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)

D3.1 Extracting Emergent Metadata Statistics and Network Metrics in Social Tagging Systems

---

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

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for this deliverable: University of Koblenz-Landau

Executive Summary

The objective of this deliverable is the set up of several protocols of data analysis to be performed on the raw datasets delivered by WP1, particularly focusing on the graph point of view. A data analysis protocol is defined by:

(a) indicating suitable quantitative measures on the raw data sets;

(b) acquiring software tools for performing the measures;

(c) extracting relevant statistical information characterizing the analyzed datasets.

In this deliverable, we will especially concentrate on analyzing the properties of the graph used to represent folksonomies. Informations about how emergent metadata statistics in folksonomies may be analyzed is available in D4.1 (TAGora, 2007) while the analysis of clustering social properties in folksonomies will be covered in detail by D3.2 at a later stage of the project.

We will start with developing a common notion of folksonomies whose main constituents are the users, tags and resources. Together with the tag assignments of a user to a resource they form a three-mode network or hypergraph. Because most of the already available measures are defined for one-mode and two-mode networks we also provide suitable projections for transforming the folksonomy hypergraph to such networks.

Further, we will analyze the emergence of common semantics by exploring trends in the folksonomy. We present a technique for analyzing the evolution of topic-specific trends. The approach is based on our *FolkRank* algorithm, a differential adaptation of the PageRank algorithm to the tri-partite hypergraph structure of a folksonomy.

After that, we will identify relevant measures for analyzing the topology of the network or graph representation of folksonomies, i.e. its structural properties. The analysis of topological properties is especially known from the area of social network analysis (SNA). Typical examples of SNA measures that are to be described in this deliverable are the clustering coefficient and the characteristic path length in a graph. The original measures were designed in the case of one-mode graphs, i.e. graphs with only one kind of nodes. We are proposing adapted measures that can be directly applied to the three-mode graph of folksonomies.

We use further relevant measures borrowed from complex network theory, like the cumulative strength distribution and the the average nearest neighbor strength, to analyze the topological properties of the folksonomy hypergraph and the tag co-occurrence network, which we get by projecting the hypergraph to a one-mode network. We will show that folksonomy graphs exhibit small-world properties and that the structure of the tag co-occurrence graph is not only influenced by the underlying tag distributions but that it also reflects the semantics of the tags.

The small-world property is well-known from social network analysis and it means that one can reach from every node in a graph with very few links any other node in the graph. The small-world property of folksonomies directly influences the browsing experience of a user because it means that the user can start any node (e.g. a specific resource) and reach each other resource in the system with very few clicks by simply browsing e.g. the tags attached to the resource.

With the help of the analysis of the tag co-occurrence graph we were successful in spotting and detecting spamming activities by some (pseudo)-users in systems like `del.icio.us` and `Bibsonomy`. The spam detection is thus a direct benefit for the users of tagging systems, which arises from the measures described in this deliverable.

We analyzed the structure of the semantic space defined by a given set of resources and discovered the existence of well defined communities of resources corresponding to semantically separated tag clouds. On the other hand these communities also correspond to communities of users. This points in the direction of an effective cooperation among the users to efficiently map the semantic space with meaningful and, whenever possibly, not overlapping set of tags. A preliminary study is also presented, where we aim at bridging the FolkRank procedure with community detection algorithms in networks.

The semantics contained in the tag co-occurrence graph of folksonomies can also be exploited in a second way by the T-ORG system that helps a user to automatically assign tags and resources to several predefined categories like *location* or *person*. By categorizing tags and organizing them into hierarchies it is possible to improve the browsing experience of the users.

In this deliverable we were not only successful in identifying relevant features of the systems under study, but also in exploiting a set of measures that might provide future improvements in the design of large tagging systems.

Contents

List of Figures

Chapter 1

Introduction

TAGora is a project sponsored by the Future and Emerging Technologies program of the European Community (IST-034721) focusing on the semiotic dynamics of online social communities. A new paradigm is quickly gaining impact in large-scale information systems: Folksonomies. In applications like Flickr, Connotea, Citeulike, Del.icio.us, etc. people no longer make passive use of online resources - they take on an active role and enrich resources with semantically meaningful information. Such information consists of terminology (or "tags") freely associated by each user to resources and is shared with users of the online community. Despite its intrinsic anarchist nature, the dynamics of this terminology system spontaneously leads to patterns of terminology common to the whole community or to subgroups of it. Surprisingly, this emergent and evolving semiotic system provides a very efficient navigation system through a large, complex and heterogeneous sea of information.

Our project proposes visionary and high risk research aimed at giving a scientific foundation to these developments, so contributing to the growth of the new field of Semiotic Dynamics. Semiotic Dynamics studies how semiotic relations can originate, spread, and evolve over time in populations, by combining recent advances in linguistics and cognitive science with methodological and theoretical tools of complex systems and computer science.

The project aims at exploiting the unique opportunity offered by the availability of enormous amount of data. This goal will be achieved through: (a) a systematic and rigorous gathering of data that will be made publicly available to the consortium and to the scientific community; (b) designing and implementing innovative tools and procedures for data analysis and mining; (c) constructing suitable modeling schemes which will be implemented in extensive numerical simulations. We aim in this way at providing a virtuous feedback between data collection, analysis, modeling, simulations and (whenever possible) theoretical constructions, with the final goal to understand and engineer the Semiotic Dynamics of on line social systems.

1.1 The challenges

To successfully navigate one's way in a sea of information, one needs a system of fixed points, a coordinate system and maps. In most current day large-scale information systems (e.g. enterprise-resource planning systems, knowledge management systems) the coordinate system is given by some kind of ontology, the fixed points are given by annotations that describe documents or other data, while the map is produced by connecting ontology and annotations.

Unfortunately, this system becomes less and less workable as it concerns ever larger-scale information systems. On the one hand, the top down architecture of the system does not respond flexibly enough to the needs of the users who plow the sea of information. On the other hand, anarchical approaches mostly do not work either, because without organization of ontologies and annotations one ends up with plenty of information systems that live completely independently

from each other and cannot be joined for useful exploitation. For instance the distributed creation of taxonomies and tags and the multiplicity of conceptual schemata generate the well known problem of semantic interoperability. One solution is to standardize. The different users of a collective information system could all agree a priori to use the same taxonomies to structure their data and to use the same conceptual schemata for their data and meta-data. The tags in the owner taxonomies can then act as a shared communication protocol between peers. Unfortunately such a standardization approach is unlikely to work for truly open-ended collective information systems in rapidly changing domains like music file sharing, picture exchange, medical imaging, scientific papers, etc.

1.2 The opportunity

This project aims to contribute to the economic development of the Community by advancing the state of the art in Complexity Science as relevant to IT and by exploring highly novel IT applications based on Science of Complexity. In particular, the project takes at its starting point that there has been a massive increase in the amount of autonomy and information flow and in the number of people that are participating in IT processes. Moreover there is a pressing need that information systems become ever more adaptive to user needs and rapidly expanding infrastructures. Consequently, there is a much higher interdependence of actors than in the past and various properties observed in complex systems (such as self-organized criticality) are now also observed in information systems. In particular, the focus of this project bears direct relevance to fundamental and applicative problems that are currently attracting the attention of researchers from different fields, ranging from multi-agent systems to knowledge-management, from the semantic web to peer-to-peer content distribution. This is possible because of two general traits explicitly addressed and investigated by this project:

1.2.1 Expected impact

The project includes partners who have a proven record of excellence in developing and publishing research results in their respective scientific areas. One of the primary objectives of this project is to bring together two communities: researchers in various domains of complex systems and researchers in various areas of Information Technology (computer science, web technologies, ubiquitous computing), in particular those facing the challenge of Semiotic Dynamics in on line social communities. The impact that we seek on the scientific community is potentially enormous because it goes beyond the specific scientific objectives and technologies focused on in this project. We want to foster a general movement towards the interrelation of complex systems and Information Technology by showing successful examples of cooperation and by posing and solving concrete non-trivial problems. It is only by seeing clear examples that we expect the scientific research community to follow suit. These clear successful research examples will be documented and presented through regular scientific communication channels so that they are accessible to the community at large.

1.3 Structure of this Deliverable

In this deliverable we will set up of several protocols for the analysis of the raw datasets delivered by WP1. A data analysis protocol is defined by (i) indicating suitable quantitative measures on the raw data sets, (ii) acquiring software tools for performing the measures and (iii) extracting relevant statistical information characterizing the analyzed datasets. In this deliverable, we will especially concentrate on analyzing the topological properties of the graph used for representing folk-

sonomies. Information about how emergent metadata statistics can be analyzed for folksonomies is available in (TAGora, 2007). The analysis of clustering social properties in folksonomies will be covered in detail by D3.2 at a later stage of the project.

In chapter 2 we will start with developing a common notion of folksonomies whose main constituents are the users, tags and resources. Together with the tag assignments of a user to a resource they form a three-mode network or hypergraph. Because most of the already available measures are defined for one-mode and two-mode networks we also provide suitable projections for transforming the folksonomy hypergraph to such networks.

In chapter 3, we will identify relevant measures for analyzing the topology of the network or graph representation of folksonomies, i.e. its structural properties. The analysis of topological properties is especially known from the areas of complex networks (Albert and Barabási, 2001; Borner et al., 2007; Dorogovtsev and Mendes, 2003; Newman, 2003; Pastor-Satorras and Vespignani, 2004) and social network analysis (SNA). Typical examples of such measures are the clustering coefficient and the characteristic path length in a graph, which will be described in this deliverable. We will furthermore describe relevant measures borrowed from complex network theory, like the cumulative strength distribution and the the average nearest neighbor strength. The measures will be used for showing that folksonomy hypergraphs exhibit small-world properties. As a side effect, the statistical measures identified atypical spurious contributions, as the activity of spam users. In such a case, this allows to easily filter away the corresponding undesired information.

More interestingly, the structure of the tag co-occurrence graph is reflects the semantics of tags. In fact, trivial frequency effects, related to the linguistic nature of tags, are responsible only partially for the scale-free character of the graph. More subtle properties, as is assortative nature, are strongly semantics dependent.

Semantics surely represent the most interesting information hidden in the metadata contributed by the users. The chance to extract a commonly shared tag semantics is one of the most appealing possibility offered by folksonomy systems. As a first step in this direction, we analyzed the structure of the semantic space defined by a given set of resources and found the existence of well defined communities of resources corresponding to semantically separated tag clouds. We do this applying recent spectral techniques for cluster identification.

In chapter 4, we will analyze the emergence of common semantics from a different point of view, i.e. by exploring trends in the folksonomy. We present a technique for analyzing the evolution of topic-specific trends. The approach is based on our *FolkRank* algorithm (Hotho et al., 2006a), a differential adaptation of the PageRank algorithm (Brin and Page, 1998) to the tri-partite hypergraph structure of a folksonomy.

The two approaches, trend detection and the cluster identification, are not necessarily so different. In a preliminary study, we show that Folkrank and a known cluster detection algorithm (MCL) give comparable results.

Always in the direction of bridging the gap between folksonomies and semantic web, we will describe in chapter 5 a system called T-ORG. T-ORG automatically classifies the resources of a tagging system into predefined categories and thus can provide a better browsing experience to the user. To this aim, the T-ORG system classifies the tags into categories using its pattern library, categories extracted from a given ontology and Google search results.

In appendix A, we will provide a list of tools that can be used to perform the above described measures and that can help to exploit some of the features underlying the tagging systems.

1.4 Dissemination of the Results

Parts of this deliverable have been published as follows:

- Chapter 3 is partially based on (Schmitz et al., 2007), while Section 3.4 is part of a work submitted to ECCS 2007, *European Conference on Complex Systems* — Dresden October 1-5.
- Chapter 2 is partially based on (Hotho et al., 2006a) and (Schmitz et al., 2006). Section 4.4 was presented as a talk at the DPG Conference, *Section Dynamics and Statistical Physics*, *Ranking and Community detection in unweighted networks*, Spring Meeting of the German Physical Society (DPG), Regensburg (Germany), March 26-30th, 2007.
- Chapter 4 is partially based on (Hotho et al., 2006b). The work has been done in collaboration with the EU FP6 IP 'Nepomuk – Social Semantic Desktop'.
- Chapter 5 is partially based on (Abbasi et al., 2007).

Chapter 2

Folksonomy structure

Social resource sharing systems are web-based systems that allow users to upload their resources, and to label them with arbitrary words, so-called *tags*. The systems can be distinguished according to what kind of resources are supported. Flickr, for instance, allows the sharing of photos, del.icio.us the sharing of bookmarks, CiteULike[1] and Connotea[2] the sharing of bibliographic references, and 43Things[3] even the sharing of goals in private life. The *BibSonomy*,[4] system, which is further developed in the context of TAGora, allows to share simultaneously bookmarks and BibTex entries (see Fig. 2.1).

The activity of users interacting with a collaborative tagging system consists of either navigating the existing body of resources by using tags, or of adding new resources. In order to add a new resource into the system the user is prompted for a reference to the resource and a set of tags to associate with it. Thus, the basic unit of information in a collaborative tagging system is a `({tags}, user, resource)` triple, referred to as *post*. Further, each post can be split in multiple tag assignments (TAS), according on the number of tags in it. A post typically contains a *temporal marker* recording the (physical) time of the tagging event, so that temporal ordering can be preserved in storing and retrieving posts.

The collection of all tag assignments of a user are called his *personomy*, the collection of all personomies constitutes the *folksonomy*. The user can explore his personomy, as well as the personomies of the other users, in all dimensions: for a given user one can see all resources he had uploaded, together with the tags he had assigned to them (see Fig. 2.1); when clicking on a resource one sees which other users have uploaded this resource and how they tagged it; and when clicking on a tag one sees who assigned it to which resources.

2.1 Folksonomies – A Definition

A folksonomy describes the users, resources, and tags, and the user-based assignment of tags to resources. We present here a formal definition of folksonomies, which is also underlying the BibSonomy system. In Fig. 2.2 one can see the hypergraph representation of a folksonomy as it is defined in Def. 1.

Definition 1 *A* folksonomy *is a tuple* $\mathbb{F} := (U, T, R, Y, \prec)$ *where*

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

[1] `http://www.citeulike.org/`
[2] `http://www.connotea.org/`
[3] `http://www.43things.com/`
[4] `http://www.bibsonomy.org`

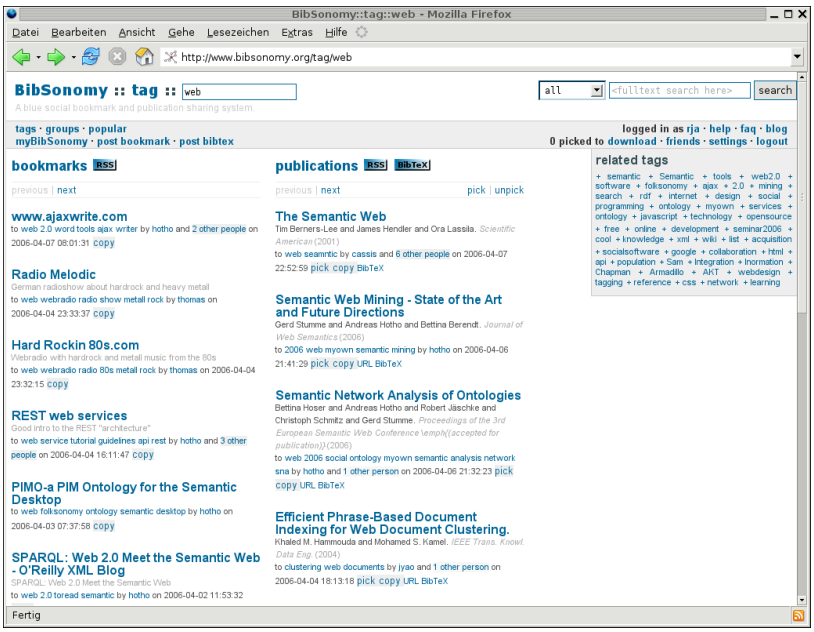

Figure 2.1: Bibsonomy displays bookmarks and BibTeX based bibliographic references simultaneously.

- $\prec$ *is a user-specific subtag/supertag-relation, i. e.,* $\prec \subseteq U \times T \times T$*, called* subtag/supertag relation.

The personomy $\mathbb{P}_u$ *of a given user* $u \in U$ *is the restriction of* $\mathbb{F}$ *to* u*, i. e.,* $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$ *with* $I_u := \{(t,r) \in T \times R \mid (u,t,r) \in Y\}$*,* $T_u := \pi_1(I_u)$*,* $R_u := \pi_2(I_u)$*, and* $\prec_u := \{(t_1,t_2) \in T \times T \mid (u,t_1,t_2) \in \prec\}$*, where* π_i *denotes the projection on the* i*th dimension.*

Users are typically described by their user ID, and tags may be arbitrary strings. What is considered as a resource depends on the type of system. For instance, in del.icio.us, the resources are URLs, and in flickr, the resources are pictures. From an implementation point of view, resources are internally represented by some ID.

In this paper, we do not make use of the subtag/supertag relation for sake of simplicity. I. e., $\prec = \emptyset$, and we will simply note a folksonomy as a quadruple $\mathbb{F} := (U, T, R, Y)$. This structure is known in Formal Concept Analysis (Ganter and Wille, 1999; Wille, 1982) as a *triadic context* (Lehmann and Wille, 1995; Stumme, 2005). An equivalent view on folksonomy data is that of a tripartite (undirected) hypergraph $G = (V, E)$, where $V = U \dot{\cup} T \dot{\cup} R$ is the set of nodes, and $E = \{\{u,t,r\} \mid (u,t,r) \in Y\}$ is the set of hyperedges.

Certain computations are not applicable on hypergraphs. A transformation into a simple undirected graph $G^\nabla = (V, E^\nabla)$ is realized by splitting up each hyperedge into three weighted edges connecting the respective user, tag, and resource such that the weight of an edge is the number of original hyperedges between the connected nodes. $E^\nabla = E^\nabla_{UT} \cup E^\nabla_{UR} \cup E^\nabla_{TR}$ with $E^\nabla_{UT} = \{((u,t), |E^*|) \mid E^* = \{(u,t,r) \in Y\},\ r \in R\}$ and E^∇_{UR}, E^∇_{TR} analogously. The nodes remain unchanged.

2.2 Projections

As most algorithms for Social Network Analysis do not allow for three-mode data, we will use several projections to two- and one-mode data. Several such projections have already been introduced in (Lehmann and Wille, 1995). In (Stumme, 2005), a more complete approach has been

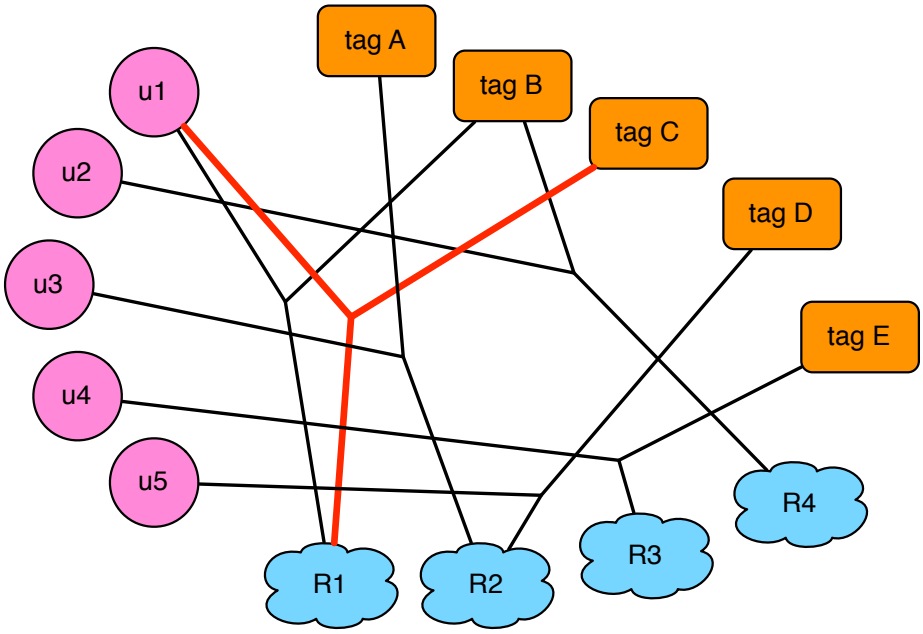

Figure 2.2: Hypergraph representation of a folksonomy.

provided, which we here adapt slightly to the association rule mining scenario.

