Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D4.2: Interim report on models and simulators

Period covered: from 01/06/2007 to 31/05/2008
Start date of project: June 1st, 2006
Due date of deliverable: July 15th, 2008
Distribution: Public

Date of preparation: 31/05/2008
Duration: 36 months
Actual submission date: June 20th, 2008
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: PHYS-SAPIENZA
Lead contractor for this deliverable: PHYS-SAPIENZA

Contents

List of Figures

Chapter 1

Introduction

1.1 Structure of the deliverable

This deliverable reports the theoretical and modeling activity carried on during the second year of the project. The structure of the deliverable is the following:

- This introduction contains a brief summary of the formal definition of folksonomy and a survey of the modeling strategy typical of complex systems study.

- The second chapter focus on the main measures performed on the real data-sets. This is divided in two parts: i) stream measures, mainly tag frequency distributions and dictionary growth curves; ii) study on the tag co-occurrence networks. Some measures performed on texts are reported too, as a comparison with folksonomy.

- The third and fourth chapters report the main theoretical studies. In the third chapter a model (Dellschaft and Staab, 2008) is presented, which refines previous models for tag streams. The main aim here is to model the tagging activity of an average user, in order to recover synthetic tag streams statistically similar to the one observed in real systems. The novelty introduced in the model is an effective way to simulate and model the background knowledge of the user. On the other hand, the exposition of the user to other users' activity influences his/her activity: consequently attention has been paid at the role of the interface. The model introduces several parameters, discussing their meaning, role and effect in the resulting stream. The evaluation section shows how the model manages to reproduce, simultaneously, several different features observed in real systems, notably the frequency rank distribution shape and the sub-linear growth of the tag dictionary size.

- The fourth chapter describes a model (Cattuto et al., 2008a) recently introduced in order to describe and understand the structure of the tag co-occurrence network. It is shown how a simple schematic modeling of an abstract semantic space, representing an hidden, shared concept map, allows the simulation of synthetic posts, whose corresponding tag co-occurrence network very closely mimics the statistics of real co-occurrence network. Moreover, the associated stream reproduces both the frequency rank distribution and the dictionary growth curves observed in real systems. The model allows the study of more refined statistical measures, more sensible to hidden semantic correlations in the tagging data.

- The fifth and last chapter is devoted to a brief description of control strategies for applications, inspired by the modeling activity.

1.2 Formal definition of folksonomy

In social resource sharing systems users can upload resources and assign arbitrary words, the so-called tags. The tags can later be used for retrieving specific resources from the system. The different systems can be distinguished by the kind of resources which can be uploaded. For example, Del.icio.us allows the sharing of bookmarks with other users, in Flickr the user can upload photos and the Bibsonomy system allows users to share bookmarks and BibTeX entries.

The collection of all users, resources and tags as well as the assignments of tags to resources are called folksonomy. We formally define a folksonomy as follows (cf. (Schmitz et al., 2006) and (Baldassarri et al., 2007)):

Definition 1 *A folksonomy $\mathbb{F}$ is a tuple $\mathbb{F} := (U, T, R, Y, pt)$ 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)*
- *pt is a function $pt : Y \to n$ which assigns to each tag assignment of Y a temporal marker $n \in \mathbb{N}$. It corresponds to the time at which a user assigned a tag to the resource.*

The tag assignments can be grouped into several postings. A posting contains all tag assignments made by the same user to the same resource at the same time. The temporal marker of the post is equal with the temporal marker of any of the contained tag assignments.

The temporal marker associated with the tag assignments or posts allows their temporal ordering into a stream of tagging events.

1.3 Modeling strategies

The main aim of WP4 is that of exploiting methods, concepts, tools and techniques from statistical physics and complex systems science to analyse different data streams in the project, in order to find correlations and develop models of semiotic, human, collaborative processes at play.

This is achieved through a specific workflow consisting in (i) Discovering emergent features, (ii) Modeling fundamental mechanisms and (iiia) Feedback on measures and observations, as well as (iiib) Inspire new control strategies for systems. In other words, this workflow can result in a virtuous loop, where measures inspire models, model analysis suggests new measures and observations, which in turn allow the evaluation and refinement of models. Once the a satisfactory level of agreement between theory and experiments is achieved, the theoretical description can suggest and inspire control strategies and directions for improving systems.

The discover of emergent features is based on the measure of statistically significant quantities, paying particular attention to the generality and the specificities of the quantity in study, i.e. considering similar measures on different data streams in the same system (e.g. tag co-occurrence and resource tag stream, see below) or comparing the same quantities in different systems (i.e. Flickr, delicious, bibsonomy) or in conceptually different contexts (e.g. tag streams vs. text streams in books).

Models definition is inspired to an "Occam's razor" principle, where the number of free, tunable parameters is kept as low as possible in order to recover qualitative and, when possible, quantitative agreement with the statistical features observed on real data-sets. Models capture essential

mechanisms and "universal" features, with the twofold aim of allowing an analytical treatment and to be useful in describing and understand similar systems.

As described in the project proposal, we can extract data from the system in study either in the form of streams or in form of networks. In the first case, the temporal evolution of the system is explicitly taken in account and the focus of the models is on the user tagging activity, even if considered in an average way. In the second case, more attention is devoted to "structural" correlations in the data: measuring, modeling and understanding these correlations could shed light on the user cognitive dynamics and inspire methods and control strategies for managing, navigating and mining the data.

Chapter 2

Emergent features for tagging systems

First, we will discuss the stream view of folksonomies, and in the following section we'll report some analysis on a special projection of the tri-partite folksonomy network, the tag co-occurrence network.

2.1 Stream view

The analysis of the streams allows for making interesting observations of the ongoing dynamics in tagging systems and the underlying mechanisms. It helps for better understanding the potential and the limits of social tagging systems.

Examples of tag stream analysis are amongst others available in (Golder and Huberman, 2006), (Halpin et al., 2007) and (Cattuto, 2006; Cattuto et al., 2007b).

There are two characteristic properties of tag streams on which we will concentrate our analysis:

Dictionary Growth The first characteristic property which can be observed in co-occurrence streams is the growth of the number of distinct tags. This is the analogous of the growth of dictionary size in texts, called Heaps' law (Heaps, 1978).

Frequency-Rank Distribution Another characteristic property of tag streams is the distribution of the occurrence probabilities of the distinct tags in a tag stream. This is the analogous of the Zipf's law in texts (Zipf, 1949).

Distribution of inter-arrivaltimes Here we study the temporal patterns of users' tagging activity by measuring the probability distribution of inter-arrivaltimes, i.e. the distribution of times elapsed between the occurrences of the same tag.

Tag-tag correlations We study the temporal self correlations of tag occurrences inside the stream, by counting how many identical tags occur at a fixed temporal distance.

We will distinguish between the characteristic properties of co-occurrence tag streams, resource tag streams and user tag streams. They will be compared to properties of the human language that are already known from the analysis of natural language texts.

2.1.1 Co-occurrence Tag Streams

A co-occurrence tag stream only contains tags that co-occur with another specific tag in the same posting.

In Fig. 2.1 one can see a sub-linear growth rate for the dictionary size of three tag streams, the streams of tags co-occurring respectively with the tag *ajax*, *blog* and *xml*: in all the three cases, at

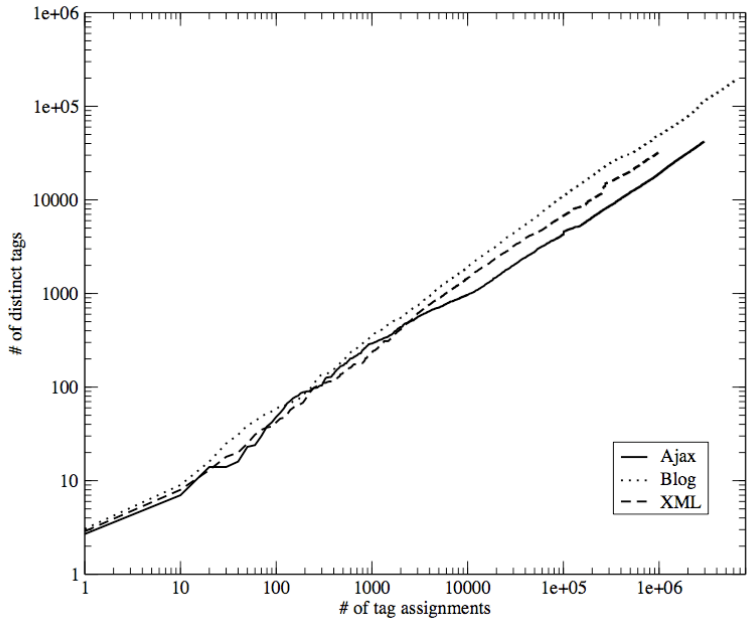

Figure 2.1: Growth of the number of distinct tags in three tag streams, as a function of the length of the stream.

the beginning the growth rate is higher and decays with the age of the stream but doesn't stop at a certain limit. This growth pattern can also be observed in the global system as well as on the level of resource streams and reminds of a power-law (Cattuto et al., 2007a).

Fig. 2.2 shows the frequency-rank distribution for our three tag streams. The graph is shown in log-log scale. One can see that again they remind of a power-law decay for the medium to less frequently used tags and a flattened slope for the most frequently used tags (cf. (Cattuto et al., 2007b)).

It remains the question in how far the slope of the frequency-rank distribution of tags in a tag stream are characteristic for the tagging systems or whether it reflects a property that is already known from natural language texts. For example, for large text corpora it is known that their word frequencies exhibit a power-law behavior for all words, not only for the medium and less frequently used words. For example, in (Montemurro and Zanette, 2002) they analyzed a large corpus comprising 2,606 books in English. The resulting frequency-rank graph of the words showed a power law behavior.

However it seems that the strict power-law behavior is restricted to general text corpora only. If one analyzes corpora which are centered around a certain topic one can observe a behavior similar to that of co-occurrence tag streams. This is demonstrated in Fig. 2.3 that shows the frequency-rank distribution of text corpora crawled from the Web. They only contain texts related to one of the three topics of our co-occurrence tag streams, i. e. they were crawled by downloading all resource URLs that are contained in the streams. For example, the *ajax* Web corpus contains the 71,525 resources from the *ajax* tag stream.

As one can see in Fig. 2.3, the frequency-rank distributions of the three Web corpora show a slope for the most frequently used words that is similar to the slope in Fig. 2.2. The difference is that in the Web corpora the flattened slope can be observed for words with ranks between 1 and 1,000 while for tag streams the flattened slope is typically observable for tags with ranks between 1 and 100. It thus seems to be characteristic for tag streams that the flattened slope occurs at much lower ranks compared to text corpora. Furthermore, one can observe an overall lower number of different tags in the co-occurrence streams than in the corresponding web corpus. For example,

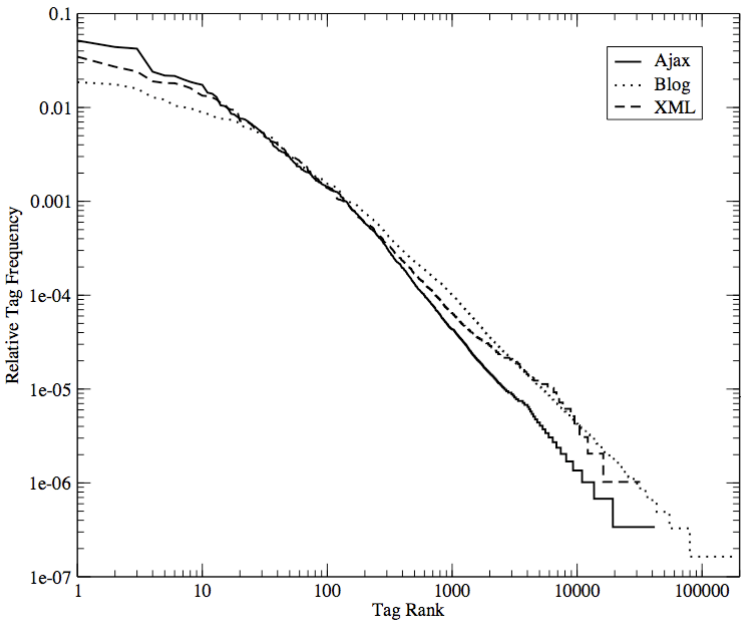

Figure 2.2: Frequency-rank distribution of tags in three tag streams.

the *ajax* stream contains 41,898 different tags while the *ajax* web corpus contains approximately 290,000 different words.

