Logsonomy — Social Information Retrieval With Logdata

Beate Krause ⋆
krause@cs.uni-kassel.de

Robert Jäschke ⋆ ‡
jaeschke@cs.uni-kassel.de

Andreas Hotho ⋆
hotho@cs.uni-kassel.de

Gerd Stumme ⋆ ‡
stumme@cs.uni-kassel.de

⋆ Knowledge & Data Engineering Group, University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany
‡ Research Center L3S, Appelstr. 9a, 30167 Hannover, Germany

ABSTRACT

Social bookmarking systems constitute an established part of the Web 2.0. In such systems users describe bookmarks by keywords called tags. The structure behind these social systems, called *folksonomies*, can be viewed as a tripartite hypergraph of user, tag and resource nodes. This underlying network shows specific structural properties that explain its growth and the possibility of serendipitous exploration.

Today's search engines represent the gateway to retrieve information from the World Wide Web. Short queries typically consisting of two to three words describe a user's information need. In response to the displayed results of the search engine, users click on the links of the result page as they expect the answer to be of relevance.

This clickdata can be represented as a folksonomy in which queries are descriptions of clicked URLs. The resulting network structure, which we will term *logsonomy* is very similar to the one of folksonomies. In order to find out about its properties, we analyze the topological characteristics of the tripartite hypergraph of queries, users and bookmarks on a large snapshot of del.icio.us and on query logs of two large search engines. All of the three datasets show small world properties. The tagging behavior of users, which is explained by preferential attachment of the tags in social bookmark systems, is reflected in the distribution of single query words in search engines. We can conclude that the clicking behaviour of search engine users based on the displayed search results and the tagging behaviour of social bookmarking users is driven by similar dynamics.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Online Information Services—*Web-based services*; H.2.8 [**Information Systems**]: Database Applications—*Data Mining*

General Terms

Experimentation, Measurement

Keywords

Search Engine, Folksonomy, Query Log Analysis, Logsonomy

http://dx.doi.org/10.1145/1379092.1379123

1. INTRODUCTION

Folksonomies are complex systems consisting of user-defined labels added to web content such as bookmarks, videos or photographs by different users. In contrast to classical search engines, which index the web and offer a simple user interface to search in this index, a folksonomy can be explored in different dimensions taking users, tags and resources into account. A further, fundamental difference consists in the way a folksonomy's and a web search engine's index is created: While search engines automatically crawl the web, the content of a folksonomy is determined by its users. As a consequence, the content selection and retrieval in folksonomies is a social process, in which users decide about relevance.

User relevance feedback is integrated into search engine ranking algorithms as well. The feedback is extracted from log files which track a user's click history. However, as the evolvement of social bookmarking systems or recommendation systems on popular websites such as Amazon have shown, web searchers are not only interested in a ranked list of search results, but they like to explore community content as well.

In this paper we discuss the realization of such "search communities" within search engines by building an anonymized folksonomy similar to the del.icio.us social bookmarking system from search engine logdata. As logdata contain queries, clicks and session IDs, the classical dimensions of a folksonomy can be reflected: Queries or query terms represent tags, session IDs correspond to users, and the URLs clicked by users can be considered as the resources they tagged with the query terms. Search engine users can then browse this data along the well known folksonomy dimensions of tags, users, and resources.

A search engine folksonomy, which we will call *logsonomy* in the sequel, brings a variety of features to search engines. Partly discussed in blogs [18] one can picture users adding additional tags to their pages to have them higher

ranked. Temporal aspects can be introduced by incorporating a fourth dimension and showing popular tags, users or resources at a certain time. Finally, search engine users may interact with each other, commenting and copying search results of each other.

Logsonomies open a wide field of exploration. What kind of semantics can we extract from logsonomies? Is the serendipitous discovery of information also possible in logsonomies? How does the structure of logsonomies differ from folksonomies? In this paper, we address these questions by analyzing the topological properties of two logsonomy datasets and comparing our findings to a social bookmarking system. In previous work [6], it was shown that folksonomies exhibit specific network characteristics (e.g. small world properties, power laws, and long tail degree distributions). These characteristics help to explain why people are fascinated from this structure: A small world leads to short ways between users, resources and tags, which allows for finding interesting resources by browsing the system randomly. High clustering coefficients show dense neighbourhoods which are tracked by the formation of communities around different topics. Finally, cooccurrence graphs show the building of user enabled shared semantics.

By looking at a logsonomy graph's components we find that logsonomies collapse in more disconnected components than folksonomies do. In contrast, small world properties considering the shortest path length and the clustering coefficient, compared to random graphs and del.icio.us, can be confirmed, and finally, the strength of each node expresses similar tagging semantics as folksonomies do. Most of the differences in topological structure can be explained by the differences in user behaviour and the creation of metadata in both systems. Overall, we think that our findings strenghten the idea that clickdata can enable social information retrieval and serve as a basis for further analysis.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. In Section 3 we introduce a formal model of a folksonomy and show how to adapt click data to fit this model. Section 4 discusses the topological structure of logsonomies, Section 5 analyses the tag-tag–cooccurence graph to explain the emergent semantics in the logsonomy, and in Section 6 we discuss our vision of future uses of logsonomies.

2. RELATED WORK

In this part we will review related work on extracting social information from logdata, practical approaches of integrating social features into search engines as well as research on the analysis of social networks our analysis is based on.