2.2.1 Two-Mode Projections

As we want to analyze all facets of the folksonomy, we want to allow to use any (combination) of the three sets U, T, and R as the set of objects – on which the support is computed – at some point in time, depending on the task on hand. Therefore, we will not fix the roles of the three sets in advance. Instead, we consider a triadic context as symmetric structure, where all three sets are of equal importance. For easier handling, we therefore denote the folksonomy $\mathbb{F} := (U, T, R, Y)$ alternatively by $\mathbb{F} := (X_1, X_2, X_3, Y)$ in the following.

We determine the set of objects – i. e., the set on which the support will be counted – by a permutation σ on the set $\{1, 2, 3\}$. The choice of a permutation indicates, together with one of the aggregation modes '$\mathfrak{D}$', '$\mathfrak{W}$', '$\exists n$' with $n \in \mathbb{N}$, and '$\forall$', on which formal context $\mathbb{K} := (G, M, I)$ the association rules are computed.

- $\mathbb{K}^{\sigma,\mathfrak{D}} := (X_{\sigma(1)} \times X_{\sigma(3)}, X_{\sigma(2)}, I)$ with $((x_{\sigma(1)}, x_{\sigma(3)}), x_{\sigma(2)}) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.
- $\mathbb{K}^{\sigma,\mathfrak{W}} := (X_{\sigma(1)}, X_{\sigma(2)} \times X_{\sigma(3)}, I)$ with $(x_{\sigma(1)}, (x_{\sigma(2)}, x_{\sigma(3)})) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.
- $\mathbb{K}^{\sigma,\exists n} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if there exist n different $x_{\sigma(3)} \in X_{\sigma(3)}$ with $(x_1, x_2, x_3) \in Y$.
- $\mathbb{K}^{\sigma,\forall} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if for all $x_{\sigma(3)} \in X_{\sigma(3)}$ holds $(x_1, x_2, x_3) \in Y$. This mode is equivalent to '$\exists n$' with $n = |X_{\sigma(3)}|$.

These projections are complemented by the following way to 'cut slices' out of the folksonomy. A slice is obtained by selecting one dimension (out of user/tag/resource), and then fixing in this dimension one particular instance.

- Let $x := x_{\sigma(3)} \in X_{\sigma(3)}$. $\mathbb{K}^{\sigma,x} := (X_{\sigma(1)}, X_{\sigma(2)}, I)$ with $(x_{\sigma(1)}, x_{\sigma(2)}) \in I$ if and only if $(x_1, x_2, x_3) \in Y$.

Concluding, the two-mode projection is defined by the following two selections:

- a permutation $\sigma \in S_3$, and
- the choice of one of the aggregation modes $\mathfrak{H}$, $\mathfrak{W}$, $\exists n$ with $n \in \mathbb{N}$, $\forall$, or $x \in U \dot{\cup} T \dot{\cup} R$.

2.2.2 One-Mode Projections

Many interesting features can also be derived from one-mode projections, also called *co-occurrence networks*. We consider two types of projections:

$$CO_{T/UR}(t_i, t_j) := \{(u, r) \in U \times R \mid (u, t_i, r) \in Y \wedge (u, t_j, r) \in Y\}, \tag{2.1}$$

The projections $CO_{U/TR}$ and $CO_{R/UT}$ are defined analogously.

$$CO_{T/R}(t_i, t_j) := \{r \in R \mid (u, t_i, r) \in Y \wedge (u', t_j, r) \in Y \wedge u, u' \in Y\}, \tag{2.2}$$

The projections $CO_{T/U}$, $CO_{R/U}$, $CO_{R/T}$, $CO_{U/T}$ and $CO_{U/R}$ are defined analogously.

$$ES_U(t) = \{(u, \omega) \mid ((u, t), \omega) \in E^{\nabla}\}$$

Chapter 3

Network Analysis

In this chapter, we deal with the topological analysis of the network or graph representation of folksonomies, i.e. its structural properties. The analysis of topological properties is especially known from the area of social network analysis (SNA). Typical examples of SNA measures are the different centrality measures (e.g. degree and betweenness centrality), the clustering coefficient or the characteristic path length in a graph. A good introduction into SNA and its relevant measures is available in e.g. (Wasserman and Faust, 1995).

Because of the availability of good introductory literature we do not want to give a complete overview of all the available SNA measures. Instead, we will concentrate on the most relevant measures which help us to explain e.g. why folksonomies can still be efficiently browsed even if the system grows to millions of nodes. This phenomenon is related to the small-world property of graphs. In section 3.2 we will propose adapted characteristic path length and clustering coefficient measures which can directly be applied on three-mode graphs. They will be used for showing that folksonomy graphs indeed exhibit small-world properties.

Furthermore, in section 3.3 we will adapt and apply measures known from complex network theory. They will be used for examining tag co-occurrence graphs which one gets by projecting the three-mode folksonomy graph into a one-mode graph (cf. section 2.2.2). With the help of the cumulative strength distribution and the average nearest neighbor strength we will discuss in how far the structure of tag co-occurrence graphs can be simply explained by the underlying tag distribution or whether it is also influenced by the semantics of the corresponding tags.

3.1 Datasets

In the experimental part of this chapter we will use the following two datasets.

3.1.1 del.icio.us Dataset

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

We obtained a folksonomy with $|U| = 75,242$ users, $|T| = 533,191$ tags and $|R| = 3,158,297$ resources, related by in total $|Y| = 17,362,212$ tag assignments (TAS). In addition, we gener-

ated monthly dumps from the timestamps associated with posts, so that 14 snapshots in monthly intervals from June 15th, 2004 through July 15th, 2005 are available.

3.1.2 BibSonomy Dataset

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

3.2 Small Worlds in Three-Mode-Networks

The notion of a *small world* has been introduced in a seminal paper by Milgram (Milgram, 1967). Milgram tried to verify in a practical experiment that, with a high probability, any two given persons within the United States would be connected through a relatively short chain of mutual acquaintances. Recently, the term "small world" has been defined more precisely as a network having a small characteristic path length comparable to that of a (regular or Erdős) random graph, while at the same time exhibiting a large degree of clustering (Watts, 1999) (which a random graph does not). These networks show some interesting properties: while nodes are typically located in densely-knit clusters, there are still long-range connections to other parts of the network, so that information can spread quickly. At the same time, the networks are robust against random node failures. Since the coining of the term "small world", many networks, including social and biological as well as man-made, engineered ones, have been shown to exhibit small-world properties.

In this section, we will define the notions of characteristic path length and clustering coefficient in tripartite hypergraphs such as folksonomies, and apply these to the data sets introduced in Section 3.1 in order to demonstrate that these graphs do indeed exhibit small world properties.

3.2.1 Characteristic Path Length

The *characteristic path length* of a graph (Watts, 1999) describes the average length of a shortest path between two random nodes in the graph. If the characteristic path length is small, few hops will be necessary, on average, to get from a particular node in the graph to any other node.

As folksonomies are triadic structures of *(tag, user, resource)* assignments, the user interface of such a folksonomy system will typically allow the user to jump from a given tag to (a) any resource associated with that tag, or (b) any user who uses that tag, and conversely for users and resources. Thus, the effort of getting from one node in the folksonomy to another can be measured by counting the *hyper*edges in shortest paths between the two.

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

Following Watts (Watts, 1999), we define $\bar{d}_v$ as the mean of $d(v, u)$ over all $u \in (T \cup U \cup R) - \{v\}$, and call the median of the $\bar{d}_v$ over all $v \in T \cup U \cup R$ the *characteristic path length* L of the folksonomy.

In Section 3.2.3, we will analyze the characteristic path length on our datasets.

3.2.2 Clustering Coefficients

Clustering or transitivity in a network means that two neighbors of a given node are likely to be directly connected as well, thus indicating that the network is locally dense around each node.

To measure the amount of clustering around a given node v, Watts (Watts, 1999) has defined a clustering coefficient γ_v (for normal, non-hyper-graphs). The clustering coefficient of a graph is γ_v averaged over all nodes v.
Watts (Watts, 1999, p. 33) defines the clustering coefficient γ_v as follows ($\Gamma_v = \Gamma(v)$ denotes the neighborhood of v):

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

Note that Watts combines two aspects which are *not* equivalent in the case of three-mode data. The first one is: how many of the possible edges around a node do actually occur, i. e. does the neighborhood of the given vertex approach a clique? On the other hand, the second aspect is that of neighbors of a given node being connected themselves.
Following the two motivations of Watts, we thus define two different clustering coefficients for three-mode data:

Cliquishness: From this point of view, the clustering coefficient of a node is high iff many of the possible edges in its neighborhood are present.

More formally: Consider a resource r. Then the following tags T_r and users U_r are connected to r: $T_r = \{t \in T \mid \exists u : (t,u,r) \in Y\}$, $U_r = \{u \in U \mid \exists t : (t,u,r) \in Y\}$. Furthermore, let $tu_r := \{(t,u) \in T \times U \mid (t,u,r) \in Y\}$ the (tag, user) pairs occurring with r.

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

$$\gamma_{cl}(r) = \frac{|tu_r|}{|T_r| \cdot |U_r|} \tag{3.1}$$

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

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

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

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

$$\gamma_{co}(r) := \frac{|\widetilde{tu_r}|}{|tu_r|} \tag{3.2}$$

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

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

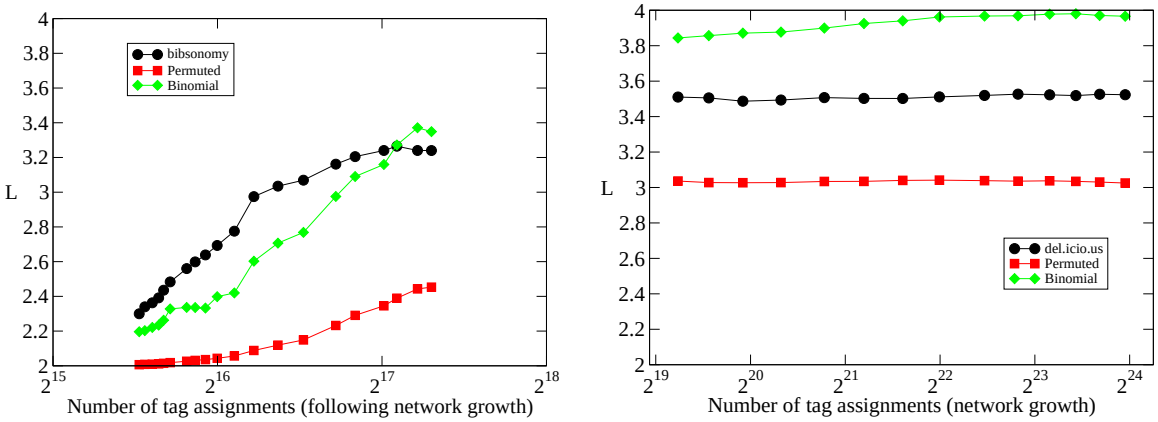

Figure 3.1: Characteristic path length for the `bibsonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graph have been obtained as a function of the number of nodes of the networks (not shown). Note how the characteristic path length takes quite similar low values, typical of small world networks, for all graphs.

It was shown in (Schmitz et al., 2007) that the clustering coefficients γ_{cl} and γ_{co} do indeed capture different characteristics of the graph and are not intrinsically related.

3.2.3 Experiments

Setup

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

Binomial: These graphs are generated similar to an Erdős random graph $G(n, M)$ (Bollobas, 2001). T, U, R are taken from the observed folksonomies. $|Y|$ many hyperedges are then created by picking the three endpoints of each edge from uniform distributions over T, U, and R, resp.

Permuted: These graphs are created by using T, U, R from the observed folksonomy. The tagging relation Y is created by taking the TAS from the original graph and permuting each dimension of Y independently (using a Knuth Shuffle (Knuth, 1981)), thus creating a random graph with the same degree sequence as the observed folksonomy.

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

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

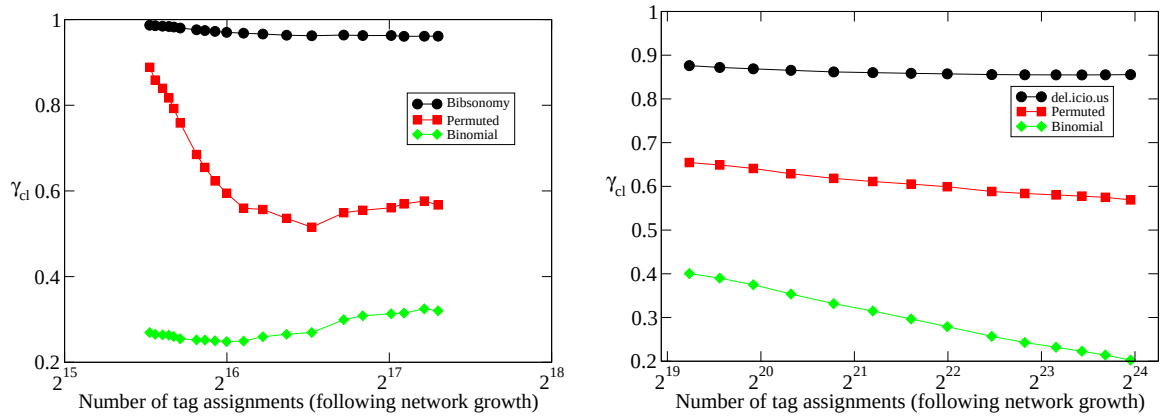

Figure 3.2: Cliquishness of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). The cliquishness for the folksonomy networks takes quite high values, higher than the corresponding random graph (permuted and binomial).

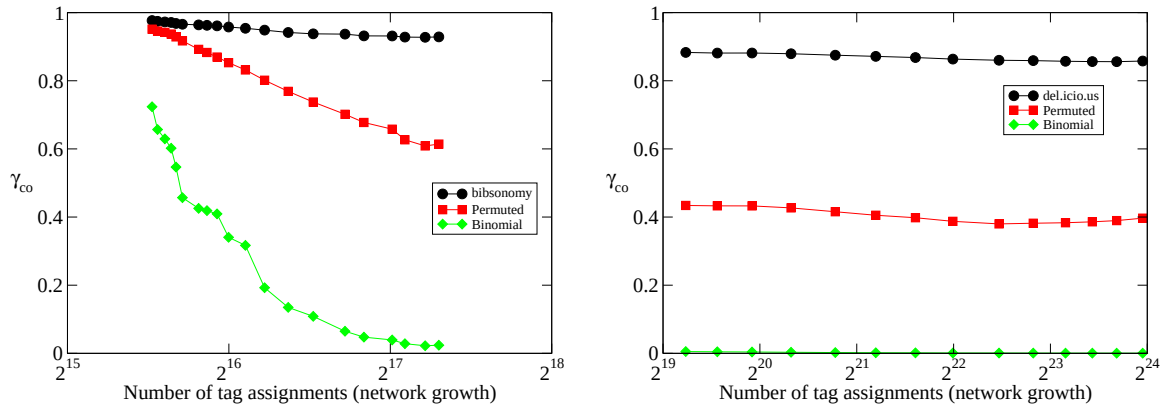

Figure 3.3: Connectedness/Transitivity of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). As in the case of cliquishness, the values of connectedness/transitivity are very high for the folksonomy networks, at odds with the corresponding random graphs (permuted and binomial).

Observations

Figures 3.1– 3.3 show the results for the clustering coefficients and the characteristic path lengths for both datasets, plotted against the number $|Y|$ of tag assignments for the respective monthly snapshots.

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

Del.icio.us In the del.icio.us dataset (Figures 3.2 and 3.3, right hand sides), it can be seen that both clustering coefficients are extremely high at about 0.86, much higher than those for the permuted and binomial random graphs. This could be an indication of coherence in the tagging behavior: if, for example, a given set of tags is attached to a certain kind of resources, users do so consistently.

On the other hand, the characteristic path lengths (Figure 3.1, right) are considerably smaller than for the random binomial graphs, though not as small as for the permuted setting. Interestingly, the path length has remained almost constant at about 3.5 while the number of nodes has grown about twenty-fold in the observation period.

As explained in Section 3.2.1, in practice this means that on average, every user, tag, or resource within del.ici.us can be reached within 3.5 mouse clicks from any given del.icio.us page. This might help to explain why the concept of serendipitous discovery (Mathes, 2004a) of contents plays such a large role in the folksonomy community – even if the folksonomy grows to millions of nodes, everything in it is still reachable within few hyperlinks.

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

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

Figures 3.2 and 3.3 (left) show that clustering is extremely high at $\gamma_{cl} \approx 0.96$ and $\gamma_{co} \approx 0.93$ – even more so than in the del.icio.us data.

At the same time, Figure 3.1 shows that the characteristic path lengths are somewhat larger, but at least comparable to those of the binomial graph.

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

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

3.3 Networks of Tag Co-Occurrence

In order to investigate the emergent semantic properties of the folksonomy, we focus on the relations of co-occurrence among tags. Since the process of tagging is inclusive (Golder and Huberman, 2005), and large overlap often exists among resources marked with different tags, the relations of co-occurrence among tags expose the semantic aspects underling collaborative tagging, such as homonymy, synonymy, hierarchical relations among tags and so on.

The simplest way to study tag co-occurrence at the global level is to define a network of tags,

where two tags t_1 and t_2 are linked if there exists a post where they have been associated by a user with the same resource. A link weight can be introduced by defining the weight of the link between t_1 and t_2, $t_2 \neq t_1$, as the number of posts where they appear together. Formally, the set of posts $CO_{T/UR}$ where two tags t_1 and t_2 co-occur can be defined as follows (cf. 2.2.2)

$$CO_{T/UR}(t_1, t_2) := \{(u, r) \in U \times R \mid (t_1, u, r) \in Y \wedge (t_2, u, r) \in Y\}\,, \tag{3.3}$$

The link weight is then defined as $w(t_1, t_2) := |CO_{T/UR}(t_1, t_2)|$. The above link strength defines on $T \times T$ a symmetric similarity matrix which is analogous to the usual adjacency matrix in graph theory. The strength s_t of a node t is defined as

$$s_t := \sum_{t' \neq t} w(t, t')\,. \tag{3.4}$$

3.3.1 Cumulative Strength Distribution

A first statistical characterization of the network of tags is afforded by the cumulative strength distribution $P_>(s)$, defined as the probability of observing a strength in excess of s. These distributions are displayed for *del.icio.us* and *BibSonomy* in Figs. 3.4 and 3.5, respectively. This is a standard measure in complex network theory and plays the same role of the degree distribution in unweighted networks. We observe that $P_>(s)$ is a fat-tailed distribution for both folksonomies: this is related to a lack of characteristic scale for node strengths and is one of the typical fingerprints of an underlying complex dynamics of interacting human agents (Vazquez et al., 2006; Vazquez, 2005). A coarse indicator such as $P_>(s)$, despite its simplicity, is able to point out anomalous activity (i.e. spam) within the investigated folksonomies, as discussed in the captions of Figs. 3.4 and 3.5. Quite interestingly, on filtering out these undesired (and probably automatically generated) contributions, the probability distributions for *del.icio.us* and *BibSonomy* become rather similar, even though the two systems under study are dramatically different in terms of user base, size and age.