2.1.2 Resource Tag Streams

Here we consider a different kind of tag stream, i.e. we will look at the frequency-rank distributions of streams which only contain the tags assigned to a specific resource. In Fig. 2.4 one can see the frequency-rank distributions of three different resources in the Del.icio.us data set. For `http://www.netvibes.com/` and `http://www.googleguide.com/` one can see a sharp drop in the frequencies of tags between rank 7 and 10. For the resource stream of `http://www.pandora.com/` this drop is not as sharp but still present. Pandora and Netvibes are among the most heavily tagged resources in Del.icio.us while Googleguide represents a less famous web site in Del.icio.us.

In (Halpin et al., 2007), this drop artifact was observed for almost all sites in a data set of 500 heavily tagged sites from Del.icio.us. Thus, it is highly improbable that the drop only represents noise in the data set although there is a wider variety between the slopes of the single resource streams than what we observed for the co-occurrence streams. Instead of noise it is likely "a consistent effect of the way tagging is performed" (cf. (Halpin et al., 2007)).

Two possible explanations are offered in (Halpin et al., 2007): (1) It may be related to a cognitive effect during tagging (e. g. based on the average number of tags contained in a posting) or (2) it may be an artifact of the user interface specific to Del.icio.us. In Section 3.3 we will show, that with the help of our model we can explain the effect as being likely an artifact of the Del.icio.us user interface.

2.1.3 Distribution of inter-arrival times

Aim of this section is to briefly present our work on the temporal patterns of users' tagging activity and show that the statistical properties of inter-arrival times between subsequent tagging events cannot be explained without taking into account correlation in users' behaviors. This will show

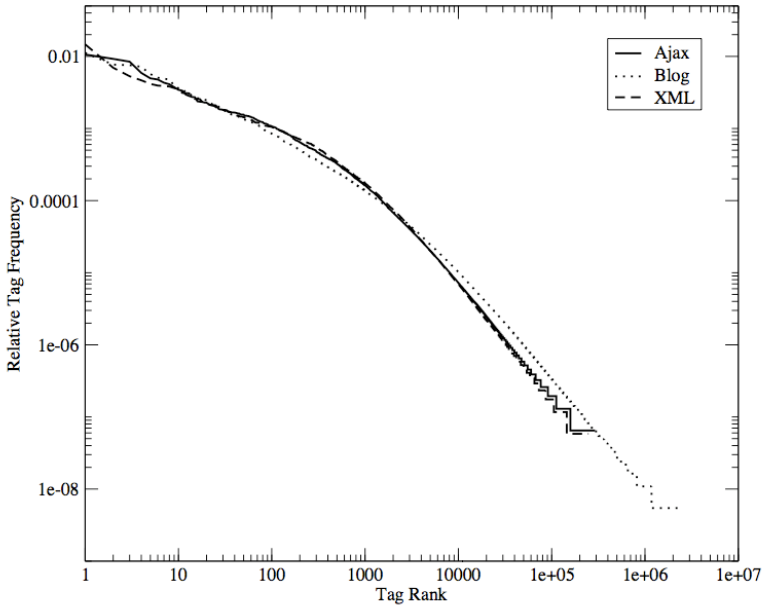

Figure 2.3: Frequency-rank distribution of words contained in three topic centric corpora.

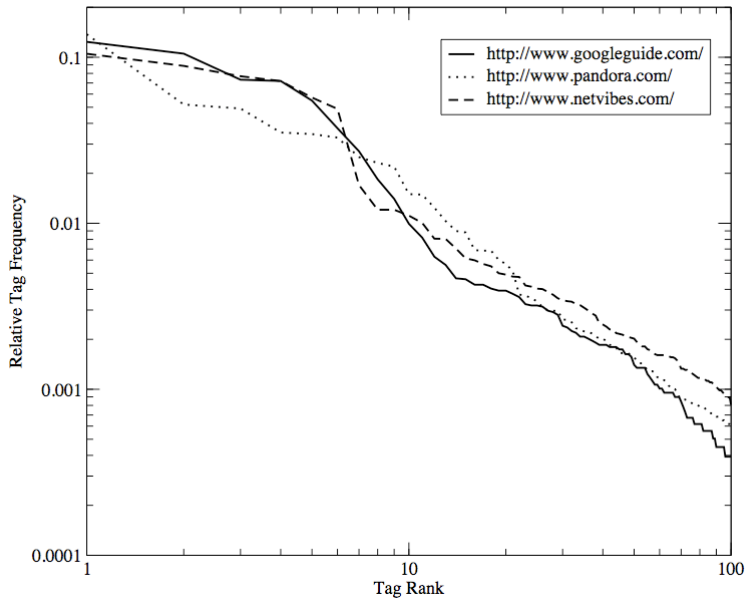

Figure 2.4: Frequency-rank distribution for the 100 most frequently assigned tags for three resource streams from a Del.icio.us data set.

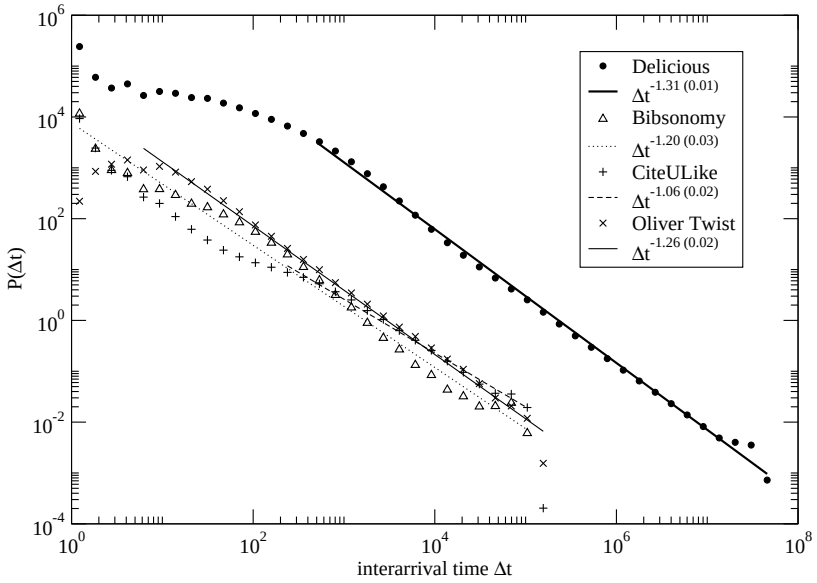

Figure 2.5: Tag inter-arrival time distribution in collaborative tagging communities compared with word inter-arrival times in Dickens' novel *Oliver Twist*.

that social interaction in collaborative tagging communities shapes the evolution of folksonomies. We will observe a consensus formation process involving the usage of a small number of tags for a given resources through a numerical and analytical analysis of some well-known folksonomy data-sets.

Correlations in the behavior of user collaborating in tagging resources online can be studied by inspecting the temporal statistics of tag usage. Time, here, is discrete and measured as number of successive posts. For example, one can study the inter-arrival time of tags, that is, the time length occurring between two subsequent tag assignments involving the same tag. If users behave independently, tags are added with a constant probability at each time unit. Accordingly, the arrival of tag would be described by a Poissonian process, where each occurrence is uncorrelated from the previous one. In this case, inter-arrival times are distributed according to an exponential distribution with a well-defined average inter-arrival time given by $1/f$, where f is the tag frequency. By contrast, observed individual tag inter-arrival time distribution shows that inter-arrival times span over all time scales, with a fat-tailed distribution, as shown in Fig. 2.5 The number of inter-arrival times of time length t, computed over all tags, is a power law $W(t) \propto t^{-\gamma}$, with $\gamma \simeq 1.3$ in different tagging systems.

The bursty behavior of tagging activities is not in itself a signature that complexity arises due to the interaction of users. A clearer sign of user cooperation can be found by analyzing the temporal pattern corresponding to individual resources. Inter-arrival times t between subsequent tagging of the same resource are distributed according to a power-law with a sharp cut-off for large values of t going to infinity for less tagged resources, as displayed in Fig. 2.6. Since a user cannot tag a resource twice, the fact that individual resource are tagged in "avalanches" depends on the contribution of many users. By contrast, if users were tagging independently one from each other, t should be distributed as an exponential random variable, as happens for Poissonian processes. The individual resource inter-arrival distribution can be analyzed as done above for tags, showing that resources are tagged in bursts spanning all time length scales.

However, this is not yet a proof of cooperation among online users. In fact, bursts of attention may arise by both a direct mutual influence between users one on each other; otherwise, users may independently be influenced by the same sources of information and news, where attention bursts may originate without any interaction among them. The stream of tag assignment involving a given

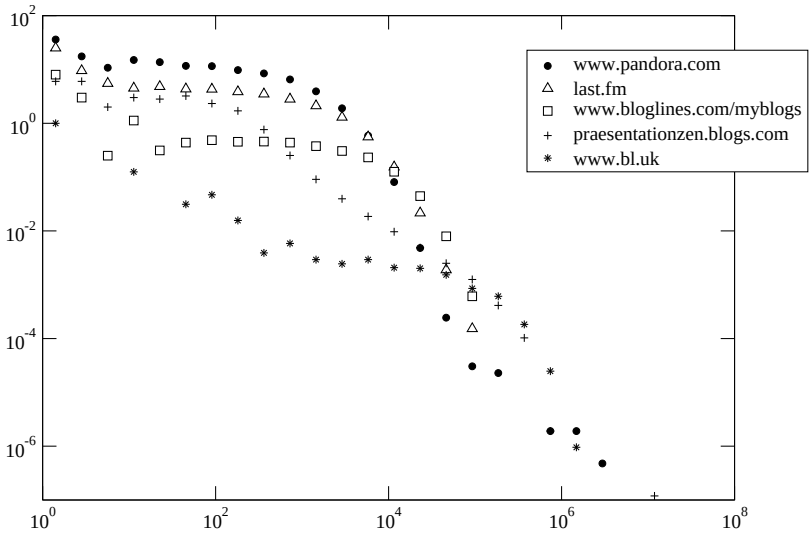

Figure 2.6: The distribution of inter-arrival times between subsequent tag assignments involving the same resource.

resource, though, carries a clearer evidence of users interaction. By plotting the number of distinct tags, i.e. the vocabulary, used for a resource as a function of the number of tag assignments to it, one observes a sub-linear vocabulary growth: so, the pace at which new tags are introduced by users to describe a resource decreases with time, and new tags are introduced less and lesser. In other words, users tend to employ the same tags used by previous peers when describing the same resources. Fig. 2.7 shows that the sub-linear relation between tag assignments and number of distinct tags involving a single resource holds for the large majority of them. Interestingly, this relation is not respected by spam bookmarks, that is, by tag assignments violating of the collective agreement about tag semantical organization. As other signatures of complex features, this relation may reveal useful in methods of spam detection.

2.1.4 Tag-tag correlations

Another possible indicator of user correlated activity is the tag-tag correlation defined as:

$$C(\Delta t) = \frac{1}{T - \Delta t} \sum_{t=1}^{t=T-\Delta t} \delta(\text{tag}(t + \Delta t), \text{tag}(t)) \,,$$

where $\delta(\text{tag}(t + \Delta t), \text{tag}(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t + \Delta t$. The same definition of $C(\Delta t)$ may be extended to any kind of stream and therefore can be used in the frame of fiction texts written by a given author. The curve representing this quantity in the case of Dickens' novel *Oliver Twist* is shown in Fig. 2.8.

It is clear that the stream of words originated from the same author displays a correlation at short times. In fact, by reshuffling the whole text, i.e. by removing correlations by hand, a flat correlation occurs. The correlation displayed by the text is of logarithmic (or equivalently small power) type.

