, January. 2007.
- [8] H. A. Simon, "On a Class of Skew Distribution Functions", *Biometrika*, vol. 42, pp. 425-440, December 1, 1955.
- [9] F. Suchanek, M. Vojnovic and D. Gunawardena, "Social tags: Meaning and suggestions," in *CIKM '08*, pp. 223-232, 2008.
- [10] A. Iamnitchi, M. Ripeanu and I. Foster, "Small-world file-sharing communities," in *INFOCOM'04*, pp. 952-963, 2004.
- [11] X. Li, L. Guo and Y. Zhao, "Tag-based social interest discovery," in *WWW '08*, 2008, pp. 675-684.
- [12] C. Cattuto, et al. "Network Properties of Folksonomies," *AI Communications Journal*, 2007.
- [13] A. Hotho et al., "BibSonomy: A social bookmark and publication sharing system," in *Proc. of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, 2006.
- [14] B. Krause et al. "Logsonomy - social information retrieval with logdata," in *HT '08*, pp. 157-166.
- [15] Y. Yanbe et al., "Can social bookmarking enhance search in the web?," in *JCDL '07*, pp. 107-116, 2007.
- [16] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, vol. 30, pp. 107-117, 4. 1998.
- [17] D. Zhou et al. "Exploring social annotations for information retrieval," in *WWW '08*, 2008, pp. 715-724.
- [18] P. Jaccard, "The Distribution of the Flora in the Alpine Zone", *New Phytologist*, vol. 11, pp. 37-50, 1912.
- [19] E. Chi, P. Pirolli and S. Lam, "Aspects of augmented social cognition: Social information foraging and social search," in *Online Communities and Social Computing*, pp.60-69, 2007.
- [20] D. Helbing (ed), *Managing Complexity: Insights, Concepts, Applications*. Springer, 2007,
- [21] D. M. Blei et al., "Hierarchical topic models and the nested chinese restaurant process," in *NIPS*, pp. 2003.
- [22] T. Pedersen et al., "WordNet: : Similarity - Measuring the Relatedness of Concepts." pp. 1024-1025, 2004.
- [23] C. Leacock and M. Chodorow, "Combining local context with WordNet similarity for word sense identification," in *WordNet: A Lexical Reference System and its Application*, 1998.
- [24] M. Warin, H. Oxhammar and M. Volk, "Enriching an ontology with WordNet based on similarity measures," in *Proc. of the MEANING-2005 Workshop*, 2005,
- [25] M. Dubuisson and A. Jain, "A modified hausdorff distance for object matching", In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pp. 566-568 vol.1, 1994.
- [26] R. Kumar et al., "Structure and evolution of online social networks," in *KDD '06*, pp. 611-617, 2006.