Small World Folksonomies: Clustering in Tri-Partite Hypergraphs

Christoph Schmitz

November 30, 2006

Abstract

Many recent Web 2.0 resource sharing applications can be subsumed under the "folksonomy" moniker. Regardless of the type of resource shared, all of these share a common structure describing the assignment of tags to resources by users.

In this report, we generalize the notions of *clustering* and *characteristic path length* which play a major role in the current research on networks, where they are used to describe the small-world effects on many observable network datasets. To that end, we show that the notion of clustering has two facets which are not equivalent in the generalized setting.

The new measures are evaluated on two large-scale folksonomy datasets from resource sharing systems on the web.

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.

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

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.

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. [11, 5] 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 [7], discovering trends in the tagging behaviour of users [4, 8], or learning taxonomic relations from tags [6, 16, 15, 12].

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" [14], using concepts from social network analysis, graph theory, as well as statistical physics; see [14] 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 [19]; for particular kinds of networks, such as bipartite [10] or weighted [1] 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.

2.3 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" [14], using concepts from social network analysis, graph theory, as well as statistical physics; see [14] for an overview.

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 [17].

The notions of clustering coefficient and characteristic path length as indicators for small world networks have been introduced by Watts and Strogatz [19]; for particular kinds of networks, such as bipartite [10] or weighted [1] 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.

Work has been done also on the complex network of WikiPedia [3] 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 [7], 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.

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

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 snapshots 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 the author is 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 snapshots 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 [13]. 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 [18] (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 [18] 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.

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

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 [18], 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 5, 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 [18] 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 [18, 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.

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 $\gamma_{co}(\mathbb{F}) < \gamma_{co}(\mathbb{G})$.

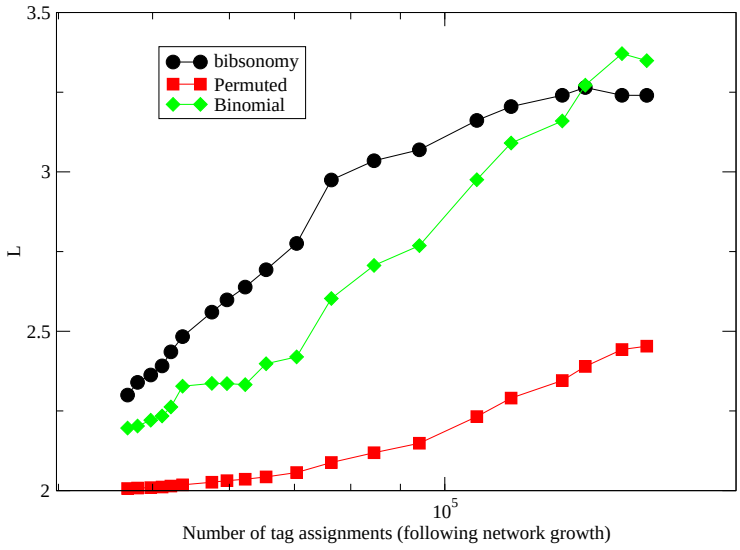

Figure 1: Characteristic Path Length for the BibSonomy dataset

5 Experiments

5.1 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 $G(n, M)$ [2]. 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 [9]), 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