The question is now whether the tagging activity of a single user displays the same type of correlations as in texts. To check this point we studied the correlations for a set of delicious users and discovered that there exist two different type of users. There are users who started their tagging activity by importing their local bookmarks into del.icio.us and users who started their activity from the scratch. In Fig. 2.9 we show the tag-tag correlation curve for a typical user who did not import his bookmarks and compare it with the corresponding word-word correlation curve of Dickens' *Oliver Twist*. It is rather surprising to observe almost the same behavior at small times. Instead, the

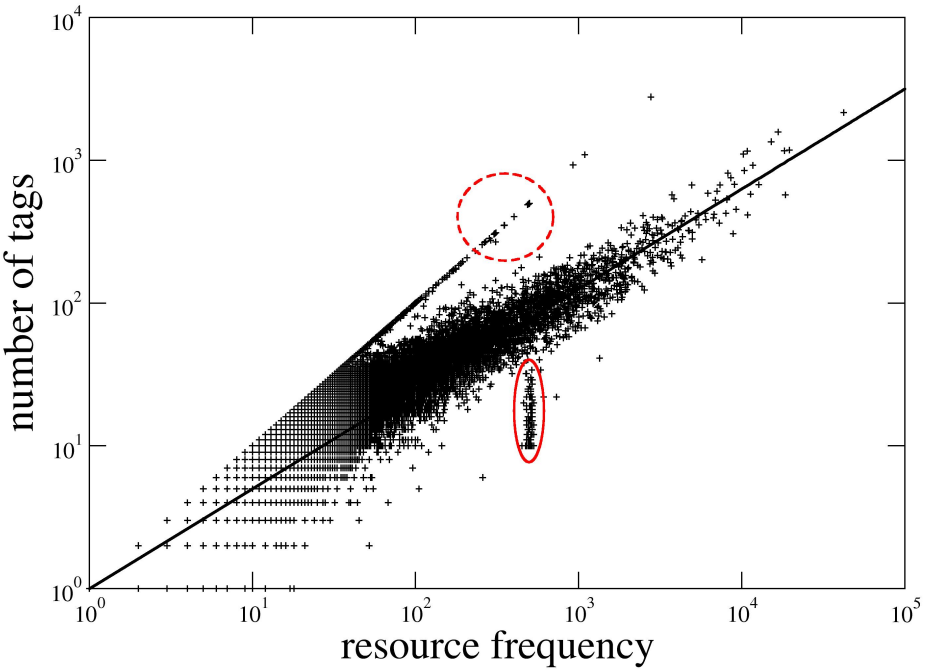

Figure 2.7: The number of distinct tags associated to a resource as a function of the number of tag assignments involving a resource, for the social tagging website del.icio.us. The resources inside the red ellipses correspond to spam posts of two kinds. In the dashed ellipse, spam formed by a single post with many (up to thousands) tags. In the solid line ellipse, spam made of repeated posts made by a few users tagging resources belonging to the same spam domain pcwash.spyware.com. See section 5.4.

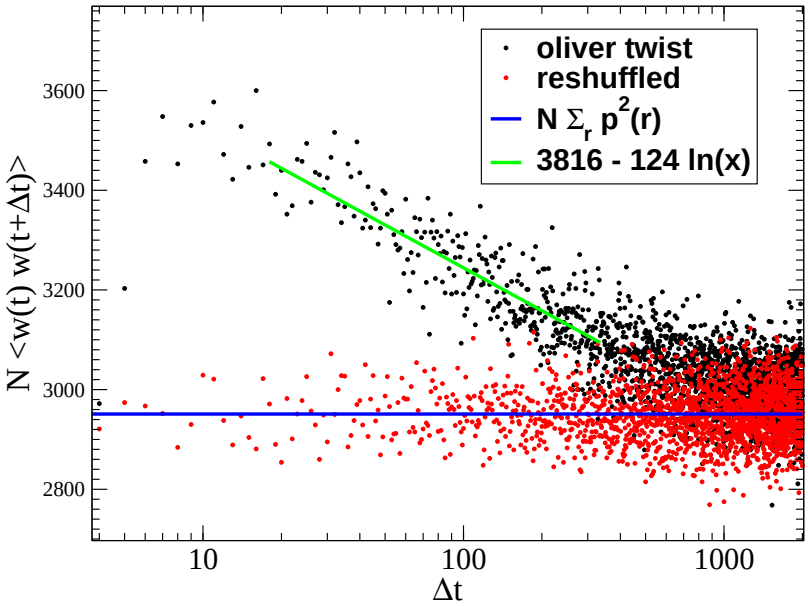

Figure 2.8: Word-word correlation in Dickens' novel *Oliver Twist* compared to the same quantity calculated for a reshuffled version of the text. The straight horizontal line is the asymptotic theoretic limit of an uncorrelated poissonian process with words extracted independently from the measured word frequency distribution.

correlation curve referred to users who uploaded their bookmarks shows no particular interesting features: a great number of tags where uploaded virtually at the same time, so that it is not possible to discern the correct time order of them corresponding to the logical reasoning of the user.

Last question we pose is whether the stream of tags related to the same resource shares the same correlation behavior of the previous two analyzed streams. In Fig. 2.10 we show such correlations for the resource `http://www.flickr.com`. Here, the correlation looks rather flat, apart the usual fluctuations, but remains systematically higher with respect to the reshuffled version of the stream, which by definition contains no correlations.

We ascribe this behavior to a co-operation of users, possibly tuned by external events, yielding to bursty tagging activity. This result was also confirmed in the finding of the previous section.

2.1.5 Tag length analysis and comparison with texts

Additionally, we report here a quick, unpublished statistical comparison between tag streams and texts. In particular we consider the length of tags, i.e. their number of characters, compared with the length of words in (english) texts. Although the analysis was formerly motivated by a check of null model for Zipf distributions (Mandelbrot, 1953; Mitzenmacher, 2004), it reveals some interesting features.

In the left part of Fig. 2.11 (top panel) we show the number of distinct tags of a given length (i.e. the number of characters) in Del.icio.us and Flickr. At odds with the same quantity measured on the words used in "Oliver Twist" (right of Fig. 2.11), a fatter tail is observed, denoting a larger number of very long tags, with respect to long words. This could be easily explained since tags are often composed of several words (i.e. *socialbookmarking* is a popular tag), but it could also be the effect of interface misuse (some users separate more tags with the wrong characters, as the tag *book,elvis,music,wine,free,posters,text* which is considered as a single tag by the system) or spam. Note also that the measure in Del.icio.us and Flickr does not show qualitative differences between the two systems.

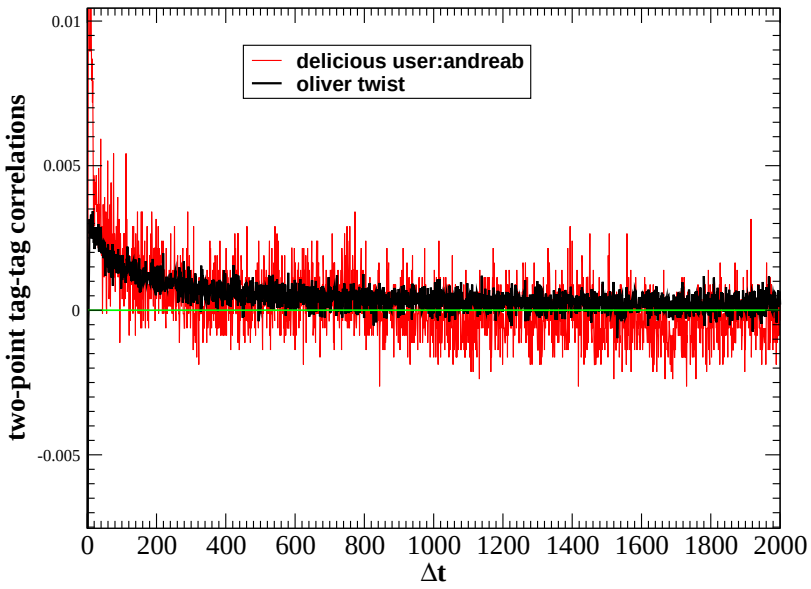

Figure 2.9: Tag-Tag correlation for the del.icio.us user *AndreaB* compared with the word-word correlation of Dickens' novel *Oliver Twist*. Curves where shifted in order to share the same asymptotic poissonian value.

In the left bottom panel of Fig. 2.11 is reported the overall frequency of tags of a given length (i.e. again the number of characters). Here a more complex scenario appears and the curves looks as the superposition of several exponential decays (continuous lines are tentative fitting in the corresponding ranges), with different characteristic lengths. This is very different with respect to words in texts, as shown by the curve in the right of Fig. 2.11, where a more regular and slow decrease is displayed. This result could be expected, beside the (mis)use of long tags as before, to the common use of acronyms in tagging.

Finally, let note that, as expected, both measures on tags and words show non trivial features, as opposed by what could be predicted by simplistic (null) models based on very basic informational schemes ("monkey typing" models (Baldassarri et al., 2007; Mandelbrot, 1953; Mitzenmacher, 2004)): i.e. an exponential increase in the number of distinct tags and an exponential decrease of frequency (both as a function of tag lengths).

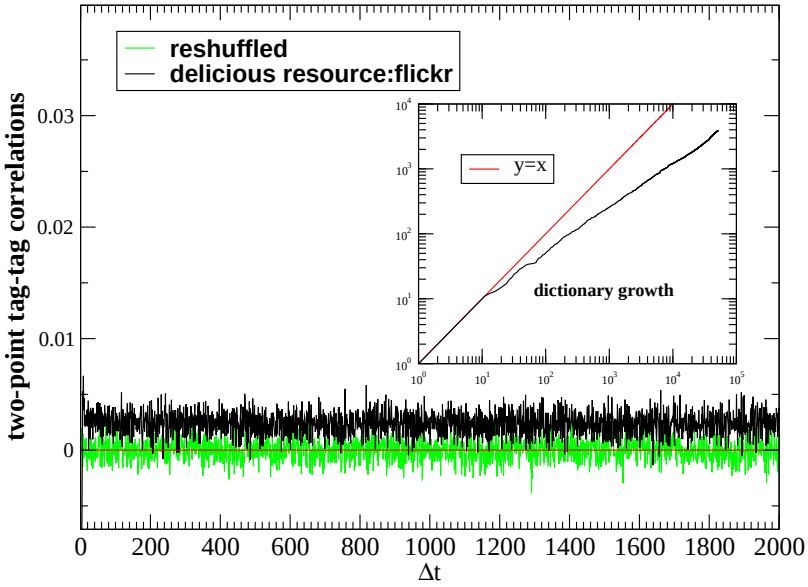

Figure 2.10: Tag-Tag correlation for the resource `http://www.flickr.com` compared with the corresponding reshuffled stream of tags, which contains no correlations.

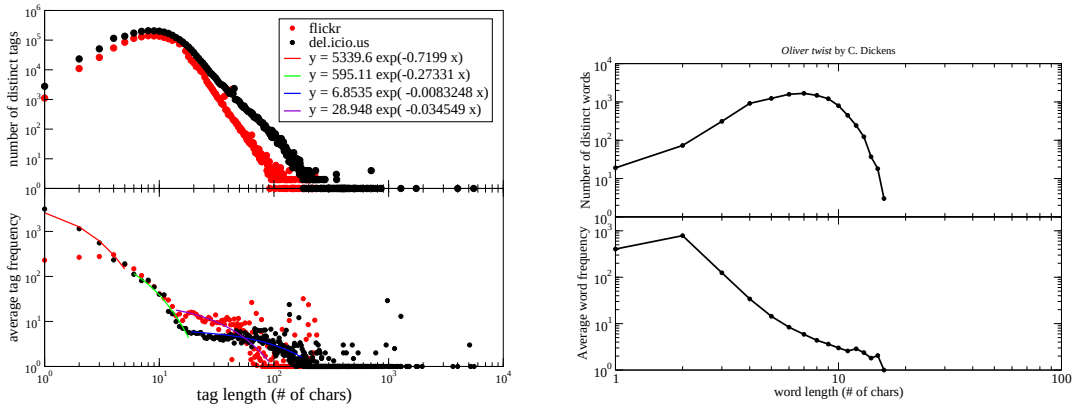

Figure 2.11: Left: Analysis of tag length in folksonomies. Top: Number of distinct tags of a given length (number of characters). Bottom: Frequency of tags of a given length. Right: Analysis of word length for texts (Oliver Twist).

2.2 Network view

2.2.1 Co-occurrence network

The tag co-occurrence network is a special projection of the full tri-partite folksonomy network, which encodes semantic correlation between tags. The formal definition of this network is based on the concept of *post*, that is a single user annotation, i.e. a triple of the form (u, r, P), where $u \in U$ is a user identificator, $r \in R$ is the unique identificator of a resource, and $P = \{t_1, t_2, \dots\}$ is a set of tags $t_i \in T$. Now, given a set of posts, we create an undirected and weighted network where nodes are tags and two tags t_1 and t_2 are connected by an edge if and only if there exists one post in which they were used in conjunction. The weight w_{ij} of an edge between tags t_i and t_j can be naturally defined as the number of distinct posts where t_i and t_j co-occur.

Each node i of the network is first characterized by its degree k_i (number of links). Moreover, since each link i, j carries a weight w_{ij}, an important measure of a node's importance is given by its strength s_i, defined as

$$s_i = \sum_{j \in \mathcal{V}(i)} w_{ij} \ , \tag{2.1}$$

where the sum runs over the set $\mathcal{V}(i)$ of i's neighbours. This quantity naturally generalizes the degree by measuring the strength of vertices in terms of the total weight of their connections.

