Network Properties of Folksonomies

Ciro Cattuto [c,b] Christoph Schmitz [a]
Andrea Baldassarri [b] Vito D. P. Servedio [b,c]
Vittorio Loreto [b,c] Andreas Hotho [a]
Miranda Grahl [a] Gerd Stumme [a]
[a] *Knowledge & Data Engineering Group*
Dept. of Mathematics and Computer Science
Univ. of Kassel
Wilhelmshöher Allee 73
D-34121 Kassel, Germany
{*lastname*}*@cs.uni-kassel.de*
[b] *Dipartimento di Fisica, Università di Roma "La Sapienza"*
P.le A. Moro, 2,
I-00185 Roma, Italy
{*firstname.lastname*}*@roma1.infn.it*
[c] *Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi*
Compendio Viminale
I-00184 Roma, Italy

Social resource sharing systems like YouTube and del.icio.us have acquired a large number of users within the last few years. They provide rich resources for data analysis, information retrieval, and knowledge discovery applications. A first step towards this end is to gain better insights into content and structure of these systems. In this paper, we will analyse the main network characteristics of two of these systems. We consider their underlying data structures – so-called folksonomies – as tri-partite hypergraphs, and adapt classical network measures like characteristic path length and clustering coefficient to them.

Subsequently, we introduce a network of tag co-occurrence and investigate some of its statistical properties, focusing on correlations in node connectivity and pointing out features that reflect emergent semantics within the folksonomy. We show that simple statistical indicators unambiguously spot non-social behavior such as spam.

1. Introduction

A new family of so-called "Web 2.0" applications is currently emerging on the Web. These include user-centric publishing and knowledge management platforms like Wikis, Blogs, and social resource sharing systems. In this paper, we focus on resource sharing systems, which all use the same kind of lightweight knowledge representation, called *folksonomy*. The word 'folksonomy' is a blend of the words 'taxonomy' and 'folk', and stands for conceptual structures created by the people.

Resource sharing systems, such as YouTube[1] or del.icio.us,[2] have acquired large numbers of users (from discussions on the del.icio.us mailing list, one can approximate the number of users on del.icio.us to be several hundreds of thousands) within less than three years. The reason for their immediate success is the fact that no specific skills are needed for participating, and that these tools yield immediate benefit for each individual user (e.g. organizing ones bookmarks in a browser-independent, persistent fashion) without too much overhead. Large numbers of users have created huge amounts of information within a very short period of time.

In this paper, we will investigate the growing network structure of folksonomies over time from different viewpoints, using two datasets from running systems as examples.

Firstly, we investigate the network structure of folksonomies much on the same line as the developments in the research area of complex networks. To that end, we will adapt measures for so-called "small world networks" which have been used on a wide variety of graphs in recent years, to the particular tripartite structure of folksonomies and show that folksonomies do indeed exhibit a small world structure.

Secondly, beyond the analysis of the whole hypergraph, we also consider specific projections of it by narrowing the scope and focusing on particular features of the structure. We analyze in particular the tag co-occurrence network and study its properties. This is a weighted network where

[1] `http://www.youtube.com/`
[2] `http://del.icio.us`

AI Communications
ISSN 0921-7126, IOS Press.

each tag is a node and links are drawn between a pair of tags whenever the two tags co-occur in the same post and the weight is given by the number of different posts where that pair appears. This tag co-occurrence network can be used to get insights into the tagging behaviour of users and to detect anomalies, e. g. those inflicted by spammers.

The remainder of the paper is structured as follows: In Section 2, we discuss related work. Section 3 introduces two large scale folksonomy datasets which our analyses will be based on. Section 4 introduces quantitative measures for the network properties for the tripartite structure of a folksonomies. Section 5 examines a projection of the tripartite graph by studying the structure of the tag co-occurrence network. Finally in Section 6 we draw some conclusions and highlight open issues.

2. Related Work

2.1. Folksonomies and Folksonomy Mining

As the field of folksonomies is a young one, there are relatively few scientific publications about this topic. Refs. [19,9] provide a general overview of folksonomies, their structure, and provide some insights into their dynamics.

More recently, particular aspects of folksonomies have been elaborated in more detail, e.g. ranking of contents [12], discovering trends in the tagging behaviour of users [7,13], or learning taxonomic relations from tags [10,27,26,20,14].

2.2. Small World Networks

The graph-theoretic notions of Section 4 are derived from those developed in an emerging area of research which has been called "the new science of networks" [23], using concepts from social network analysis, graph theory, as well as statistical physics; see [23] for an overview.

In particular, the notions of clustering coefficient and characteristic path length as indicators for small world networks have been introduced by Watts and Strogatz [31]; for particular kinds of networks, such as bipartite [18] or weighted [2] graphs, variants of those measures have been devised. To the best of our knowledge, no versions of these measures for tripartite hypergraphs such as folksonomies, or hypergraphs in general, have been proposed previously.

Networks related to folksonomy, in line with other different human based social or technological networks, possess a lot of other peculiar characteristics. The most striking of them is probably the observation that the degree of nodes, i. e., the number of links connected to a node, follows a fat tailed distribution index of a complex interaction between human agents [29]. Work has been done also on the complex network of WikiPedia [4] where links also possess a specific direction.

The notion of a *small world* has been introduced in a seminal paper by Milgram [21]. Milgram tried to verify in a practical experiment that, with a high probability, any two given persons within the United States would be connected through a relatively short chain of mutual acquaintances. Recently, the term "small world" has been defined more precisely as a network having a small characteristic path length comparable to that of a (regular or Erdős) random graph, while at the same time exhibiting a large degree of clustering [30] (which a random graph does not). These networks show some interesting properties: while nodes are typically located in densely-knit clusters, there are still long-range connections to other parts of the network, so that information can spread quickly. At the same time, the networks are robust against random node failures. Since the coining of the term "small world", many networks, including social and biological as well as man-made, engineered ones, have been shown to exhibit small-world properties. We will show in the remainder of this paper that folksonomies have a small world structure.

3. Folksonomy Datasets

In this section, we will introduce the formal notation used in the remainder of the paper, as well as the two large scale data sets that we will discuss in the following sections.

3.1. Folksonomy Notation

In the following, we briefly recapitulate the formal notation for folksonomies introduced in [12], which we will use in the remainder of the paper.[3]

A *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y)$ where

[3] We use the simplified version without personomies or hierarchical relations between tags here.

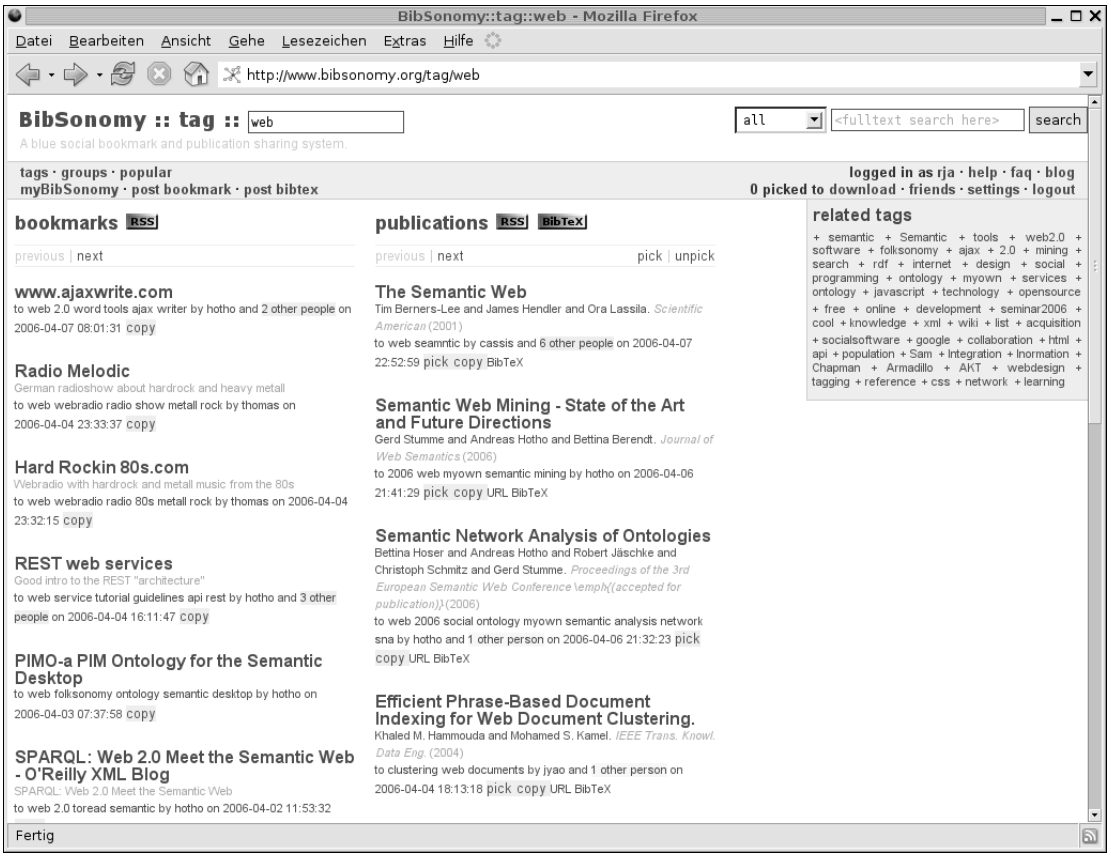

Fig. 1. BibSonomy displays bookmarks and BibTeX based bibliographic references simultaneously.

– U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., and
– Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$, called *tag assignments* (*TAS* for short).