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

In Figs. 3.4 and 3.5 we show that by performing this shuffling operation (blue dots) the distribution is only marginally affected. Far from being obvious, this shows that the global frequencies of tags – and not their co-occurrence relations – are the main factors shaping the distribution $P_>(s)$. In other words, the fat-tailed nature of $P_>(s)$ is induced by the distribution of tag frequencies, which has been known to be fat-tailed (Cattuto et al., 2007; Golder and Huberman, 2005), in analogy to Zipf's law (also observed in human languages).

3.3.2 Average Nearest Neighbor Strength

In order to probe deeper into the structure of the co-occurrence network and recognize the contribution of semantics, we need to compute observables more sensitive to correlations and to the local structure of the network. To this end, a useful quantity studied in complex networks is the nearest neighbor connectivity. As defined above, the sum of all weights $s_i = \sum_j w_{ij}$ of links connected to a given node i is called strength of node i. Now, given a node i with strength s_i, we

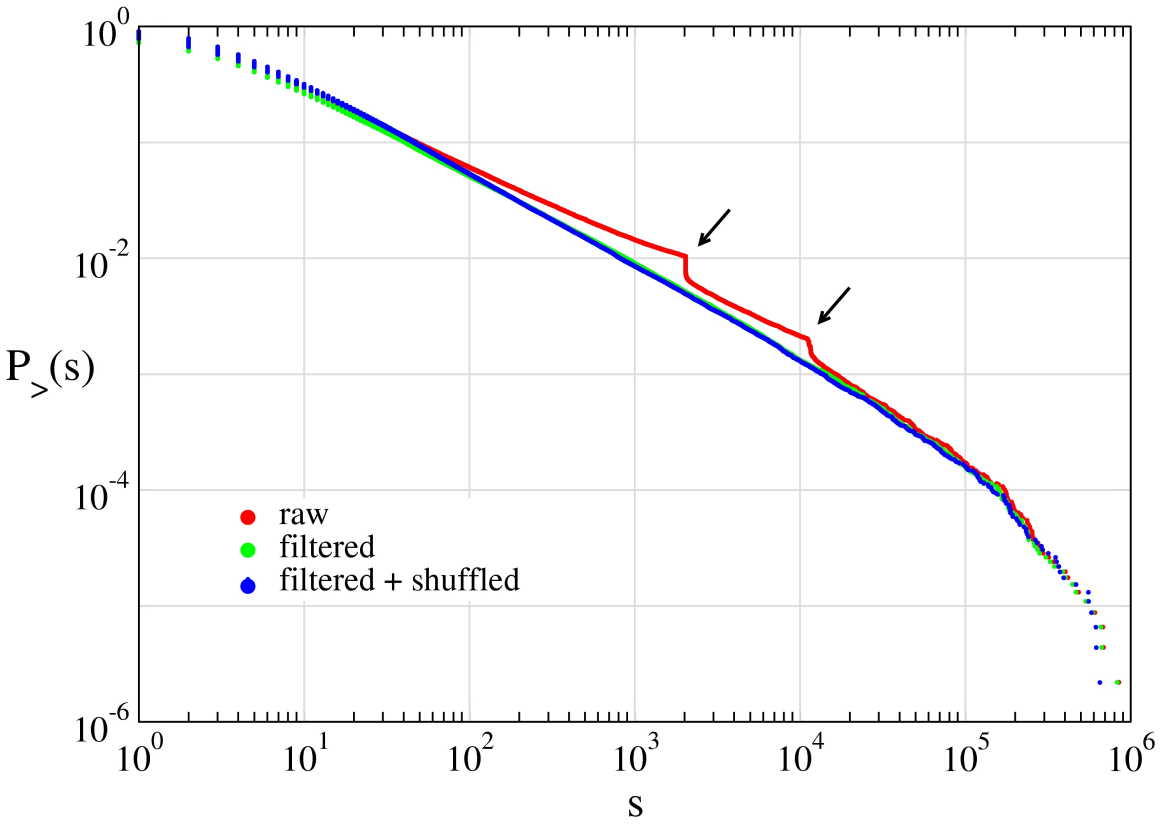

Figure 3.4: Cumulative strength distribution for the network of co-occurrence of tags in *del.icio.us*. $P_>(s)$ is the probability of having a node with strength in excess of s. Red dots correspond to the whole co-occurrence network. The two steps indicated by arrows correspond to an excess of link with a specific weight and can be related to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps (green dots). Shuffling the tags contained in posts (blue dots) does not affect significantly the cumulated weight distribution. This proves that such a distribution is uniquely determined by tag frequencies within the folksonomy, and not by the semantics of co-occurrence.

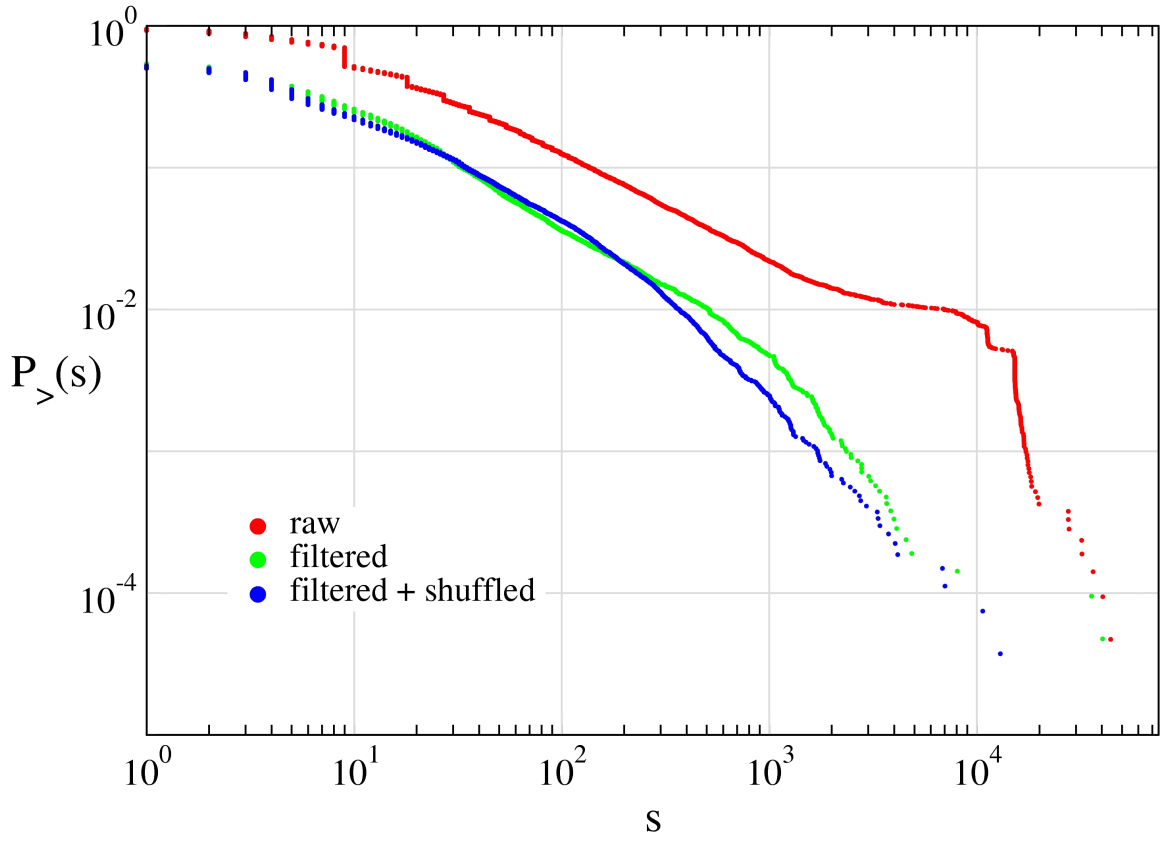

Figure 3.5: Cumulative strength distribution $P_>(s)$ for the network of co-occurrence of tags in *BibSonomy* (see also Fig. 3.4). Red dots correspond to the whole co-occurrence network. The irregular behavior for high strengths can be linked to spamming activity: identified spam in *BibSonomy* consists of posts with a large number of tags, as well as a large number of posts with exactly 10 tags, injected by a small group of spammers. Both types of spam were identified by inspecting the distribution of the number of tags per post. Excluding the above posts from the analysis (green dots), the distribution becomes smooth and similar to the filtered one observed for del.icio.us. Similarly, shuffling the tags contained in posts (blue dots) has a small effect on the cumulated weight distribution.

define its average nearest-neighbor strength as:

$$S_{nn}(s_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} s_j \,, \tag{3.5}$$

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

The average nearest neighbor strength $S_{nn}(s)$, as a function of the node strength s, provides information on correlations among the strength of nodes and therefore is also known in literature as node nearest-neighbor strength correlation $S_{nn}(s)$ (Barrat et al., 2004). When referred to unweighted networks, i.e. where all existing links have unit strength, S_{nn} is able to discriminate between technological networks, where $S_{nn}(s)$ is a decreasing function of the strength s, and social networks, where, on the contrary, $S_{nn}(s)$ displays an increasing behavior. These two networks with opposite behaviors are commonly referred to as disassortative and assortative mixing networks, respectively (Capocci and Colaiori, 2005; Newman, 2002).

Figs. 3.6 and 3.7 display our results for *del.icio.us* and *BibSonomy*, respectively. In the figures, each dot correspond to a node of the network (i.e. a tag), with its strength s as the abscissa and the average strength of its neighbors S_{nn} as the ordinate. Both quantities span several orders of magnitude, hence we use a logarithmic scale along both axes to display the global features of the scatter plot. This is related to the fat-tailed behavior observed for the strength distribution $P_>(s)$, which is in fact recovered by projecting the data points along the s-axis and computing the cumulative distribution.

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

Overall, the plots for *del.icio.us* and *BibSonomy* look quite similar, and this suggests that the features we report here are generally representative of collaborative tagging systems. An assortative region (S_{nn} roughly increasing with s) is observed for low values of the strength s, while disassortative behavior (S_{nn} decreasing with s) is visible for high values of s. As we have already done for the probability distribution $P_>(s)$, we can highlight the contribution of semantics by randomly shuffling tags in TAS entries (blue dots in Fig. 3.6 and 3.7). In this case, shuffling the tags (blue points) affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the network of co-occurrence beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics. Interestingly, the main effect seems to be the disappearance of points in the assortative (low strength) region of the plot, possibly identifying this region as the one exposing semantically relevant connections between tags. Notice, for example, the disappearance of a whole cloud of points at the top-left of Fig. 3.7: those points represent nodes (tags) with low strength that are attached preferentially to nodes of high strength. Similarly, in Fig. 3.6, the highly populated region with s roughly ranging between 10 and a few thousands also disappears when tag shuffling is applied. Those data points also represent low-strength nodes (tags) preferentially connected with higher-strength nodes (tags). Such properties are commonly found in hierarchically organized networks, and could be related to an underlying hierarchical organization of tags (Heymann and

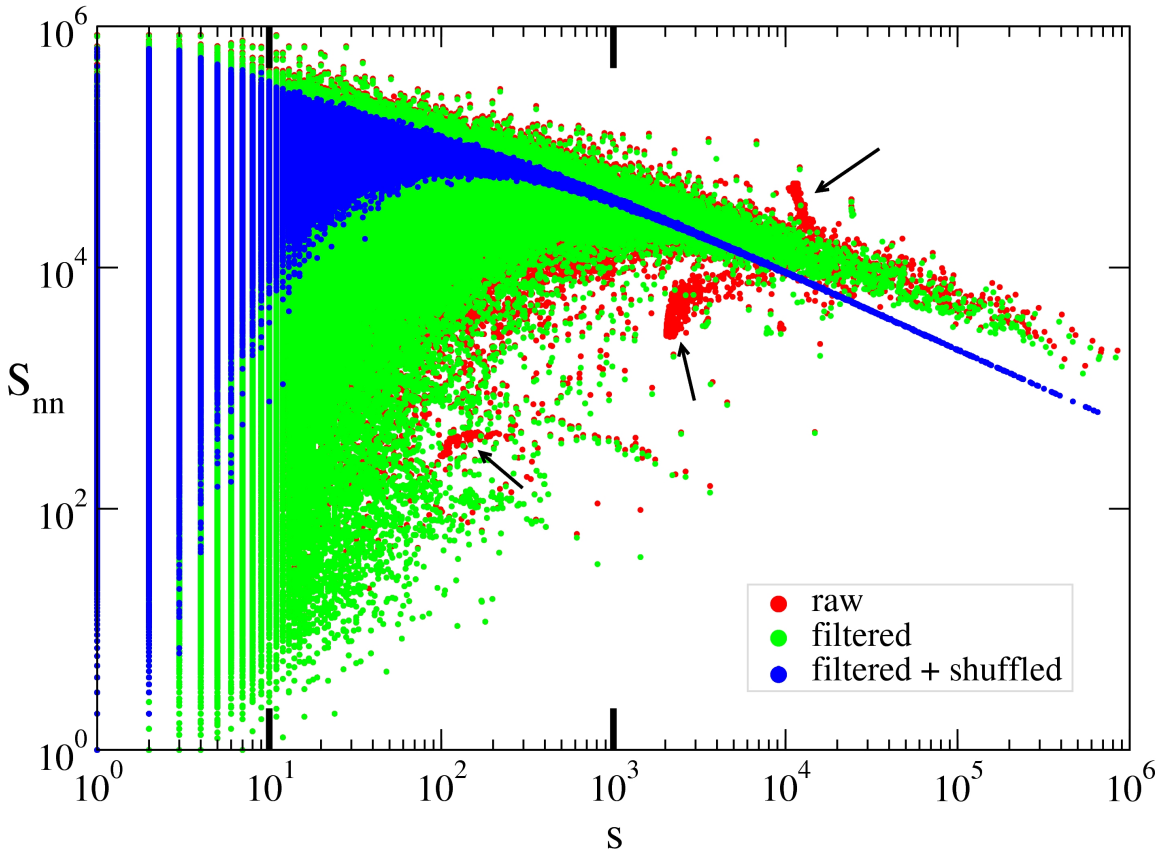

Figure 3.6: Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in *del.icio.us*. Red dots correspond to the whole co-occurrence network. Assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. A few clusters (indicated by arrows) stand out from the main cloud of data points. As in Fig. 3.4, such anomalies correspond to spamming activity and can be removed by filtering out posts containing an excessive number of tags (green dots). In this case, shuffling the tags (blue points) affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the network of co-occurrence beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics.

Garcia-Molina, 2006).

3.4 Cooperative structuring of folksonomy semantic space

In this section we focus on a different aspect of a folksonomy, namely the cooperative aspects of the incoherent user activity. We define in particular a semantic space as composed by a given set of resources and check whether the selfish tagging activity is able to structure such a space in a semantically meaningful way. Here meaningful structure means a partition of the semantic space made up of groups of resources for which the associated tag cloud points to a well defined area or topic. These groups of resources could also identify in principle groups or communities of users sharing the same interests.

In order to perform this analysis we have considered a dataset composed by two different sets, each composed by 200 resources (this number can be easily increased). The two sets are chosen in such a way that the first one only includes resources for which all posts include the tag `design`

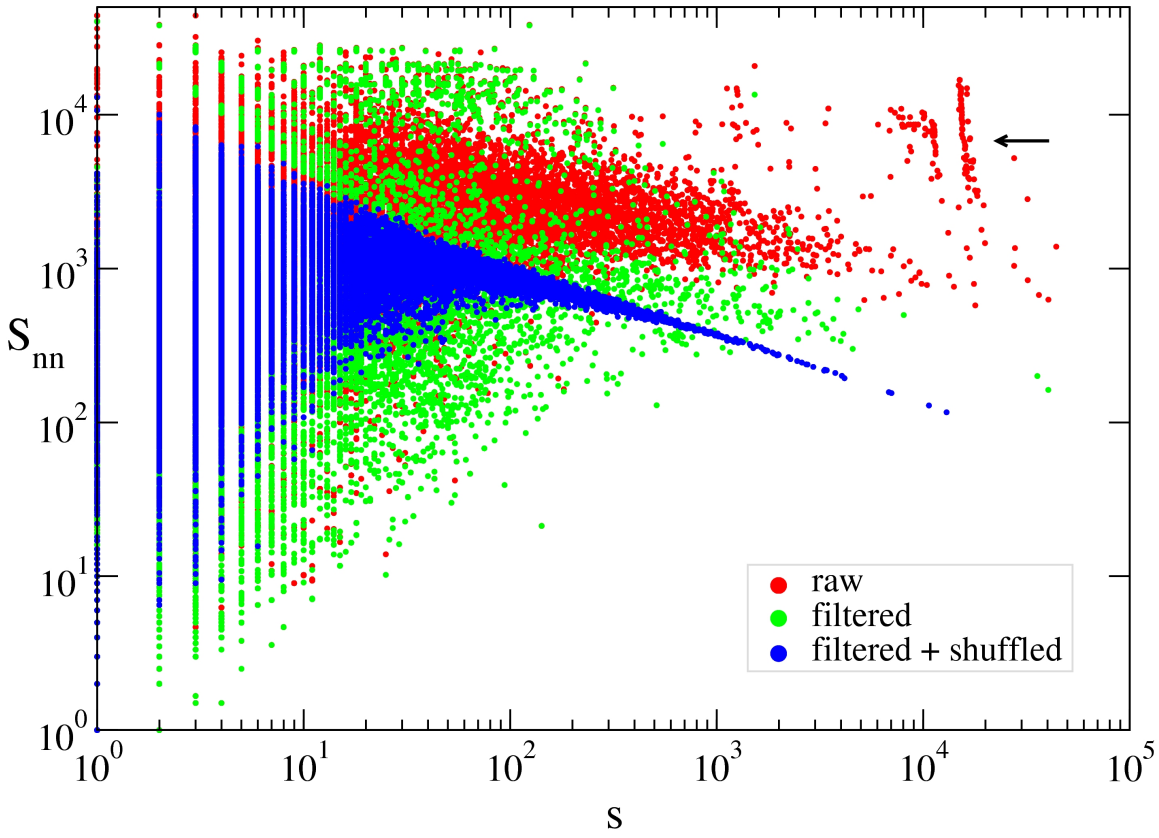

Figure 3.7: Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in *BibSonomy*. Red dots correspond to the whole co-occurrence network. The scatter plot is qualitatively very similar to the one reported in Fig. 3.6 for *del.icio.us*: assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. Again, a few clusters (indicated by arrows) stand out from the main cloud of data points and their presence can be linked to spamming activity. They disappear when we filter out posts containing an excessive number of tags (green dots). Shuffling the tags (blue dots) has the same effect as in Fig. 3.6, and the same observations apply.

while the second one only includes resources for which all posts include the tag `politics`. The idea was to artificially construct a dataset with at least two well separated semantic regions. For each resource in the dataset the whole stream of posts is available, i.e. for each resources we have the whole temporal stream of users-tags associations $[U, \{\text{tags}\}]$.

The next step has been that of constructing a similarity matrix among resources. To this end we define the level of similarity between two generic resources R_1 and R_2 using a TF/IDF weighting procedure (Salton and McGill, 1983). The TF/IDF weight (Term Frequency Inverse Document Frequency) is a often used in information retrieval and text mining where it represents a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In our implementation the tags are playing the role of words. Given two resources, one defines with T_1 and T_2 the set of tags associated with R_1 and R_2, respectively, the Union set as $U := T_1 \cup T_2$, and the Intersection set as $K := T_1 \cap T_2$.

Figure 3.8: Sets of tags considered in evaluating the strength w of Eq. 3.6: T_1 (blue) and T_2 (yellow) are the set of tags associated with resources 1 and 2, respectively. $K := T_1 \cap T_2$ is the set of tags shared by the two resources.