The study of the global properties of the tagging system, and in particular of the global co-occurrence network, is of interest but mixes potentially many different phenomena. We therefore consider a narrower semantic context, defined as the set of posts containing one given tag.

In Figures 2.12 and 2.13 we show in particular the main properties of the co-occurrence network formed from all the posts containing a certain tag t^*. Broad distributions and non-trivial correlations are observed for both topological observables and edge weights. While Figures 2.12 and 2.13 present data for a particular tag, all the measured features are robust from one tag to another within one tagging system, and also across the tagging systems we investigated.

In the following we describe in details the measures performed.

A first characterization of a network's properties is obtained by the statistical distributions of the nodes' degree and strength, and the distributions of link weights: $P(k)$, $P(s)$, $P(w)$ (see Fig. 2.12). Moreover, it is customary to investigate the average strength $s(k)$ of vertices with degree k (see inset in Fig. 2.12):

$$s(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} s_i \tag{2.2}$$

where N_k is the number of nodes of degree k and δ_{k,k_i} is the Kronecker symbol, taking value 1 if $k_i = k$ and 0 otherwise.

In order to shed light on a network's topological correlations, two main quantities are customarily measured. The clustering coefficient c_i of a node i measures the local cohesiveness around this node (Watts, 1999). It is defined as the ratio of the number of links between the k_i neighbours of i and the maximum number of such links, $k_i(k_i - 1)/2$. The clustering spectrum (shown in the bottom panel of Fig. 2.13) measures the average clustering coefficient of nodes of degree k, according to

$$C(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} c_i \tag{2.3}$$

Moreover, correlations between the degrees of neighbouring nodes are conveniently measured by the average nearest neighbours degree of a vertex i, $k_{nn,i} = \frac{1}{k_i} \sum_{j \in \mathcal{V}(i)} k_j$, and the average degree of the nearest neighbours, $k_{nn}(k)$, for vertices of degree k (Pastor-Satorras et al., 2001)

(shown in the top frame of Fig. 2.13)

$$k_{nn}(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} k_{nn,i}. \tag{2.4}$$

In the absence of correlations between degrees of neighbouring vertices, $k_{nn}(k)$ is a constant. An increasing behaviour of $k_{nn}(k)$ corresponds to the fact that vertices with high degree have a larger probability to be connected with large degree vertices (assortative mixing). On the contrary, a decreasing behavior of $k_{nn}(k)$ defines a disassortative mixing, in the sense that high degree vertices have a majority of neighbours with low degree, while the opposite holds for low degree vertices (Newman, 2002).

These quantities have been generalized to weighted networks (Barrat et al., 2004). The weighted clustering coefficient of a node i,

$$c^w(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h \in \mathcal{V}(i),\, j \in \mathcal{V}(h)} \frac{(w_{ij} + w_{ih})}{2} , \tag{2.5}$$

considers not only the presence of triangles in the neighbourhood of i, but also their total relative edge weights with respect to the vertex's strength. The weighted clustering spectrum $C^w(k)$ is the weighted clustering coefficient averaged over all vertices with degree k (see bottom frame of Fig. 2.13). If the weighted clustering is larger than the clustering coefficient, triangles are more likely formed by edges with larger weights and thus carry a strong signification for the network.

Similarly, the weighted average nearest neighbours degree is defined as

$$k^w_{nn,i} = \frac{1}{s_i} \sum_{j \in \mathcal{V}(i)} w_{ij} k_j \ . \tag{2.6}$$

This quantity performs a local weighted average of the nearest neighbour degrees according to the normalized weight of the connecting edges, w_{ij}/s_i, measuring the effective affinity to connect with high or low degree neighbours according to the magnitude of the actual interactions. The average of $k^w_{nn,i}$ over all vertices with degree k, $k^w_{nn}(k)$, marks the weighted assortative or disassortative properties considering the actual interactions among the system's elements (see again the top plate of Fig. 2.13).

These observed features are emergent characteristics of the uncoordinated action of a user community, which call for a rationalisation and for a modeling framework.

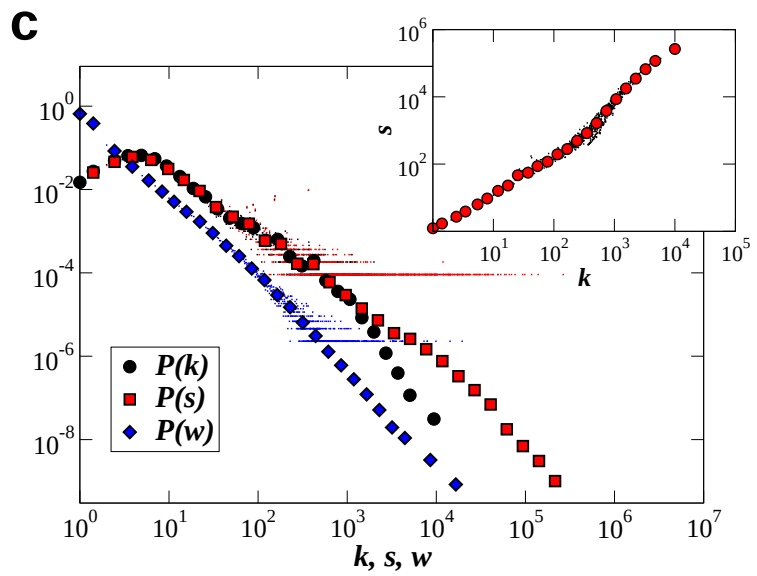

Figure 2.12: Broad distributions of degrees k, strengths s and weights w are observed. The inset shows the average strength of nodes of degree k, with a superlinear growth at large k. Both raw and logarithmically binned data are shown.

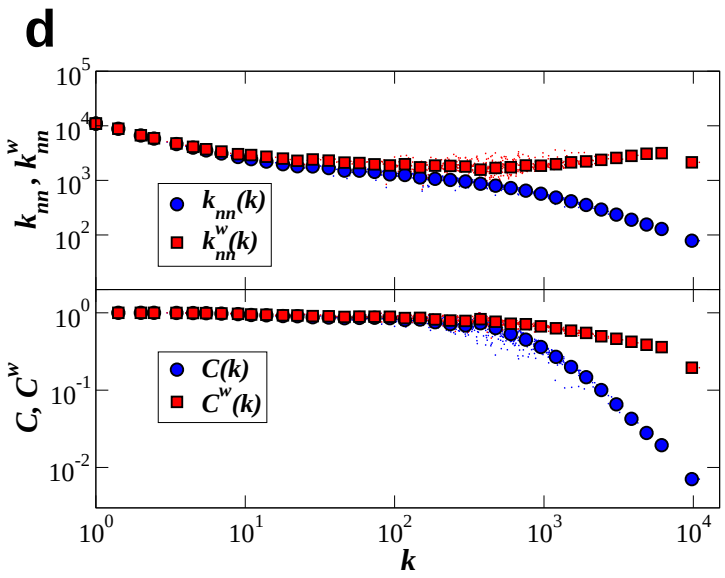

Figure 2.13: Weighted (k^w_{nn}) and unweighted (k_{nn}) average degree of nearest neighbors (top), and weighted (C^w) and unweighted (C) average clustering coefficients of nodes of degree k. k_{nn} displays a disassortative trend, and a strong clustering is observed. At small k, the weights are close to 1 ($s(k) \sim k$, see inset of B), and $k^w_{nn} \sim k_{nn}$, $C^w \sim C$. At large k instead, $k^w_{nn} > k_{nn}$ and $C^w > C$, showing that large weights are preferentially connecting nodes with large degree: large degree nodes are joined by links of large weight, i.e. they co-occur frequently together. Both raw and logarithmically binned data are shown.

Chapter 3

Models for tag streams and user behavior

3.1 Related work

Previous work has already developed models aimed at simulating the dynamic behavior of folksonomies. In order to come up with models close to the reality of observed folksonomies, the authors of this line of research have formalized assumptions about the tagging behavior of users into dynamic stochastic models. Two basic factors influence users during assigning tags to a resource: On the one hand, a user selects a tag based on his background knowledge and the content of the resource. On the other hand, he is exposed to previous tag assignments of other users that he may simply imitate in order to reduce the effort required for assigning own tags.

One of the earliest work in this field is the work of Golder and Huberman in (Golder and Huberman, 2006). They only simulated the imitation of previous tag assignments in their model and have not modeled the influence coming from the background knowledge. In practice their model is a variation of the stochastic Polya urn model originally described in (Eggenberger and Polya, 1923). The model is best explained by the metaphor of an urn containing balls with different colors. In each step of the simulation, a ball is selected from the urn and then it is put back together with a second ball of the same color. After a large number of draws the fraction of the balls with the same color stabilizes but the fractions converge to random limits in each run of the simulation.

Transfered to the simulation of tag streams, the balls contained in the urn are the tags previousy appeared in the stream, and the extraction decides the new tag to append to the stream. While in the original Polya urn model no new tags are invented, Golder and Huberman introduced the invention of new tags, with a low probability rate p. The modified model corresponds to the Simon model described in (Simon, 1955).

The model proposed by Golder and Huberman successfully explains the emergence of stable tag frequencies which they observed for resource tag streams. But it has several shortcomings: For example, it ignores the influence coming from the background knowledge of users. Furthermore, it does not reproduce the characteristic frequency-rank distributions of co-occurrence and resource tag streams that we observed in Section 2.1. Instead, it leads to a plain power-law distribution of the tag frequencies without the deviations for the most frequently used tags. Moreover, the model does not explain the typical growth of the set of distinct tags, since the constant rate of invention p leads to a linear growth of the distinct tags and not to the typical continuous, but declining growth that is shown in Section 2.1 in Fig. 2.1.

In (Cattuto, 2006; Cattuto et al., 2007b), Cattuto et al. propose a further variation of the Simon model. It takes the order of the tags in the stream into account. Like the previous models, it simulates the imitation of previous tag assignments but instead of imitating all previous tag assignments with the same probability it introduces a kind of long-term memory. It provides a fat-tailed access to

	frequency rank		
	co-occ. stream	res. stream	dictionary growth
Polya urn	○	○	fixed size
Simon model	○	○	linear
Cattuto et al. model	+	○	linear
Halpin et al. model	○	○	linear
This model	+	+	sub-linear

Table 3.1: Rating of tag stream simulation models and how they reproduce the different properties of tag streams ($+ = good, \circ = medium$). More details are available in the text.

the previous tag assignments, i. e. the probability of selecting a tag assignment located x steps into the past is given by a function $Q_t(x)$ that returns a power-law distribution of the probabilities. The Yule-Simon model with long term memory from (Cattuto, 2006; Cattuto et al., 2007b) successfully reproduces the characteristic slope of the frequency-rank distribution of co-occurrence tag streams but it fails to explain the distribution in resource tag streams as well as the decaying growth of the set of distinct tags because it leads to a linear growth.

The most recent model for simulating the evolution of tag streams is proposed in (Halpin et al., 2007). It is the first model which does not only simulate the imitation of previous tag assignments but it also selects tags based on their information value. The information value of a tag is 1 if it can be used for only selecting appropriate resources. A tag has an information value of 0 if it either leads to the selection of no or all resources in a tagging system. In (Halpin et al., 2007), Halpin et al. empirically estimate the information value of a tag by retrieving the number of webpages that are returned by a search in Del.icio.us with the tag. Besides of the selection based on the information value, the model also simulates the imitation of previous tag assignments. It corresponds to a preferential attachment or Polya urn model. Overall, the proposed model leads to a plain power-law distribution of the tag frequencies and to a linear growth of the set of distinct tags. It thus only partially reproduces the frequency-rank distributions in co-occurrence and resource tag streams and it is not successful in reproducing the decaying dictionary growth.

In Tab. 3.1 the different models described in this section are summarized and it is shown in how far they are able to reproduce the characteristic properties from Section 2.1. The basic assumption in all models is that users imitate the previous tag assignments of other users in one or the other variation of a Polya urn or Simon model. With exception of the Yule-Simon model with memory (Cattuto, 2006; Cattuto et al., 2007b) all other models can only explain a power-law like distribution of the tag frequencies and not the deviation for the most frequent tags in co-occurrence and resource tag streams. Only the Yule-Simon model with memory (Cattuto, 2006; Cattuto et al., 2007b) can explain the deviation for co-occurrence streams. But the most obvious flaw of all previous model is their inability to explain the characteristic sub-linear dictionary growth. Instead of the continuous, but decaying growth of the set of distinct tags they all lead to a linear growth or, even worse, assume a fixed vocabulary size.