A first consideration of the tripartite structure of query logs was presented by Zhang and Dong in [21], where an algorithm to rank resources based on the relationships among users, queries and resources of a search engine's log is proposed. In [3], Baeza-Yates and Tiberi proposed to present query-logs as an implicit folksonomy where queries can be seen as tags associated to documents clicked by people making those queries. The authors extracted semantic relations between queries from a query-click bipartite graph where nodes are queries and an edge between nodes exists when at least one equal URL has been clicked after submitting the queries. Our work differs in the formal underlying folksonomy model. Furthermore, we study various topological characteristics of the tripartite graph of user, resource and query nodes, while [3] only focus on a bipartite graph. Our tag-tag–co-occurrence analysis is related to their graph analysis but we consider a strength analysis, while they extract semantic relations between queries. Overall, to the best of our knowledge, a comparison between a real folksonomy dataset and logsonomy datasets as presented in this work was not carried out before.

Several popular search engines have integrated social services. This includes social bookmarking services where users can explicitly assign bookmarks to share them with other search engine users.[1] Furthermore, individual search history information is provided: Users can browse through clicked pages of the past, view their top searches and most frequently visited pages.[2] Comprehensive statistics about the overall search activities are provided by tools such as Google Trends[3], Yahoo Buzz![4] or Ask IQ.[5] Most of the statistical information is derived from query logs. To the best of our knowledge, these query logs have not been transformed to a folksonomy alike search experience before. Search engine providers do not detail to which extent click data is used to improve search, but none is currently providing a folksonomy-style navigation of query logs. A major reason can be seen in privacy considerations which would need to be addressed carefully [1].

Social network topology features. The graph-theoretic notions our analysis is based on are defined in [19, 7]. Cattuto et al. [6] extend the small world characteristics to the tripartite graph. The analysis of structural properties of social networks has been addressed by a number of studies. For example, in [2] topological characteristics of social networking services are described taking degree distribution, clustering properties, degree correlation and evolution over time into consideration. In [10] the structure of internal cooperate blogs is analyzed by Kolari et al. to improve information retrieval and expert finding in companies.

Key studies along the structure and dynamics of social tagging systems are [5, 8, 11, 12]. Major findings include the power law distribution of tags, the evolution of a vocabulary growth over time and the small world properties of the underlying graph. While the representation of clickdata in form of a folksonomy has not been realized before, clickdata was represented as a bipartite graph using queries and URLs as nodes by [4, 20]. An analysis of the bipartite clickdata graph was conducted by Shi in [17] on the AOL data set which we considered also in our study. An analysis of clickdata as tripartite hypergraph as well as a comparison to folksonomy properties has not been carried out so far.

3. LOGSONOMIES

This section will introduce the concept of a *logsonomy*. After a brief definition referring to the folksonomy model, necessary adaptations of search engine query log data will be discussed.

3.1 Formal Model of a Folksonomy

Following [9] we define a *folksonomy* as a tuple $\mathbb{F} := (U, T, R, Y)$ where

[1] http://www.google.com/s2/sharing/stuff
[2] http://www.google.com/history
[3] http://www.google.com/trends
[4] http://buzz.yahoo.com
[5] http://sp.ask.com/en/docs/iq/iq.shtml

- 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., $Y \subseteq U \times T \times R$, whose elements are called tag assignments (TAS for short).

For convenience we also define, for all $u \in U$ and $r \in R$, $\text{tags}(u, r) := \{t \in T \mid (u, t, r) \in Y\}$, i. e., $\text{tags}(u, r)$ is the set of all tags that user u has assigned to resource r. The set of all *posts* of the folksonomy is $P := \{(u, S, r) \mid u \in U, r \in R, S = \text{tags}(u, r), S \neq \emptyset\}$. Thus, each *post* consists of a user, a resource and all tags that the user has assigned to the resource.

Another perspective of this structure is that of a tripartite, undirected hypergraph $G = (V, E)$, where $V = U \dot\cup T \dot\cup R$ is the disjoint union of the sets of users, tags and resources, and every hyperedge (u, t, r) connects exactly one tag, one user, and one resource.

3.2 Adaptation to Search Engine Query Logs: Logsonomies

Let us now consider the query log of a search engine. To map it to the three dimensions of a folksonomy, we set

- U to be the set of *users* of the search engine. Depending on how users in logs are tracked, a user is represented either by an anonymized user ID, or by a session ID.
- T to be the set of *queries* the users gave to the search engine (where one query either results in one tag, or will be split at whitespaces into several tags).
- R to be the set of *URLs* which have been clicked by the search engine users.

In a logsonomy, we assume an association between t, u and r when a user u clicked on a resource r of a result set after having submitted a query t (eventually with other terms). The resulting relation $Y \subseteq U \times T \times R$ corresponds to the tag assignments in a folksonomy.

We call the resulting structure a *logsonomy*, since it resembles the formal model of a folksonomy described above. Additionally, the process of creating a logsonomy shows similarities. The user describes an information need in terms of a query. He or she then restricts the result set of the search engine by clicking on those URLs whose snippets indicate that the website has some relation to the query. These querying and clicking combinations result in the logsonomy. However, logsonomies differ from folksonomies in some important points which may effect the resulting structure of the graph:

- Users experience a bias towards clicking top results in a result list. In query log analysis these clicks are usually discounted. To construct a logsonomy, this bias may be integrated by introducing weights for the hyperedges.
- While tagging a specific resource can be seen as an indicator for relevance, users may click on a resource to check if the result is important and then decide that it is not important. However, the act of clicking already indicates an association between query and resource in our case.
- A user might click on a link of a query result list because it is interesting to him for other reasons than the query.
- A user may click on a resource several times in response to the same query when repeating search after some time. This information is lost when constructing the logsonomy, since TAS are not weighted.
- In logsonomies, a tag is created with a search click. Composed queries are thus another intentional creative process to describe the underlying resources.
- Queries are processed by search engines leaving open to which extend the terms influence the search results.
- When a resource never comes up for search, it cannot be tagged as such.
- Session IDs (in the MSN case) differ from a typical user. They are probably more coherent. We will analyse the differences between users and sessions in 4.1.