Further we denote with f_t^1, f_t^2 the frequencies of occurrence of the tag t in T_1 and T_2, respectively, while f_t will denote the frequency of the tag t in the whole corpus of tags associated to all the resources. With these definitions, the strength of the link between R_1 and R_2 is given by:

$$w_{R_1,R_2} = \frac{\sum_{t\in T_1\cap T_2} \frac{min(f_t^1 . f_t^2)}{f_t}}{\sum_{t\in T_1\cap T_2} \frac{max(f_t^1 . f_t^2)}{f_t} + \sum_{t\in T_1 - K} \frac{f_t^1}{f_t} + \sum_{t\in T_2 - K} \frac{f_t^2}{f_t}}. \tag{3.6}$$

Fig. 3.9 reports the histograms of the strengths among all the resource pairs for three different corpora of resources: the subset of resources sharing the tag `design`, the subset of resources sharing the tag `politics` and the whole of resources. Notice that the global frequency f_t of a given tag t depends on the corpus chosen for the analysis. From the figure it is evident that the logarithm scale is the best suited to visualize the full range of strength variability.

In order to investigate the existence of underlying structures in the space of the resources we proceed as follows. First of all we transform the matrix in order to optimize the dynamical range of strength variability. Since the logarithmic scale gives a good representation of the strength variability we considered a matrix where each element is raised to a very small power (in order to avoid the pathologies of the logarithm in the neighborhood of zero). The matrix w' we will use from

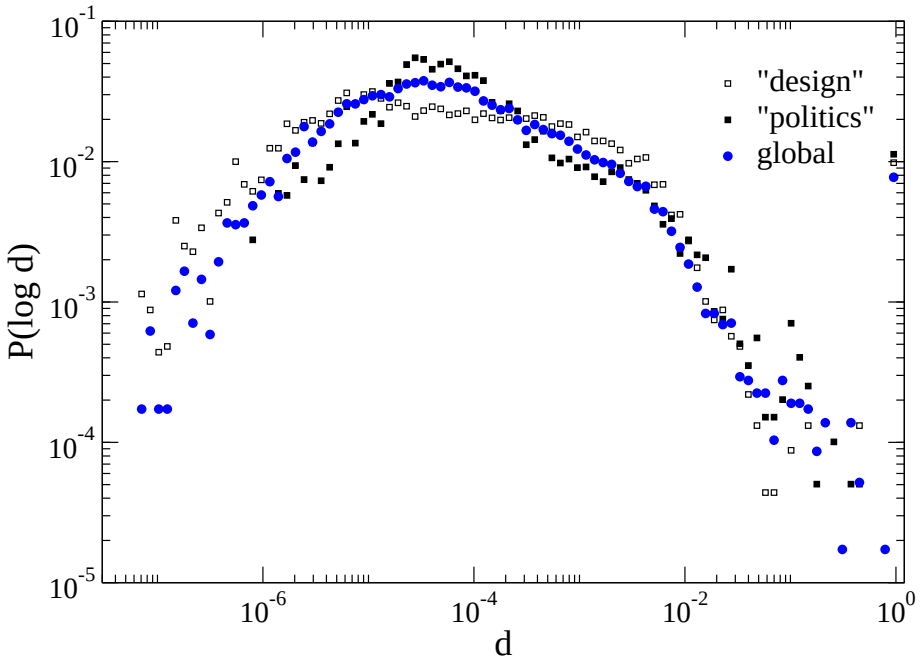

Figure 3.9: Probability distributions of link strengths. The logarithmically-binned histogram of the link strengths for all pairs of resources within a given set is displayed for three sets of resources: empty squares correspond to resources tagged with *design*, filled squares correspond to resources tagged with *politics*, and blue circles correspond to the union of the above sets. It is important to observe that the variability range of strengths spans several orders of magnitude, so that a non-linear function of link strengths becomes necessary in order to capture the full dynamic range of strength values.

now onwards has then components:

$$w'_{R_1,R_2} = (w_{R_1,R_2})^{0.1} \quad \forall \ pairs R_1, R_2 \in U. \tag{3.7}$$

Figure 3.10 reports the full matrix of strengths between pairs of resources $w\prime_{R_1,R_2}$ for the full set of 400 resources. The resources are randomly ordered and no structures are evident in this representation.

The problem we have to tackle now is that of finding the sequence of row and column permutations of the matrix of strengths, that permits to visually spot the presence of semantic communities of resources. The goal will be to obtain a matrix with a clear visible block structure on its main diagonal. One possible way to approach this problem is to construct an auxiliary matrix and use information deduced from its spectral properties to rearrange row and columns of the original matrix. The quantity we consider is the matrix

$$Q = S - W \tag{3.8}$$

where $W = w'$ except for the main diagonal where W presents all zeros and S is the diagonal matrix with each element on the main diagonal equal to the sum of the corresponding row of W, i.e. $S_{ij} = \delta_{ij} \sum_j W_{ij}$.

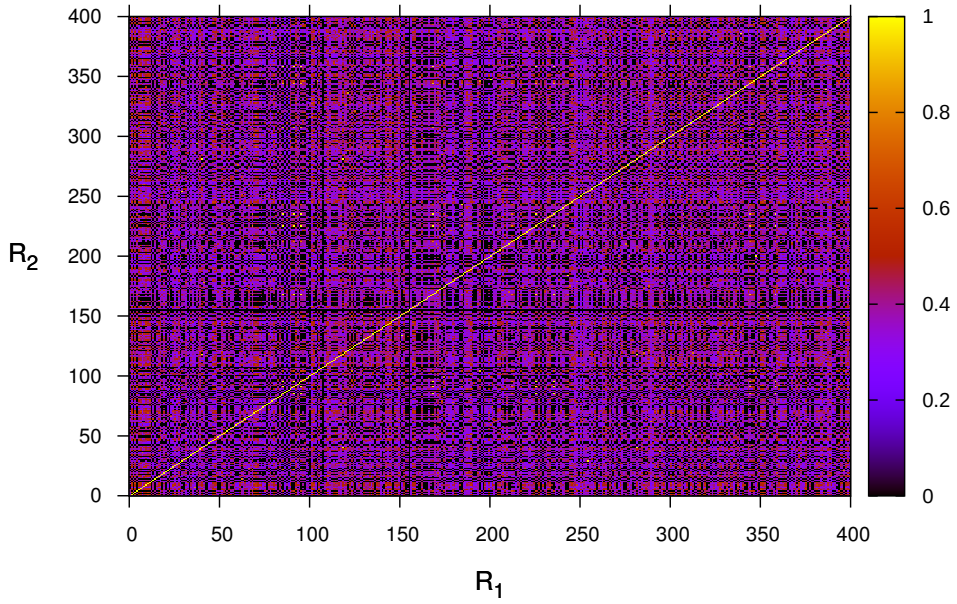

Figure 3.10: Matrix w' of link strengths (see Eq.3.7) for the global set of 400 randomly ordered resources. Except for the bright diagonal, whose elements are identically equal to 1 because of the normalization property of the strength w, the matrix appears featureless.

The matrix Q is non negative and resembles the Laplacian matrix of graph theory. As shown in (Capocci et al., 2005; Newman, 2006), studying its spectral properties can reveal the community structure present in complex networks.

The main idea is to consider the lowest eigenvalues. Firstly, for the very construction, there is a zero eigenvalue corresponding to an eigenvector with constant component values, i.e. a trivial constant eigenvector. Now consider the simplest case, where the matrix Q could be written with exactly two non zero blocks on its main diagonal (i.e. with two clearly separated semantic communities). In this case two eigenvectors with zero eigenvalues are present. Each one of these eigenvectors signals the presence of a community. In fact their components would be zero for the coordinates corresponding to a community (i.e. a diagonal block) and different from zero, but constant, for the other community.

The possible presence of small terms mixing the two blocks would remove the degeneracy of the null eigenvector. Still, the coordinates of the eigenvectors with the smallest eigenvalues will reveal the community structures, even if some trivial eigenvectors, similar to the constant one, may still be present.

Given the set of these non trivial eigenvectors, a very simple way to identify the communities consists in plotting on a multidimensional plot their coordinates. Each axis reports the values of the components of the eigenvectors. In particular each point has coordinates equal to the homologous components for the eigenvectors considered. In this kind of plots communities will emerge as well defined cluster of points. The components involved in each clusters identify the elements belonging to a given community.

Once identified the communities, it is interesting to permute the indexes of the original matrix W such that the components of the same community get close. The corresponding matrix should appear roughly made by diagonal blocks, even if some mixing terms would signals an overlap

between communities (blocks).

Figure 3.11 reports the set of all the eigenvalues of Q sorted by their numerical value. As expected, we observe the existence of a null eigenvalue, corresponding to the trivial constant eigenvector. Next largest eigenvalue is actually very small and corresponds to a trivial eigenvector too, since its components are almost constant. Higher eigenvalues corresponds to non-trivial eigenvectors.

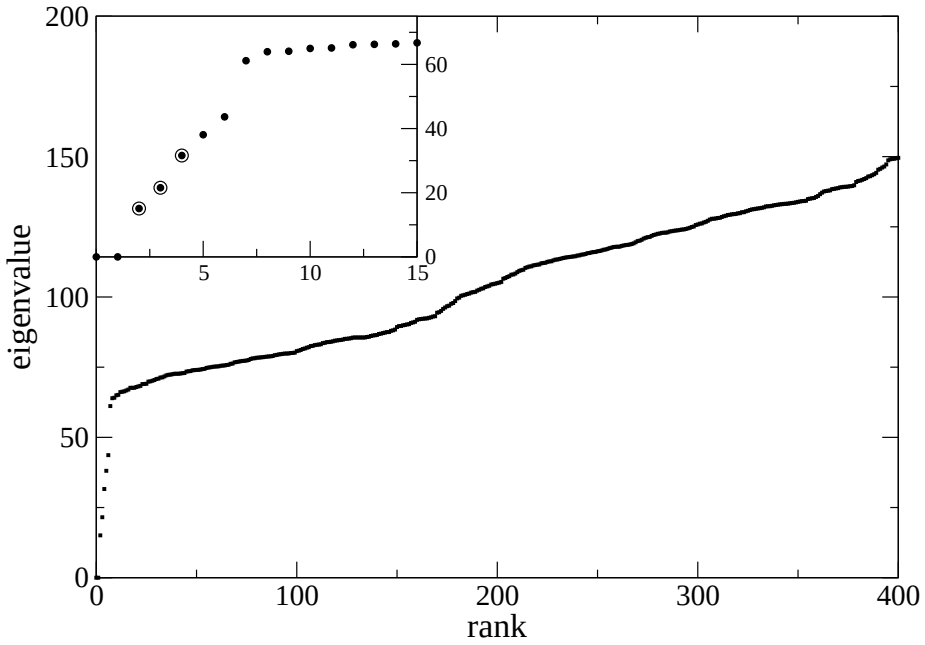

Figure 3.11: Eigenvalues of the matrix Q. Resource communities correspond to non-trivial eigenvalues of the spectrum, such as the ones visible on the leftmost side of the plot, and in the inset. The three eigenvalues marked in the inset correspond to the eigenvectors plotted in Fig. 3.12.

Figure 3.12 reports a 3-dimensional scatter plot illustrating the structure of the three eigenvectors corresponding to the first three non-trivial eigenvalues (i.e. third, fourth and fifth, see Fig.3.11). Each axis report the values of the components of the eigenvectors corresponding to the third, fourth and fifth eigenvalues. In particular each point has coordinates equal to the homologous components for the three non-trivial eigenvectors considered. It is evident the existence of at least 5 well defined communities corresponding also to the five well separated non-zero eigenvalues (see fig. 3.11). It is actually barely visible a sixth community corresponding to the sixth non-trivial eigenvalue.

Once we have diagonalized the matrix Q the permutation of indexes necessary to sort the component values of these eigenvectors determine the desired ordering of rows and columns in the original matrix W. By performing this reordering it is possible to visualize the matrix of strengths of Fig. 3.10 in a way to make it maximally diagonal. Fig. 3.13 reports the reordered matrix.

An interesting question is now whether the communities we have found through the diagonalization of the matrix Q correspond to semantically separated area in the space of resources. In order to check this point we associate to each community its corresponding tag cloud and we plot it by considering only the 30 most frequent tags. Fig. 3.14 reports the six tag clouds (ordered for decreasing number of resources) where the font of each tag is proportional to the logarithm of the

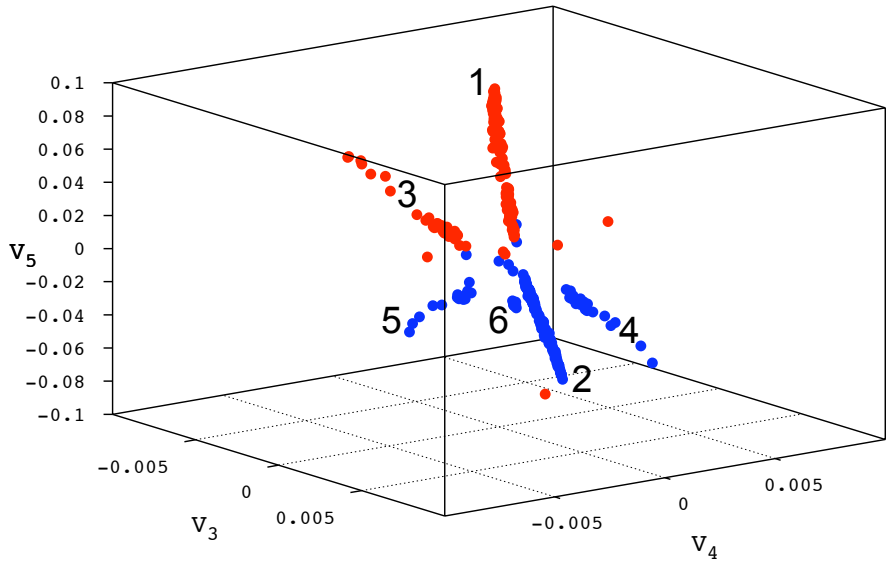

Figure 3.12: Eigenvectors of the matrix Q. The scatter plot displays the component values of the first three non-trivial eigenvectors of the matrix (V_3, V_4, V_5, also marked in Fig. 3.11). The scatter plot is parametric in the component index. Five or six clusters are apparent, corresponding to the non-trivial eigenvalues of the matrix. Each cluster, marked with a numeric label, defines a community of "similar" resources (in terms of tags). Blue and red points correspond to resources tagged with *design* and *politics*, respectively. It is important to note that our approach clearly separates the two communities, as well as highlighting few more finer-grained structures. Tag clouds for the identified communities are shown in Fig. 3.14.

tag frequency.

Despite the intrinsic difficulty of identifying the semantic area defined by a given tag cloud, it is possible to associate, at least to the four main largest communities, well defined areas. In particular the first community could be associated to humor in politics, the second one to visual design, the third one to news in politics and the fourth one to web design. It is evident how the a-priori selfish activity of the users is give raise to an effective cooperative behaviors able to structure the semantic space of the resources with semantically meaningful set of tags.

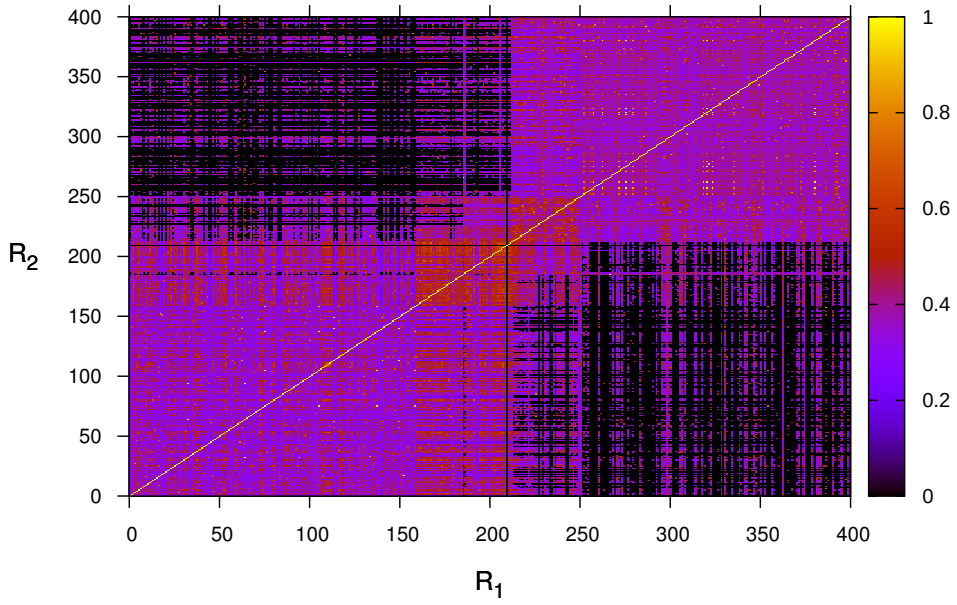

Figure 3.13: Matrix w' of link strengths (see Eq.3.7) for our set of 400 resources. Here the resource indices are ordered by community membership (the sequence of communities along the axes is $2,\ 4,\ 6,\ 5,\ 3,\ 1$, see Fig. 3.14). In striking contrast with Fig.3.10, the permutation of indices we employed clearly exposes the community structure of our set of resources: two main regions of high-similarity, corresponding to blue/red rectangles at the top-right and bottom-left of the matrix correspond, respectively, to resources tagged with *design* and *politics*. On top of this, our approach also reveals the presence of finer-grained community structures within the above communities (red rectangular regions towards the center of the matrix). On direct inspection, such communities of resources turn out to have a rather well defined semantic characterization in terms of tags, as shown by the tag clouds in Fig.3.14.

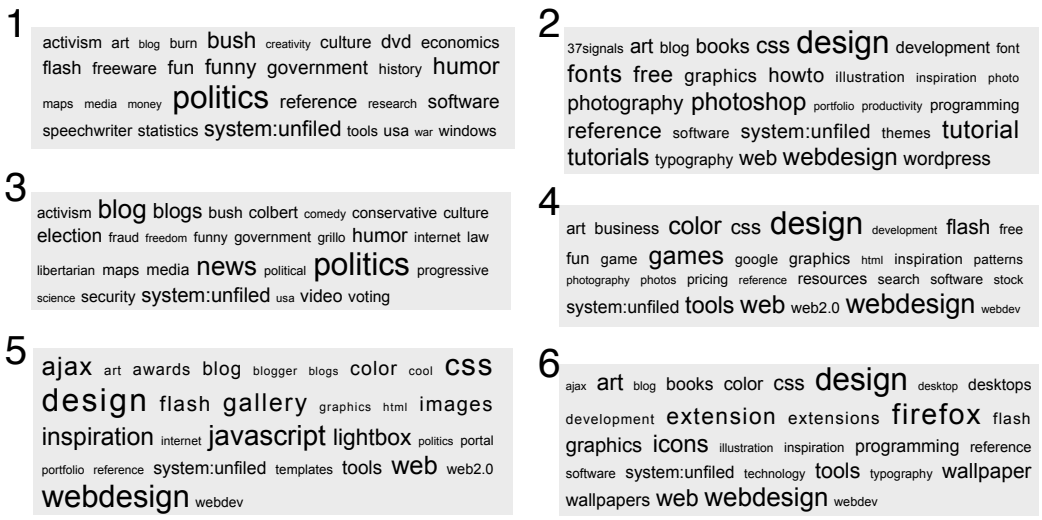

Figure 3.14: Tag clouds for the 6 resource communities identified by our analysis(see Fig. 3.12), ordered by decreasing community size. Each tag cloud shows the 30 topmost frequent tags associated with resources belonging to a given community. Within tag clouds, as usual, the size of text labels increases with the logarithm of the frequency of the corresponding tag. The first two communities (the largest ones) correspond to the main division between resources tagged with *politics* and *design*, respectively. Notice how each tag clouds is strongly characterized by only one of the above two tags. In addition to discriminating the above two main communities, our approach also identifies additional and unexpected communities. On inspecting the corresponding tag clouds, one can recognize a rather well-defined semantic connotation pertaining to each community, as discussed in the main text.