3.2 An Epistemic Dynamic Model for Tagging Systems

In the following, we will propose a dynamic stochastic model which simulates the evolution of tag streams based on assumptions about the factors that influence users assigning tags. By modeling the influence factors at the micro-level of the single tag assignments we try to understand how the previously described characteristic properties on the macroscopic level of tag streams emerge.

Our model consists of two components, each simulating one of the two factors that have an influence on the tag assignments of a user. We will motivate and discuss the five parameters of our simulation model in detail.

3.2.1 Imitation

The simulation of a tag stream always starts with an empty stream. Then, in each step of the simulation, it is simulated with probability I that one of the previous tag assignments is imitated. With probability BK, the user selects an appropriate tag from his background knowledge about the resource (see Section 3.2.2).

Usually, not the whole previous tag stream is accessible. For example during posting a new resource to Del.icio.us the user sees a set of recommended tags which is the intersection between the user's tags and all tags already assigned to the resource (*recommended tags*). Furthermore, a user can see all tags that he previously attached to other resources (*your tags*) and the 7 most popular tags of the resource (*popular tags*). By clicking on any of the tags the user can easily assign it to the resource.

For keeping our model as simple as possible, we will restrict it to only simulating the imitation of the most popular tags. For this purpose, we introduce two further parameters n and h which can be used for restricting the access to the previous tag stream.

The parameter n represents the number of popular tags a user has access to. In case of simulating resource streams, n will correspond to the number of popular tags shown by the Del.icio.us interface (i. e. $n = 7$). In case of co-occurrence streams n will be larger because the union of the popular tags of all resources that are aggregated in the co-occurrence stream will be depicted over time.

Furthermore, the parameter h can be used for restricting the number of previous tag assignments which are used for determining the n most popular tags. For example, for $n = 7$ and $h = 300$ only the 7 most popular tags during the last 300 tag assignments can be imitated. The probability of selecting the concrete tag t from the n tags is then proportional to how often t was used during the last h tag assignments.

By combining n and h we get an effect that is comparable to the fat-tailed access of the Yule-Simon model with memory (cf. (Cattuto, 2006; Cattuto et al., 2007b) and Section 3.1). Obviously, the imitation of previous tag assignments will never add a previously unknown tag to the stream. Thus, it will only have an effect on the frequency-rank distribution but not on the dictionary growth.

3.2.2 Background Knowledge

With probability BK, it is simulated that a tag is selected from the background knowledge about the content of the resource. It corresponds to selecting an appropriate natural language word from the active vocabulary of the user. Each word has assigned a certain probability with which it gets selected.

Obviously, it is not possible to get the active vocabulary of each individual user. Instead we will simulate the active vocabulary of an average user with the help of the Web corpora (see Section 2.1 and Fig. 2.3). The probability of selecting a specific word t corresponds to the probability with which t occurs in the Web corpus. For example, for simulating the *ajax* tag stream we used the probabilities found in the *ajax* Web corpus.

The background knowledge will influence the frequency-rank distribution as well as the dictionary growth. The simulation of the background knowledge adds tags with a certain probability (i. e. the frequency-rank distribution is influenced) but it may also add previously unknown tags if a tag gets selected the first time from the background knowledge (i. e. the dictionary growth is influenced).

3.3 Model evaluation

To determine how well our model fits to the real life tag streams, we investigate to which extent the 5 parameters of our model may be fixed in a way such that our model can consistently predict the behavior of the various observed distributions.

First, we will verify whether the selection from the background knowledge plays a role during tag assignment and whether the selection probabilities of words from the background knowledge have a frequency-rank distribution similar to the word occurrence probabilities in text corpora. Only if both assumptions are correct, we will be able to reproduce the characteristic dictionary growth rate in real tag streams with the help of our model because only these two components have an influence on the simulated dictionary growth rate.

For testing our hypothesis we have simulated several tag streams with the help of our model and compared their dictionary growth with the dictionary growth in real tag streams. During the experiments, we fixed $p(w|t)$ to the word occurrence probabilities in the *ajax* web corpus. Furthermore, we set n to a value of 50 and h to a value of 1000. Finally, we tested several values for I and BK.

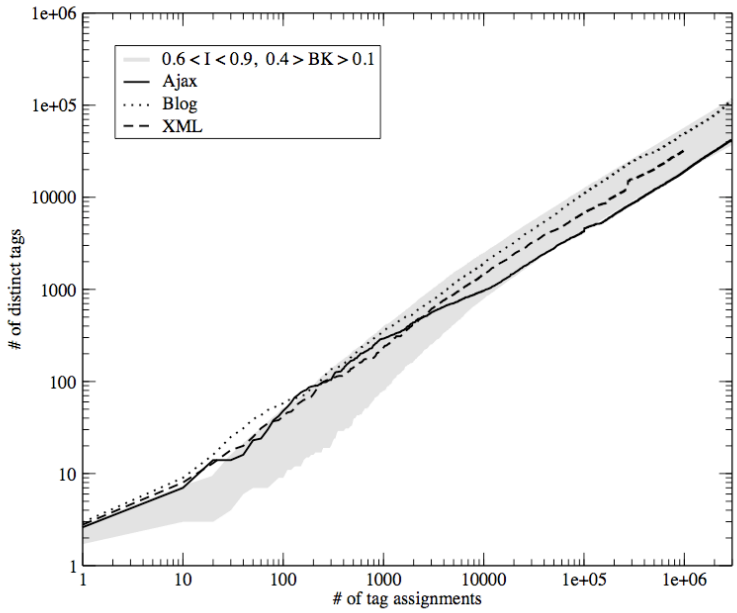

Figure 3.1: Comparison between the growth of the set of distinct tags in simulated tag streams and the growth in the *ajax*, *blog* and *xml* tag stream. The lower bound of the gray area corresponds to a growth simulated with $I = 0.9$ and the upper bound corresponds to a simulated growth with $I = 0.6$.

Fig. 3.1 one can see that the growth rates of the simulated tag streams and the *ajax* stream have the same exponent. These findings have also be confirmed by looking at the growth rates of other two tag streams.

This confirms that background knowledge is suitable for modeling the dictionary growth rate in real tag streams. Furthermore, it confirms that the word occurrence frequencies in text corpora are a suitable approximation of the tag selection probabilities from the background knowledge.

For what concern the frequency-rank distribution of our streams, let first note that without the influence of imitating previous tag assignments and having only the influence of the background knowledge we would observe frequency-rank distributions for streams that are comparable to that of $p(w|t)$ or $p(w|r)$. Thus, only if the imitation is correctly modeled we will be able to reproduce the

characteristic frequency-rank distributions of co-occurrence and resource tag streams. This would show that imitation also plays an important role for real users during tag assignment.

Co-occurrence Tag Streams

Based on our previous findings about the dictionary growth we can already predict the values for I, BK (and $p(w|t)$) that will lead to realistic frequency-rank distributions. For predicting appropriate values for n we will use the findings in (Cattuto, 2006; Cattuto et al., 2007b) about the semantic breadth of the *ajax*, *blog* and *xml* co-occurrence streams.

In (Golder and Huberman, 2006), they experimentally estimated the semantic breadth of the three co-occurrence streams we are using. For example, for the *blog* stream they estimate that it contains approximately 100 tags that are perceived by users as those tags characterizing the topic. In our case, this number can be used for approximating how many tags will be usually contained in the union of the top ranked tags of the different resource streams aggregated in the co-occurrence stream of *blog*. Because *ajax* and *xml* are narrower topics we will use for their simulation $n = 50$.

For the value of the parameter h, we do not have any hint that would be a realistic value according to the Del.icio.us tagging interface. But in order to verify its influence on the frequency-rank distribution we will use two good guesses for values that seem to be appropriate if such an influence exists. For this purpose, we will use $h = 500$ and $h = 1000$. Furthermore, with $h = 10000$ we will use a very high value that corresponds to seeing more or less the complete previous tag stream because after so many tag assignments the top ranked tags have reached a stable frequency and thus further increasing h will not lead to another ranking of tags. Thus, if the influence of h exists we would expect for $h = 500$ and $h = 1000$ to get realistic frequency-rank distributions and for $h = 10000$ a frequency-rank distribution that is significantly different to that of real co-occurrence tag streams.

In order to be comparable with the frequency-rank distributions of the real co-occurrence streams we will also have to simulate the same number of tag assignments, e. g. for comparing the simulation with the real *ajax* stream we will also have to simulate a stream with 2.95 million tag assignments.

Looking at the frequency-rank distributions of the streams simulated with the above mentioned parameters and varying value for h, one can see that a high value of h leads to a sharp drop in the frequency-rank distributions while with lowering h it gets closer to the distribution in the background knowledge. Moreover, the effect of h is increased with a higher probability of imitating a previous tag assignment. The results confirm that the parameter h is relevant for explaining the frequency-rank distributions because if there isn't such an influence we would get a slope that is similar to that for $h = 10000$, i. e. there would be a sharp drop. Furthermore, it seems to be more appropriate to simulate co-occurrence streams with $h = 1000$ instead of $h = 500$.

Looking at the influence of the parameter n on the slope of the frequency-rank distribution, i. e. how it is influenced by a different semantic breadth represented in the simulated co-occurrence stream, one can see that with an increased semantic breadth we predict more flattened slopes of the frequency-rank distribution. This effect is further increased with a higher probability of imitating a previous tag assignment.

In Fig. 3.2, we show the frequency-rank distribution of simulations that were done with the previously predicted parameter values. Each of the simulated slopes is directly compared with the corresponding frequency-rank distribution of the real co-occurrence streams. In all three cases, we have successfully reproduced the frequency-rank distribution of the real co-occurrence streams with the help of the parameter values that we predicted based on the assumptions in our model. Furthermore, we used for each of the distributions the same parameter values for BK and $p(w|t)$ with which we also reproduced their dictionary growth rate.

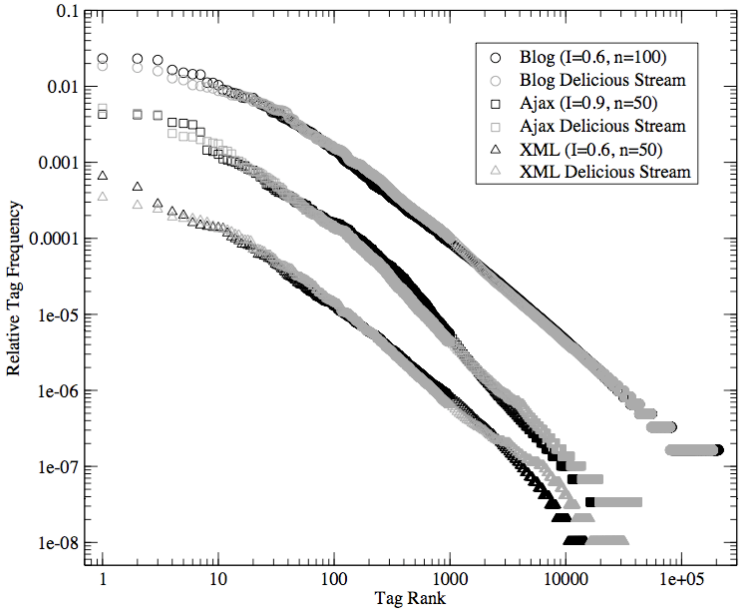

Figure 3.2: Comparison between the frequency-rank distributions of three co-occurrence tag streams and corresponding simulated tag streams. I was set to 60% probability for simulating the *blog* and *xml* stream to take their higher dictionary growth rate into account (cf. Fig. 3.1). In order to be able to show all six curves in one graph, we shifted down the relative tag frequencies of the *ajax* and *xml* stream by respectively one and two decades so that they don't overlap each other.

Resource Tag Streams

For the simulation of resource tag streams we will have to use other parameter values. According to our assumptions made in the model and the previous findings and 3.3 we would predict that the following parameter values have to be used for simulating resource tag streams:

- I and BK: Like for the dictionary growth and the frequency-rank distribution in co-occurrence streams we would expect that for I values in the range between 60% and 90% are also appropriate for simulating resource tag streams because the probabilities of this basic user decision during tag assignment are independent from co-occurrence or resource streams.
- $p(w|r)$: In opposite to $p(w|t)$, we will not get a good approximation of $p(w|r)$ by the word occurrence probabilities in r because the content of a single resource r will not contain enough words for getting stable frequencies and ranks. Thus, we will in the following approximate $p(w|r)$ with an appropriate topic specific distribution $p(w|t)$.
- n: For this parameter we will use $n = 7$ because in the Del.icio.us tagging interface the user sees the 7 most popular tags for imitation.
- h: In a single resource stream less previous tag assignments will be used for determining the rankings than in co-occurrence streams because the latter are aggregations of several resource streams. Thus, we would expect a value for h that is less than that used for simulating co-occurrence streams. Thus, in our experiments we will use $h = 300$.