Another view on this kind of data is that of a 3-regular, tripartite hypergraph, in which the node set is partitioned into three disjoint sets: $V = T \cup U \cup R$, and every hyperedge $\{t, u, r\}$ consists of exactly one tag, one user, and one resource. In Formal Concept Analysis [8], such data structures are called *triadic context* [17].

Sometimes it is convenient to consider all tag assignments of a given user to a given resource. We call this aggregation of TAS of a user u to a resource r a *post* $P(u, r) := \{(t, u, r) \in Y \mid t \in T\}$.

3.2. del.icio.us Dataset

For our experiments, we collected data from the del.ico.us system in the following way. Initially we used `wget` starting from the start page of del.ico.us to obtain nearly 6,900 users and 700 tags as a starting set. Out of this dataset we extracted all users and resources (i. e., del.icio.us' MD5-hashed URLs). From July 27 to 30, 2005, we downloaded in a recursive manner user pages to get new resources, and resource pages to get new users. Furthermore we monitored the del.icio.us start page to gather additional users and resources. This way we collected a list of several thousand usernames which we used for accessing the first 10,000 resources each user had tagged. From the collected data we finally took the user files to extract resources, tags, dates, descriptions, extended descriptions, and the corresponding username.

We obtained a folksonomy with $|U| = 75,242$ users, $|T| = 533,191$ tags and $|R| = 3,158,297$ resources, related by in total $|Y| = 17,362,212$ tag assignments. In addition, we generated monthly dumps from the timestamps associated with posts, so that 14 snapshots in monthly intervals from

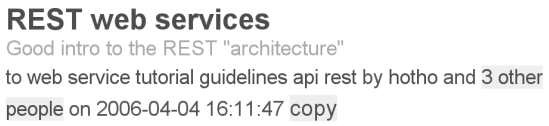

Fig. 2. detail showing a single bookmark post

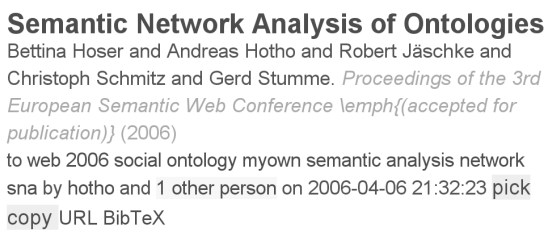

Fig. 3. detail showing a single publication post

June 15th, 2004 through July 15th, 2005 are available.

3.3. BibSonomy Dataset

As some of the authors are involved in the folksonomy site BibSonomy [11],[4] a second dataset from that system could be obtained directly from a database dump.

BibSonomy allows users to share bookmarks (i. e., URLs) as well as publication references. The data model of the publication part is based on BibTeX [24], a popular literature management system for LaTeX [16].

A typical list of posts is depicted in Figure 1 which shows bookmark and publication posts containing the tag *web*. The page is divided into four parts: the header (showing information such as the current page and path, navigation links and search boxes), two lists of posts – one for bookmarks and one for publications – each sorted by date in descending order, and a list of tags related to the posts. This scheme holds for all pages showing posts and allows for navigation in all dimensions of the folksonomy.

A detailed view of one bookmark post from the list in Figure 1 can be seen in Figure 2. The first line shows in bold the title of the bookmark which has the URL of the bookmark as underlying hyperlink. The second line shows an optional description the user can assign to every post. The last two lines belong together and show detailed information: first, all the tags the user has assigned to this post (*web, service, tutorial, guidelines* and *api*), second, the user name of that user (*hotho*) followed by a note, how many users tagged that specific resource. These parts have underlying hyperlinks, leading to the corresponding tag pages of the user (`/user/hotho/web`, `/user/hotho/service`, ...), the users page (`/user/hotho`) and a page showing all four posts (i. e., the one of user *hotho* and those of the three other users) of this resource (`/url/$r$`). The last part shows the posting date and time followed by links for actions the user can do with this post – depending on if this is his own (*edit*, *delete*) or another user's post (*copy*).

The structure of a publication post displayed in BibSonomy is very similar, as seen in Figure 3. The first line shows again the title of the post, which equals the title of the publication in BibTeX. It has an underlying link leading to a page which shows detailed information on that post. This line is followed by the authors or editors of the publication, as well as journal or book title and the year. The next lines show the tags assigned to this post by the user, whose user name comes next followed by a note how many people tagged this publication. As described for bookmark posts, these parts link to the respective pages. After the date and time the user posted this entry follow the actions the user can do, which in this case include picking the entry for later download, copying it, accessing the URL of the entry or viewing the BibTeX source code.

As with the del.icio.us dataset, we created a dump of the system, and calculated monthly snapshots, based on the timestamps. This resulted in 20 datasets. The most recent one, from July 31st, 2006, contains data from $|U| = 428$ users, $|T| = 13,108$ tags, $|R| = 47,538$ resources, connected by $|Y| = 161,438$ tag assignments.

4. Small Worlds in Three-Mode Networks

As expected, the tagging behavior in del.icio.us displays a fat-tailed distribution: Fig. 4 shows the fraction of tags, users, and resources, respectively, occurring in a given number of TAS. We observe that the probability distributions for tags and resources display a rather clean power-law tail, while the distribution for users features a different behavior for small frequencies. This suggests the existence of two classes of users, with very active users

[4] `http://www.bibsonomy.org`

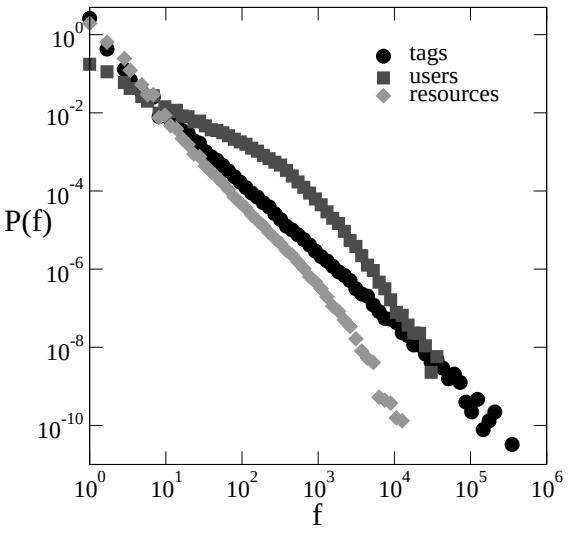

Fig. 4. Probability distribution of the frequency of occurrence of tags, users and resources in del.icio.us. For tags (circles), the abscissa of each point corresponds to a given frequency, and its ordinate is the fraction of tags that occur with the selected frequency. The same holds for users (squares) and resources (diamonds).

(less than a thousandth of the total) following a different scaling as compared to the vast majority of less active users.

We will now investigate whether folksonomies feature small-world properties. To this end, we will define the notions of characteristic path length and clustering coefficient in tripartite hypergraphs such as folksonomies, and apply these to the datasets introduced in Section 3.

4.1. Characteristic Path Length

The *characteristic path length* of a graph [30] describes the average length of a shortest path between two random nodes in the graph. If the characteristic path length is small, few hops will be neccessary, on average, to get from a particular node in the graph to any other node.

As folksonomies are triadic structures of *(tag, user, resource)* assignments, the user interface of such a folksonomy system will typically allow the user to jump from a given tag to (a) any resource associated with that tag, or to (b) any user who uses that tag, and vice versa for users and resources. Thus, the effort of getting from one node in the folksonomy to another can be measured by counting the *hyper*edges in shortest paths between the two. Here a path is defined as a sequence of *hyper*edges such that each hyperedge shares at least the user or the resource or the tag with the following hyperedge.

More precisely, let $v_1, v_2 \in T \cup U \cup R$ be two nodes in the folksonomy, and $(t_0, u_0, r_0), \ldots, (t_n, u_n, r_n)$ a minimal sequence of TAS such that, for all k with $0 \leq k < n$, $(t_k = t_{k+1}) \vee (u_k = u_{k+1}) \vee (r_k = r_{k+1})$, and $v_1 \in \{t_0, u_0, r_0\}, v_2 \in \{t_n, u_n, r_n\}$. Then we call $d(v_1, v_2) := n$ the *distance* of v_1 and v_2. We compute path lengths within connected components only.

Following Watts [30], we define $\bar{d}_v$ as the mean of $d(v, u)$ over all $u \in (T \cup U \cup R) - \{v\}$, and call the median of the $\bar{d}_v$ over all $v \in T \cup U \cup R$ the *characteristic path length* L of the folksonomy.

In Section 4.3, we will analyse the characteristic path length on our datasets. As computing the characteristic path length is prohibitively expensive for graphs of the size encountered here, we sampled 200 nodes randomly from each graph and computed the path lengths from each of those nodes to all others in the folksonomy using breadth-first search.

4.2. Clustering Coefficients

Clustering or transitivity in a network means that two neighbors of a given node are likely to be directly connected as well, thus indicating that the network is locally dense around each node. To measure the amount of clustering around a given node v, Watts [30] has defined a clustering coefficient γ_v (for normal, non-hyper-graphs). The clustering coefficient of a graph is γ_v averaged over all nodes v.

Watts [30, p. 33] defines the clustering coefficient γ_v as follows ($\Gamma_v = \Gamma(v)$ denotes the neighborhood of v):

> Hence γ_v is simply the net fraction of those possible edges that actually occur in the real Γ_v. In terms of a social-network analogy, γ_v is the degree to which a person's acquaintances are acquainted with each other and so measures the *cliquishness* of v's friendship network. Equivalently, γ_v is the probability that two vertices in $\Gamma(v)$ will be connected.