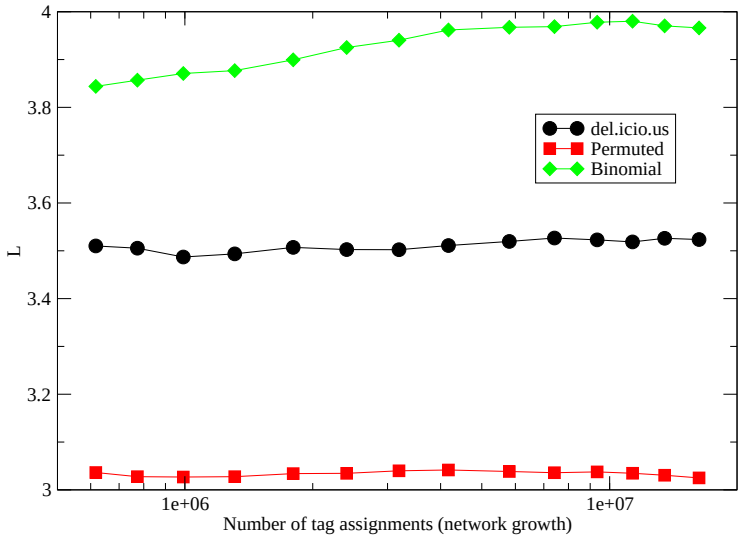

Figure 2: Characteristic Path Length for the del.icio.us dataset

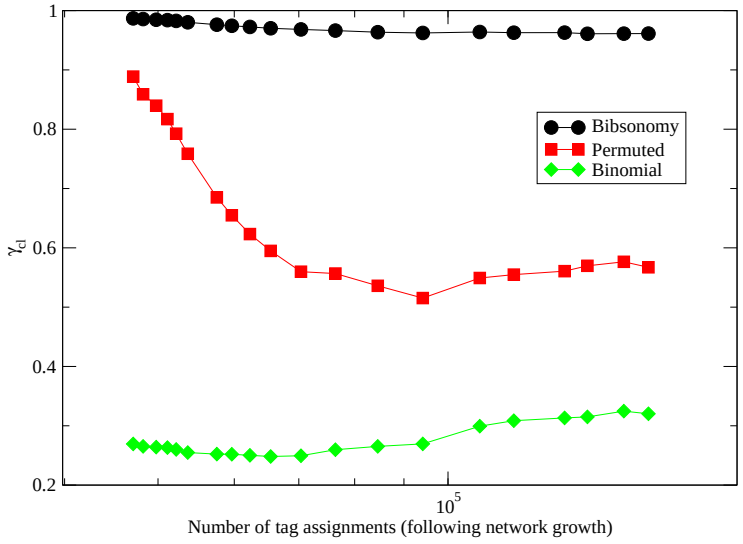

Figure 3: Cliquishness γ_{cl} of the BibSonomy dataset

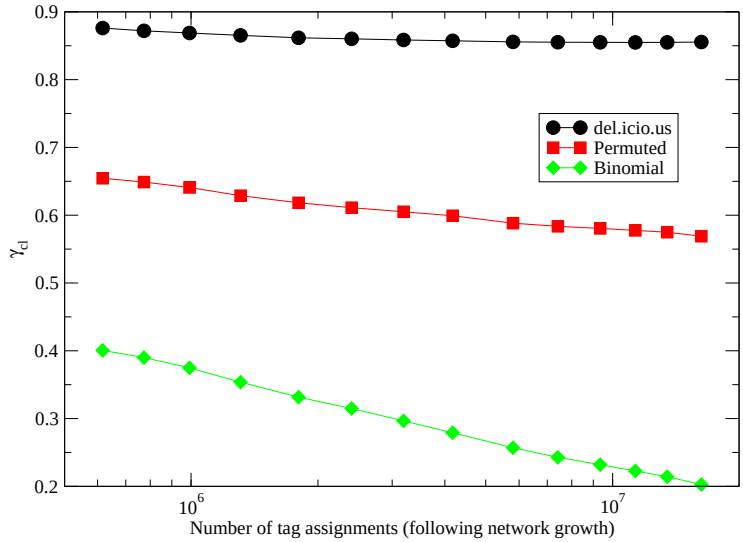

Figure 4: Cliquishness γ_{cl} of the del.icio.us dataset

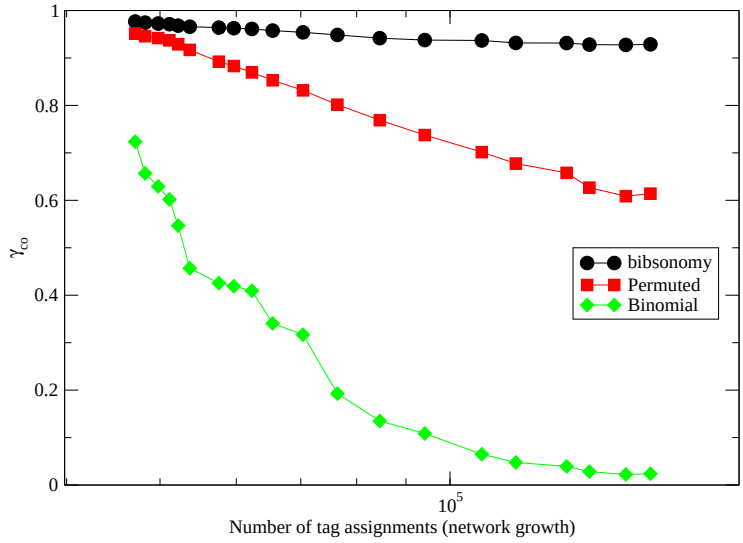

Figure 5: Transitivity γ_{co} of the BibSonomy dataset

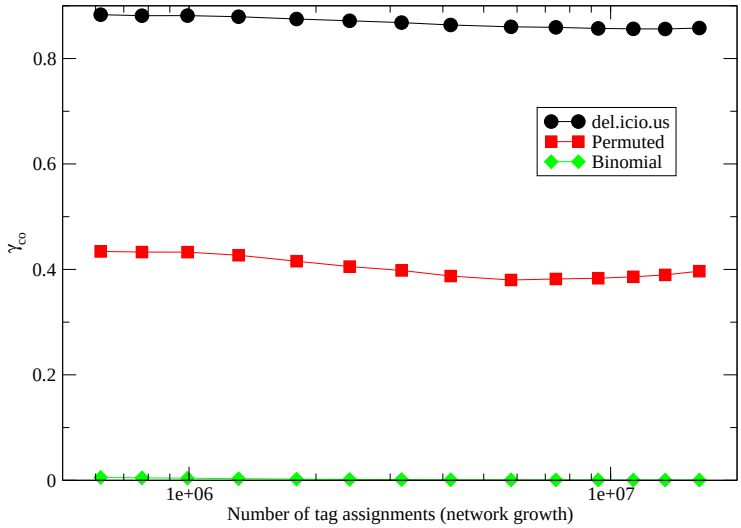

Figure 6: Transitivity γ_{co} of the del.icio.us dataset

to ensure consistency. The presented values are the arithmetic means over the runs; the deviations across the runs were negligible in all experiments.

5.2 Observations

Figures 1–6 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. As the number of tag assignments grows superlinearly over the months, we plot the time axis on a logarithmic scale.

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.

5.2.1 Del.icio.us

In the del.icio.us dataset (Figures 4 and 6), 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 2) 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 [11] 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.

5.2.2 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 3 and 5 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 just 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, current 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.

6 Summary and Outlook

In this paper, we have introduced measures for clustering and characteristic path length which are suitable for tripartite hypergraphs such as those used as the underlying data structure in folksonomy systems.

We 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. According to the usual definition, these observations show that the folksonomies under consideration exhibit a small world structure.

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

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

[15] 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.

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

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

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

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