Because of the unexact approximation of $p(w|r)$ and because of the wider variance in the frequency-rank distributions in real resource streams we will not be able to exactly reproduce the

frequency-rank distribution of a single resource stream. Instead, we can only expect to reproduce the most characteristic property of resource streams, namely the sharper drop between rank 7 and 10 in the frequency-rank distribution that we observed in Section 2.1.2 and that was also observed by Halpin et al. in (Halpin et al., 2007).

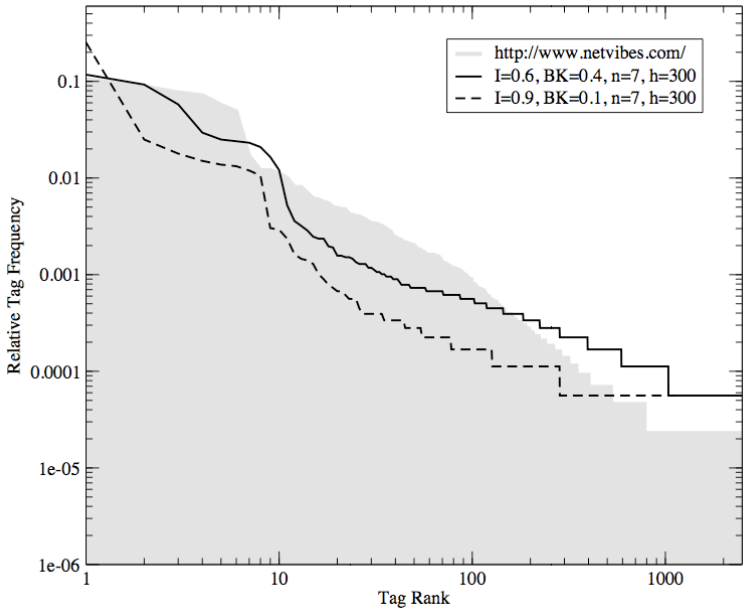

Figure 3.3: Comparison between the frequency-rank distributions of the resource tag stream for http://www.netvibes.com/ and of tag streams simulated with $n = 7$. As values for I we used 60% and 90% probability. The distribution $p(w|\texttt{netvibes.com})$ is approximated by $p(w|\texttt{ajax})$.

In Fig. 3.3, we show the frequency-rank distribution that we predicted with our model for the tag stream of the resource http://www.netvibes.com/. The website of Netvibes.com makes heavy use of AJAX technology and two most often used tags are *web2.0* and *ajax*. Thus, we approximated the distribution of $p(w|netvibes.com)$ with the distribution of $p(w|ajax)$.

As it was expected, we do not get an overlap between the simulated and the real tag stream that is as good as for the co-occurrence tag streams in Section 3.3 but we can still see the most characteristic sharp drop in the frequency rank distribution between rank 7 and 10. Thus, it is likely that this sharp drop can be explained with the restriction of the Del.icio.us tagging interface where the user only sees the 7 most frequent tags during posting a new bookmark. Nevertheless, only further experiments with a more accurate approximation of $p(w|r)$ may bring further evidence for this hypothesis.

3.4 Conclusion and future work

Our generative model of folksonomies that can be used for simulating the evolution of tag streams. It distinguishes between two basic ways a user may assign tags: (1) he may imitate a previous tag assignment from the stream or (2) he may choose a word from his active vocabulary that is related to the content of the resource. In an evaluation, we have successfully shown that both factors play an important role for real users during tag assignment. The evaluation suggests that the imitation rate during tag assignment will be between 60% and 90%.

Furthermore, with the tag growth and the frequency-rank distribution in co-occurrence streams as well as with the frequency-rank distribution in resource streams we have identified three different

characteristic properties of tag streams. During our evaluation we have also shown that our model is the first one known in the literature that can be used for consistently predicting all three of the characteristic properties (cf. Tab. 3.1). Especially, it is the first model that successfully reproduces the sub-linear tag growth.

In the future, we will do a more in depth evaluation of the model and of the characteristic properties of tag streams. For example, it still has to be evaluated in how far the observed frequency-rank distributions really obey to a power-law or whether they are better explained by other distributions with a long tail like log-normal distributions (cf. (Clauset et al., 2007)). Furthermore, we plan to extend the evaluation to additional tag streams in order to get more generalizable results.

There is a lot of ongoing research in which the co-occurrence probabilities of tags are exploited, e. g. for finding relationships between tags or for learning tag clusters. But our research suggests that the co-occurrence probabilities do not only express semantics coming from the background knowledge of users but that they are also influenced by the random process of imitating previous tag assignments.

Especially the restriction to imitating only the top-N tags makes the co-occurrence probabilities in a tagging system dependent on the previous state of the system. In this case, our model predicts that the tag co-occurrence probabilities for the same resource (e. g. the URL http://www.netvibes.com/) in two independent tagging systems with comparable user communities will not converge to the same values due to the random imitation process. This will also influence the ranking of a tag attached to the same resource in independent but comparable tagging systems, i. e. there will be a deviation of the concrete rankings in one tagging system from the mean rankings in all tagging systems. Our model can be used for calculating the expected typical deviation in dependency on the imitation rate.

Chapter 4

Modeling the co-occurrence network

4.1 Social annotation as an exploration of a semantic space

Correlations between tag occurrences are (at least partially) an externalization of the relations between the corresponding meanings (Solé et al., 2008) and have been used to infer formal representations of knowledge from social annotations (Heymann and Garcia-Molina, 2006). At the same time no modeling framework exists which can naturally account for them while reproducing their network structure. We show in particular that the idea of social exploration of a semantic space has more than a metaphorical value, and actually leads to a framework that can predict fine observables of tag co-occurrence networks as well as robust stylized facts of collaborative tagging systems.

The fundamental idea underlying our approach, illustrated in the left diagram of Fig. 4.1, is that a post corresponds to a random walk (RW) of the user in a "semantic space": starting from a given tag, the user adds other tags, going from one tag to another by semantic association. It is then natural to picture the semantic space as network-like, with nodes representing tags and links representing the possibility of a semantic link. In this framework, the vocabulary co-occurring with a tag is associated with the ensemble of nodes reached by successive random walks starting from a given node, and its size with the number of *distinct* visited nodes, $N_{distinct}$, which grows as a function of the number of performed random walks n_{RW}. Empirical evidence on the distribution of post lengths (see the right plot in Fig. 4.1) suggests to consider random walks of random lengths, distributed according to a broad law.

4.2 Model evaluation

First, analytical and numerical investigations show that sub-linear power-law-like growths of $N_{distinct}$ are then generically observed, mimicking the Heaps' law observed in tagging systems. However, vocabulary growth is only one aspect of the dynamics of tagging systems. Networks of co-occurrence carry much more detailed signatures that present very specific features (Figs. 2.12 and 2.13). Interestingly, our approach allows to construct *synthetic* co-occurrence networks: we associate to each random walk a clique formed by the nodes visited (see Fig. 4.1), and consider the union of the n_{RW} such cliques. Moreover, each link i, j built through this procedure receives a weight equal to the number of times nodes i and j appear together in a random walk. This construction mimics precisely the obtention of the empirical co-occurrence network, and Figs. 4.3 and 4.4 show the striking similarity between the characteristics of such synthetic networks and the data of Figs. 2.12 and 2.13.

Figure 4.4 in particular explores a highly non-trivial correlation between weights and topology in both real and artificial co-occurrence networks, namely how the weight w_{ij} of a link is correlated with its extremities' degrees k_i and k_j. The shape of the curve can be understood within our

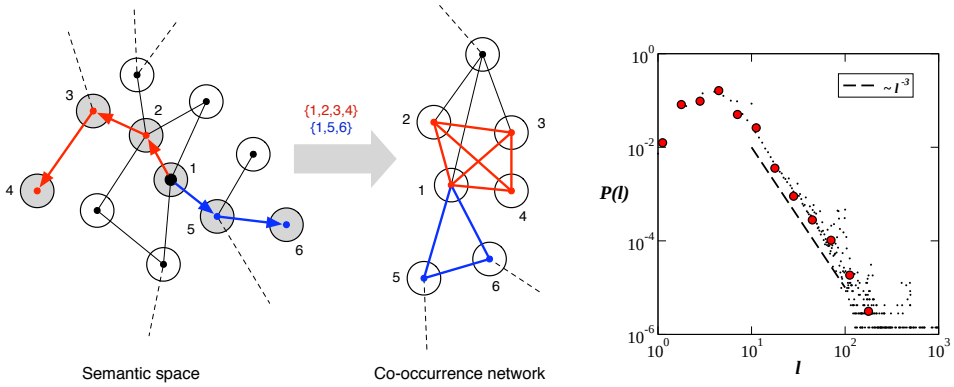

Figure 4.1: Left: Illustration of the proposed mechanism of social annotation. The semantic space is pictured as a network in which nodes represent tags and a link corresponds to the possibility of a semantic association between tags. A post is then represented as a random walk on the network. Successive random walks starting from the same node allow the exploration of the network associated with a tag (here pictured as node 1). The artificial co-occurrence network is built by creating a clique between all nodes visited by a random walk. Right: empirical distribution of posts' lengths $P(l)$. A power-law decay $\sim l^{-3}$ (dashed line) is observed.

framework. First, the broad distribution in l seems to be responsible for the plateau ~ 1 at small values of $k_i k_j$, since it corresponds to long RW that occur rarely and visit nodes that will be typically visited a very small number of times (hence small weights). Moreover, w_{ij} displays a power-law behaviour $\sim (k_i k_j)^a$ at large weights. Denoting by f_i the number of times node i is visited, $w_{ij} \sim f_i f_j$ in a mean-field approximation that neglects correlations. On the other hand, k_i is by definition the number of distinct nodes visited together with node i. Restricting the random walks to the only processes that visit i, it is reasonable to assume that such sampling preserves Heaps' law, so that $k_i \propto f_i^{\alpha}$, where α is the growth exponent for the global process. This leads to $w_{ij} \sim (k_i k_j)^a$ with $a = 1/\alpha$. Since $\alpha \simeq .7 - .8$, we obtain a close to $1.3 - 1.5$, consistently with the numerics. Strikingly, the similarity between empirical and synthetic co-occurrence networks carries through to other, even more subtle observables.

4.3 Subtle observables

In a weighted network, the similarity of two nodes i_1 and i_2 can be defined as

$$\text{sim}(i_1, i_2) \equiv \sum_j \frac{w_{i_1 j} w_{i_2 j}}{\sqrt{\sum_\ell w_{i_1 \ell}^2 \sum_\ell w_{i_2 \ell}^2}} , \tag{4.1}$$

which is simply the scalar product of the vectors of normalized weights of nodes i_1 and i_2. In Ref. (Cattuto et al., 2008b), it has been shown that this kind of distributional similarity contains non-trivial semantic information that can be used to detect synonymy relations between tags, or to uncover "concepts" from social annotations. This quantity indeed measures the similarities between neighbourhoods of nodes, and is therefore a correlation of a somehow higher order than the ones previously presented (clustering or assortativity properties).