Note that Watts combines two aspects which are *not* equivalent in the case of three-mode data. The first one is: how many of the possible edges around a node do actually occur, i. e., does the neighbor-

hood of the given vertex approach a clique? The second aspect is that of neighbors of a given node being connected themselves.

Following the two motivations of Watts, we thus define two different clustering coefficients for three-mode data:

Cliquishness: From this point of view, the clustering coefficient of a node is high iff many of the possible edges in its neighborhood are present. More formally: Consider a resource r. Then the following tags T_r and users U_r are connected to r: $T_r = \{t \in T \mid \exists u : (t, u, r) \in Y\}$, $U_r = \{u \in U \mid \exists t : (t, u, r) \in Y\}$. Furthermore, let $tu_r := \{(t, u) \in T \times U \mid (t, u, r) \in Y\}$, i. e., the (tag, user) pairs occurring with r.

If the neighborhood of r was maximally cliquish, all of the pairs from $T_r \times U_r$ would occur in tu_r. So we define the clustering coefficient $\gamma_{cl}(r)$ as:

$$\gamma_{cl}(r) = \frac{|tu_r|}{|T_r| \cdot |U_r|} \quad \in [0, 1] \tag{1}$$

i. e., the fraction of possible pairs present in the neighborhood. A high $\gamma_{cl}(r)$ would indicate, for example, that many of the users related to a resource r assign overlapping sets of tags to it.

The same definition of γ_{cl} stated here for resources can be made symmetrically for tags and users.

Connectedness (Transitivity): The other point of view follows the notion that the clustering around a node is high iff many nodes in the neighborhood of the node were connected even if that node was not present.

In the case of folksonomies: consider a resource r. Let $\widetilde{tu_r} := \{(t, u) \in T \times U \mid (t, u, r) \in Y \wedge \exists \tilde{r} \neq r : (t, u, \tilde{r}) \in Y\}$, i. e., the (tag, user) pairs from tu_r that also occur with some other resource than r. Then we define:

$$\gamma_{co}(r) := \frac{|\widetilde{tu_r}|}{|tu_r|} \quad \in [0, 1] \tag{2}$$

i. e., the fraction of r's neighbor pairs that would remain connected if r were deleted. γ_{co} indicates to what extent the surroundings of the resource r contain "singleton" combinations *(user, tag)* that only occur once.

Again, the definition works the same for tags and users, and the clustering coefficients for the whole folksonomy are defined as the arithmetic mean over the nodes.

The following example demonstrates that the clustering coefficients γ_{cl} and γ_{co} do indeed capture different characteristics of the graph and are not intrinsically related. One might suspect that there is a systematic connection between the two, such as $\gamma_{cl}(r) < \gamma_{cl}(s) \Rightarrow \gamma_{co}(r) < \gamma_{co}(s)$ for nodes $r, s \in T \cup U \cup R$, or similarly, on the level of the whole folksonomy, $\gamma_{co}(\mathbb{F}) < \gamma_{co}(\mathbb{G}) \Rightarrow \gamma_{cl}(\mathbb{F}) < \gamma_{cl}(\mathbb{G})$.

However, this is not the case: consider a folksonomy $\mathbb{F}$ with tag assignments $Y_1 = \{(t_1, u_2, r_2), (t_1, u_1, r_1), (t_1, u_1, r_2), (t_1, u_2, r_1), (t_1, u_3, r_3), (t_2, u_3, r_3), (t_2, u_4, r_4)\}$. Here we have $\gamma_{cl}(t_1) \approx 0.556 > \gamma_{cl}(t_2) = 0.5$, but $\gamma_{co}(t_1) = 0.2 < \gamma_{co}(t_2) = 0.5$.

Also, there is no monotonic connection when considering the folksonomy as a whole. For the whole folksonomy $\mathbb{F}$, we have $\gamma_{cl}(\mathbb{F}) \approx 0.906$, $\gamma_{co}(\mathbb{F}) \approx 0.470$.

Considering a second folksonomy $\mathbb{G}$ with tag assignments $Y_2 = \{(t_1, u_1, r_1), (t_1, u_1, r_3), (t_1, u_2, r_2), (t_1, u_3, r_2), (t_2, u_1, r_2), (t_2, u_2, r_1), (t_2, u_2, r_2), (t_2, u_2, r_3), (t_3, u_1, r_2), (t_3, u_2, r_2)\}$, we see that $\gamma_{cl}(\mathbb{G}) = 0.642, \gamma_{co}(\mathbb{G}) = 0.669$, thus $\gamma_{cl}(\mathbb{F}) > \gamma_{cl}(\mathbb{G})$ while $\gamma_{co}(\mathbb{F}) < \gamma_{co}(\mathbb{G})$.

4.3. Experiments

4.3.1. Setup

In order to check whether our observed folksonomy graphs exhibit small world characteristics, we compared the characteristic path lengths and clustering coefficients with random graphs of a size equal in all dimensions T, U, and R as well as Y to the respective folksonomy under consideration.

Two kinds of random graphs are used for comparison:

Binomial: These graphs are generated similar to an Erdős random graph $G(n, M)$ [3], where n is the number of nodes and M is the number of edges. Adapting the construction of $G(n, M)$ to the structure of folksonomies, T, U, R are taken as nodes from the observed folksonomies, and $|Y|$ many hyperedges are then created by picking the three endpoints of each edge from uniform distributions over T, U, and R, respectively, leading to a binomial distribution of degrees over the nodes.

Permuted: These graphs are created by using T, U, R from the observed folksonomy. The tagging relation Y is created by taking the TAS from the original graph and permuting each dimension of Y independently (using a Knuth Shuffle [15]), thus creating a random graph with the same degree sequence as the observed folksonomy.

As stated above, the computation of the characteristic path length is prohibitively expensive for graphs of our size. As for the del.icio.us and BibSonomy datasets, we sampled 200 nodes randomly from each graph and computed the path lengths from each of those nodes to all others in the folksonomy.

Although we did not take any specific measures to keep the graphs connected, almost all nodes lie in a giant connected component for all data sets. The largest number of nodes disconnected from the giant component we encountered were 351 out of 1,539,326 nodes (.02% of the nodes) for the actual del.icio.us data in month 9, and 88 out of 58,879 (.15% of the nodes) for the BibSonomy data in month 19. The random graphs showed even fewer disconnected nodes, the maximum numbers being 18 nodes out of 951,513 (.002%, del.icio.us, month 7, permuted) and 6 out of 60,984 (.01%, BibSonomy, month20, permuted).

For all experiments involving randomness (i. e., those on the random graphs as well as the sampling for characteristic path lengths), 20 runs were performed to ensure consistency. The presented values are the arithmetic means over the runs; the deviations across the runs were negligible in all experiments.

4.3.2. First Observations

Figures 5–7 show the results for the clustering coefficients and the characteristic path lengths for both datasets, plotted against the number $|Y|$ of tag assignments for the respective monthly snapshots.

Both folksonomy datasets under consideration exhibit the small world characteristics as defined at the beginning of this section. Their clustering coefficients are extremely high, while the characteristic path lengths are comparable to (BibSonomy) or even considerably lower (del.icio.us) than those of the binomial random graphs.

del.icio.us. In the del.icio.us dataset (Figures 6 and 7, right hand sides), it can be seen that both clustering coefficients are extremely high at about 0.86, much higher than those for the permuted and binomial random graphs. This could be an indication of coherence in the tagging behaviour: if, for example, a given set of tags is attached to a certain kind of resources, users do so consistently.

On the other hand, the characteristic path lengths (Figure 5, right) are considerably smaller than for the random binomial graphs, though not as small as for the permuted setting. The comparison with the random binomial graph shows the small world behavior of the human tagging activity. Our interpretation of the comparison with the permuted setting is that the latter maintains the structural features of the human tagging behavior, while introducing additional links between personomies of otherwise unrelated users; leading them thus out of their 'caveman world' [30].

Interestingly, the path length has remained almost constant at about 3.5 while the number of nodes has grown about twentyfold in the observation period. As explained in Section 4.1, in practice this means that on average, every user, tag, or resource within del.icio.us can be reached within 3.5 mouse clicks from any given del.icio.us page. This might help to explain why the concept of serendipitous discovery [19] of contents plays such a large role in the folksonomy community – even if the folksonomy grows to millions of nodes, everything in it is still reachable within few hyperlinks.

BibSonomy. As the BibSonomy system is rather young, it contains roughly two orders of magnitude fewer tags, users, resources, and TAS than the del.icio.us dataset.

On the other hand, the values show the same tendencies as in the del.icio.us experiments.

Figures 6 and 7 (left) show that clustering is extremely high at $\gamma_{cl} \approx 0.96$ and $\gamma_{co} \approx 0.93$ – even more so than in the del.icio.us data.

At the same time, Figure 5 shows that the characteristic path lengths are somewhat larger, but at least comparable to those of the binomial graph.