The described differences may lead to a different underlying topological structure regardless of the similar nature of the overall process. We will focus on a comparison of the major properties of the underlying graph and will not specifically investigate the influence of the discussed differences on this results. However, in future work we want to further consider these differences to get a better understanding of querying and tagging dynamics.

3.3 Datasets

We use three datasets in our study: two click datasets obtained from commercial search engines (MSN and AOL), and one dataset from the social bookmarking system del.icio.us.

The MSN dataset consists of about 15 million queries submitted in 7,470,915 different sessions which were tracked from the MSN search engine users in the United States in May 2006. The dataset was provided as part of the "Microsoft Live Labs: Accelerating Search in Academic Research" award in 2006[6].

We transformed the data to obtain two logsonomy datasets. In the first, the set of tags is the set of complete queries, the set of users is the set of sessions and the set of resources is the set of clicked URLs. Thus, a click on a URL r after submitting the query q within a session s results in the triple (s, q, r) of Y. To make the dataset comparable to the AOL dataset, we reduced the URLs to host only URLs, i. e., we removed the path of each URL leaving only the host name. In the following, we refer to this dataset as *MSN complete queries*. For the second dataset, we also considered host only URLs. Additionally, we decomposed each query q at whitespace positions into single terms $(q_1, \ldots, q_k)$ and collected the triples (s, q_i, r) (for $i \in \{1, \ldots, k\}$) in Y instead of (s, q, r). This splitting shall better resemble the tags added to resources in folksonomies which typically are single words. As we removed stopwords, a minor fraction of users (1,375) and URLs (282) disappeared because of their relation to a query consisting only of stopwords. The second dataset is called *MSN split queries* in the sequel.

The AOL data is a snapshot of queries from March, 1st to May, 31st 2006. The dataset consists of 657,426 unique

[6]http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx

user IDs, 10,154,742 unique queries, and 19,442,629 clickthrough events [15]. Analogously to the MSN dataset, we transformed the data into two different datasets (called *AOL complete queries* and *AOL split queries* resp.)[7].

To compare the logsonomy structure to a folksonomy, we also used a social bookmarking dataset from del.icio.us containing posts from 81,992 users up to July, 31st 2005. Again, we have two datasets: one consisting of full URLs to be comparable to prior work on folksonomies [6], and one reduced to the host part of the URL only to be comparable to the logsonomy datasets. The sizes of the different datasets are presented in Table 1.

4. TOPOLOGICAL STRUCTURE

To analyse the structural properties of logsonomies, we consider various network measures, adapted to the tripartite structure of our data.

4.1 Degree distribution

We start our analysis by looking at the degree distribution. A degree of a node in a tripartite graph reflects the number of hyperedges, (e. g., the triples (u, t, r)) which contain the specific node.

It has been shown, that the distribution of the degree of nodes for tags and resources in a folksonomy follows a power law distribution [8], $P(k) \sim k^{-\gamma}$, where k is the node degree and γ the exponent of the distribution. A power law distribution implies that a very high number of nodes have few links to other nodes and very few nodes have a very large number of links [7]. Here we examine if this property is maintained in a logsonomy graph.

Figure 1 shows the distributions of users for the different datasets. The distributions are plotted using a log-log scale — power law distributions would show up in such plots as a straight line. For users neither in the query log data nor in the del.icio.us data a power law distribution is reflected. While the curve of the AOL users shows a progression similar to the one of del.icio.us, the curve for the MSN users exhibits a steeper gradient. This is probably due to the nature of sessions representing the users in this dataset: though long-term cookies to track users exist in MSN, sessions have a shorter life time as opposed to unique, timeless user IDs. The probability of being strongly interlinked is therefore lower.

For all datasets, the distribution of the resources are surprisingly similar to each other (cf. Figure 2). This may be an indicator that interests in folksonomies and search engines considering the generality/specificity of content is similarily distributed, i. e., there exist few URLs that are of high interest to many users (authorities), and many specific URLs that are of interest to individuals only.

Finally, the distributions of queries and tags are plotted in Figure 3. When splitting queries into single tags, the distribution is very similar to the tag distribution of del.icio.us. The datasets containing complete queries as nodes show a steeper distribution than the other datasets. We attribute this difference to the fact that full queries have less overlap across and within users.

To conclude, the distribution between tags and queries as well as resources is very similar. This is an indicator that

[7]We used unique user IDs, because session IDs were not included in the AOL dataset.

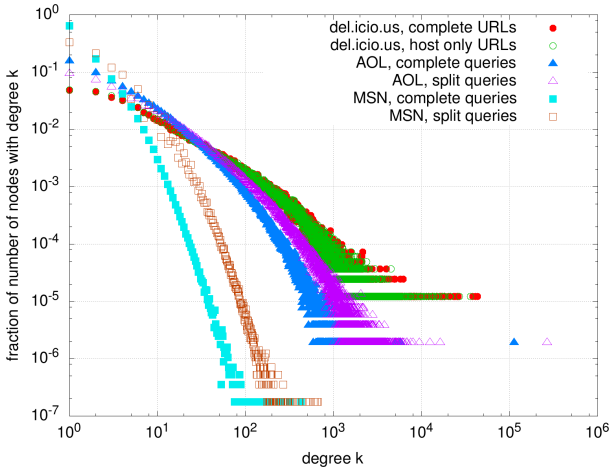


Figure 1: Degree distribution of user nodes. The x-axis shows the degree k, the y-axis the fraction of users with this degree.