In Fig. 4.2 we report the histograms of pair-wise similarities between nodes in various co-occurrence networks. We first present the data for the tag co-occurrence network of various tags of del.icio.us and BibSonomy: a clearly skewed character of the distributions is observed, with a peak for low values of the similarities. The data are quite similar for the various tags investigated

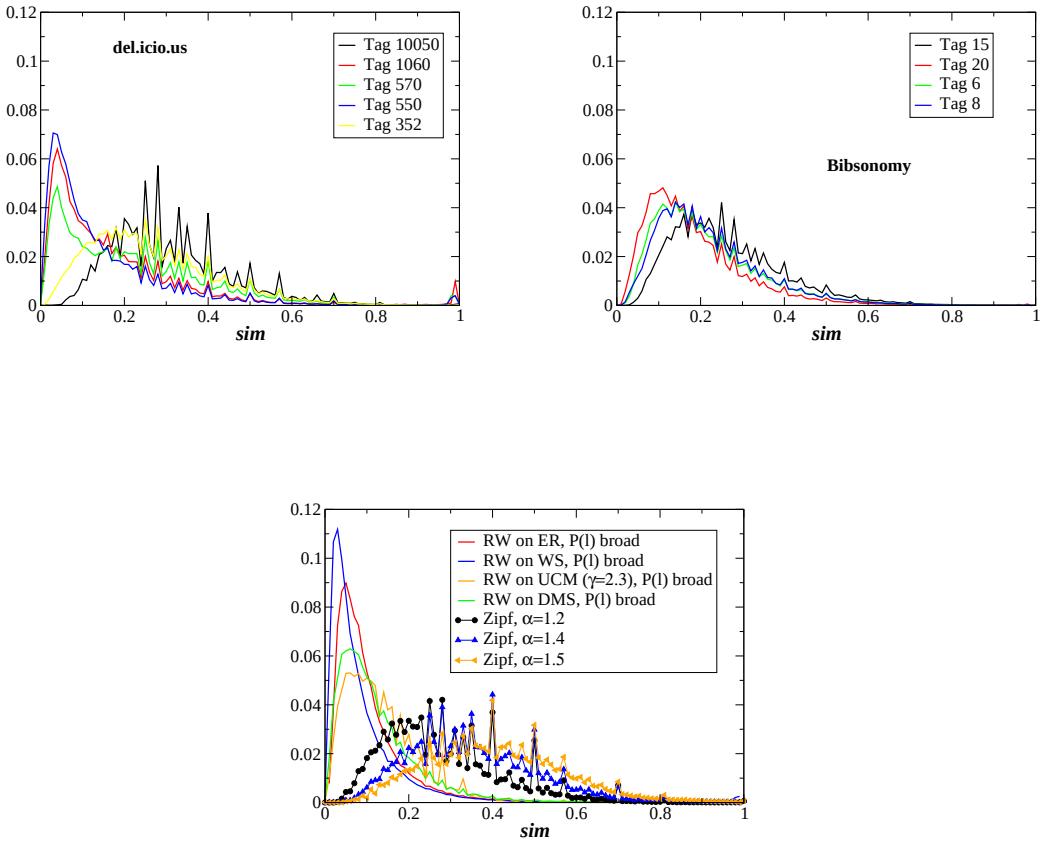

Figure 4.2: Distributions of cosine similarities for real and synthetic co-occurrence networks. For del.icio.us and BibSonomy, the tag number represents its popularity rank in the database. Two types of processes are considered for building synthetic co-occurrence networks: random walks (RW) on various types of networks (ER: Erdős-Rényi random graph; WS: Watts-Strogatz network; UCM: uncorrelated configuration model (Catanzaro et al., 2005) with broad degree distribution $P(k) \sim k^{-\gamma}$; DMS: highly clustered scale-free network with degree distribution $P(k) \sim k^{-3}$ and artificial posts built from a list of tags whose a priori frequencies follow a Zipf's law of exponent α (symbols).

in BibSonomy, while the peak can be more or less broad for del.icio.us. Figure 4.2 also displays the similarity histograms observed for the networks constructed according to the mechanism proposed in the main text, i.e. random walks of broadly distributed lengths, performed on underlying networks with various properties. A very similar behaviour is observed, with a (more or less pronounced) peak at low values of sim. For comparison, we also display the histogram of similarities for networks constructed from artificial tags as follows: (i) we start with a list of tags whose *a priori* frequencies follow a Zipf's law of exponent α; (ii) we construct artificial posts of length l distributed as l^{-3} by choosing at random l tags with probability proportional to their a priori frequency (the first tag of each post is always the same, since for the real data we are considering the posts containing all a given tag); (iii) we build the corresponding co-occurrence network. Clearly, this null model (which contains the Zipf's law as an ingredient, in contrast to our mechanism) does not contain any semantic information as the tags are used without any correlations. As shown in Fig. 4.2, the distribution of similarities is then indeed very different, and in particular it is much less skewed.

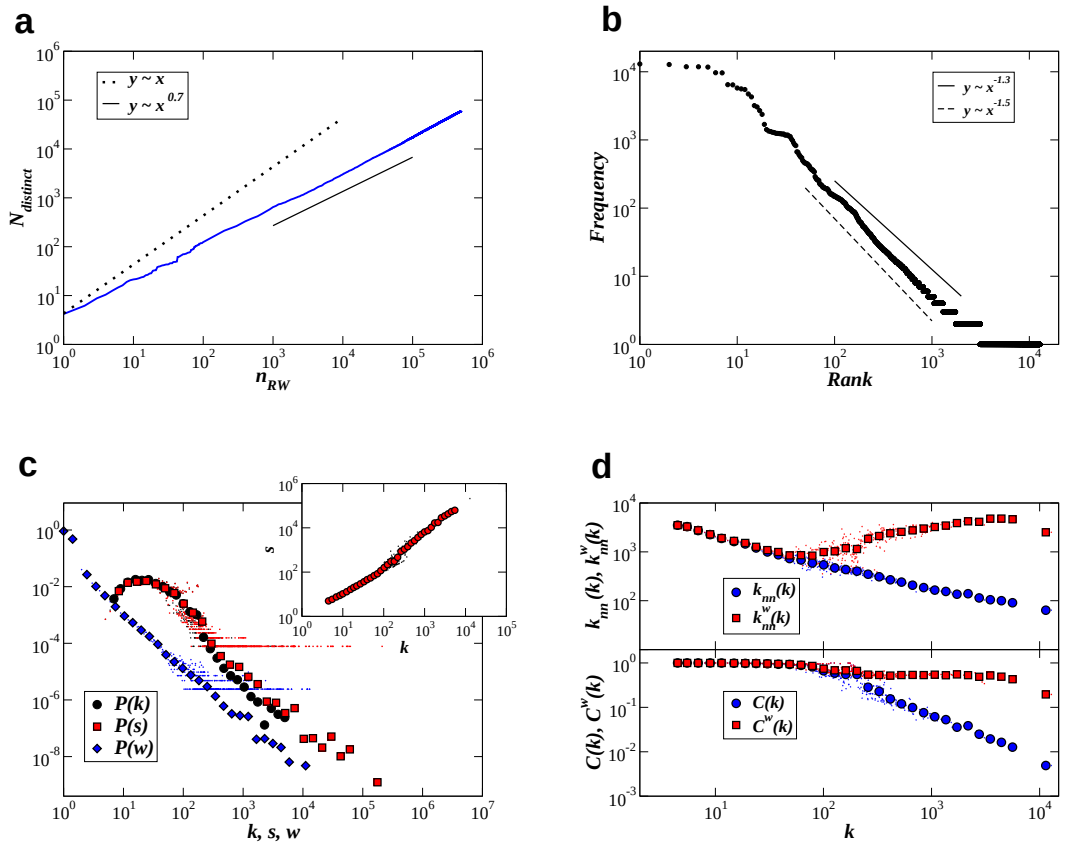

Figure 4.3: Synthetic data produced through the proposed mechanism. a) Growth of the number of distinct visited sites as a function of the number of random walks performed on a Watts-Strogatz network of size $5 \cdot 10^4$ nodes and average degree 8, rewiring probability $p = 0.1$. Each random walk has a random length l taken from a distribution $P(l) \sim l^{-3}$. The dotted line corresponds to a linear growth law while the continuous line is a power-law growth with exponent 0.7. b) Frequency-rank plot. The continuous and dashed line have slope -1.3 and -1.5, respectively. c) and d) Properties of the synthetic co-occurrence network obtained for $n_{RW} = 5 \cdot 10^4$, to be compared with the empirical data of Figs. 2.12 and 2.13.

4.4 Model robustness

The data shown in Figs. 4.3 and 4.4 correspond to a particular example of underlying network, a Watts-Strogatz network (Watts, 1999), taken as a cartoon for the semantic space: starting from a ring of N vertices in which each vertex is symmetrically connected to its $2m$ nearest neighbours (m vertices clockwise and counterclockwise), for every vertex, each edge connected to a clockwise neighbour is rewired with probability p, and preserved with probability $1-p$. The rewiring connects the edge's endpoint to a randomly chosen vertex, avoiding self-connections, and thus creating shortcuts between distant parts of the ring. For $1/N \ll p \ll 1$, a network with large number of triangles (due to the initial ring structure) and small diameter (thanks to the shortcuts) is obtained. The Watts-Strogatz model is a special example of network with large transitivity and at the same time small-world properties, i.e. short distances between nodes. However, we investigated also the dependence of the synthetic network properties on the structure of the semantic space and on the other parameters, such as n_{RW} or the distribution of the random walk lengths. Interestingly, we find an overall extremely robust behaviour for the diverse synthetic networks, showing that the proposed mechanism reproduces the empirical data without any need for strong hypothesis on the

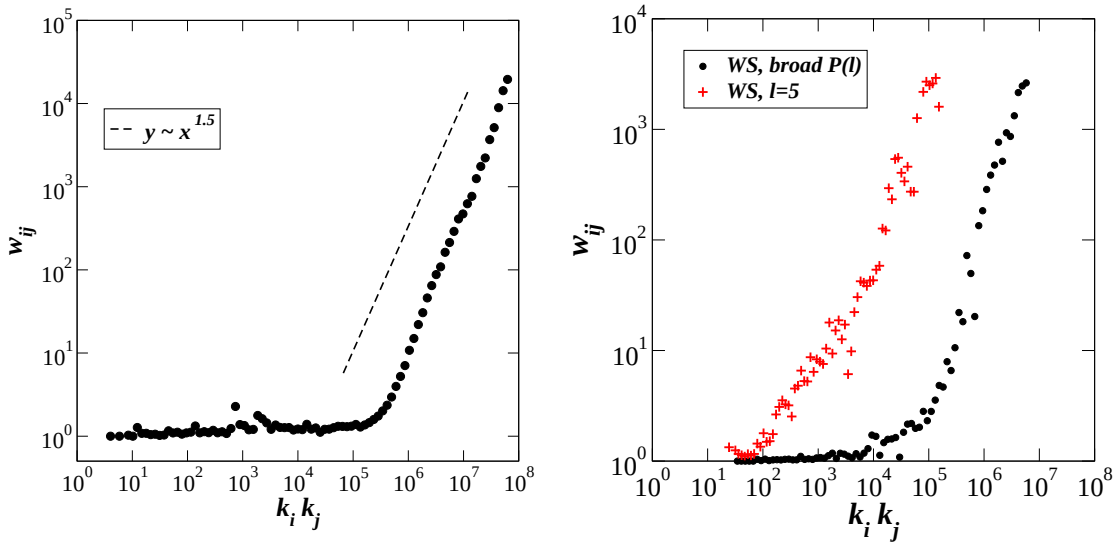

Figure 4.4: Correlations between the weights of the links in the co-occurrence networks and the degrees of the links' endpoints, as measured by plotting the weight w_{ij} of a link i, j versus the product of the degrees $k_i k_j$. Left: co-occurrence network of the tag "Folksonomy" of del.icio.us; each green dot corresponds to a link; the black circles represent the average over all links i, j with given product $k_i k_j$. Right: synthetic co-occurrence networks obtained from $n_{RW} = 5 \cdot 10^4$ random walks performed on a Watts-Strogatz network of 10^5 nodes. The black circles correspond to random walks of random lengths distributed according to $P(l) \sim l^{-3}$, and the red crosses to fixed length random walks ($l = 5$).

semantic space structure.

In order to check the robustness of the data analysis, we have considered several tags from the two analyzed datasets, del.icio.us and BibSonomy. For each dataset, the tag number corresponds to its popularity ranking, where the popularity is simply given by the number of posts in which it appears. For each tag, we perform the same analysis as in the main text, considering only the posts which contain this tag. We measure the growth of the vocabulary associated with this tag as a function of the number of posts containing it, and we always obtain sublinear power-law growths, as the one observed in real systems. We also build the co-occurrence network of each tag, and characterize these networks using the measures detailed above. All networks display the same properties observed in real systems, both at the topological level and for the weighted quantities.

4.5 Conclusions

Investigating the interplay of human and technological factors in user-driven systems is crucial to understand the evolution and the potential impact that these techno-social systems will have on our societies. Here we have shown that the main properties of information networks stemming from social annotations can be captured by regarding the process of social annotation as a collective exploration of a semantic space, modeled as a graph, by means of a series of random walks. These simple assumptions turn out to be remarkably robust, yielding network structures that are almost independent from the specific choice of the underlying graph and from the properties of the random walks. The mechanism introduced here captures sophisticated features that quantify the correlations among tags, as well as the evolution of the emerging data structures, matching experimental data from different systems. These results represent an important step towards a more comprehensive understanding and control of the technological and social aspects of user-

driven information networks.

Chapter 5

Control strategies

5.1 Roadmap leading from modeling activity to control strategies

Modeling activity has focused mainly on the study of tag streams and co-occurrence networks. This could shed light on the average user behavior and could possibly lead to improvement in system interfaces. As a first example, we report, in the following, the implementation in bibsonomy of a new scheme for suggestion and navigation of tags. This scheme is based on the systematic analysis and study of several similarity measures (Cattuto et al., 2008b,c; Markines et al., 2008), which led – in particular – to the identification of measures which spot in a realiable way "synonym" or "similar" tags (for more details, see the section on semantic grounding of D3.3).