There is considerably more fluctuation in the values measured for BibSonomy due to the fact that the system started only briefly before our observation period. Thus, in that smaller folksonomy, small changes, such as the appearance of a new user with a somewhat different behaviour, had

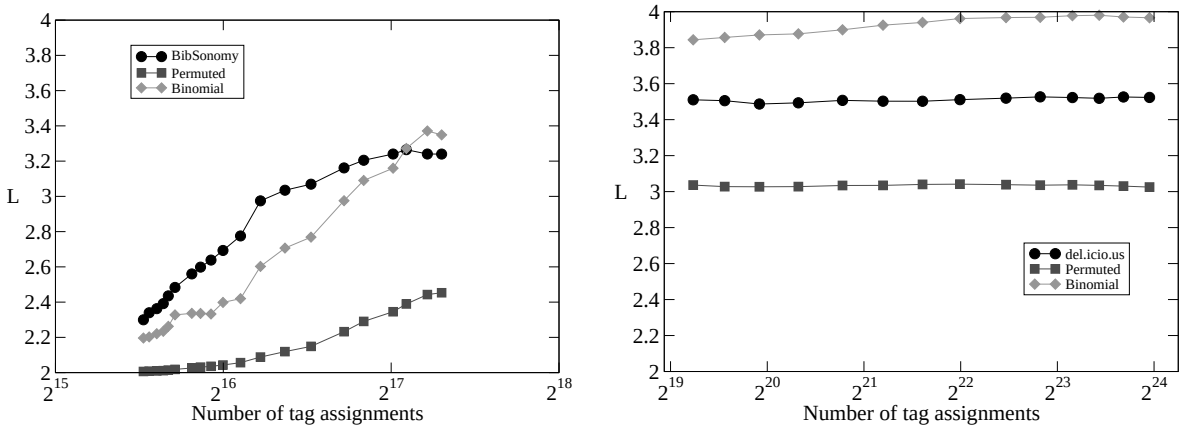

Fig. 5. Characteristic path length for the BibSonomy folksonomy (left) and the del.icio.us folksonomy (right) as a function of network growth, measured as the total number of tagging events in the folksonomy. The path length for random graphs of corresponding size is also shown (permuted and binomial, see text). Note how the charateristic path length takes quite similar low values, typical of small world networks, for all graphs.

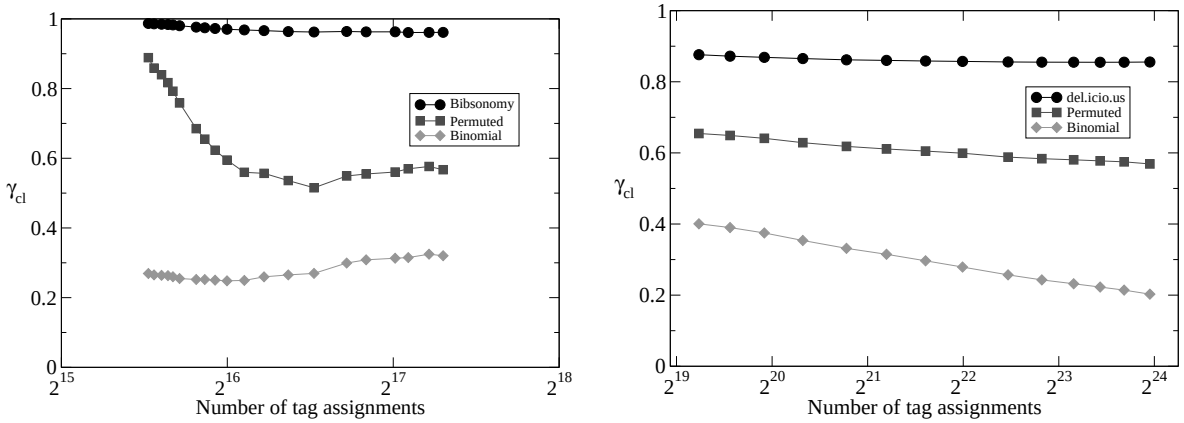

Fig. 6. Cliquishness for the BibSonomy folksonomy (left) and the del.icio.us folksonomy (right) as a function of network growth, measured as the total number of tagging events in the folksonomy. The cliquishness for random graphs of corresponding size is also shown (permuted and binomial). The cliquishness for the folksonomy networks takes rather high values, higher than the corresponding random graphs.

more impact on the values measured in our experiments.

Furthermore, many BibSonomy users are early adopters of the system, many of which know each other personally, work in the same field of interest, and have previous experience with folksonomy systems. This might also account for the very high amount of clustering.

4.4. Characteristic Path Length for Tags

Figure 5 demonstrated that the characteristic path length L of the two folksonomies under consideration grows comparably to that of the respective "binomial" random folksonomies. As the number of resources $|R|$ dominates the numbers of tags $|T|$ and users $|U|$ by almost one and two orders of magnitude, resp., L is heavily influenced by the characteristic path length for resources.

In order to get an insight into the behaviour of tags in that respect, we computed the characteristic path length as described in 4.1, but this time taking only the values $\bar{d}_t$ for tags $t \in T$ into account for L.

Figure 8 shows the growth of L for tags in the BibSonomy and del.icio.us folksonomies. Interestingly, the average path length for tags in the BibSonomy dataset is much larger than that for the

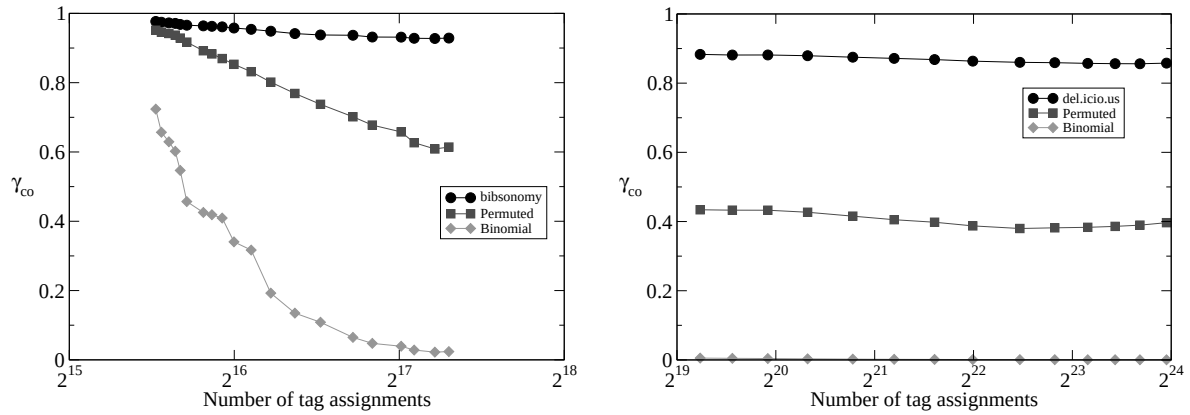

Fig. 7. Connectedness/Transitivity for the BibSonomy folksonomy (left) and the del.icio.us folksonomy (right) as a function of network growth, measured as the total number of tagging events in the folksonomy. The connectedness/transitivity for random graphs of corresponding size is also shown (permuted and binomial). As in the case of cliquishness, the values of the connectedness/transitivity are very high for the folksonomy networks, at odds with the corresponding random graphs.

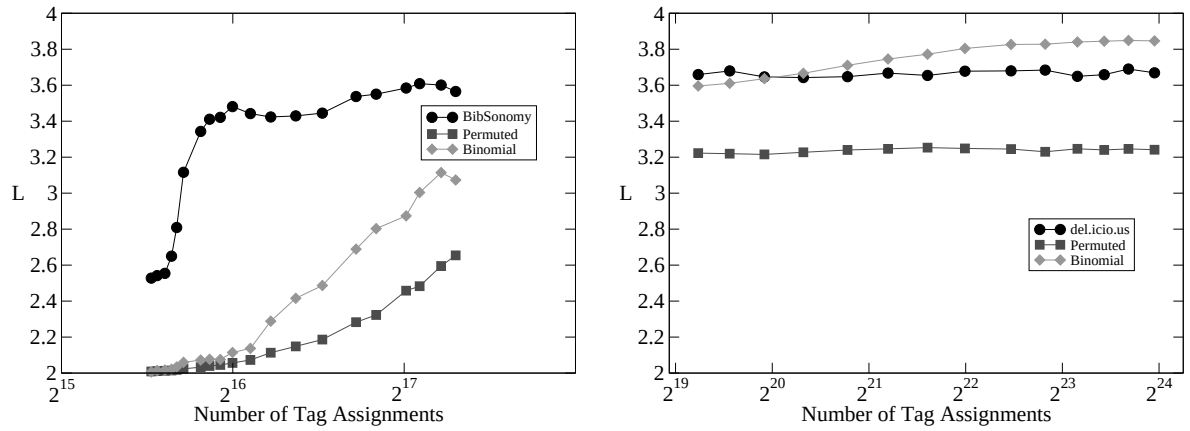

Fig. 8. Characteristic path length L computed by considering tags only, in BibSonomy (left) and del.icio.us (right). The figures show that the characteristic path length restricted to tags is almost constant, starting from a very early point in the growth of the folksonomy.

random folksonomies and rises to about 3.5 to 3.6 very early in the life of BibSonomy, but then remains almost constant. In the del.icio.us folksonomy, which is considerably larger than the latter one, the characteristic path length for tags still remains almost the same at about 3.7.

Our interpretation is that even a small number of early folksonomy users introduces a considerable amount of idiosyncratic vocabulary, large parts of which are rather distant from the rest of the folksonomy. Interestingly, even in the larger del.icio.us folksonomy, the average tag is still farther away from the rest of the folksonomy at $L \approx 3.7$ as opposed to the $L \approx 3.5$ from Figure 5 which is largely dominated by resources. This is surprising, as the average tag occurs in about 9 times as many tag assignments as the average resource.

4.5. A Closer Look on del.icio.us

We will conclude this section by a closer look on how the characteristic path length, the cliquishness, and the connectedness are distributed over the users, tags, and resources in del.icio.us.

