Network Properties of Folksonomies

Christoph Schmitz[3] Miranda Grahl[3]

Andreas Hotho[3] Gerd Stumme[3]

[3]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

Ciro Cattuto[2,1] Andrea Baldassarri[1]

Vittorio Loreto[1,2] Vito D.P. Servedio[1,2]

[1]Dipartimento di Fisica, Università di Roma "La Sapienza"
P.le A. Moro, 2, I-00185 Roma, Italy
[2]Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi
Compendio Viminale, 00184 Roma, Italy
{firstname.lastname}@roma1.infn.it

ABSTRACT

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 the 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.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; G.2.2 [**Mathematics of Computing**]: Graph Theory

Keywords

folksonomies, semiotics, semiotic dynamics, small worlds

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 an area of research called "the new science of 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

WWW2007, May 8–12, 2007, Banff, Canada.

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

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 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. [15, 8] 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 [11], discovering trends in the tagging behaviour of users [7, 12], or learning taxonomic relations from tags [9, 21, 20, 16].

2.2 Complex 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", using concepts from social network analysis, graph theory, as well as statistical physics; see [19] 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 [25]; for particular kinds of networks, such as bipartite [14] 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 lot of other peculiar characteristics. Probably the most striking is 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 [23]. Work has been done also on the complex network of WikiPedia [4] where links also posses a specific direction.

3. FOLKSONOMY DATA SETS

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 [11], which we will use in the remainder of the paper.[3]

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

- 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.

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 montly dumps from the timestamps associated with posts, so that 14 snapshots in monthly intervals from 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[4], a second dataset from that system could be obtained directly from a database dump.

As with the del.icio.us dataset, we created monthly dumps from the timestamps, resulting 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

The notion of a *small world* has been introduced in a seminal paper by Milgram [17]. Milgram tried to verify in a practical experiment that, with a high probability, any two given persons within the United States would be connected

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

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

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 [24] (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.

In this section, we will define the notions of characteristic path length and clustering coefficient in tripartite hypergraphs such as folksonomies, and apply these to the data sets introduced in Section 3 in order to demonstrate that these graphs do indeed exhibit small world properties.

4.1 Characteristic Path Length

The *characteristic path length* of a graph [24] 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 (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.

More precisely, let $v_1, v_2 \in T \cup U \cup R$ be 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 $(t_k = t_{k+1}) \vee (u_k = u_{k+1}) \vee (r_k = r_{k+1})$ for $0 \leq k < n$ and $v_1 \in \{t_0, u_0, r_0\}, v_2 \in \{t_n, u_n, r_n\}$. Then we call $d(v_1, v_2) := k$ the *distance* of v_1 and v_2.

Following Watts [24], 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.

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 [24] 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 [24, 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 neighborhood of the given vertex approach a clique? On the other hand, 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 $tu_r := \{(t, u) \in T \times U \mid (t, u, r) \in Y\}$ 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:

$$\gamma_{cl}(r) = \frac{|tu_r|}{|T_r| \cdot |U_r|} \tag{1}$$

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

$$\gamma_{co}(r) := \frac{|\widetilde{tu_r}|}{|tu_r|} \tag{2}$$

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.

Refer to the appendix for an example demonstrating that the clustering coefficients γ_{cl} and γ_{co} do indeed capture different characteristics of the graph and are not intrinsically related.

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]. T, U, R are taken from the observed folksonomies. $|Y|$ many hyperedges are then created by picking the three endpoints of each edge from uniform distributions over T, U, and R, resp.

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 [13]), thus creating a random graph with the same degree sequence as the observed folksonomy.

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.

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 Observations

Figures 1– 3 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.

4.3.2.1 Del.icio.us.

In the del.icio.us dataset (Figures 2 and 3, 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 1, right) are considerably smaller than for the random binomial graphs, though not as small as for the permuted setting. 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.ici.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 [15] 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.

4.3.2.2 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 2 and 3 (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 1 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 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.