Chapter 4

Trend Detection in Folksonomies

In this chapter, we will analyze this emergence of common semantics by exploring trends in the folksonomy. Since the structure of a folksonomy is symmetric with respect to the dimensions 'user', 'tag', and 'resource', we can apply the same approach to study upcoming users, upcoming tags, and upcoming resources. We present a technique for analyzing the evolution of topic-specific trends. Our approach is based on our *FolkRank* algorithm (Hotho et al., 2006a), a differential adaptation of the PageRank algorithm (Brin and Page, 1998) to the tri-partite hypergraph structure of a folksonomy. Compared to pure co-occurrence counting, FolkRank takes also into account elements that are related to the focus of interest with respect to the underlying graph/folksonomy. In particular, FolkRank ranks synonyms higher, which usually do not occur in the same bookmark posting together.

With FolkRank, we compute topic-specific rankings on users, tags, and resources. In a second step, we can then compare these rankings for snapshots of the system at different points in time. We can discover both the absolute rankings (who is in the Top Ten?) and winners and losers (who rose/fell most?).

The contributions of this work are:

Ranking in folksonomies. We describe a general ranking scheme for folksonomy data. The scheme allows in particular for topic-specific ranking.

Trend detection. We introduce a trend detection measure which allows to determine which tags, users, or resources have been gaining or losing in popularity in a given time interval. Again, this measure allows to focus on specific topics.

Application to arbitrary folksonomy data. As the ranking is solely based on the graph structure of the folksonomy – which is resource-independent – we can also apply it to any kind of resources, including in particular multimedia objects, but also office documents which typically do not have a hyperlink structure per se. It can even be applied to an arbitrary mixture of these content types. Actually, the content of the tagged resources will not have to be accessible in order to manage them in a folksonomy system.

Evaluation. We have applied our method to a large-scale dataset from an actual folksonomy system.

The chapter is organized as follows. In the next section, we describe our ranking and trend detection approach. In Section 4.2, we apply the approach to a large-scale dataset, a one-year snapshot of the del.icio.us system and Section 4.3 discusses related work.

4.1 Trend Detection in Folksonomies

For discovering trends in a social resource sharing system, we will need snapshots of its folksonomy at different points of time. For each snapshot, we will need a ranking, such that we can compare the rankings of consecutive snapshots. As we also want to discover topic-specific trends, we will additionally need a ranking method that allows to focus on the specific topic. We will make use of our search and ranking algorithm *FolkRank* (Hotho et al., 2006a) which we summarize below.

4.1.1 Ranking

In this section we recall the principles of the *FolkRank* algorithm that we developed for supporting Google-like search in folksonomy-based systems. It is inspired by the seminal PageRank algorithm (Brin and Page, 1998).

Because of the different nature of folksonomies compared to the web graph (undirected triadic hyperedges instead of directed binary edges), PageRank cannot be applied directly on folksonomies. In order to employ a weight-spreading ranking scheme on folksonomies, we will overcome this problem in two steps. First, we transform the hypergraph into an undirected graph. Then we apply a differential ranking approach that deals with the skewed structure of the network and the undirectedness of folksonomies.

Folksonomy-Adapted Pagerank.

First we convert the folksonomy $\mathbb{F} = (U, T, R, Y)$ into an *un*directed tri-partite graph $G_{\mathbb{F}} = (V, E)$ as follows.

1. The set V of nodes of the graph consists of the disjoint union of the sets of tags, users and resources: $V := U \dot\cup T \dot\cup R$. (The tripartite structure of the graph can be exploited later for an efficient storage of the adjacency matrix and the implementation of the weight-spreading iteration in the FolkRank algorithm.)

2. All co-occurrences of tags and users, users and resources, tags and resources become edges between the respective nodes: $E := \{\{u, t\} \mid \exists r \in R : (u, t, r) \in Y\} \cup \{\{t, r\} \mid \exists u \in U : (u, t, r) \in Y\} \cup \{\{u, r\} \mid \exists t \in T : (u, t, r) \in Y\}$.

The original formulation of PageRank (Brin and Page, 1998) reflects the idea that a page is important if there many pages linking to it, and if those pages are important themselves. The distribution of weights can thus be described as the fixed point of a weight passing scheme on the web graph. This idea was extended in a similar fashion to bipartite subgraphs of the web in HITS (Kleinberg, 1999) and to n-ary directed graphs in (Xi et al., 2004). We employ the same underlying principle for our ranking scheme in folksonomies. The basic notion is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users, thus we have a graph of vertices which are mutually reinforcing each other by spreading their weights.

Like PageRank, we employ the random surfer model, a notion of importance for web pages that is based on the idea that an idealized random web surfer normally follows hyperlinks, but from time to time jumps to a new webpage without following a link. This results in the following definition. The rank of the vertices of the graph are the entries in the fixed point $\vec{w}$ of the weight spreading computation

$$\vec{w} \leftarrow dA\vec{w} + (1 - d)\vec{p} \ , \tag{4.1}$$

where $\vec{w}$ is a weight vector with one entry for each web page, A is a row-stochastic version of the adjacency matrix of the graph $G_{\mathbb{F}}$ defined above, $\vec{p}$ is the random surfer component that outweighs the loss of weight in dangling links, and $d \in [0, 1]$ is determining the influence of $\vec{p}$. Usually, one will choose $\vec{p} = \mathbf{1}$, i. e., the vector composed by 1's, to achieve uniform damping. In order to compute personalized PageRanks, however, $\vec{p}$ can be used to express user preferences by giving a higher weight to the components which represent the user's preferred web pages. If $||\vec{w}||_1 = ||\vec{p}||_1$,[1] the weight in the system will remain constant.

As the graph $G_{\mathbb{F}}$ is undirected, most of the weight that went through an edge at moment t will flow back at $t+1$. The results are thus rather similar (but not identical, due to the damping) to a ranking that is simply based on edge degrees. The reason for applying the more expensive PageRank approach nonetheless is that its random surfer vector allows for topic-specific ranking.

FolkRank — Topic-Specific Ranking.

As the graph $G_{\mathbb{F}}$ that we created in the previous step is undirected, we face the problem that an application of the original PageRank would result in weights that flow in one direction of an edge and then 'swash back' along the same edge in the next iteration, so that one would basically rank the nodes in the folksonomy by their degree distribution. This makes it very difficult for other nodes than those with high edge degree to become highly ranked, no matter what the preference vector is.

This problem is solved by the *differential* approach in FolkRank, which computes a personalized ranking of the elements in a folksonomy as follows:

1. The preference vector $\vec{p}$ is used to determine the topic. It may have any distribution of weights, as long as $||\vec{w}||_1 = ||\vec{p}||_1$ holds. Typically a single entry or a small set of entries is set to a higher value, and the remaining weight is equally distributed over the other entries. Since the structure of folksonomies is symmetric, we can define a topic by giving a higher value to either one or more tags and/or one or more users and/or one or more resources.
2. Let $\vec{w_0}$ be the fixed point from Equation (4.1) with $d = 1$.
3. Let $\vec{w_1}$ be the fixed point from Equation (4.1) with $d < 1$. In our experiments, we set $d = 0.85$.
4. $\vec{w} := \vec{w_1} - \vec{w_0}$ is the final weight vector.

Thus, we compute the winners and losers of the mutual reinforcement of nodes when a user preference is given, compared to the baseline without a preference vector. We call the resulting weight $\vec{w}[x]$ of an element x of the folksonomy the *FolkRank* of x. In (Hotho et al., 2006a) we showed that $\vec{w}$ provides indeed valuable results on a large-scale real-world dataset while $\vec{w_1}$ provides an unstructured mix of topic-relevant elements with elements having high edge degree.

4.1.2 Trend Detection

In order to analyze the trends around a specific topic, we first have to describe the topic by defining the preference vector $\vec{p}$. Then we compute, for each point in time $t \in \{0, \ldots, n\}$, the rank vector $\vec{w_t}$ within the folksonomy $\mathbb{F}_t$ which consists of all tag assignments performed before t.[2]

We select then from the resulting rank vectors those entries which are assigned to one of the three dimensions 'tags', 'users', and 'resources' — depending on where we want to see rising and falling elements. Else an analysis would be difficult, since users have higher weights than tags, which in their turn have higher weights than resources, due to the different sizes of the sets U, T, and R.

[1] . . . and if there are no rank sinks – but this holds trivially in our graph $G_{\mathbb{F}}$.

[2] If no entries were deleted, $\mathbb{F}_{t+1}$ contains thus $\mathbb{F}_t$, for all t.

As the total weight in the system will differ at different points of time because of new tags, users, and resources, we normalize at last each rank vector such that its largest value equals 1. This allows to compare rankings from different points in time. If the preference vector has only one distinguished element, then this element is the one with the highest value in the resulting weight vector. The closer another entry is to this value, the more important is its associated element to the topic. By plotting the values of the Top 10 or Top 20 over time, one can thus discover the rise and fall of the most popular elements. Figure 4.1 shows such a plot for the del.icio.us users which are most important for the topic 'music', while Figure 4.2 shows the tags which are most important for the topic 'politics'. How these diagrams are to be read, and what the most important findings are, will be described in detail in the next section.

Going a step further, we may not only be interested in the most important elements, but also in those where the increase or decrease of rank is the steepest. To this end, we have developed the following *popularity change* measure, which allows for detecting topic-specific trends.

Assume x is a tag, user or resource of the folksonomy $\mathbb{F}$, i.e. $x \in U \cup T \cup R$. (In the following, we assume it is a resource; the same methods apply symmetrically for tags and users.) Similar to the relative change used for word occurrences in (Kleinberg, 2006), we define the *popularity change* $pc_{t_0 \to t_1}(x)$ of x from t_0 to t_1 as follows.

At times $t_0 < t_1$, let the resource x be ranked at position r_0 and r_1, respectively, in the descending weight order of the FolkRank computation. Let n_0 and n_1 be the sizes $|R|$ of the resource dimension at times t_0, t_1. The popularity change is defined as

$$pc_{t_0 \to t_1}(x) := \left(\frac{r_0}{n_0} - \frac{r_1}{n_1} \right) \log_{10} \left(\frac{n_1}{r_1} \right) \tag{4.2}$$

(where elements not present at time t_i are treated as being positioned at $r_i = n_i + 1$). Here, the fractions in the first term indicate the relative positions of x at the given times, $1/n_i$ being the best (i.e. having maximum FolkRank) and 1 being the worst. The second term discounts the change with respect to the relative position where the change took place: to get from a top 90 % position to a top 80 % one would be considered three times easier than to get from the top 0.09 % to the top 0.08 %.

Combined with a topic-directed FolkRank computation, we use this measure of a change in popularity to get an insight into what are the trends in a certain community in the folksonomy. We point out the winning and losing elements of the folksonomy in a given time interval.

4.2 Experiments

4.2.1 Evaluation of Popularity Change in del.icio.us

In order to evaluate our approach, we have analyzed the popular social bookmarking system del.icio.us.[3] Del.icio.us is a server-based system with a simple-to-use interface that allows users to organize and share bookmarks on the internet. The resources del.icio.us is pointing to cover various formats (text, audio, video, etc.). In particular, the system is not restricted to a single type (like photos in Flickr). As discussed above, our approach is specially suited for this situation. In addition to the URL, del.icio.us allows to store a description, an extended description, and tags (i. e., arbitrary labels). Del.icio.us is online for a sufficiently long time (since May 2002) to allow for extracting significant time series.

For our experiments, we collected data from the del.icio.us system between July 27 and July 30, 2005 in the following way. Initially we used `wget` starting from the start page of del.icio.us to obtain nearly 6900 users and 700 tags as a starting set. Out of this dataset we extracted all users

[3] `http://del.icio.us`

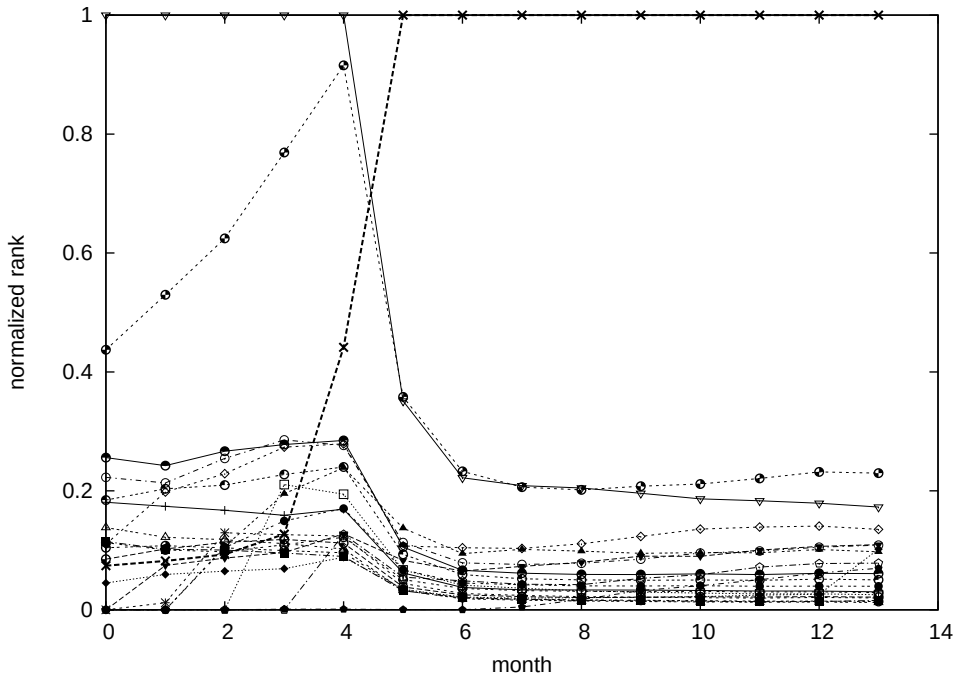

Figure 4.1: Evolution of the ranking of users related to 'music'. User names are omitted for privacy reasons.

and resources (i. e., del.icio.us' MD5-hashed urls). We downloaded in a recursive manner user pages to get new resources and resource pages to get new users. Furthermore we monitored the del.icio.us start page to gather additional users and resources. This way we collected a list of several thousand usernames which we used for accessing the first 10000 resources each user had tagged. From the collected data we finally took the user files to extract resources, tags, dates, descriptions, extended descriptions, and the corresponding username.

We obtained a folksonomy with $|U| = 75,242$ users, $|T| = 533,191$ tags and $|R| = 3,158,297$ resources, related by in total $|Y| = 17,362,212$ tag assignments. We created monthly snapshots as follows. $\mathbb{F}_0$ contains all tag assignments performed on or before June 15, 2004, together with the involved users, tags, and resources; $\mathbb{F}_1$ all tag assignments performed on or before July 15, 2004, together with the involved users, tags, and resources; and so on until $\mathbb{F}_{13}$ which contains all tag assignments performed on or before July 15, 2005, together with the involved tags, users, and resources.

Figure 4.1 shows the evolution of the ranking of all users tags that were among the Top 10 in at least one month for the topic 'music'. The diagram was obtained with $d = 0.85$, and the preference vector $\vec{p}$ set such that the tag 'music' gets 50% of the overall preference, the rest is spread uniformly as described above. The user names have been omitted for privacy reasons. The diagram shows three outstanding users. The first one could keep the top position for the first four months, followed by a steep fall. Another user could approach him steadily during the first four months, followed by almost the same fall. The fall of both was caused by the steep rise of a new user, which also shadowed the rankings of all other users related to 'music'. A detailed analysis of this user's data in the system revealed us that he posted more than 5500 bookmarks, 85 % of which tagged with 'music'. In total he used only about 100 tags. The 5500 bookmarks account for about 2% of *all* occurrences of 'music' in the system (with more than 70.000 users in the system at that time), and are about 3.5 times as many as those of the second user for that tag.

Figure 4.2 shows the evolution of all tags that were among the Top Ten in at least one month for the topic 'politics'. The line for the topic 'politics' itself can't be seen, as it has a rank of 1. The diagram was obtained with $d = 0.85$, and the preference vector $\vec{p}$ set such that the tag 'politics'

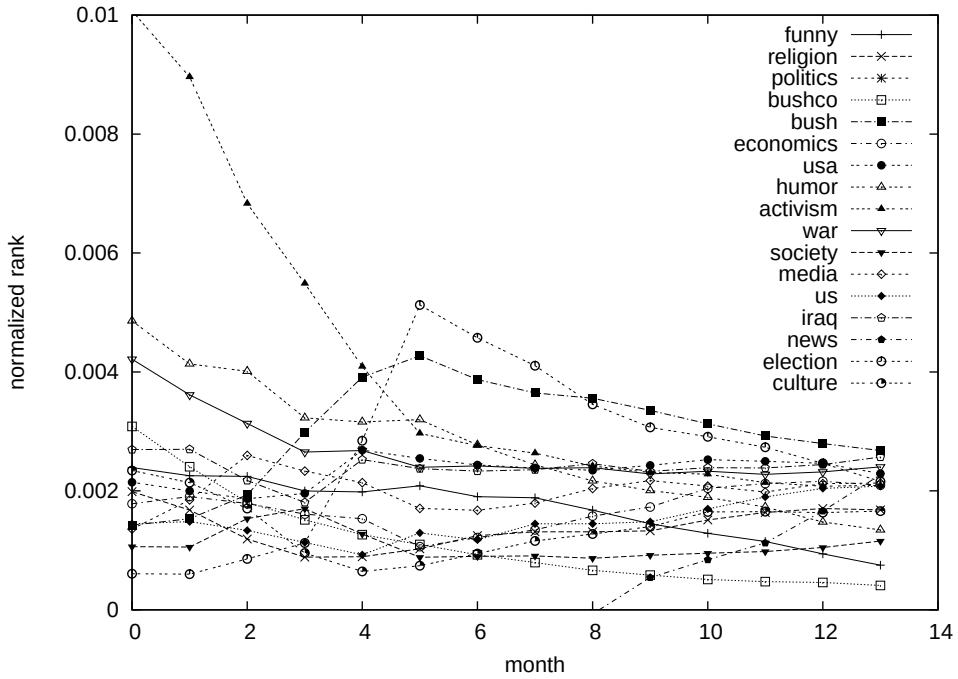

Figure 4.2: Evolution of the ranking of tags related to 'politics' over time. 'Politics' has value 1.0 due to normalization and is left out for clarity of the presentation of the other values.

gets 50% of the overall preference, while the rest is spread uniformly over the other tags, users and resources. The diagram shows that the early users of del.icio.us were more critical/idealistic, as they used tags like 'activism', 'humor', 'war', and 'bushco'[4]. With increasing time, the popularity of these tags faded, and the tags turned to a more uniform distribution, as the closing lines at the right of the figure indicate. In particular one can discover the rise of the tags 'bush' and 'election', both having a peak around the election day, November 2nd, 2004, and remaining on a high level afterwards. Within the analysis of the topic 'technology' (not displayed due to space reasons), we have discovered a similar trend: The early adopters of del.icio.us used the tag 'technology' together with tags like 'culture', 'society' or 'apple', while later tags like 'gadgets', 'news' or 'future' rise, converging towards more mainstream topics.

Both Figures 4.1 and 4.2 show that there is a change of structure in autumn 2004 (month 4 in the diagrams). This is supported by Figure 4.3 showing the development of the top resources. Analyzing possible reasons for this change in behavior, one indicator is that the number of elements passed in month 4 the threshold of 10.000 users, 70.000 tags, and 500.000 resources. Apparently, with this number of users, one reaches a critical mass which modifies the inherent behavior of such a system. Figure 4.3 shows the rank of those resources which where among the top 5 at the beginning or the end. Our hypothesis that the del.icio.us community changes significantly at month 4 is supported by two observations: specific topics of the beginning, such as web design, see a decline, while on the other hand, mainstream pages gain rapidly, such as Slashdot, as well as pages concerned with folksonomies per se.

Finally, we have analyzed those resources that were the strongest winners and losers within specific topics, according to the popularity change measure defined in Section 4.1.2, to automatically identify trends within certain topics in del.icio.us. Our aim was to discover trends in the Semantic Web community in the month around the European Semantic Web Conference (ESWC) 2005.