On the other hand, the study of the co-occurrence network, shed light about its semantic structure and its dynamics. This will afford (as we have already demonstrated in a real-world application) enhanced navigation systems, more advanced forms of social classification of resources, as well as better tools to fight non-social behavior and spam.

5.2 Social similarity and semantic distance in folksonomies

An outcome of TAGora that has immediate impact on applications is represented by our research on social similarity and semantic distance in folksonomies. The definition of such notions, so far, has been rather *ad-hoc*, with no systematic classification or characterization of distance available to guide the choices of application developers.

In the context of TAGora, the similarity between nodes in a foksonomy (tags, resources, users) as well as different measures of semantic distance between tags have been explored in a series of recent works (Cattuto et al., 2008b,c; Markines et al., 2008). The main contributions of these works are:

- a methodology for evaluating measures of tag similarity by means of semantic grounding in formal representation of knowledge;
- a formal framework for the systematic definition of similarity measures based on different schemes of projection and aggregation;
- a first exploration of the tradeoffs between accuracy and computational scalability, as well as the first definition of a distance which is *both* accurate and scalable.

In an application context, the above results can be used to guide the choice of a measure of similarity (or relatedness) as a function of the task at hand. Specifically, we are now able to suggest folksonomy-based measures of tag similarity that are best suited for the following tasks:

- **folksonomy navigation.** A semantic characterization of the relations between tags allows to devise new aids to navigation: the user interface can expose tags that bear a *controlled* semantic relation to the tag the user is currently browsing. This allows the user to navigate the folksonomy in controlled "directions", for example moving from one tag to a more general one, or from one tag to "similar" tags (meant as synonyms, or functionally equivalent tags). Biases due to pure frequency/popularity effects can be equally controlled, allowing the discovery of low-frequency tags with strong semantic proximity to the original one. Section 5.3 below details a working implementation of similarity-based navigation aids in BibSonomy.

- **concept hierarchy.** both FolkRank and co-occurrence similarity yield tags that are more general than the original one. These measures provide valuable input for algorithms extracting taxonomic relationships between tags.

- **tag recommendation.** the applicability of both FolkRank and co-occurrence for tag recommendations was demonstrated in (Jäschke et al., 2007). Both measures allow for recommendations by straightforward modifications. Our evaluation showed that FolkRank delivers superior and more personalized results than co-occurrence.

- **query expansion.** Our analysis suggests that resource or tag context similarity could be used to discover synonyms and – together with some string edit distance – spelling variants of the tags in a user query. The original tag query can be expanded by using the tags obtained by these measures.

- **discovery of multi-word lexemes.** Depending on the allowed tag delimiters, it can happen that multi-word lexemes end up as several tags. We observed that FolkRank is best to discover these cases.

We believe that these first steps pave the way to making folksonomy-based measures less of a craft and more of a science, gaining the better control that a deeper understanding implies.

5.3 Implementation of Similar Tags in BibSonomy

In order to make available parts of the results of Refs. (Cattuto et al., 2008b,c) to BibSonomy users, we have implemented the cosine tag relatedness measure based on the vector space of the $10\,000$ most frequently used tags. For its computation, it is necessary to store a vector representation of each tag, expressing its co-occurrence profile with the top $10\,000$ tags. Between all pairs of these vectors, we compute the cosine similarity in order to find the most similar tags for each tag. Even in a medium-scale tagging system like BibSonomy with currently almost $100\,000$ distinct tags, performing these tasks efficiently is challenging. To summarize, we were concerned especially with the following aspects:

- *Efficient updating of the tag-tag co-occurrence information:* As we need co-occurrence counts for several purposes, BibSonomy already contains a built-in mechanism which maintains a system-wide tag-tag co-occurrence matrix. The core idea hereby to maintain efficiency is to keep track which co-occurrence counts might have changed (namely those of tags assigned to updated or newly added posts), and then to relocate the recomputation of these counts to a background process. As this mechanism was already available and fitted our needs, we chose to build our implementation on top of that.

- *Efficient computation of the cosine similarity:* Having the co-occurrence information readily avaible, we represent each tag as a sparse vector. In order to find the most similar tag for a given tag, we have to compute the cosine between its vector representation and all other tag

Figure 5.1: The BibSonomy web page for a tag ("ir", in this case) displays "similar" tags (as defined in Refs. (Cattuto et al., 2008b,c)) as a navigation aid.

vectors in the system. Even with the mentioned sparse representation, online computation proved to be computationally too expensive. This is why we decided once more to relocate this computation to a periodical background process. It builds the tag vectors, computes the pairwise similarity between them, and writes back the similarity information into the system. This process is currently scheduled once a day, and takes roughly 40 minutes to finish. In the running system, finding the most similar tag for a given tag is then reduced to a simple lookup in the similarity table.

For each tag, we display its 10 most similar tags (computed as described above) on its respective tag page, e.g. `http://www.bibsonomy.org/tag/java`. The similar tags are found in the lower right corner of the user interface, right beneath the related tags (Fig. 5.1). This way, we hope to enable the user different kinds of navigation through the available content.

5.4 Spam detection

The annotation of web sites in social bookmarking systems has become a popular way to manage and find information on the web. With the growing popularity of social bookmarking systems, spammers discovered this kind of service as a playground for their activities. A first reference to the spam detection prob- lem in folksonomies is given in (Cattuto et al., 2007c) (see also (Koutrika et al., 2007) and (Heymann et al., 2007)).

The community structure of such systems is one reason for their attractiveness for spammers: recent post pages, popular pages or specific tag pages can be manipulated easily. Usually, spammers pursue two goals: On the one hand, they place links in the system to attract people to advertising sites. On the other hand, they increase the PageRank of their sites by placing links in as many popular web 2.0 sites as possible, in order to increase their visibility in Google and other search engines.

Usual counter-measures like captchas are not efficient enough to effectively prevent the misuse of the system. One possibility to approach this problem is to classify users as spammers resp. non-spammers by machine learning algorithms.

At the moment, spam detection strategies has been introduced in bibsonomy, but they are limited to a classical machine learning approach, as described in the following section.

However, we have shown in (Cattuto et al., 2007c) how spam activity can be revealed as global statistical features emerging from the data. An other, recently investigated example has been shown in Fig. 2.7. We foresee that statistical observations and the knowledge and understanding coming from the modeling activity, would surely contribute to improve strategies of spam detection.

5.4.1 Spam Detection in BibSonomy

The classical approach in machine learning is to determine relevant features that describe the system's users, train different classifiers with the selected features and choose the one with the most promising evaluation results. We have transfered this approach to a social bookmarking setting to identify spammers. We have presented features considering the topological, semantic and profile-based information which people make public when using the system. The data-set used is a snapshot of the social bookmarking system BibSonomy and was built over the course of several months when cleaning the system from spam. Based on our features, we have learned a large set of different classification models and compare their performance. Our results represent the groundwork for a first application in BibSonomy and for the building of more elaborate spam detection mechanisms.

This work has been presented at the Fourth International Workshop on Adversarial Information Retrieval on the Web 2008 (Krause et al., 2008).

Spam Detection on BibSonomy is also a dissemination activity of the TAGora project. This year, we organize the Discovery Challenge of the ECML/PKDD conference, see `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`. The challenge presents two tasks in the new area of social bookmarking. One task covers spam detection and the other covers tag recommendations. As we are hosting the social bookmark and publication sharing system BibSonomy, we are able to provide a data-set of BibSonomy for the challenge. A training data-set for both tasks is provided at the beginning of the competition. The test data-set will be released 48 hours before the final deadline. The presentation of the results will take place at the ECML/PKDD workshop where the top teams are invited to present their approaches and results.

For the challenge, we have set up a mailing list rsdc08@cs.uni-kassel.de . We will use the list to distribute news about the challenge or other important information. Furthermore, the list can be used to clarify questions about the data-set and the different tasks. As the welcome message on the list contains information about how to access the data-set, subscribing to this list is essential to participate in the challenge.

In the last year, we were able to collect data of more than 2,000 active users and more than 25,000 spammers by manually labeling spammers and non-spammers. The provided data-set consists of these users and of all their posts. This includes all public information such as the URL, the description and all tags of the post. The goal of the challenge is to learn a model which predicts whether a user is a spammer or not. In order to detect spammers as early as possible, the model should make good predictions for a user when he submits his first post.

All participants can use the training data-set to fit the model. The training data-set contains flags that identify users as spammers or non-spammers. The test data-set will have the same format as the training data-set and can be downloaded two days before the end of the competition. All participants must send one file containing one line, for each user, composed by the user number and a confidence value separated by a tab. The higher the confidence value, the higher the probability that the user is a spammer. If no prediction is provided we assume the user is not a spammer. The evaluation criterion for this challenge is the AUC (the Area Under the ROC Curve) value. We compare the submitted spammer predictions of the participants with the manually assigned labels on a user basis.

Bibliography

A. Baldassarri, C. Cattuto, K. Dellschaft, V. Loreto, V. Servedio, and G. Stumme. Theoretical tools for modeling and analyzing collaborative social tagging systems – a stream view. *Deliverable D4.1, TAGora Project*, 2007. URL http://www.tagora-project.eu.

A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752, 2004. doi: 10.1073/pnas.0400087101. URL http://www.pnas.org/cgi/content/abstract/101/11/3747.

M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71:027103, 2005.

C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. 2007a. URL http://arxiv.org/abs/0704.3316.

C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto. The collective dynamics of social annotation. 2008a.

Ciro Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46(0):33–37, 2006. URL http://dx.doi.org/10.1140/epjcd/s2006-03-004-4.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007b. URL http://www.pnas.org/cgi/content/short/104/5/1461.

Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, , Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on "Network Analysis in Natural Sciences and Engineering"*, 20(4):245–262, 2007c. ISSN 0921-7126. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2007/cattuto2007network.pdf.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (ECAI2008)*, 7 2008b. URL http://arxiv.org/abs/0805.2045. http://arxiv.org/abs/0805.2045.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems, 2008c.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. 2007. URL http://arxiv.org/abs/0706.1062.

Klaas Dellschaft and Steffen Staab. An epistemic dynamic model for tagging systems. In *HYPERTEXT 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, 2008.

F. Eggenberger and G. Polya. Uber die statistik verketteter vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, 1:279–289, 1923.

Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006. doi: 10.1177/0165551506062337. URL http://arxiv.org/abs/cs.DL/0508082.

H. Halpin, V. Robu, and H. Shepard. The complex dynamics of collaborative tagging. In *In Proceedings of 16th World Wide Web Conference WWW-2007*, 2007.

H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978. ISBN 0123357500.

P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford InfoLab, 2006.

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007. ISSN 1089-7801. doi: http://dx.doi.org/10.1109/MIC.2007.125. URL http://portal.acm.org/citation.cfm?id=1304062.1304547&coll=GUIDE&dl=.

Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514. Springer, 2007. ISBN 978-3-540-74975-2.

Georgia Koutrika, Frans Adjie Effendi, Zoltán Gyöngyi, Paul Heymann, and Hector Garcia-Molina. Combating spam in tagging systems. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 57–64, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-732-2. doi: http://doi.acm.org/10.1145/1244408.1244420. URL http://portal.acm.org/citation.cfm?id=1244408.1244420#.

Beate Krause, Andreas Hotho, and Gerd Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008. URL http://airweb.cse.lehigh.edu/2008/submissions/krause_2008_anti_social_tagger.pdf.

B. Mandelbrot. An infromational theory of the statistical structure of language. *Communication theory*, 486, 1953.

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Social similarity, 2008.

Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

Marcelo A. Montemurro and D. Zanette. Frequency-rank distribution of words in large text samples: phenomenology and models. *Glottometrics*, 4:87–99, 2002.

M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0205405.

R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.*, 87:258701, 2001.

Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, editors, *Data Science and Classification. Proceedings of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Heidelberg, July 2006. Springer. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/schmitz2006mining.pdf.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425, 1955.

Ricard V. Solé, Bernat Corominas, Sergi Valverde, and Luc Steels. Language networks: their structure, function and evolution. *Trends in Cognitive Sciences*, 2008. URL http://www.isrl.uiuc.edu/~amag/langev/paper/sole05languageNetworks.html.

D. J. Watts. *Small-worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, NJ (USA), 1999. URL http://www.amazon.com/Small-Worlds-Duncan-J-Watts/dp/0691005419.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA (USA), 1949.