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 [8], 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 t_1 and t_2 are linked if there exists a post where they have been associated by a user with the same resource. A link weight can be introduced by defining the weight of the link between t_1 and t_2, $t_2 \neq t_1$, as the number of posts where they appear together. Formally, we define the set $W(t_1, t_2)$ as

$$W(t_1, t_2) := \{(u, r) \in U \times R \mid (t_1, u, r) \in Y \wedge (t_2, u, r) \in Y\}\,, \tag{3}$$

and define the link weight $w(t_1, t_2) := |W(t_1, t_2)|$. 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

$$s_t := \sum_{t' \neq t} w(t, t')\,. \tag{4}$$

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. 4 and 5, 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

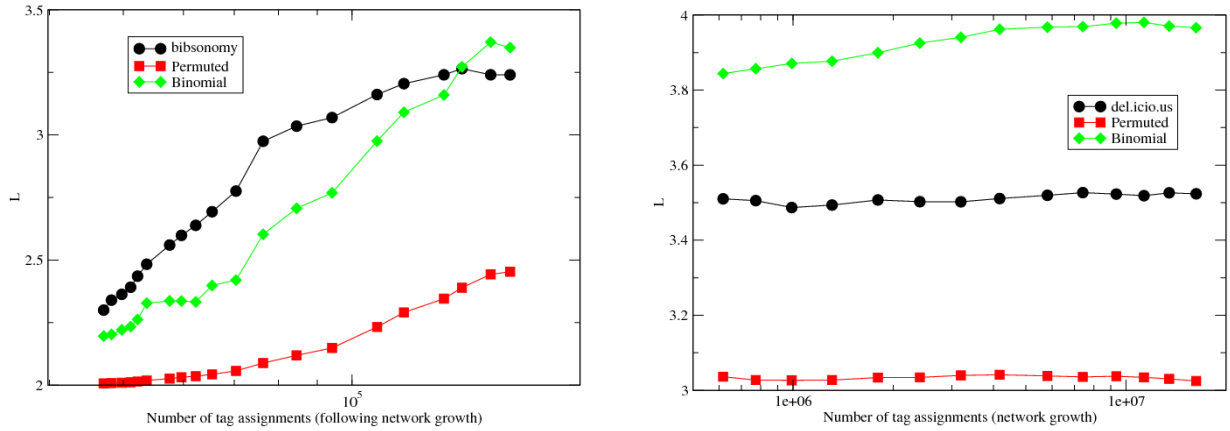

Figure 1: Characteristic path length for the `bibsonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graph have been obtained as a function of the number of nodes of the networks (not shown). Note how the charateristic path length takes quite similar low values, typical of small world networks, for all graphs.

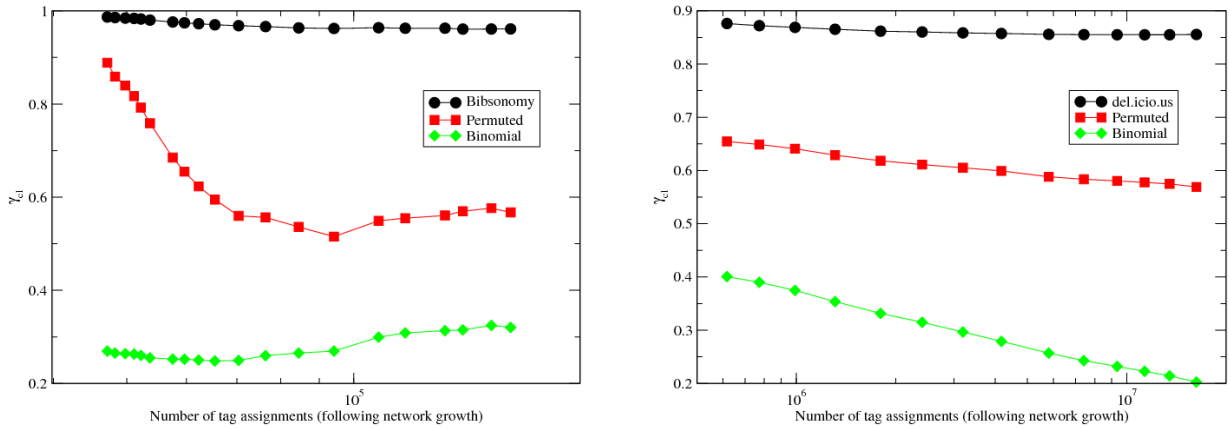

Figure 2: Cliquishness of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). The cliquishness for the folksonomy networks takes quite high values, higher than the corresponding random graph (permuted and binomial).