For the computation, we took those resources that ended up in the Top 100 in the June 2005

[4] In del.icio.us, 'bushco' was used for tagging webpages about the interference of politics and economics in the U. S. administration.

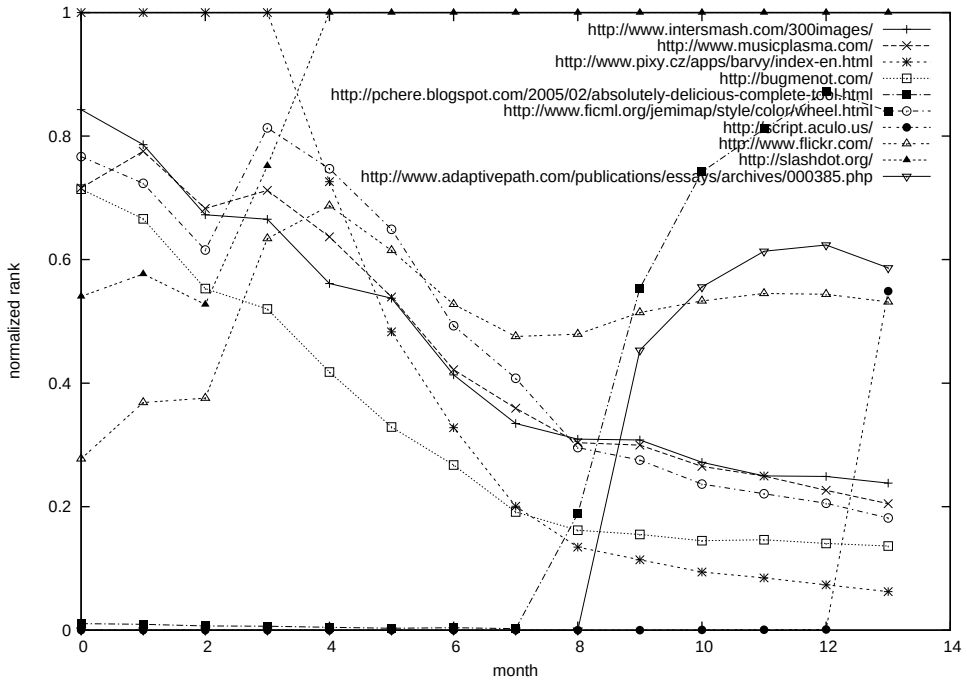

Figure 4.3: Evolution of the global ranking of resources, without specific preference vector.

ranking for the preference vector highlighting the tags ‘semantic_web’, ‘semantic’, ‘web’, and ‘semanticweb’, since the top results e. g. in searches are typically the ones attracting the most users. For these 100 URLs, we computed the popularity change coefficient from May 15 to June 15. Table 4.1 shows the 20 ULRs with the highest popularity change.

The top winner (#1) is a site about shallow semantic markup in XHTML, which was obviously first discovered by the community during the period under consideration and made it to the 39th position out of 2.2M resources; the corresponding line in Table 4.1 shows that the FolkRank value and the position in May is undefined for this resource, while the rank in June is 0.13065 and the position in this ranking is 39. Among the followers are articles about the Semantic Web and folksonomies (e. g. #2, #3, #5, #8, #9, #16), pages about new Semantic Web projects (#4, #15, #17, #19), or events such as the Scripting workshop (#7) that took place together with the ESWC conference during the period under consideration, introducing new Semantic Web projects. Note that while the #1 page leaped from nowhere to the 39th position out of 2.2 million entries, the popularity change measure still honors movements at the top of the ranking: the Piggy Bank site (which is an important semantic web project that has been promoted at the ESWC conference), improving from 21 to 1 in the period, still gets into the top 15 winners.

Together, the results of the FolkRank computation and the popularity change measure presented in this section can thus be used to get an insight into the structure and development of communities in folksonomy systems, independent of and across different media types.

4.2.2 Comparison with the Interestingness of Dubinko et. al.

Closest to the approach of this paper is the visualization of Dubinko et. al. (Dubinko et al., 2006). We tried to get an insight into how our FolkRank compares to the interestingness of (Dubinko et al., 2006). In that paper, the authors introduce an efficient way of mining large-scale folksonomy data sets for frequent tags in given time intervals. A measure of *interestingness*, is introduced and computed for a sliding one-day window over a Flickr dataset. Similar to the TF/IDF measure from Information Retrieval, the interestingness is defined as $\text{Int}(o, I) = \sum_{i \in I} \gamma(o, i)/(C + \gamma(o))$, where

Table 4.1: Popularity Change from May 15 to June 15, 2005.

#	URL	Pop.Chg.	May		June	
			Rank	Pos	Rank	Pos
1	http://mezzoblue.com/downloads/markupguide/	4.822604	undef	undef	0.13065	39
2	http://www.betaversion.org/~stefano/linotype/news/89/	4.515983	undef	undef	0.08296	79
3	http://shirky.com/writings/ontology_overrated.html	0.073704	0.00866	28598	0.39329	4
4	http://simile.mit.edu/piggy-bank/index.html	0.000805	0.05160	377	0.18740	24
5	http://www.dlib.org/dlib/april05/hammond/04hammond.html	0.000183	0.08831	142	0.09532	61
6	http://www.w3.org/2004/02/skos/	0.000175	0.08282	155	0.08369	78
7	http://www.semanticscripting.org/SFSW2005/	0.000134	0.09427	124	0.09055	67
8	http://www.scientificamerican.com/article.cfm?...	0.000133	0.08396	152	0.07208	97
9	http://jena.hpl.hp.com/~stecay/papers/xmleur...	0.000129	0.09979	111	0.09990	56
10	http://www.tantek.com/presentations/2004ete...	0.000112	0.09047	137	0.07407	92
11	http://users.bestweb.net/~sowa/peirce/ontometa.htm	0.000111	0.10273	106	0.09550	60
12	http://www.sciam.com/print_version.cfm?articleID=...	0.000101	0.09608	121	0.08178	81
13	http://www.xml.com/pub/a/2001/01/24/rdf.html	0.000089	0.09391	127	0.07314	94
14	http://developers.technorati.com/wiki/hCalendar	0.000071	0.09748	117	0.07389	93
15	http://simile.mit.edu/piggy-bank/	0.000057	0.29151	21	1.00000	1
16	http://en.wikipedia.org/wiki/Semantic_web	0.000033	0.10472	102	0.07186	98
17	http://www.semanticplanet.com/	0.000025	0.10893	91	0.07510	90
18	http://pchere.blogspot.com/2005/02/absolute...	0.000023	0.18154	41	0.13729	35
19	http://swoogle.umbc.edu/	0.000022	0.16785	48	0.12429	43
20	http://www.scientificamerican.com/print_ve...	0.000022	0.13367	68	0.09142	66

$\gamma(o, i)$ is the number of occurrences of object o in time interval i out of a larger interval I, and $\gamma(o)$ is the total number of occurrences of o. As the interestingness is based on a count of occurrences of items[5] in a given interval, it does not allow for an easy integration of topic-specific rankings. Thus, one obtains a ranking of one particular tag (user, resource), which does not generalize to related elements of the folksonomy.

We computed the equivalent of Figure 4.3 for the interestingness measure, i.e., we show the rankings for those resources that were within the Top 5 for any of the months. As our time window was one month, we used $C = 1500$ instead of $C = 50$ as in the original paper which used a one-day window. For lack of space and because the diagram did not yield any clear structure, we omit the diagram and summarize the findings.

The top resources were more volatile than in our method. I. e., in our approach, ten different resources made up the top five over all months. In the interestingness computation, there were 70 resources, i.e. each month had a new top five; Table 4.2 shows the top resource for each month. This indicates that the interestingness is more sensitive to momentary changes in the folksonomy than the FolkRank, and makes it harder to discover long- and medium-term trends. In the top resources, there were few general interest pages such as Slashdot or Flickr. Instead, there were more sites that seemed to be popular at one particular moment in time, but to fade soon afterwards. Figure 4.4 presents those four resources out of the 70 that overlap with Figure 4.3. It can be seen that while the interestingness shows some more jitter, the results have the same general direction for both computations.

We conclude that the interestingness, while more scalable and lending itself to a sliding-window visualization as in (Dubinko et al., 2006) due to its computational properties, lacks the dampening and generalizing effect of the FolkRank computation, so that it is more useful for short-term observations on particular folksonomy elements.

4.3 Related Work

There are currently only very few scientific publications about folksonomy-based web collaboration systems. Among the rare exceptions are (Dubinko et al., 2006) as discussed above, (Hammond et al., 2005) and (Lund et al., 2005) who provide good overviews of social bookmarking tools with special emphasis on folksonomies, and (Mathes, 2004b) who discusses strengths and limitations

[5] In (Dubinko et al., 2006), only tags are evaluated. Still, the method can be applied symmetrically to users and resources.

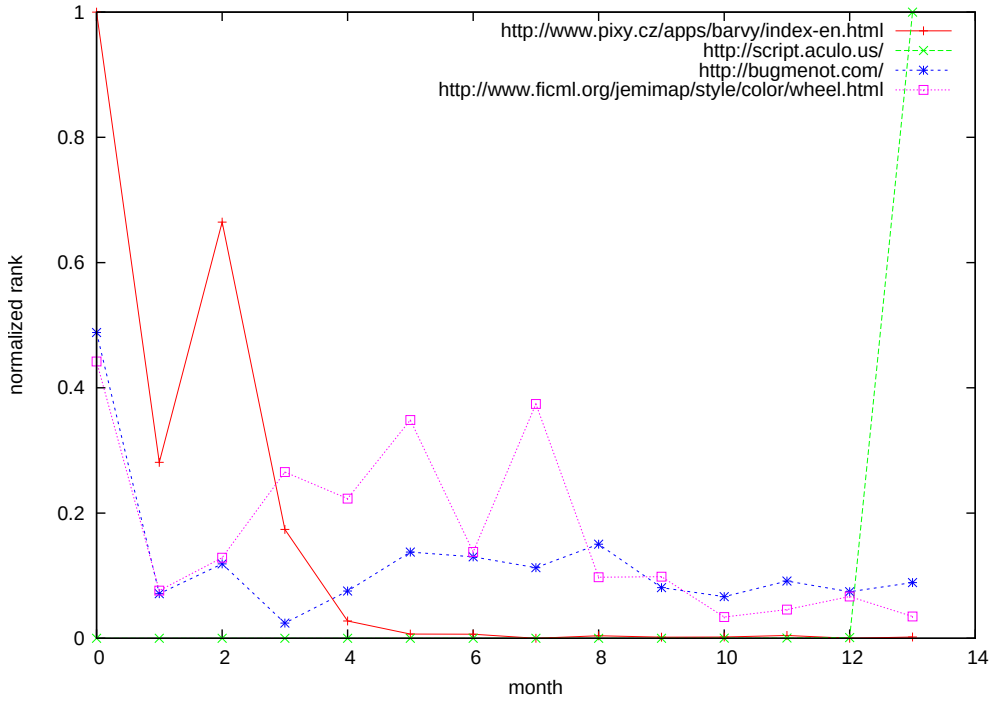

Figure 4.4: Evolution of the interestingness values of those resources which overlap with Figure 4.3; graph plotted the same way as Figure 4.3.

Table 4.2: Top resources for each month according to the interestingness defined in (Dubinko et al., 2006)

Month	Resource	Int'ness
0	http://www.pixy.cz/apps/barvy/index-en.html	0.1937
1	http://craphound.com/msftdrm.txt	0.0970
2	http://extensions.roachfiend.com/howto.html	0.1339
3	http://richard.jones.name/google-hacks/gmail-filesystem/gmail-filesystem.html	0.1983
4	http://37signals.com/papers/introtopatterns/	0.2150
5	http://www.fuckthesouth.com/	0.1898
6	http://www.supermemo.com/articles/sleep.htm	0.2585
7	http://www.returnofdesign.com/spectacle/specials/colors.html	0.2958
8	http://www.hertzmann.com/articles/2005/fables/	0.4117
9	http://fontleech.com/	0.4906
10	http://pro.html.it/esempio/nifty/	0.6511
11	http://www.alvit.de/vf/en/essential-. . . -developers.html	0.5678
12	http://www.newscientist.com/channel/being-human/mg18625011.900	0.6222
13	http://script.aculo.us/	0.8478

of folksonomies. The main discussion on folksonomies and related topics is currently only going on mailing lists, e.g. (Connotea). In (Mika, 2005), Mika defines a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Besides calculating measures like the clustering coefficient, (local) betweenness centrality or the network constraint on the extracted one-mode network, Mika uses co-occurrence techniques for clustering the concept network.

There are several systems working on top of del.icio.us to explore the underlying folksonomy. CollaborativeRank[6] provides ranked search results on top of del.icio.us bookmarks. The ranking takes into account, how early someone bookmarked an URL and how many people followed him or her. Other systems show popular sites (Populicious[7]) or focus on graphical representations (Cloudalicious[8], Grafolicious[9]) of statistics about del.icio.us.

The tool Ontocopi described in (Alani et al., 2003) performs what is called Ontology Network Analysis for initially populating an organizational memory. Several network analysis methods are applied

[6] http://collabrank.org/
[7] http://populicio.us/
[8] http://cloudalicio.us/
[9] http://www.neuroticweb.com/recursos/del.icio.us-graphs/

to an already populated ontology to extract important objects. In particular, a PageRank-like (Brin and Page, 1998) algorithm is used to find communities of practice within individuals represented in the ontology. OntoRank (Ding et al., 2005) uses a PagesRank-like approach on the RDF graph to rank search results within Swoogle, a search engine for ontologies.

Along the same line, in (Hoser et al., 2006), we have presented a technique for analyzing ontologies that considers not only the first eigenvector (as PageRank and Ontocopi do), but the full eigensystem of the adjacency matrix of the ontology.

In (Amitay et al., 2004), the evolution of the web graph over time is analyzed. The application of the proposed method lies in the improved detection of current real-life trends in search engines. In comparison to our work, they base their approach on counting timestamped links on pages returned by web searches on given topics, while our contribution infers communities around given users, sites, or topics from the structure of the web graph itself. The algorithm of (Amitay et al., 2004) can currently not be applied to folksonomies, as there exist no folksonomy search engines yet.

Kleinberg (Kleinberg, 2006) summarizes several different approaches to analyze online information streams over time. He distinguishes between three methods to detect trends: using the normalized absolute change, relative change and a probabilistic model. The popularity gradient that we introduced in Section 4.1.2 is related to the second approach, but differs insofar as it allows for the discovery of *topic-specific* trends, and that we honor steep rises more if they occur higher in the ranking, where the text mining scenario described in (Kleinberg, 2006) requires focusing on words that are neither too frequent nor too infrequent.

4.4 Markov CLustering algorithm (MCL)

Site ranking algorithms, as for instance the PageRank algorithm (Brin and Page, 1998), use topological information embedded in a directed network to infer the relative importance of nodes. In chapter 4 and particularly in section 4.1.1, we introduced a node ranking procedure for folksonomies, the FolkRank algorithm (Hotho et al., 2006a). In contrast with PageRank, FolkRank provides useful informations also in the case of undirected networks. Taking on a different perspective, community detection algorithms can be employed to attempt the detection of relation similarities at a higher level. We showed an example of a community identification procedure in section 3.4. A yet different procedure is the Markov Clustering algorithm (MCL), in which a renormalization-like scheme is used in order to detect communities of nodes in weighted networks. In this section, we analyze the commonalities of the two approaches. In particular we begin to address the relationships between ranking and community identification in folksonomies.

4.4.1 MCL at work

MCL (van Dongen, 2000) is a procedure based on iterative operations on the adjacency matrix associated with a graph. We recall that the adjacency matrix A of a network is the square matrix of dimensions n – with n equal to the order of the graph, i.e. the number of nodes in the graph – that has binary-valued entries $A_{ij} = 1$ if there exist a link connecting nodes i and j, and $A_{ij} = 0$ otherwise. In the case where links connecting nodes have a weight associated with them, then A_{ij} equals the weight w_{ij} of the link connecting nodes i and j. The MCL algorithms proceeds as follows:

1. The adjacency matrix is normalized so that, for each column, the sum of its entries is equal to unity; in this way one builds the so-called normal matrix N of the graph, known also as transition matrix; Although the normal matrix is no more symmetric, its eigenvalues are real;

2. The normal matrix gets squared, i.e. a new matrix $M = N^2$ is built;

3. Each element of the matrix M, which is also a transition matrix with normalized columns, is then raised to a power $\gamma > 1$; this step has the effect of reducing the importance of the smaller column elements with respect to the larger ones.

4. the resulting contracted matrix is normalized and the procedure of point 2 is repeated until convergence is reached.

The physical interpretation of the MCL procedure is quite natural in terms of random walkers moving on the network. Each element N_{ij} of the normal matrix can be viewed as the probability that a random walker sitting on node j jumps to one of its nearest neighbors. This probability is by construction proportional to the link weight w_{ij}. Each element of the squared matrix N_{ij}^2, in turn, gives the probability that the random walker jumps to the second neighbor i of node j in two steps. The successive contraction has the effect to dump all those less probable paths. As a result of the convergence, the walker will tend to stay inside (if existing) islands of nodes, which one can interpret as communities. A sketch of the MCL algorithm at work is displayed in Fig. 4.5.

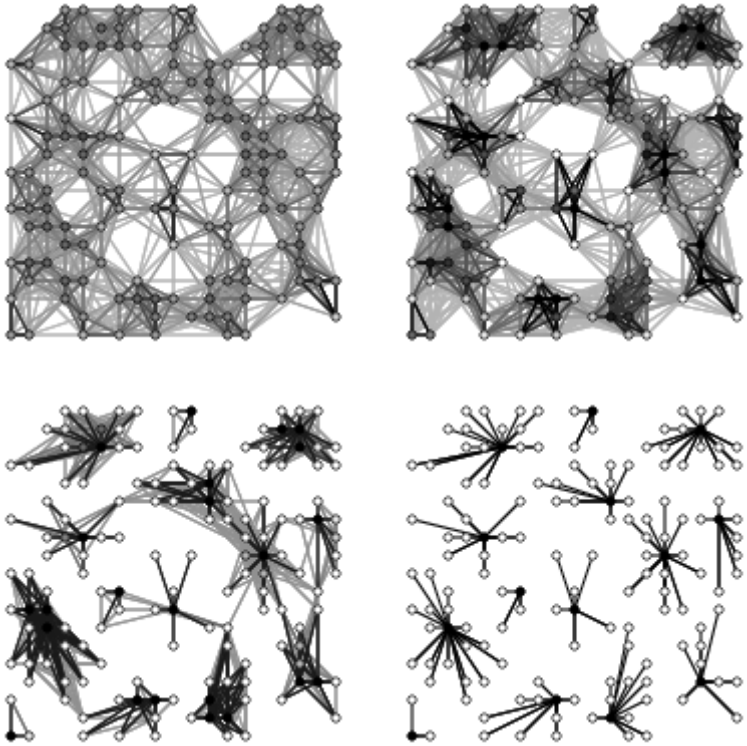

Figure 4.5: Illustration of a typical MCL procedure applied to an unweighted network. Here the contraction parameter γ has the default value $\gamma = 2$. The original network is displayed in the upper left frame, while the result of the procedure, after convergence, is the disjoint, star-like graph shown in the lower right frame. This image was taken from the official MCL web site `http://micans.org/mcl/`. For more information, see also the appendix A.7.