To this end, we have computed the co-occurence graphs for the three dimensions users, tags, and resources. More formally, the co-occurence graph for the tags has the set T of tags as vertices; and two tags t_1 and t_2 are connected by an undirected edge, iff there is at least one resource r and one

user u such that $(u, t_1, r), (u, t_2, r) \in Y$. The characteristic path length and clustering coefficients of the (non-hyper) co-occurrence graphs are shown in the left diagrams of Figures 9 and 10. The characteristic path length was approximated by taking a 200-node sample, and for the clustering coefficient the approximation from [25] was used with a precision of $\epsilon = 10^{-3}$ and a probability of 0.99.

The left diagram of Figure 9 shows the characteristic path lengths of the three co-occurrence graphs.[5] The result is as expected: the set of resources is almost an order of magnitude larger than the set of tags, which is about the same ratio larger than the set of users. The larger graphs have higher characteristic path lengths.

The right diagram of Figure 9 shows the different contributions of tags, users, and resources to the del.icio.us path length shown in the right diagrams of Figures 5 and 8. For computing the values, the random nodes have been drawn only from the respective classes. The low path length for the user nodes indicates that personomies (defined as the set of TAS associated with a single given user) are a structural element in a folksonomy: Consider the extreme case that all personomies are completely disjoint. Then the users are the central nodes in their connected component (which equals their personomy), and have thus shorter characteristic path lengths in average.

The characteristic path lengths of the tags and resources in the right diagram are reversed compared to the left diagram. This is likely to be due to the fact that users tend to invent new, personal tags – which are further away from the core of the folksonomy – whereas there is less divergence of the URLs to be included in the system.

Figure 10 shows the clustering coefficient of del.icio.us for the three co-occurrence graphs (left) and connectedness of the hypergraph by dimension (right). Both diagrams show that the neighborhoods around tags and resources are denser than around users. This is likely to stem from the fact that users usually have different interests. An interesting observation is that the user curve decreases over time in the left diagram, while it increases in the right one. Both effects result from the increasing number of neighbors over time. The clustering coefficient decreases because fewer and fewer neighbors are connected to each other when the size of the neighborhood increases. γ_{co}, on the other hand, increases over time, as – with growing neighborhoods – it becomes more likely that, for a given TAS, another user has assigned exactly the same tag to the same resource. This indicates that, although each user invents new, personal tags, a form of consensus grows with time on the vocabulary associated to the same resource, i.e. a common semantics emerges.

[5] Note that the three values are measured in three different graphs.

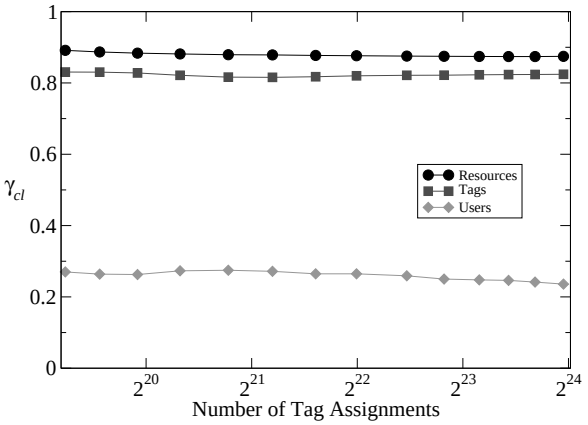

Fig. 11. Cliquishness of del.icio.us for the three dimensions in the hypergraph. The diagram shows that the cliquishness for tags and resources is high – indicating that if, e.g., a resource is given certain tags and tagged by certain users, many of the possible combinations of those tags and users are likely to occur. On the other hand, the cliquishness for users is considerably lower, indicating different fields of interest for each user.

Figure 11 shows that the cliquishness for tags and resources is high – indicating that if, e.g., a resource is given certain tags and tagged by certain users, many of the possible combinations of those tags and users are likely to occur, i. e., there is a natural set of tags which seem appropriate for a given resource, and vice versa, for a given tag, the users using that tag agree to a large extent on which resources should be tagged with it. On the other hand, the cliquishness for users is considerably lower. This demonstrates that, other than tags and resources, users typically have different fields of interest and thus are connected to elements of the other dimensions which will not necessarily occur in many of the possible combinations.

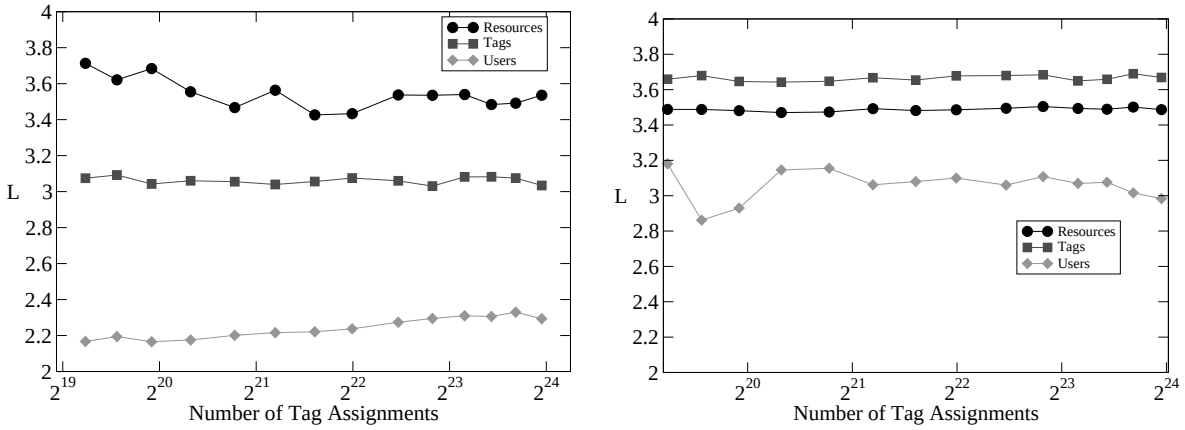

Fig. 9. Characteristic path lengths of del.icio.us in the three co-occurrence graphs (left) and in the three modes of the hypergraph (right). Left: The differences reflect the different sizes of the co-occurrence graphs. Larger graphs have higher characteristic path length. Right: the curves display the separate contributions of tags, users and resources to the del.icio.us path length shown in the right diagrams of Figures 5 and 8. The low characteristic path lengths for user nodes reflect the structural importance of personomies in the hypergraph.

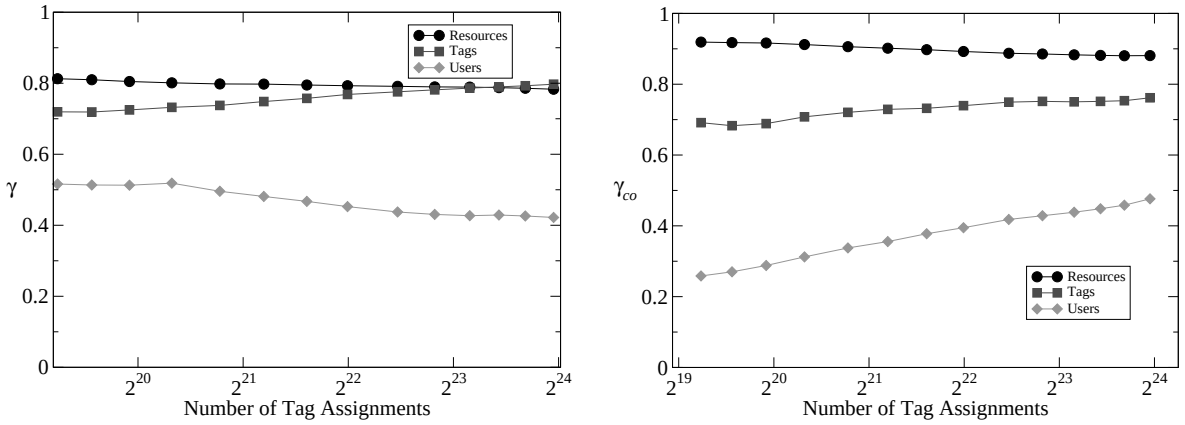

Fig. 10. Clustering coefficient of del.icio.us for the three co-occurrence graphs (left) and connectedness of the hypergraph for the three modes (right). Both diagrams show that the neighborhoods around tags and resources are denser than around users. The user curve in the left diagram decreases over time, while it increases in the right one.

5. Networks of Tag Co-occurrence

In order to investigate the emergent semantic properties of the folksonomy, we focus on the relations of co-occurrence among tags. Since the process of tagging is inclusive [9], and large overlap often exists among resources marked with different tags, the relations of co-occurrence among tags expose the semantic aspects underling collaborative tagging, such as homonymy, synonymy, hierarchical relations among tags and so on.

The simplest way to study tag co-occurrence at the global level is to define a network of tags, where two tags i and j are linked if there exists a post where they have been associated by a user with the same resource. A link weight for two nodes i and $j(\neq i)$ can be introduced and defined as the number of posts where they appear together. Formally, we define $W(i,j)$, i. e., the set of posts where i and j appear together, as

$$W(i,j) := \{(u,r) \in U \times R \mid [(i,u,r) \in Y] \wedge [(j,u,r) \in Y]\}\,, \quad (3)$$

and define the link weight $w(i,j) := |W(i,j)|$. This is thus a weighted version of the co-occurrence graph defined in Section 4.5.