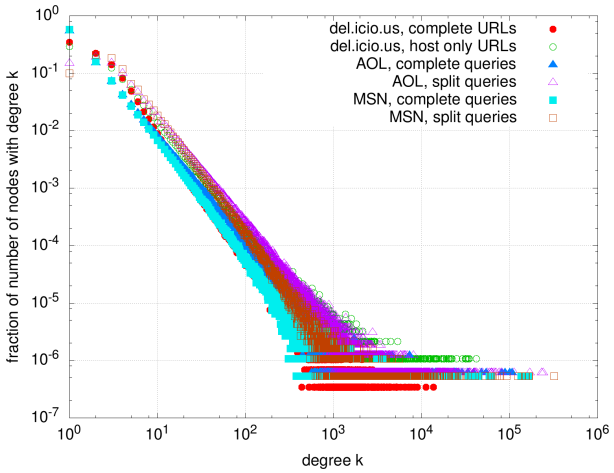


Figure 2: Degree distribution of resource nodes. The x-axis shows the degree k, the y-axis the fraction of resources with this degree.

both systems share a common distribution of the used vocabulary and show a similar tagging and clicking behaviour.

4.2 Connected components

A *connected component* in a tripartite, undirected graph represents a maximal connected subgraph where two nodes are part of the component if there exists a path between them. The size of a connected component is defined as the number of its nodes. According to [14], a *giant connected component* (GCC) is the largest component which scales linearly with the size of the whole graph after a certain percolation threshold is exceeded. The rest of the network, e. g., separate, finite connected components, are called the *disconnected components* (DC). With the presence of giant components networks can be described as a "unit" organism [7]. Figure 4 shows that in all six datasets the GCC comprehends most of the existing nodes. For instance, in del.icio.us with host only URLs the size of the GCC is 1,446,888. As the dataset contains in total 1,447,093 nodes, the GCC covers

Table 1: Datasets

dataset	$\lvert T\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$
del.icio.us host only URLs	430,526	81,992	934,575	14,730,683
del.icio.us complete URLs	430,526	81,992	2,913,354	16,217,222
AOL complete queries	4,811,436	519,250	1,620,034	14,427,759
AOL split queries	1,074,640	519,203	1,619,871	34,500,590
MSN complete queries	3,545,310	5,680,615	1,861,010	10,880,140
MSN split queries	902,210	5,679,240	1,860,728	24,204,125

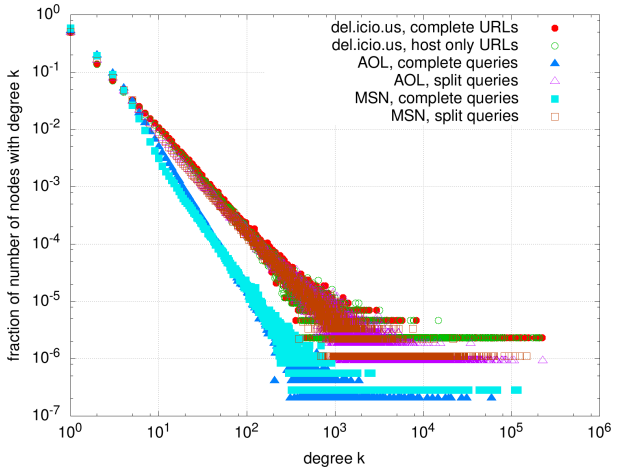


Figure 3: Degree distribution of queries/tags. Again, the degree k is plotted against the fraction of query/tag nodes with this degree.

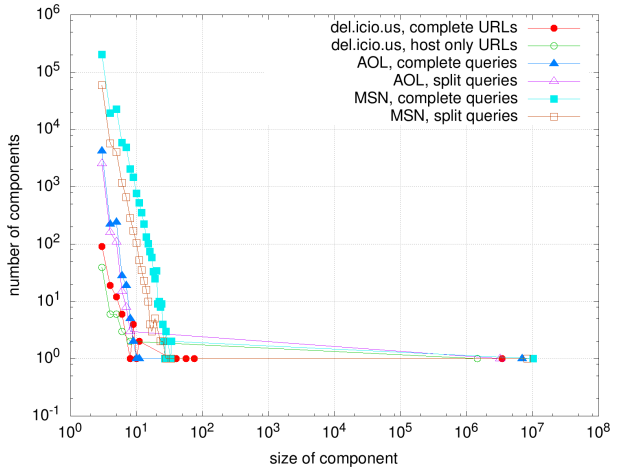


Figure 4: Number of connected components with a specific size.

99.99% of the whole hypergraph. In the AOL split query dataset the relation is similiar with 3,220,395 vs. 3,229,100 total nodes. When comparing the number of disconnected components, the logsonomies have more disconnected components than the folksonomies. Most of these components consist of singletons: a user submitted only one very specific query not submitted by anybody else and thereafter visited one URL that nobody else visited. This behaviour may result from a search query which did not deliver a relevant result and was reformulated in the following.

A comprehensive comparison of different sizes is given in Table 2.

4.3 Small-world properties

It has been shown in [6], that folksonomies exhibit small world characteristics. Small worlds have a network topology for which the degree of clustering is high like in regular networks, but the average shortest path length like in random networks [19]. In the following, we investigate to which extent these characteristics hold for logsonomies. Thereby, we follow the experiments of [6] in order to be comparable to former findings regarding folksonomy properties.

In these experiments, binomial and shuffled graphs[8] of the same size than the original folksonomy were selected to compare the original graph to random networks. For a given folksonomy (U, T, R, Y), a *binomial* random network is a logsonomy $(U, T, R, \hat{Y})$ where $\hat{Y}$ consists of $|Y|$ randomly drawn tuples from $U \times T \times R$. A *shuffled* random network is then a folksonomy $(U, T, R, \check{Y})$ where $\check{Y}$ is derived from Y by randomly shuffling all occurrences of tags in the TAS, followed by shuffling all occurrences of the resources. (For a complete shuffling, it is sufficient to shuffle any two of the three dimensions.) The binomial network has thus the same number of TAS as the original logsonomy, while the shuffled network has additionally the same degree distribution.