4.4.2 Comparison between MCL and FolkRank

Since the FolkRank algorithm, too, can be viewed in terms of random walkers on a graph, it seems rather natural to investigate how the MCL and FolkRank algorithms are related to each other. One main difference is that FolkRank only provides information about the community to which a given node belongs to (by relating it to the highest ranked node in the same community), while MCL

provide global information on the community structure of the network, at least in principle. The advantage of FolkRank is that it also provides a *ranking* of results and thus it may have an important impact in improving the navigability of folksonomies. Finally, since FolkRank only provides local information, it scalability with the size of the network is much better than the scalability of MCL, and this makes it more appealing for deployment in a production system. In order to empirically test the similarities between the two methods, we performed some preliminary experiments, described in the following. We ran MCL on a subset of the `Bibsonomy` weighted tag co-occurrence network and obtained a list of star-like clusters of tags, which were labeled according to the node (tag) lying at the center of the star. Then, we selected an arbitrary tag X ("`sciencefiction`") and computed the FolkRank value of all the tags of the network, using a preferential vector pointing towards the chosen tag X (see section 4.1.1). The idea of the experiment is to explore the relation between the computed FolkRank values and the community structure identified via MCL: For each tag cluster identified by MCL we computed the average FolkRank value of its member tags. Fig. 4.6 shows the result of this analysis, and we can clearly recognize that FolkRank is indeed correlated with the communities found by MCL, since the cluster to which the tag X ("`sciencefiction`") belongs to has the highest FolkRank value. Some of the entries in Fig. 4.6 seem rather spurious,

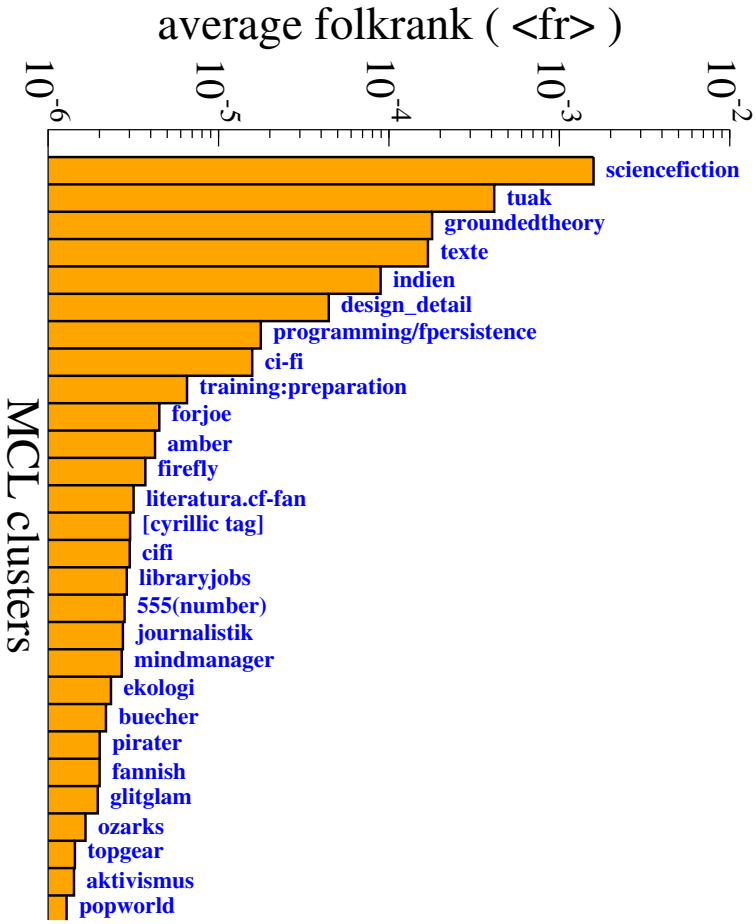

Figure 4.6: The communities of tags detected by MCL (ran with default parameters) are labeled with the tag of the node lying at their respective centers. For each cluster, the average FolkRank of its nodes is shown as a horizontal axis value.

e.g. the entry "`tuak`". In Fig. 4.7 we show that this is mainly due to an excessive importance attributed to clusters of small size.

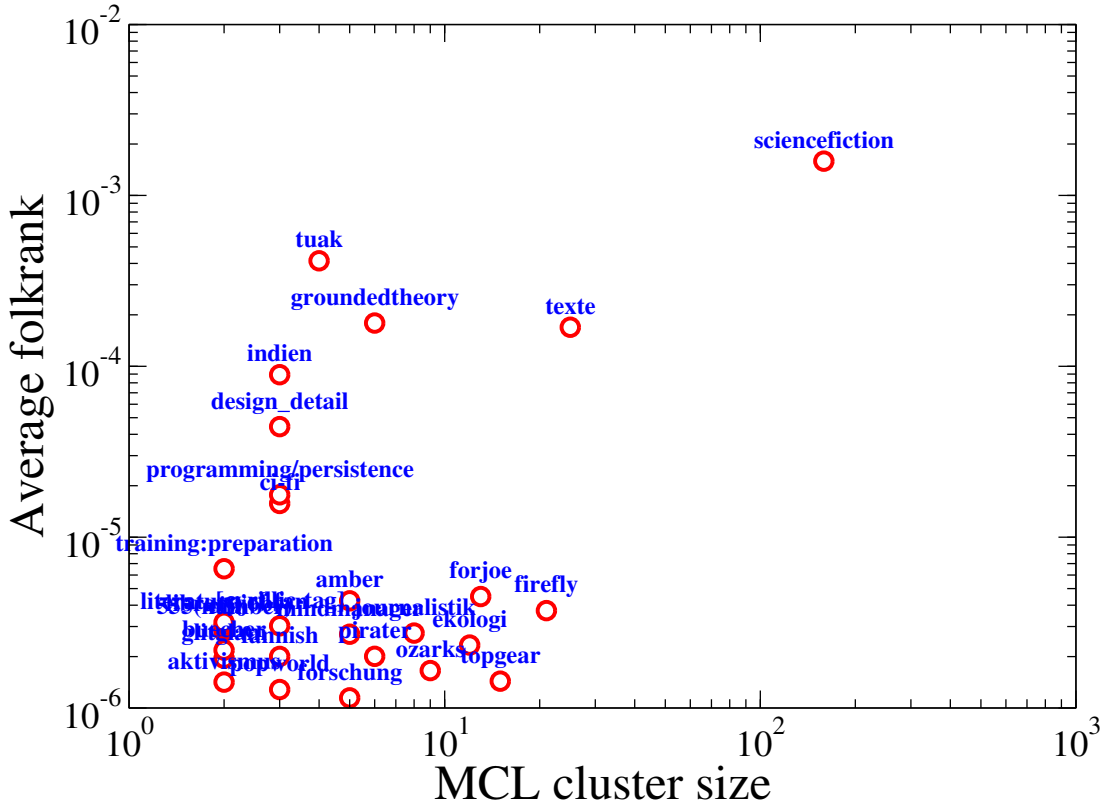

Figure 4.7: Size (number of nodes) of the communities detected by MCL, labeled according to the tag that corresponds to the central node of each cluster.

Chapter 5

Exploiting the Semantics of Tag Co-Occurrence

5.1 Automatic Organization of Tags

In this chapter we will present the Tag-Organizer system (T-ORG) which automatically classifies the resources of a tagging system into predefined categories and thus can provide a better browsing experience to the user.

The classification of resources is based on the classification of tags attached to these resources. If a resource has two tags belonging to two different categories, then the resource is classified as both of these categories. For example, if a resource has the tags “Paris” and “Peugeot” and these tags are classified as “Location” and “Vehicle” respectively, then the resource is placed in both of these categories (i. e., Location and Vehicle). Tag classification can help a user to use tags on a tagging system in a more organized way. For example, instead of representing different tags in a tag cloud, sometimes it could be more useful, if a “Tag Cloud” displays the abstract tags (i. e., categories) and when a user clicks an “abstract tag”, its subsequent tags are displayed. In such way, a user can explore different type of tags (and hence resources) available on a tagging system, which might not be possible with a simple “Tag Cloud”.

The T-ORG system exploits two different sources of information. On the one hand, it exploits the semantics of the tag co-occurrence network. For this purpose, we define four different context in 5.2.2 which basically correspond to the related tags in the co-occurrence network. The contexts can be used for disambiguating tags. For example, if the tag “Ford” means the car than the context will very likely contain further tags from the vehicle category. But if the tag “Ford” means the former US president or another person with the name then the context will likely contain tags related to the person category. T-ORG also uses on the other hand semantic background knowledge coming from ontologies found on the web. For example, such an ontology may contain hierarchical knowledge like “Paris” is a city and a city is a location.

The disambiguation by means of the tag context was then used in a further step where different tags and/or resources were assigned to several predefined categories like *location* or *person*. During the experiment we also used semantic background knowledge for improving the results. The background knowledge came from ontologies found on the web from which we exploited the hierarchical knowledge that e.g. “Paris” is a city and that a city is a location. Both sources of information (the relations in the co-occurrence network and the background knowledge coming from ontologies) are then combined by T-ORG for categorizing tags into hierarchies.

The core of T-ORG is its classification method T-KNOW (Tag classification using KNowledge On the Web). It is based on an unsupervised mechanism for classifying tags in folksonomies. T-KNOW uses Google for finding categories of tags; therefore it does not require any training and can be used for unsupervised classification of tags (like (Cimiano et al., 2005)). It classifies the tags into

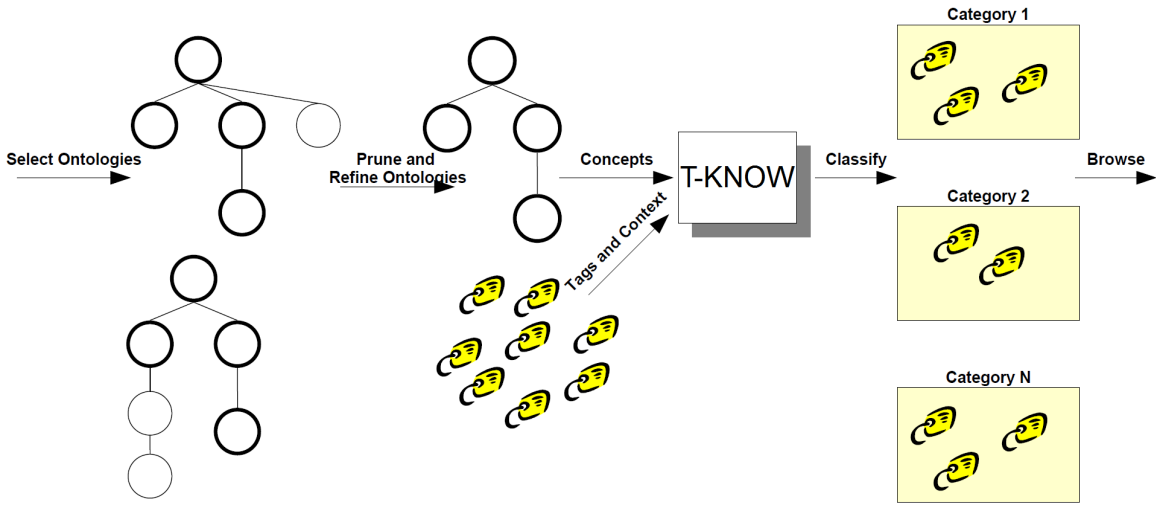

Figure 5.1: Process of T-ORG.

categories using its pattern library, categories extracted from a given ontology and Google search results. As there might be several results returned by Google against a query posed by T-KNOW, a method is required to select best results on the basis of the similarity between tagging and search results. T-KNOW uses the context of the tag to measure the similarity between Google search results and the tag. We also propose four methods of selecting the context of a tag.

5.1.1 Process of the Tag-Organizer (T-ORG)

The purpose of T-ORG is to organize resources by classifying their tags into categories. This process is done by selecting concepts from single or multiple ontologies related to the required categories and then pruning and refining these ontologies. These concepts are considered as categories into which the tags are classified. Fig. 5.1 shows the overall process of T-ORG while each step is described below.

Selecting Ontology

The user of T-ORG has to decide about the categories into which the resources are to be classified. The user selects ontologies relevant to the required categories. Concepts from these ontologies are used as categories. For example to browse through the images of vehicles at Flickr, one would select a vehicle ontology. Currently this step is done manually in T-ORG.

Pruning and Refining Ontology

After selecting ontologies, they must be pruned and refined for the desired categories. Only those concepts from these ontologies are considered which have some relation to the required categories. Unwanted concepts are pruned. Redundant and conflicting concepts are refined. Missing concepts are also added into the given ontology. For example to include the images of a “draisine”, one might have to add this concept into a given vehicle ontology. Once the ontology is pruned and refined, its concepts are used as categories. Currently this step is also done manually in T-ORG.

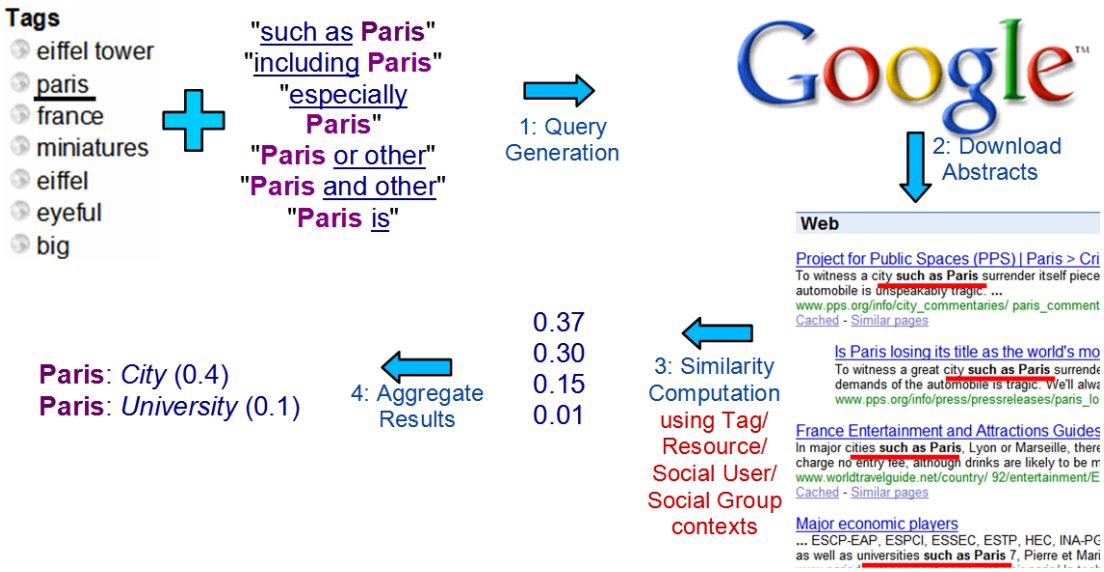

Figure 5.2: Process of T-KNOW.

Applying T-KNOW for Classifying Tags

Classifying the tags is a major step in the process of T-ORG. After the ontology is selected, pruned, refined, and categories are extracted from it, these categories and the context of tags are used for classification. Once all tags are classified into categories, each category is subsumed by its parent category, for example, every tag classified as Train, Bulldozer or Bus is finally classified as Vehicle. Section 5.2 describes the detailed process of classifying tags using T-KNOW.

Browsing the Resources

After classifying each tag, resources may be browsed according to the categories assigned to their tags. The browser may use information of resources to display them in categories, so that the user can browse particular type of resources present in these categories.

5.2 Tag classification using Knowledge On the Web (T-KNOW)

T-KNOW uses lexico-syntactic patterns and Google APIs for finding the appropriate categories of the tags. Given a list of tags and categories, T-KNOW classifies these tags into categories. It builds queries by combining linguistic patterns (Hearst Patterns (Hearts, 1992) and a few more (Cimiano et al., 2005)) and the categories and then searches these queries on Google using Google API. The process of classifying tags using T-KNOW is shown in Fig. 5.2. Following, we describe in more detail the steps shown in Fig. 5.2.

Assume the tag “Paris” is to be classified in a context as depicted in Fig. 5.3

Step 1 Queries are generated by concatenating the tag and the clues, e.g. “such as **Paris**” is a query generated by combining the clue “such as” and the tag “Paris”

Step 2 The queries are searched using the Google API and abstracts of search results are downloaded, e.g. “To witness a **city** such as *Paris* surrendered itself. . . ” is a search result abstract downloaded for the query “such as Paris”

Step 3 The similarity between each abstract and context (described in Section 5.2.1) of tag is computed, e.g. between the abstract "To witness a city such as Paris..." and context of the tag "Paris" (eiffel tower, france, miniatures...). If similarity is above a certain threshold value, then depending upon the clue used, the abstract is matched against the pattern, e.g. the abstract "To witness **city** such as *Paris*..." is matched against the Hearst pattern (Hearts, 1992) "**CONCEPT** such as (*INSTANCE*,?)+ ((and|or) INSTANCE)", where CONCEPT is the expected category and INSTANCE is the tag. Hence "City" is extracted as an expected category of the tag "Paris" from this abstract.

Step 4 The results are aggregated and the category having highest similarity with the tag's context is returned, e.g. for the tag "Paris" the category "City" is returned, because it has higher similarity than the other category e.g. "University"

5.2.1 Measuring similarity between search results and tags

There can be multiple ways for computing the similarity between the search result and the tag depending upon the context of the tag. We have proposed four methods of selecting the context of the tag. For measuring similarity between Google search result and the context of a tag, the cosine measure is computed between the bag of word representations of the abstract of the downloaded search result $\vec{a}$ and the context $\vec{C}$ of the tag t. These vectors simply represent the words and their frequencies (occurrences), and a cosine measure between them represents that how similar they are. If the cosine measure is above a certain threshold value, the result is considered for further processing. The cosine measure is calculated as

$$cos(\angle(\vec{C},\vec{a})) = \frac{\vec{C}\cdot\vec{a}}{\|\vec{C}\|\cdot\|\vec{a}\|} \tag{5.1}$$

Section 5.3.2 presents different results obtained using different threshold values and different contexts. To understand the different contexts, consider the images in Fig. 5.3. The left most image shows the "Eiffel Tower" while the image in the middle shows "Notre Dame" and the right most a cow.

Tagging system (or Folksonomy) is formally defined in section 2.1, in addition to it, we also use the set of groups G that might be found in some tagging systems (like Flickr). Users can post their resources to these groups. In the following we will define four different contexts which help in categorizing tags appropriately.

5.2.2 Context Definitions

Resource Context (R)

In order to represent a tag by its context, we here consider the case of resource context. We choose the tags that belong to the current resource except the tag (to be classified) itself. The Resource Context of tag t for resource r can be defined as

$$C_R(t,r) = \{t' \in T\backslash\{t\} \mid (u,t',r) \in Y \wedge u \in U\} \tag{5.2}$$

We are also interested in the frequency of t_i in resource r_j (in case of Flickr it is at most 1, because one tag can occur only at most once in a resource) to create a bag of words using this context. $W_R(t,r)$ represents the number of times tag t appears with resource r.

$$W_R(t,r) = \mid \{(u,t,r) \in Y \mid u \in U\} \mid \tag{5.3}$$

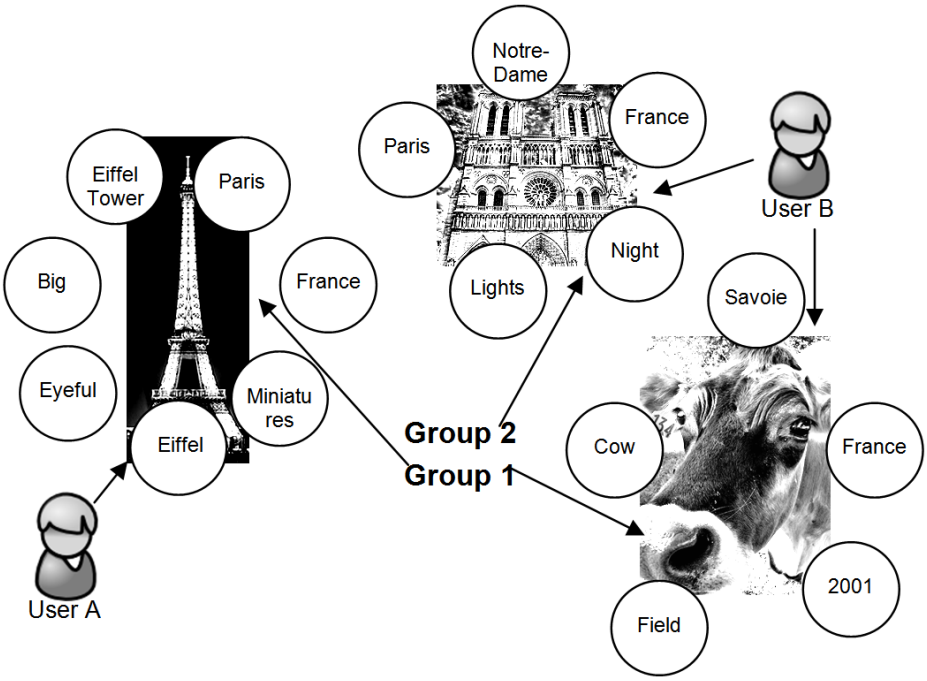

Figure 5.3: Sample Images with Tags.