The above link strength defines on $T \times T$ a symmetric similarity matrix which is analogous to

the usual adjacency matrix in graph theory. The strength s_t of a node t is defined as [1]

$$s_i := \sum_{j \neq i} w(i,j)\,. \tag{4}$$

5.1. Cumulative probability distribution of node strength

A first statistical characterization of the network of tags is afforded by the cumulative probability distribution $P_>(s)$, defined as the probability of observing a strength in excess of s. These distributions are displayed for del.icio.us and BibSonomy in Figs. 12 and 13, respectively. This is a standard measure in complex network theory and plays the same role of the degree distribution in unweighted networks. We observe that $P_>(s)$ is a fat-tailed distribution for both folksonomies: this is related to a lack of characteristic scale for node strengths and is one of the typical fingerprints of an underlying complex dynamics of interacting human agents [28,29]. A coarse indicator such as $P_>(s)$, despite its simplicity, is able to point out anomalous activity (i. e., spam) within the investigated folksonomies. (Figs. 12 and 13). For example, in Fig. 12 the black curve corresponds to the raw co-occurrence network, and the two steps indicated by arrows are related to an excess of links with a specific weight, and can be linked to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps altogether (dark gray). Quite interestingly, on filtering out these undesired (and probably automatically generated) contributions, the probability distributions for del.icio.us (Fig. 12) and BibSonomy (Fig. 13) become rather similar, even though the two systems under study are dramatically different in terms of user base, size and age.

Uncovering the detailed "microscopic" mechanism responsible for the observed distribution is a daunting task. A simple way to identify the contribution of semantics – and in general of human activity – to those distributions consists in destroying semantics altogether by randomly shuffling tags among TAS entries. In the tripartite graph view of the folksonomy, this corresponds to introducing a random permutation of the set of tags T, biunivocally mapping each tag $t \in T$ into a corresponding tag t'. Correspondingly, each hyperedge (t,u,r) is mapped into a new hyperedge (t',u,r). Each post in the original folksonomy corresponds to a new post with the same number of tags, but now the co-occurrence relations are completely different.[6]

In Figs. 12 and 13 we show that by performing this shuffling operation (light gray dots) the distribution is only marginally affected. Far from being obvious, this shows that the global frequencies of tags – and not their co-occurrence relations – are the main factors shaping the distribution $P_>(s)$. In other words, the fat-tailed nature of $P_>(s)$ is induced by the distribution of tag frequencies, which has been known to be fat-tailed [9,6], in analogy to Zipf's law (also observed in human languages).

In order to probe deeper into the structure of the co-occurrence network and recognize the contribution of semantics, we need to compute observables more sensitive to correlations and to the local structure of the network. To this end, a useful quantity studied in complex networks is the nearest neighbor connectivity. Given a node i, we define its average nearest-neighbor strength as:

$$S_{nn}(i) = \frac{1}{k_i} \sum_{j=1}^{k_i} s_j\,, \tag{5}$$

where k_i denotes the number of links with non-zero weight connected to node i, and where s_j denotes the strength of node j (see Eq. 4; the neighbors of i are written as $j = 1, \ldots, k_i$ here for ease of notation). The concept of nearest neighbor needs to be clarified here. In principle all nodes are connected to each other in a weighted graph, but in this particular context we ignore the existence of those links that have weight zero. Consequently, we consider two nodes as nearest neighbors, iff there exists a link with non-zero weight connecting them.

The relation of the average nearest neighbor strengths $S_{nn}(i)$ to the node strengths s_i provides information on correlations among the strength of nodes and therefore is also known in literature as node nearest-neighbor strength correlation $S_{nn}(s)$ [1]. When referred to unweighted networks, i. e., where all existing links have unit strength, S_{nn} is able to discriminate between technological networks, where S_{nn} and s are negatively correlated, and social networks, where, on the contrary, $S_{nn}(s)$ displays an increasing behavior. These two networks with opposite behaviors are commonly

[6] In difference to the permuted graph introduced in Section 4.3.1, we shuffle the tags only.

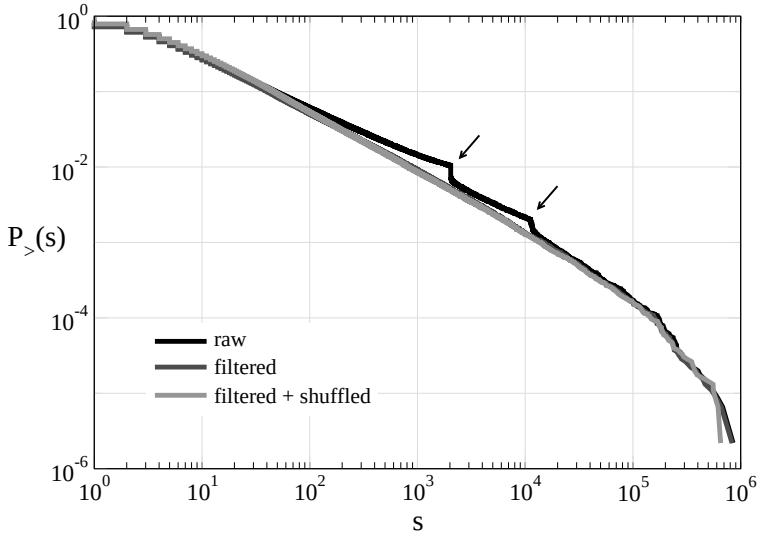

Fig. 12. Cumulative strength distribution for the network of tag co-occurrence in del.icio.us. $P_>(s)$ is the probability of having a node with strength in excess of s. The black curve corresponds to the whole co-occurrence network. The two steps indicated by arrows correspond to an excess of links with a specific weight and can be related to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps (dark gray). Shuffling the tags contained in posts (light gray) does not affect significantly the cumulated weight distribution. This proves that such a distribution is uniquely determined by tag frequencies within the folksonomy, and not by the semantics of co-occurrence.

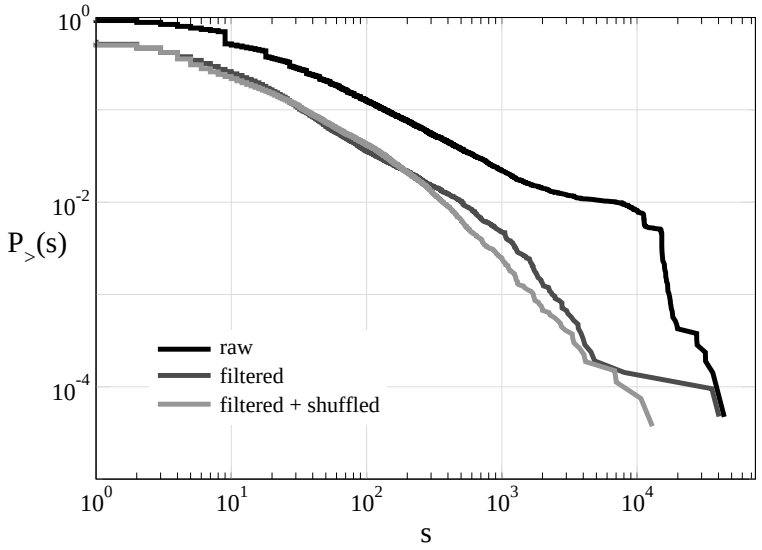

Fig. 13. Cumulative strength distribution $P_>(s)$ for the network of tag co-occurrence in BibSonomy. The black curve corresponds to the whole co-occurrence network. The irregular behavior for high strengths can be linked to spamming activity: identified spam in BibSonomy consists of posts with a large number of tags, as well as a large number of posts with exactly 10 tags injected by a small group of spammers. Excluding the above posts from the analysis (dark gray), the distribution becomes smooth and similar to the filtered one reported for del.icio.us in Fig. 12. Similarly, shuffling the tags contained in posts (light gray) has a small effect on the cumulated weight distribution.

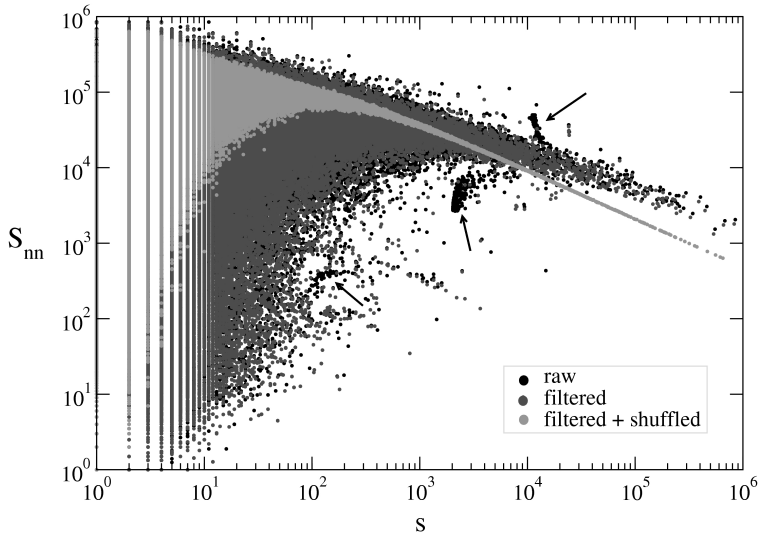

Fig. 14. Average nearest-neighbor strength S_{nn} of nodes (tags) in relation to the node (tag) strengths s, in del.icio.us. Black dots correspond to the whole co-occurrence network. Assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. A few clusters (indicated by arrows) stand out from the main cloud of data points. As in Fig. 12, such anomalies correspond to spamming activity and can be removed by filtering out posts containing an excessive number of tags (dark grey). Shuffling the tags (light grey) affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the network of co-occurrence beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics.