4.3.1 *Average shortest path length*

The average shortest path length denotes the mean distance between any two nodes in the graph. In a tripartite hypergraph, the path between any two nodes is the number of hyperedges that lie between them. The shortest path denotes the minimum number of hyperedges connecting the two nodes.

Because of complexity reasons, we approximated the average path length as follows. For each of the datasets, we computed the average path length by randomly selecting 4000 nodes and calculating the average path length of each of those nodes to all other nodes in its connected component.

Table 3 shows the average shortest path length of each dataset together with the values for the corresponding random graphs. Comparing the two del.icio.us datasets, the average shortest path length does not vary to a large extend when considering host only URLs (3.48 for the host-only–graph versus 3.59 for the graph with complete URLs). The average shortest path length of the AOL and MSN datasets with split queries are smaller than those of the datasets with complete queries. This can be explained by the higher overlap, the splitting of queries produces. As a side effect, this also leads to a mixing of contents, e. g., the term “java” in

[8] In [6] the *shuffled graphs* are called *permuted graphs*.

Table 2: Number and size of connected components

dataset	size of GCC	#components with size 10-100	#components with size < 10
del.icio.us, complete URLs	3,425,146	0	140
del.icio.us, host only URLs	1,446,888	0	56
AOL, complete queries	6,951,513	2	4,578
AOL, split queries	3,220,395	0	2,749
MSN, complete queries	10,165,911	2,363	257,952
MSN, split queries	8,207,977	258	71,195

Table 3: Average shortest path length

dataset	raw	shuffled	binomial
del.icio.us, complete URLs	3.59	3.08	3.99
del.icio.us, host only URLs	3.48	3.06	3.67
AOL, complete queries	4.11	3.81	5.76
AOL, split queries	3.62	3.20	3.90
MSN, complete queries	5.43	4.10	8.78
MSN, split queries	3.94	3.42	5.48

"java programming language" and "java island" will link to different topics. However, such wording issues also exist in folksonomies.

Compared to del.icio.us, all four datasets from MSN and AOL provide larger path lengths. Capturing the intuition of serendipitous browsing, it takes longer to reach other queries, users, or URLs within a logsonomy than it takes to jump between tags, users and resources in a folksonomy. In particular, the high values for MSN are likely to result from the fact that a user cannot bridge between different topics if he searched for them in different sessions.

Small world properties are still confirmed by the shortest path length: Comparing each logsonomy to the corresponding binomial and random graphs, the path lengths differ only slightly.

4.3.2 Cliquishness and Connectedness

The clustering coefficient characterizes the density of connections in the environment of a node. It describes the cliquishness, (i. e., *are neighbor nodes of a node also connected among each other*) and the connectedness of a node, (i. e., *would they stay acquainted if the node was removed*). In a tripartite graph, these measures are considered separately. In [6] the following definitions are introduced.

Cliquishness.

Consider a resource r. Then the following sets of 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 cliquishness coefficient $\gamma_{cl}(r)$ as:

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

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

Connectedness.

Table 4: Cliquishness

dataset	raw	shuffled	binomial
del.icio.us, complete URLs	0.86	0.55	0.20
del.icio.us, host only URLs	0.75	0.51	0.05
AOL, complete queries	0.85	0.66	0.32
AOL, split queries	0.70	0.43	0.04
MSN, complete queries	0.87	0.75	0.47
MSN, split queries	0.85	0.50	0.23

Table 5: Connectedness

dataset	raw	shuffled	binomial
del.icio.us, complete URLs	0.85	0.37	0.00
del.icio.us, host only URLs	0.83	0.32	0.00
AOL, complete queries	0.33	0.03	0.00
AOL, split queries	0.66	0.10	0.00
MSN, complete queries	0.42	0.03	0.00
MSN, split queries	0.70	0.11	0.00

Consider a resource r. Let $\widetilde{tu_r} := \{(t, u) \in |tu_r| \wedge \exists \tilde{r} \neq r : (t, u, \tilde{r}) \in Y\}$, 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.

The results in Tables 4 and 5 show the cliquishness and connectedness coefficients averaged over all nodes. One can see that the coefficients of the original del.icio.us, AOL, and MSN graphs are in general higher than the ones of the corresponding random graphs. This indicates that there is some systematic aspect in the search behaviour which is destroyed in the randomized versions. Comparing the two logsonomies to the folksonomy del.icio.us, however, the connectedness coefficients of the folksonomy exceed those of the logsonomies. With the experiments so far, we lack an explanation for this difference. The cliquishness coefficients show less distinction. This is probably because many resources (tags, users) exist which only appear in very few TAS, but which then are well connected among each other. The cliquishness coefficient of these nodes is than (close to) one.

5. STRENGTH IN THE TAG-TAG–CO-OCCURENCE GRAPH

In this section we focus on the analysis of the semantics behind the querying and clicking behavior in a logsonomy. Therefore we study the properties of the *tag-tag–co-*

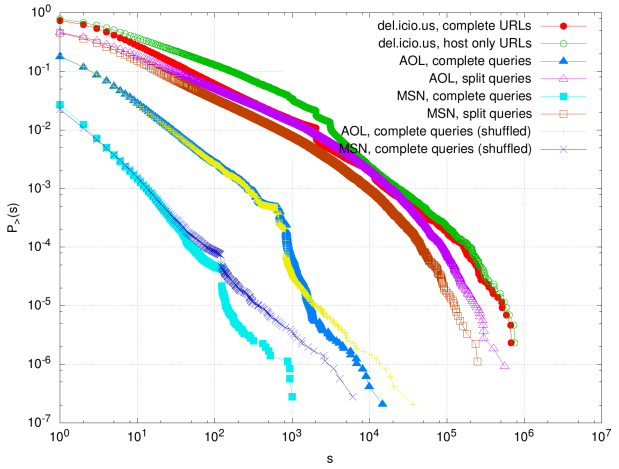


Figure 5: Cumulative strength distribution for the network of cooccurrence of tags and queries for all datasets. Split query logsonomies show a very similar distribution to the del.icio.us folksonomy.