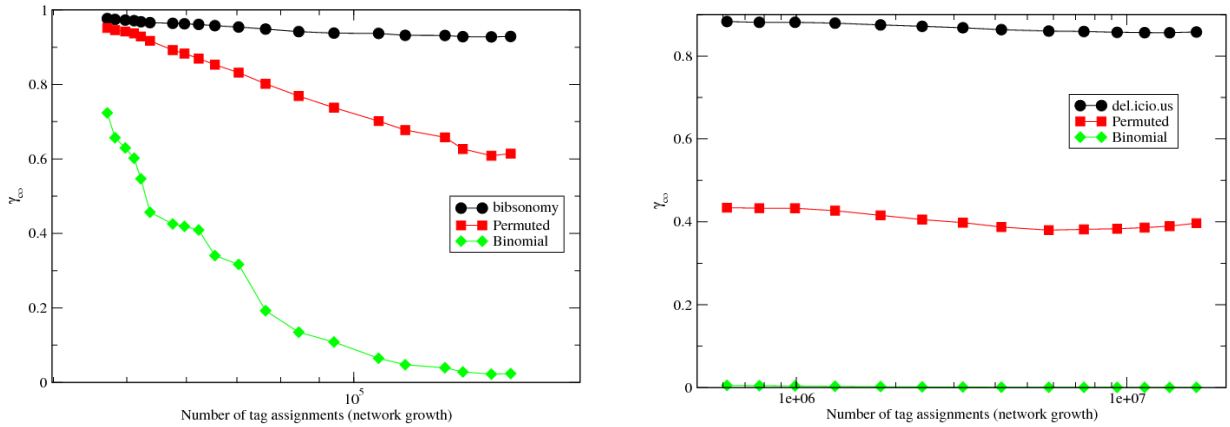

Figure 3: Connectedness/Transitivity of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). As in the case of cliquishness, the values of connectedness/transitivity are very high for the folksonomy networks, at odds with the corresponding random graphs (permutated and binomial).

node strengths and is one of the typical fingerprints of an underlying complex dynamics of interacting human agents [22, 23]. A coarse indicator such as $P_>(s)$, despite its simplicity, is able to point out anomalous activity (i.e. spam) within the investigated folksonomies, as discussed in the captions of Figs. 4 and 5. Quite interestingly, on filtering out these undesired (and probably automatically generated) contributions, the probability distributions for *del.icio.us* and *BibSonomy* 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.

In Figs. 4 and 5 we show that by performing this shuffling operation (blue 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 [8, 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. The sum of all weights $s_i = \sum_j w_{ij}$ of links connected to a given node i is called strength of node i. Given a node i with strength s_i, we define its average nearest-neighbor strength as:

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

with k_i corresponding to the number of links with non vanishing weight connected to node i. The concept of nearest neighbor needs to be clarified here. In principle all nodes are connected each other in a weighted graph, but in this particular context we ignore the existence of those links that have vanishing weight. Consequently, we consider two nodes as nearest neighbors, if exists a link between them with non vanishing weight.

The average nearest neighbor strength $S_{nn}(s)$, as a function of the node strength s, 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 $S_{nn}(s)$ is a decreasing function of the strength s, and social networks, where, on the contrary, $S_{nn}(s)$ displays an increasing behavior. These two networks with opposite behaviors are commonly referred to as disassortative and assortative mixing networks, respectively [18, 5].

Figs. 6 and 7 display our results for *del.icio.us* and *BibSonomy*, respectively. In the figures, each dot correspond 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.