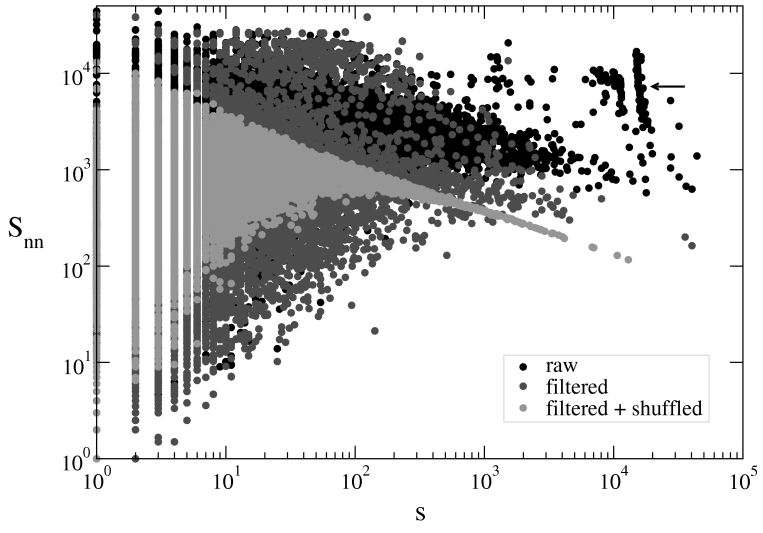

Fig. 15. Average nearest-neighbor strength S_{nn} of nodes (tags) in relation to the node (tag) strengths s, in *BibSonomy*. Black dots correspond to the whole co-occurrence network. The scatter plot is qualitatively very similar to the one reported in Fig. 14 for del.icio.us: assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. Again, a few clusters (indicated by arrows) stand out from the main cloud of data points and their presence can be linked to spamming activity. They disappear when we filter out posts containing an excessive number of tags (dar grey). Shuffling the tags (light gray) has the same effect as in Fig. 14, and the same observations apply.

referred to as disassortative and assortative mixing networks, respectively [22,5].

Figs. 14 and 15 display our results for del.icio.us and BibSonomy, respectively. In the figures, each dot corresponds to a node of the network (i. e., a tag), with its strength s as the abscissa and the average strength of its neighbors S_{nn} as the ordinate. Both quantities span several orders of magnitude, hence we use a logarithmic scale along both axes to display the global features of the scatter plot. This is related to the fat-tailed behavior observed for the strength distribution $P_>(s)$, which is in fact recovered by projecting the data points along the s-axis and computing the cumulative distribution.

The broad cloud of points in the scatter plots can also be studied by considering the distribution of S_{nn} values for nodes having a given strength s. In Figure 16 two such histograms are shown for the case of the del.icio.us data of Figure 14. Two values of the strength s were chosen, in order to show more clearly the distribution of scattered points in two qualitatively different regions of Fig. 14. The two chosen regions are only slightly affected by spam. The histogram in the upper panel of Fig. 16 ($s = 10$) displays a broad distribution, characterized by a most probable high value of S_{nn} and a broad algebraic tail for smaller values. Shuffling the tags (light grey) dramatically changes the distribution and the tail disappears, indicating that the tail behavior is not a trivial frequency effect.

In the scatter plots, the anomalous activity such as spam is more clearly detectable, and its contribution appears in the form of foreign clusters (indicated by arrows) that clearly stand out from the otherwise smooth cloud of data points, a fact that reflects the anomalous nature of their connections with the rest of the network. Excluding spam from the analysis, those clusters disappear altogether (dark gray dots). The general shape of the cloud of data points remains unchanged, even though, in the case of BibSonomy, it shifts down towards lower strengths. This happens because BibSonomy is a smaller system and spam removal has a more significant global impact on the network and the strengths of its nodes.

Overall, the plots for del.icio.us and BibSonomy look quite similar, and this suggests that the features we report here are generally representative of collaborative tagging systems. An assortative region (S_{nn} roughly increasing with s) is observed for low values of the strength s, while disassortative behavior (S_{nn} decreasing with s) is visible for high values of s. As we have already done for the probability distribution $P_>(s)$, we can highlight the contribution of semantics by randomly shuffling tags in TAS entries (light gray dots in Fig. 14 and 15). In this case, shuffling the tags affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the co-occurrence network beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics. Interestingly, the main effect seems to be the disappearance of points in the assortative (low strengh) region of the plot, possibly identifying this region as the one exposing semantically relevant connections between tags. Notice, for example, the disappearance of a whole cloud of points at the top-left of Fig. 15: those points represent nodes (tags) with low strength that are attached preferentially to nodes of high strength. Similarly, in Fig. 14, the highly populated region with s roughly ranging between 10 and a few thousands also disappears when tag shuffling is applied. Those data points also represent low-strength nodes (tags) preferentially connected with higher-strength nodes (tags). Such properties are commonly found in hierarchically organized networks, and could be related to an underlying hierarchical organization of tags [10].

5.2. Spam detection and characterization

From a semantic point of view, spam contaminates the system. Therefore the removal of spam is of primary importance if the system is to be studied from a semantic point of view. A crude way to remove spam is to cut off all posts that contain more than a certain arbitrary (but large) number of TAS, i. e., number of associated tags. The drawback of this fast method is that also semantic valuable posts with large number of tags are filtered out. An elegant alternative way to proceed in spam removal is to consider the structure of the quantity $S_{nn}(s_i)$. As already mentioned, most spam in the system is easily spotted by looking at the scattered plot in Fig. 14. In that picture, three discrete features can be ascribed to spam. In order to better understand the structure of spam, we plot the same scatter plot of Fig. 14, but raising the single link weights to a power $\gamma \in \{1.0, 0.8, 0.2\}$. The corresponding scatter plots are shown in Fig. 17. The

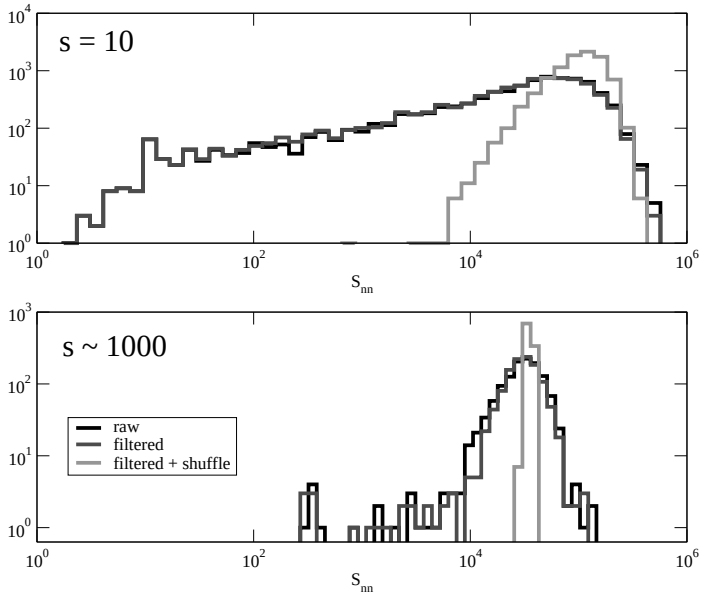

Fig. 16. Histograms of the average nearest-neighbor strength S_{nn} for nodes having a given strength s, in del.icio.us. The data are extracted from the scatter plot of Fig. 14 for two values of the strength: $s = 10$ (upper panel) and $s = 1000$ (lower panel). In the upper panel ($s = 10$) a broad distribution appears, characterized by a most probable high value of S_{nn} and a broad algebraic tail for smaller values. Shuffling the tags (light grey) dramatically changes the distribution and the tail disappears, indicating that the tail behavior is not a trivial frequency effect. In the lower panel ($s = 1000$) the tail is narrower, but the shuffling procedure gives qualitatively the same result.

value $\gamma = 1$ trivially reproduces the plot of Fig. 14, while the value $\gamma = 0$ (not shown) would reproduce the unweighted analogous topological quantity of $K_{nn}(k)$. As the value of γ decreases, the spam spots displays two different behaviors. The spots indicated with number 1 and 2 remain on the $S_{nn} = s$ line, while spot 3 merges and disappears into the main cloud. This suggests that spots 1 and 2 are posts made of a large number of TAS with quite uncommon tags in the system, such that these posts can be considered as cliques rather decoupled from the system. In fact, an isolated clique would lie always on the $S_{nn} = s$ and $K_{nn} = k$ lines. The comet like shape of those spots is due to the presence of more common tags that are present in the system. Spot number 3 is in fact strictly bound to the rest of the system; this is the reason why it disappears as the value of γ tends to 0. A detailed inspection of spots 1, 2 and 3 is provided in the caption of Fig. 17.

5.3. Strength as a function of degree

We could ask to which extent the strength of a node in the network is important with respect to the topological degree of a node. We recall that the strength of a node is the sum of the weights of the link connecting it to the network, i. e., the number of its total co-occurrences, while the degree of a node is the number of different nodes co-occurring with it. The quantity that can be analyzed is $s_i(k_i)$, i. e., the relation of the strengths of the nodes to their topological degrees. The relation between strength and degree has been already studied in the literature for some weighted networks and dependencies of the type $s(k) \approx k$ and $s(k) \approx k^{1.5}$ have been measured in the air transportation network and network of article co-authorship, respectively [1].