occurrence graph, as it mainly reflects the semantics describing the clicked URLs with respect to the queries. This graph consists of tags which are linked if they occur in the same post. More formally, $G := (T, E)$ with $E := \{(t_1, t_2) \mid \exists u \in U, \exists r \in R\colon (u, t_1, r) \in Y \wedge (u, t_2, r) \in Y\}$ defines the tag-tag–co-occurrence graph on the set T of tags. Naturally, we can add weights to the edges by counting in how many posts two tags appear together. We define the *weight* $w(t_1, t_2)$ of an edge (t_1, t_2) to be $w(t_1, t_2) := |\{(u, r) \in U \times R \mid (u, t_1, r) \in Y \wedge (u, t_2, r) \in Y\}|$. The *strength* s_t of a tag t in the graph is then defined as

$$s_t := \sum_{t' \neq t} w(t, t'). \tag{3}$$

Each of the following figures contains data from two types of datasets: the *raw* data of the datasets as described in Section 3.3, and a *shuffled* version of the datasets where we shuffled the tags[9] in the triples of the relation Y. To do so, for each triple (u, t, r) we randomly picked a tag t' and exchanged (u, t, r) in Y with (u, t', r). We picked each tag with a probability according to its degree, such that the tag degree distribution of the resulting folksonomy is identical to the original one.

One of the standard measures of complex network theory is the *cumulative strength distribution* $P_>(s)$ [6]. It specifies for a given node strength the probability that a node will exceed this strength. For the del.icio.us dataset we observe the same fat tailed distribution as in [6] (cf. Figure 5). The logsonomy with split queries for AOL as well as for MSN shows a very similar distribution to the del.icio.us folksonomy. This distribution is also not disturbed by the shuffling process on the tags which confirms that the strength distribution for both the logsonomy and folksonomy data only depends on the tag frequencies and not on their semantics — which is destroyed by the shuffling process (Due to space restrictions we did not include a figure with the shuffled distribution for all datasets.)

[9]In contrast to the shuffled versions in Section 4.3, where we shuffled all three dimensions.

We observe a different behavior for the datasets with complete queries. Queries with high strengths (above 10^2 for MSN and above 10^3 for AOL) show up less frequently than expected: these frequencies are significantly below those obtained for the shuffled versions. This can be explained by the construction process of the dataset: The probability that a user clicks on the same URL within one session but based on another query is very unlikely. The number of queries that are connected to many other queries by some (user, resource) pairs is therefore below expectation.

Next, we want to take a closer look into the co-occurrence network of the logsonomy to see whether another property holds or not. Therefore, we measure the *average nearest-neighbor strength* of tags in this graph. For that purpose we define the neighborhood N_t of a tag t to be $N_t := \{t' \mid (t, t') \in E\}$. The average nearest-neighbor strength is then defined as:

$$S_{nn}(t) = \frac{1}{|N_t|} \sum_{t' \in N_t} s_{t'}. \tag{4}$$

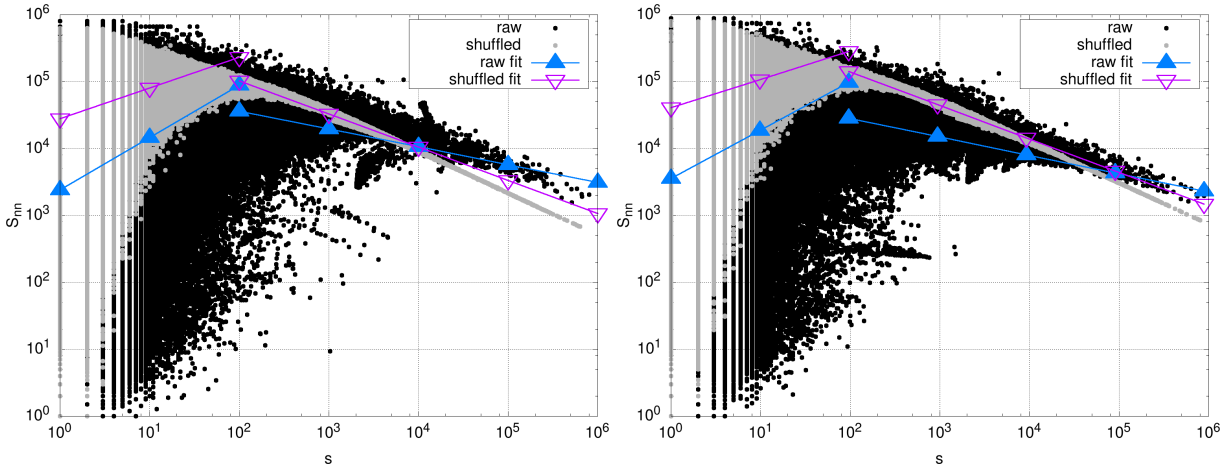
For each tag $t \in T$, we will set its average nearest-neighbor strength $S_{nn}(t)$ in relation to its own strength s_t. This relation can reveal the difference between human-produced social networks and technological artefacts [13]: a positive correlation — called *assortative mixing* — hints at social networks while a negative correlation frequently shows up in technological and biological networks.