In the scatter plots, the anomalous activity such as spam

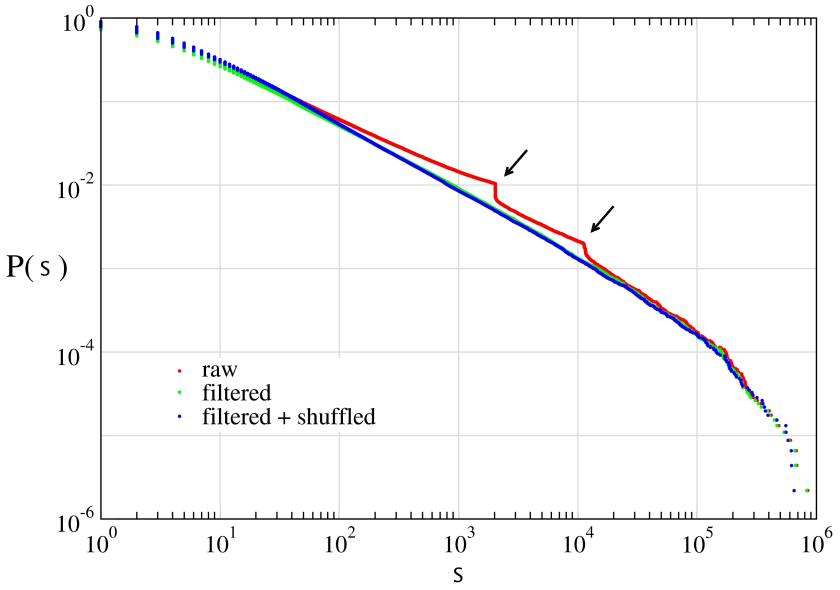

Figure 4: Cumulative strength distribution for the network of co-occurrence of tags in *del.icio.us*. $P_>(s)$ is the probability of having a node with strength in excess of s. Red dots correspond to the whole co-occurrence network. The two steps indicated by arrows correspond to an excess of link with a specific weight and can be related to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps (green dots). Shuffling the tags contained in posts (blue dots) 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.

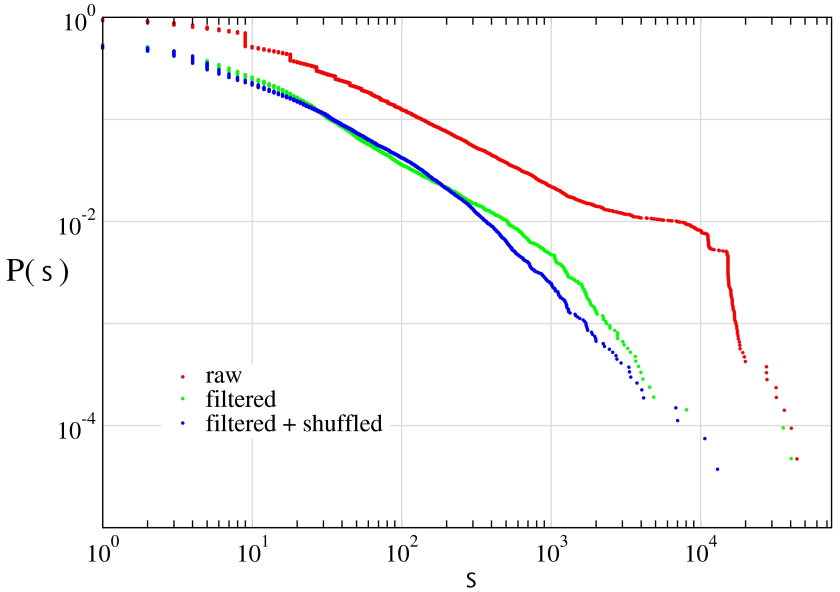

Figure 5: Cumulative strength distribution $P_>(s)$ for the network of co-occurrence of tags in *BibSonomy* (see also Fig. 4). Red dots correspond to the whole co-occurrence network. The irregular behavior for high strengths can be linked to spamming activity: identified spam in *BibSonomy* consists 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. Both types of spam were identified by inspecting the distribution of the number of tags per post. Excluding the above posts from the analysis (green dots), the distribution becomes smooth and similar to the filtered one observed for del.icio.us. Similarly, shuffling the tags contained in posts (blue dots) has a small effect on the cumulated weight distribution.

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 (green 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 (blue dots in Fig. 6 and 7). In this case, shuffling the tags (blue points) 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. 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. 7: those points represent nodes (tags) with low strength that are attached preferentially to nodes of high strength. Similarly, in Fig. 6, 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].

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 existance of complex, possibly hierarchical structures in the network of tag co-occurrence.

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.

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. Part of this research was supported by the European Union in the Tagora project (FP6-2005-34721).

7. 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 case of wikipedia, 2006.

[5] Andrea Capocci and Francesca Colaiori. Mixing properties of growing networks and the simpson's paradox, 2005.

[6] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *in press in the Proceedings of the National Academy of Sciences (PNAS)*, 2006.

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

[9] 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.

[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.