The corresponding scatter plot is shown in Fig. 18. A linear behavior of the relation (k_i, s_i) for tags i would indicate that these two statistical quantities are equivalent. Rare tags that appear only once (i. e., in one post) in the whole system must have $s(k) = k$ by definition, while we expect the strength to become more important for the most frequent tags, which occur in a lot of posts. An inspection of Fig. 18 shows three different regimes: for very low degrees $s(k) \approx k$, then an extended crossover region followed by an asymptotic regime above $k \approx 500$ featuring a pronounced power $s(k) \gtrsim k^{1.5}$. In Fig. 18 we also show the $s(k)$

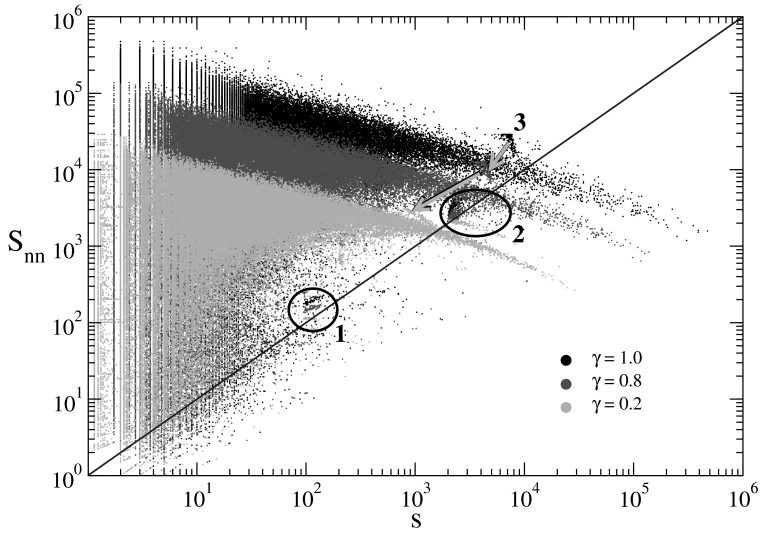

Fig. 17. Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in del.icio.us. Here S_{nn} is computed on a network obtained from the one of Fig. 14 by raising the weights of all edges to the power γ. The case $\gamma = 1$ corresponds exactly to the data shows in Fig. 14 (black dots). On decreasing the value of γ (dark grey dots and light grey dots) one can progressively lower the importance of weights, making contact with the unweighted network for $\gamma = 0$ (all weights equal to 1). This analysis allows us to further probe the nature of the spam activity corresponding to the discrete clusters visible in figure (labeled as 1, 2 and 3). Clusters 1 and 2 lie along the $S_{nn} = s$ line and changes in the value of γ don't affect their position. This means that weights play no role in their presence, i. e., clusters 1 and 2 correspond to structures whose edges have all the same weights. Isolated cliques have this property, and indeed direct inspection of the data shows that cluster 1 corresponds to a single post with over 2000 tags. Since those tags are not common in the rest of the system, this large post induces a clique-like structure in the S_{nn} vs s plot. The change in shape that occurs when γ is varied is due to the weak linking of the clique-like structure to the rest of the tag network. Cluster 2, similarly, corresponds to a set posts with about 100 tags each. For this kind of spam, the tags involved have low global frequency, and the overlap between tags belonging to different posts is low, so that each post induces its own clique-like structure. Cluster 3, conversely, corresponds to a set of posts using common tags. This creates a structure that is better linked to the main component of the network. Because of this, as γ is changed, cluster 3 tracks the displacement of the main cloud visible in the plot. For this kind of spam, the overlap between posts is almost total, and this is responsible for high weights of the corresponding links, i. e., its raised position with respect to the line $S_{nn} = s$.

behavior in the case of the shuffled system, as already described in the previous section. From the latter it is clear that the regimes of $s(k) \approx k$ and $s(k) \approx k^{1.5}$ are to be ascribed to mere frequency effects. The crossover region, which seems to feature an exponent slightly larger than 1, is not yet understood and deserves further studies.

6. Summary and Outlook

6.1. Conclusion

In this paper, we have analyzed the network structure of the folksonomies of two social resource sharing systems, del.icio.us and BibSonomy. We observed that the tripartite hypergraphs of their folksonomies are highly connected and that the relative path lengths are relatively low, facilitating thus the "serendipitous discovery" of interesting contents and users.

We subsequently introduced a weighted network of tags where link strengths are based on the frequencies of tag co-occurrence, and studied the weight distributions and connectivity correlations among nodes in this network. Our evidence is compatible with the existence of complex, possibly hierarchical structures in the network of tag co-occurrence. Our analysis and experiments evidence the statistical signature of the emergence of a shared semantics in the metadata system, anarchically negotiated by users.

Our experiments hint that spam – which becomes an increasing nuisance in social resource sharing systems – systematically shows up in the connectivity correlation properties of the weighted tag co-occurrence network.

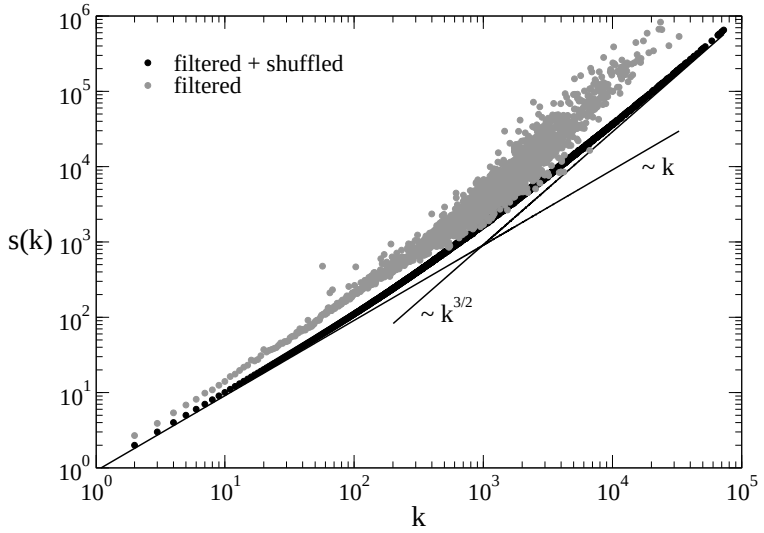

Fig. 18. Node strength S as a function of node degree k in the co-occurrence network of tags, for del.icio.us. The underlying network data are the same of Fig. 14. A two-slope regime is visible for the spam-filtered data (gray dots) as well as for the shuffled data (black dots). In the latter case, a linear regime is visible for low values of k, while a different power-law regime is visible for high k (solid lines).

6.2. Future Work

(Semi-)automatic Spam Detection. At the moment, spam handling in BibSonomy is mostly done by manual inspection and removal of offending content.

In a follow-up paper, we will turn our observations about spamming anomalies in the connectivity of tags into a spam detection mechanism for folksonomies. Using the techniques from Section 5, support for the administrators can be provided to detect spamming activities.

Identification of Communities. As the results from Section 4 suggest that the folksonomy consists of densely-connected communities, a second line of research that we are currently pursuing and that will benefit from the observations in this paper is the detection of communities.

This can be used, for example, to make those communities explicit which already exist intrinsically in a folksonomy, e. g. to provide user recommendations and support new users in browsing and exploring the system.

Acknowledgement This research has been partly supported by the TAGora project funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-034721. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of data appearing in this publication.

References

[1] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752, 2004.

[2] Marc Barthelemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, 92:178701, 2004.

[3] B. Bollobas. *Random Graphs.* Cambridge University Press, 2001.

[4] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Phys. Rev. E*, 74:036116, 2006.

[5] Andrea Capocci and Francesca Colaiori. Mixing properties of growing networks and the simpson's paradox. *Phys. Rev. E*, 74:026122, 2006.

[6] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences (PNAS)*, 104:1461–1464, 2007.

[7] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proceedings of the 15th International WWW Conference*, May 2006.

[8] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical foundations*. Springer, 1999.

[9] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[10] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[11] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A social bookmark and publication sharing system. In Aldo de Moor, Simon Polovina, and Harry Delugach, editors, *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, Aalborg, Denmark, July 2006. Aalborg University Press.

[12] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, Budva, Montenegro, June 2006.

[13] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In *Prof. First International Conference on Semantics And Digital Media Technology (SAMT)*, Athens, Greece, dec 2006.

[14] Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme. Trias - an algorithm for mining iceberg tri-lattices. Hong Kong, December 2006. (to appear).

[15] Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981.

[16] Leslie Lamport. *LaTeX: A Document Preparation System*. Addison-Wesley, 1986.

[17] F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *Lecture Notes in Computer Science*, pages 32–43. Springer, 1995.

[18] Pedro G. Lind, Marta C. Gonzalez, and Hans J. Herrmann. Cycles and clustering in bipartite networks. *Phys. Rev. E*, 72(5), nov 2005.

[19] Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[20] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.

[21] Stanley Milgram. The small world problem. *Psychology Today*, 67(1):61–67, 1967.

[22] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[23] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, NJ, USA, 2006.

[24] Oren Patashnik. BibTeXing, 1988. (Included in the BibTeX distribution).

[25] Thomas Schank and Dorothea Wagner. Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications*, 9(2):265–275, 2005.

[26] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In *Data Science and Classification: Proc. of the 10th IFCS Conf.*, Ljubljana, Slovenia, July 2006.

[27] Patrick Schmitz. Inducing ontology from Flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006.

[28] A. Vazquez, J. Gama Oliveira, Z. Dezso, K. I. Goh, I. Kondor, and A. L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006.

[29] Alexei Vazquez. Exact results for the Barabasi model of human dynamics. *Physical Review Letters*, 95:248701, 2005.

[30] Duncan J. Watts. *Small Worlds – The Dynamics of Networks between Order and Randomneess*. Princeton University Press, Princeton, New Jersey, 1999.

[31] Duncan J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.