Each of the following six figures (6(a) to 8(b)) shows the strengh of each tag versus its average strength for the raw dataset and its corresponding shuffled version. Additionally, the figures contain linear least squares fits for each dataset. Those are splitted into two regions: tags with low strength ($s_t < 10^3$) and tags with high strength ($s_t > 10^3$).

Figure 6 shows the relation between s_t and $S_{nn}(t)$ for the two del.icio.us datasets in consideration. The plot for the dataset with complete URLs (6(a)) shows that the average strength of the neighbors of tags with low strength varies strongly while for tags with higher strength the variation is much smaller, as already observed in [6]. The average nearest-neighbor strength for tags with high values of s_t and s_t itself are slightly anti-correlated. Clusters in the diagram (regions of points separated from the main point cloud, like the one for $10^3 < s_t < 10^4, 10^3 < S_{nn}(t) < 10^4$) are mainly caused by artifacts in the data (such as spam).

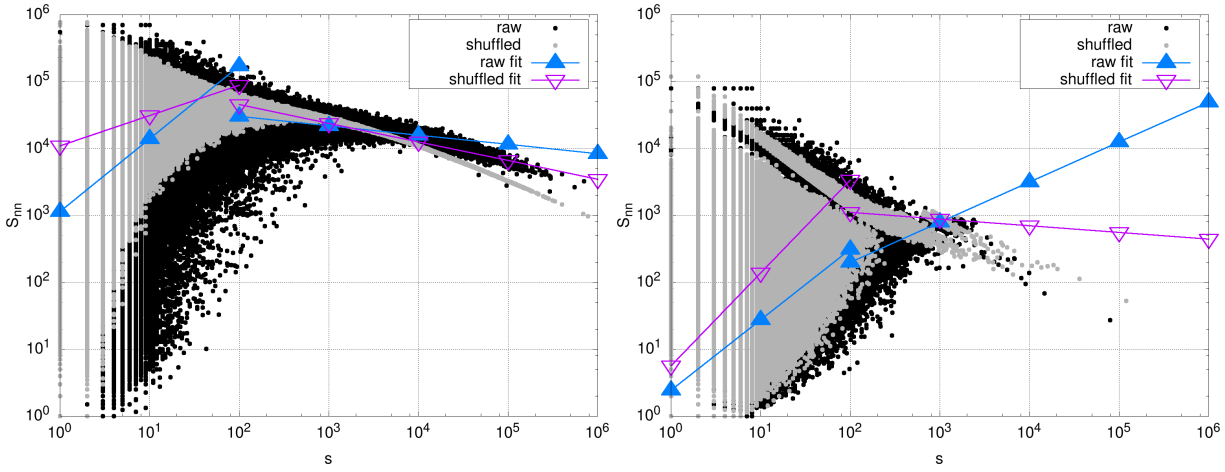
The shuffled data shows a more regular distribution of the average nearest-neighbor strength over s_t and a larger anti-correlation for higher values of s_t. Since shuffling destroys semantics inherent in the original network, the obvious difference to the raw data, especially in the low strength regions, is a strong indicator that the infrequent tags are frequently grouped together by their inherent semantics — an effect which is destroyed by shuffling [6]. Removing the paths from the URLs of the del.icio.us dataset does not change the picture much: only some clusters (dis)appear, as Figure 6(b) shows.

The strength distributions of the split versions of the AOL and MSN datasets (Figures 7(a) and 8(a)) show noticeable similarity to the behaviour in del.icio.us for both the original and the shuffled data. This supports the hypothesis that the semantics of the single words within web search engine queries provide topically organized local structures on the tag-tag–co-occurrence graph similar to the behavior in a folksonomy.



(a) complete URLs (b) host only URLs

Figure 6: Average nearest-neighbor strength S_{nn} of tags in relation to the tag strengths in del.icio.us. The distributions of both datasets is very similar: the average nearest-neighbor strength for tags with low strength varies strongly, while for tags with higher strength the variation is much smaller.



(a) split queries (b) complete queries

Figure 7: Average nearest-neighbor strength S_{nn} of tags in relation to the tag strengths in AOL. The datasets with split queries show a similar assortative and dissortative behaviour for the original and shuffled datasets. The full query dataset differs in size and shape.

The strength distributions for the complete queries of AOL and MSN (Figures 7(b) and 8(b)), on the other hand, differ substantially from the distributions of the del.icio.us data. Not only are the strengths and average nearest-neighbor strengths smaller than in del.icio.us (which is in line with the results for the cumulative strength distribution in Figure 5), but also the shape is different: it is more strongly bulged on its lower part, which results from a large number of queries with medium to high strength (around 10^2 in AOL and $10^{1.5}$ in MSN) that are connected in average to less strong queries. We assume that this structure stems from frequency effects rather than from semantically induced structures, as now the shuffled data differ only slightly from the raw data.

In the distribution for the complete AOL queries, we additionally observe — both for the raw and the shuffled data — a separated cluster on top of the distribution. We currently lack an explanation for this phenomenon.

We summarize the results of the analysis of the tag-tag–co-occurrence graph with the conclusion that the logsonomies based on split queries are closer in terms of semantical behavior to folksonomies.

6. VISION

In this paper we presented the idea of transforming a search engine query log into a "logsonomy". We analyzed the resulting graph structure to find similarities and dissimilarities to the existing folksonomy del.icio.us. We found similar user, resource and tag distributions, whereby the split query datasets are closer to the original folksonomy than the complete query datasets. We could show that both graph structures have small world properties in that they exhibit relatively short shortest path length and high clustering coefficients. Finally, the analysis of the strength in the tag-tag–co-occurrence network revealed very similar properties between folksonomies and logsonomies with split queries.