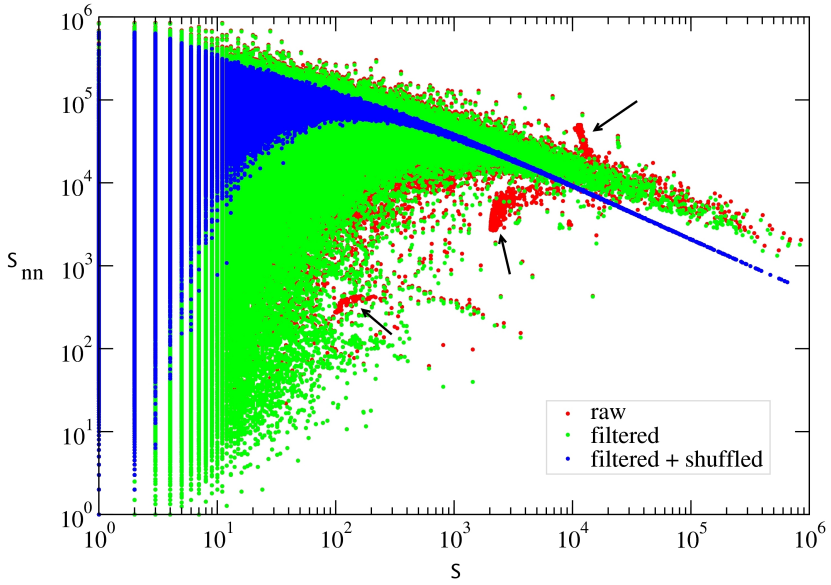

Figure 6: Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in *del.icio.us*. Red dots correspond to the whole co-occurrence network. Assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. A few clusters (indicated by arrows) stand out from the main cloud of data points. As in Fig. 4, such anomalies correspond to spamming activity and can be removed by filtering out posts containing an excessive number of tags (green dots). In this case, shuffling the tags (blue points) 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.

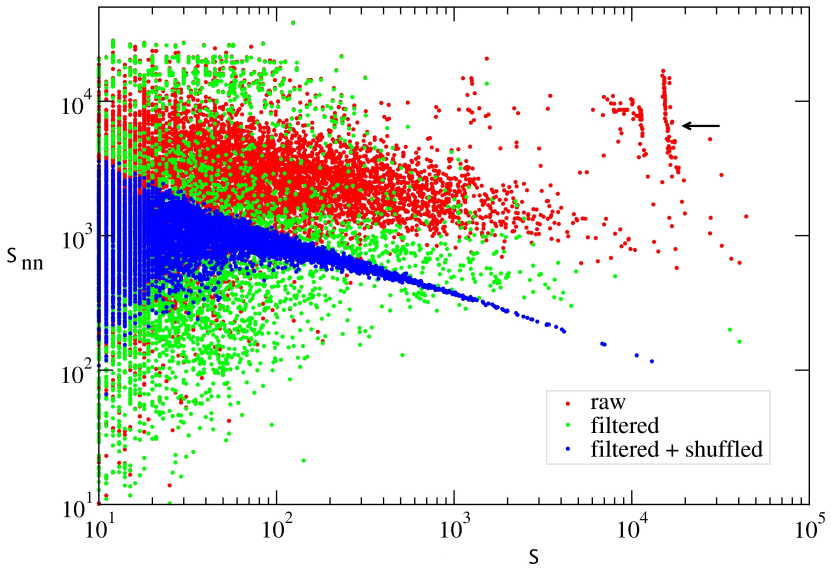

Figure 7: Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in *BibSonomy*. Red dots correspond to the whole co-occurrence network. The scatter plot is qualitatively very similar to the one reported in Fig. 6 for *del.icio.us*: assortative behavior is 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 (green dots). Shuffling the tags (blue dots) has the same effect as in Fig. 6, and the same observations apply.

[11] 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.

[12] 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.

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

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

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

[16] 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.

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

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

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

[20] 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.

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

[22] 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.

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

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

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

Appendix

The two clustering coefficients from Section 4.2 were motivated by two different views on the original definition for one-mode graphs. Still, 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, $\gamma_{co}(\mathbb{F}) < \gamma_{co}(\mathbb{G}) \Rightarrow \gamma_{cl}(\mathbb{F}) < \gamma_{cl}(\mathbb{G})$.

The following example demonstrates that 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 $\gamma_{co}(\mathbb{F}) < \gamma_{co}(\mathbb{G})$.