In general, the differences between the folksonomy and logsonomy model mentioned in Section 3.2 did not effect the graph structure of the logsonomies. Minor differences are triggered by the session IDs which do not have the same thematic overlap as user IDs have. Also, full queries show less inherent semantics than the splitted datasets do. In future work, a more thourough analysis of these differences will be interesting.

Overall, the results support our vision to merge the search engine and folksonomy worlds into one system. While some search engines already allow to store and browse search results, they do not provide folksonomy-alike navigation or the possibility to add or change tags. From a practical point of view, the following considerations are further arguments for a logsonomy implementation and its combination with a folksonomy system:

- Users could enrich visited URLs with their own tags (besides the automatically added words from the query) and the search engine could use these tags to consider such URLs for later queries — also from other users. Thus, those tags could improve the quality of the search engine.

- The popularity of folksonomy systems could increase the customer loyalty for a search engine. The community-feeling known from folksonomies could pass over to search engines.

- Search engines typically have the problem of finding new, unlinked web pages. Assumed, users store new pages in the folksonomy, the search engine could direct its crawlers better to new pages. Additionally, those URLs would have been already annotated by the user's tags — even without crawling the pages it would be possible to present them in result sets.

- As described in [16], folksonomies can assist in finding trends in society. Many social bookmarking users can be viewed as trend setters or early adopters of innovative ideas — their data is valuable for improving a search engine's topicality.

- Bookmarked URLs of the user may include pages, the search engine can not reach (intranet, password-protected pages, etc.). These pages can then be integrated into personalized search results.

However, privacy issues are very important when talking about search engine logs. They provide details of a user's life and often allow to identify the user himself [1]. Certainly, this issue needs attention when implementing a logsonomy system.

Acknowledgement. We would like to thank Eytan Adar for further insights into the AOL dataset. Part of this research was funded by the European Union in the Nepomuk (FP6-027705) and Tagora (FET-IST-034721) projects and by the Microsoft Live Labs Award "Accelerating Search in Academic Research".

7. REFERENCES

[1] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Logs Workshop at WWW2006*, 2007.

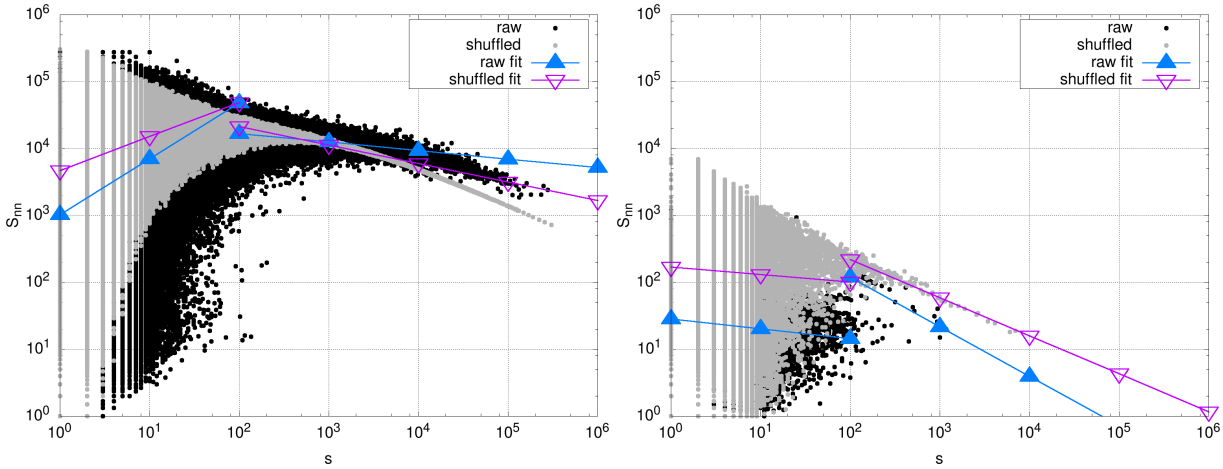
[2] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th International Conference on the World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[3] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85, New York, NY, USA, 2007. ACM.

[4] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 407–416, New York, NY, USA, 2000. ACM.

[5] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems, 2007. http://www.citebase.org/abstract?id=oai:arXiv.org:0704.3316.

[6] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Special Issue on "Network Analysis in Natural Sciences and Engineering" (to appear)*, 2007.



(a) split queries (b) complete queries

Figure 8: Average nearest-neighbor strength S_{nn} of tags in relation to the tag strengths in MSN. The datasets with split queries show a similar assortative and dissortative behaviour for the original and shuffled datasets. The full query dataset differs in size and shape.

[7] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, January 2003.

[8] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, 2006.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Heidelberg, June 2006. Springer.

[10] P. Kolari, T. Finin, Y. Yesha, Y. Yesha, K. Lyons, S. Perelgut, and J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[11] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW2006*, May 2006.

[12] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the Fourth International Semantic Web Conference (ISWC 2005)*, LNCS, pages 522–536. Springer, 2005.

[13] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, 2002.

[14] M. E. J. Newman. *Random graphs as models of networks*, pages 35–68. Wiley, first edition, 2003.

[15] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proc. 1st Intl. Conf. on Scalable Information Systems*. ACM Press New York, NY, USA, 2006.

[16] J. Röttgers. Am Ende der Flegeljahre — Das Web 2.0 wird erwachsen. *c't 25/2007*, page 148, 2007.

[17] X. Shi. Social network analysis of web search engine query logs. Technical report, University of Michigan, School of Information, University of Michigan, 2007.

[18] G. Smith. Search tagging, 2005. http://atomiq.org/archives/2005/05/search_tagging.html.

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

[20] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 118–126, New York, NY, USA, 2004. ACM.

[21] D. Zhang and Y. Dong. A novel web usage mining approach for search engines. *Computer Networks*, 39(3):303–310, june 2002.