We can get the Resource Context of a tag t of resource r using $C_R(t, r)$ and for each tag t in the Resource Context of tag t, we can get its number of occurrences in resource r using $W_R(t, r)$. We can define a bag-of-words resource context representation of a tag t appearing in resource r, i. e., by

$$B_R(t, r) = \{(t', W_R(t', r)) \mid t' \in C_R(t, r)\} \tag{5.4}$$

Note that B_T, B_{SU}, and B_{SG} can be defined in the similar manner for Tag, Social User, and Social Group contexts respectively. Consider that we want to classify the tag "Paris" of the Eiffel-Tower image in Fig. 5.3, only the tags of the Eiffel-Tower image are selected as the context. With the Eiffel-Tower image denoted as r_1 we get the following resource context: $C_R(\textit{Paris}, r_1) = \{\textit{Eiffel Tower, France, Miniatures, Eiffel, Eyeful, Big}\}$. The bag-of-words representation of the tag "Paris" of Eiffel-Tower will be $B_R(\textit{Paris}, r_1) = \{(\textit{Eiffel Tower}, 1), (\textit{France}, 1), (\textit{Miniatures}, 1), (\textit{Eiffel}, 1), (\textit{Eyeful}, 1), (\textit{Big}, 1)\}$.

Tag Context (T)

In case of Tag Context, we select all the tags joint to the resources having the tag t, except the tag t itself. Tag Context can be defined as

$$C_T(t) = \{t' \in T \backslash \{t\} \mid (u, t, r) \in Y \wedge (u', t', r) \in Y \wedge u \in U \wedge u' \in U \wedge r \in R\} \tag{5.5}$$

In parallel to equation 5.4 we define a bag of words representation using this context. We define $W_T(t, t')$ that represents the number of times tag t appears with tag t'.

$$W_T(t) = \mid \{(u', t', r) \in Y \mid (u, t, r) \in Y \wedge u \in U \wedge u' \in U \wedge r \in R\} \mid \tag{5.6}$$

We can get the Tag Context of a tag t using $C_T(t)$ and for each tag t' in the Tag Context of tag t, we can get its number of occurrences with tag t using $W_T(t, t')$.

Consider that we want to classify the tag "Paris" of the image Eiffel-Tower. All tags of images having the tag "Paris" are selected as the Tag Context except the tag "Paris" itself. In example of Fig. 5.3, Eiffel-Tower and Notre-Dame have the tag "Paris", so all the tags of the images Eiffel-Tower and Notre-Dame are added to the context of the tag "Paris" except the tag "Paris" itself, and number of occurrences of each of these tags with tag t can be calculated using W_T. Thus, $B_T(\textit{Paris}) = \{(\textit{Eiffel Tower}, 1), (\textit{France}, 2), (\textit{Miniature}, 1), (\textit{Eiffel}, 1), (\textit{Eyeful}, 1), (\textit{Big}, 1), (\textit{Notre-Dame}, 1), (\textit{Night}, 1), (\textit{Lights}, 1)\}$ is the bag-of-word representation constructed using Tag Context of the tag "Paris".

Social User Context (SU)

In case of Social User Context of a tag t, we select all the tags used by a user u, except the tag t itself. Social User Context of tag t of user u can be defined as

$$C_{SU}(t, u) = \{t' \in T \backslash \{t\} \mid (u, t', r) \in Y \wedge r \in R\} \tag{5.7}$$

In parallel to equation 5.4 we define a bag of words representation using this context. We define $W_{SU}(t, u)$ that represents the number of times tag t is used be the user u.

$$W_{SU}(t, u) = \mid \{(u, t, r) \in Y \mid r \in R\} \mid \tag{5.8}$$

Consider that we want to classify the tag "Paris" of the image Notre-Dame that belongs to user B. All tags of images that belong to the user B are selected as the context except the tag "Paris" itself. In example of Fig. 5.3, the images Notre-Dame and Cow belong to the user B, so all the tags of the images Notre-Dame and Cow are added to the context of the tag "Paris" except the tag "Paris". Thus, $B_{SU}(\textit{Paris}, B) = \{(\textit{Notre Dame}, 1), (\textit{France}, 2), (\textit{Night}, 1), (\textit{Lights}, 1), (\textit{Savoie}, 1), (\textit{2001}, 1), (\textit{Field}, 1), (\textit{Cow}, 1)\}$ is the bag-of-word representation constructed using social user context.

Social Group Context (SG)

In case of Social Group Context of tag t that is present in group(s) g, we select all the tags of all resources present in the same group g, except the tag t itself. The Social Group Context can be defined as

$$C_{SG}(t, g) = \{t' \in T \backslash \{t\} \mid (u, t', r) \in Y \wedge u \in U \wedge r \in R \wedge g \in \text{ Group}(u, r)\} \tag{5.9}$$

where $Group(u, r)$ is a function which returns the groups that contain the user u and resource r. In parallel to equation 5.4 we define a bag of words representation using this context. We define $W_{SG}(t, g)$ that represents the number of times tag t appears in the group g.

$$W_{SG}(t, g) = \mid \{(u, t, r) \in Y \mid u \in U \wedge r \in R \wedge g \in \text{ Group(u,r)}\} \mid \tag{5.10}$$

Consider that we want to classify the tag "Paris" of the image Eiffel-Tower that belongs to group 1. All tags of images present in group 1 are selected as the context except the tag "Paris" itself. In example of Fig. 5.3, the images Eiffel-Tower and Cow are present in group 1, so all the tags of the images Eiffel Tower and Cow are added to the context of the tag "Paris" except the tag "Paris" itself. $B_{SG}(\textit{Paris, Group1}) = \{(\textit{Eiffel Tower}, 1), (\textit{France}, 2), (\textit{Miniatures}, 1), (\textit{Eiffel}, 1), (\textit{Eyeful}, 1), (\textit{Big}, 1), (\textit{Savoie}, 1), (\textit{2001}, 1), (\textit{Field}, 1), (\textit{Cow}, 1)\}$ is the bag-of-word representation constructed using social group context.

5.3 Evaluation

In order to evaluate our system, we have used images, tags, user, and group information from Flickr website. We asked two persons to classify the data into four categories. We have then classified the same data set using T-KNOW in order to evaluate T-KNOW.

5.3.1 Experimental Setup

To organize tags into predefined categories, we have chosen four categories “Person”, “Location”, “Vehicle”, and “Organization”. To get ontologies related to these categories, we have searched Swoogle (Ding et al., 2004) for general purpose ontologies and used the ontology OntoSem. For this ontology, we have used concepts and sub-concepts of p1:vehicle, p1:organization, p1:place, p1:geopolitical-entity, and p1:human as categories. We have used a total of 932 concepts as categories out of this ontology.

After selecting the categories, we have gathered data from groups present at the Flickr website. Users can post their images to different groups on Flickr. One group usually contains images related to the topic of that group. For example, the vehicles group contains images of vehicles. We have searched for groups related to the topics (i) people, (ii) locations, and (iii) vehicles using the group search facility provided by Flickr, and then selected three groups from each topic. We have selected only those groups which had at least 100 images and 25 members. The groups selected were candid_celebrity, 35212032@N00 (famous people), politicians, CarDirectory, classic_cars, vehicles, PraiseAndCurseOfTheCity, signcity, and cities. Out of these groups, only the “famous people” group had 27 members and 165 images, all other groups had at least 100 members and more than 500 images. We have then randomly selected 21 images from each of these nine groups. There were a total of 1754 tags in all of these 189 images.

We asked two persons K and S (human classifiers) to classify the tags. They did not have any kind of information about this research and method. They have classified all the tags regardless of the language and spelling mistakes, which has of course affected the results of T-KNOW because T-KNOW uses English patterns for identifying categories. For example, the users have classified the tags “Russia” and “Russland” (German word for Russia) as location, whereas T-KNOW was unable to identify “Russland”, as this is not an English word and hence is not supported by the pattern library used. A spreadsheet was provided to each human classifier with resources, tags, and links to the original Flickr images, Wikipedia, and Google. For example if a user finds a tag “Essen” (a German city as well as the German word for meal) and is unable to decide about its category, he can view the image (in which this tag is present) on Flickr website, if this image is not helpful to identify the tag, he can search it in Wikipedia, and still if it unclear, then he can find it in Google. Human classifiers (K and S) agreed upon classification of only 1166 tags out of 1754 tags.

5.3.2 Results

This section contains the results obtained by classifying tags using T-KNOW with different contexts and threshold values. We have used F-measure and Cohen’s Kappa for evaluation of our method. F-measure is a common measure in information retrieval, in case of tags classification we have computed F-measure as, if $A =$ set of correct classifications by test, $B =$ set of all classifications by Gold Standard, $C =$ set of all classifications by test (In our evaluation, user K is the gold standard, and test is either user S or T-ORG) then, we define Precision, Recall, and F-measure as

$$Precision = \frac{A}{C} \tag{5.11}$$

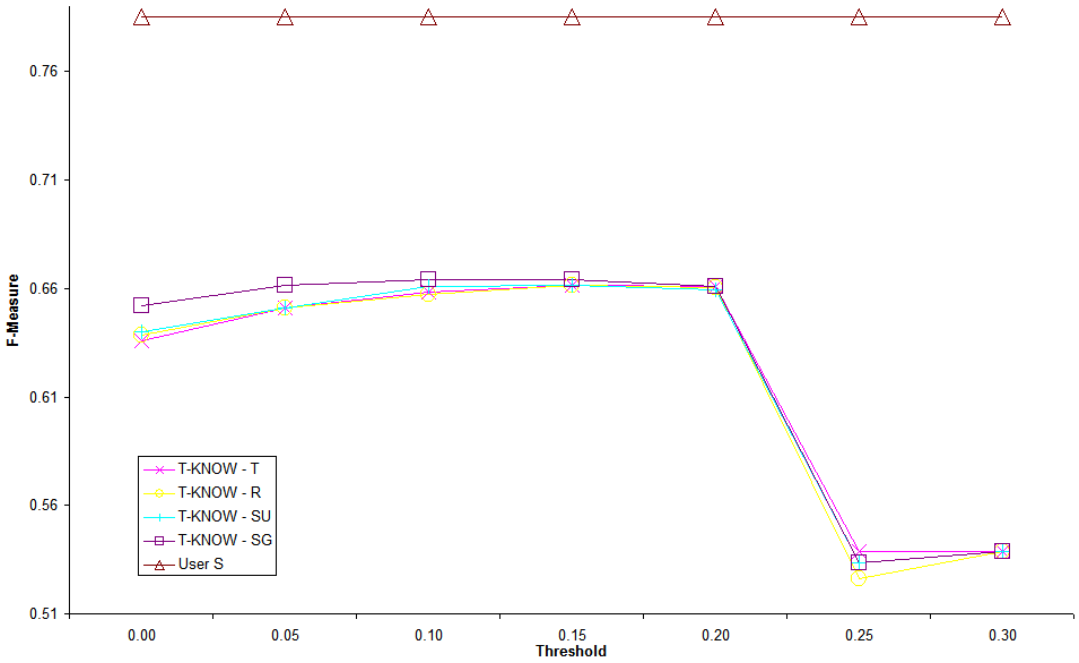

Figure 5.4: F-Measure with user K defining the gold standard.

$$Recall = \frac{A}{B} \tag{5.12}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.13}$$

Fig. 5.4 displays the F-measure with user K defining the gold standard and T-KNOW using different threshold values and contexts and it also shows the F-measure of the classification of user K and user S (shown as a constant line). Due to the possibility of classification that might occur just by chance, we have also calculated the Cohen's Kappa (Cohen, 1960) between a user's classification and the system's prediction. Cohen's Kappa is defined as

$$K = \frac{P_0 - P_c}{1 - P_c} \tag{5.14}$$

where P0 is the observed agreement between classifiers and Pc is the agreement occurred due to chance. If the two classifiers agree completely, then the value of Cohen's Kappa is 1. Fig. 5.5 shows the Kappa values of the classification of user K and T-KNOW (with different threshold values and contexts) and it also shows the Cohen's Kappa value between the classifications of user K and user S (shown as a straight line).

5.3.3 Discussion

The task of organizing resources by classifying tags in a tagging system is not trivial. It is observed that two humans classifying the same data set might not totally agree with each other, as observed in the case of humans classifiers of user K and user S, the Kappa value was 0.53, whereas this value would be 1 in case of complete agreement between classifiers.

The best F-measure obtained was 0.66 with the context Social Group (SG) at thresholds of 0.10 and 0.15 and this small advantage was stable over other thresholds except 0.25. The F-measure is affected by the problem of classification by chance. Therefore we have calculated Cohen's

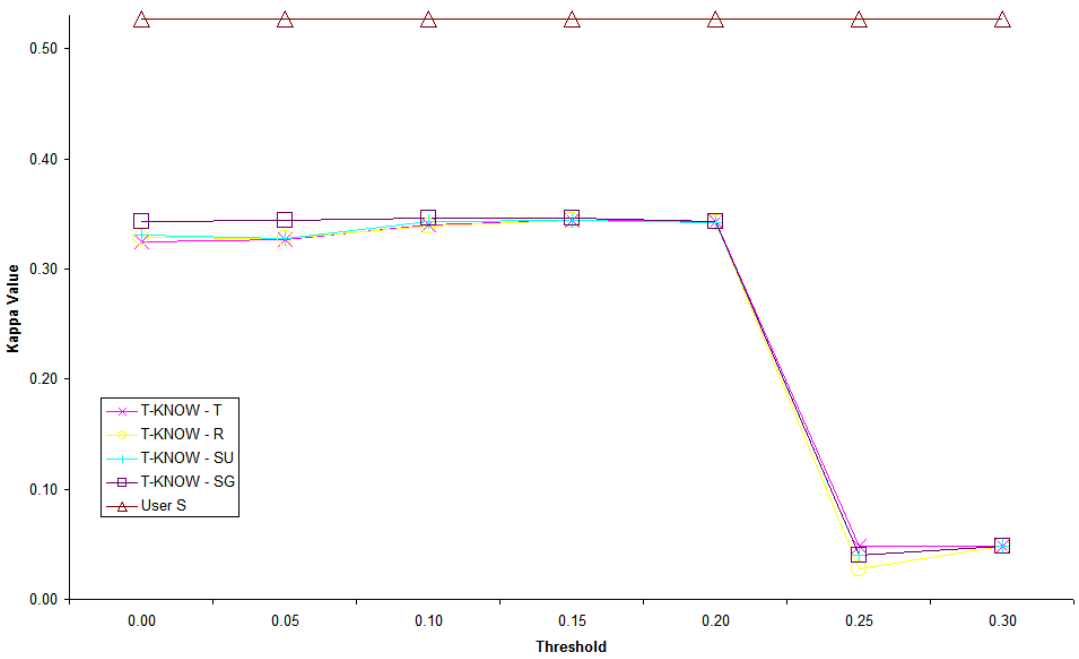

Figure 5.5: Cohen's Kappa values for classification of T-KNOW and User S with user K defining the gold standard.

Kappa (Cohen, 1960) to measure the agreement between two users and between T-KNOW and user K. The majority class ("Other" in our case) scores zero in Cohen's Kappa (Cohen, 1960). F-measure lacks this property. The Cohen's Kappa between classification of users K and S was 0.53 (shown as a straight line in fig. 5.5), which shows the disagreement between the classifications of human users. Best kappa value for gold standard (user K) was 0.35 with Social Group (SG) context and using threshold of 0.10 or 0.15.

The results show that, the different approaches for selecting a context are statistically not significantly different. Keeping in view the small difference between different approaches, Social Group (SG) context has given overall better results as compared to other contexts. This is because the tags which are chosen as context belong to the same type of resources/images (as a group mostly contains same type of resources). In case of other contexts, tags of the resources with different subjects might be selected as context, which might affect the similarity measure.

Chapter 6

Conclusion and Outlook

This deliverable described several measures for analyzing the datasets collected in WP1 of the TAGora project. In particular, we analyze the topological properties of folksonomy hypergraphs and tag co-occurrence graphs. The analysis reveals the small-world character of the network, identifies the undesired contribution of spammer users, and exhibits semantic correlations in tag co-occurrence.

A clustering technique applied for the classification of resources could reveal the shared semantics of tags, resulting from the uncoordinate activity of users.

At the same time, such an activity is dynamic in its very nature. A ranking algorithm specifically designed for folksonomy network (so called Folkrank), could give insight in this dynamics providing an effective trend detection. Incidentally, we presented a preliminary study, bridging this node ranking procedure with a known community detection algorithm in networks.

Finally, we describe an unsupervised system for the automatic classification of resources, named T-ORG. T-ORG is based on the categorization of tags, it makes use of ontologies found on the web, and, ultimately, realizes the classification using Google for associating tag to categories.

In the next year, our work will be further developed into two directions: On the one hand, we plan to further investigate clustering and social properties of folksonomies, e.g. for detecting user communities. On the other hand, we will investigate in how far it is possible to incorporate background knowledge into the data analysis.

Appendix A

Tools

This chapter offers an overview of the tools mostly used to obtain the results presented in this deliverable. It is not the goal of this chapter to give a detailed roundup of all available applications.

A.1 Net

License: Free for non-commercial use

OS: Linux only

URL: `http://pil.phys.uniroma1.it/~servedio/software.html`

Objective: Creation and Statistical Analysis of complex networks

This program was written in C language inside the TAGora project. It is continuously under development, growing as soon as new kinds of statistical analysis need to be performed on complex networks.

The input parameters of the program can be inserted either editing a plain text file, or running a graphic user interface generated using Qt3 libraries. The statistical analysis of the input network may be directly directed to the standard numerical analysis packages in common use among physicist's community. Networks may also be visualized by means of external programs as Graphviz and/or Grip directly linked to the program.

Complex networks can be imported from plain text files, specifying the list of links, which can be directed and weighted, one link in each row. Particularly complex graphs may be reduced to less number of nodes using a tunable minimum betweenness criterion. Communities may be detected by means of the spectral analysis of the graph.

The following network related statistical quantities can be analyzed:

- degree distribution
- clustering coefficient (also up to the second neighbors)
- degree correlations
- site and edge betweenness
- distribution of pair distances
- cluster dimensions of graph disjoint components

Furthermore, the program is able to generate random networks according to some of the existent graph generation models. In order to analyze the ensemble statistics associated with a model, many realizations of the model may be generated and analyzed in one shot.

A.2 Pajek

License: Free for non-commercial use

OS: Windows, Linux (via Wine)

URL: `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

Objective: Analysis and visualization of large networks

Pajek is a tool especially aimed at the analysis and visualization of large networks. Furthermore, it can be used for factorizing a large network into several smaller networks on which then more sophisticated analysis methods can be applied (cf. (Batagelj and Mrvar, 1998)). The size of the networks which can be analyzed with Pajek is only limited by the size of the available memory. Because most social networks are weakly connected, Pajek doesn't use matrices to represent a network but instead it only saves the for each vertex the connecting edges or arcs. Furthermore, it doesn't keep labels and names of nodes in extremely large networks. This makes it possible, that it can be applied on networks with more than 1 million nodes (cf. (Huisman and van Duijn, 2005)). A detailed overview of Pajek's capabilities and how to use it is available in (de Nooy et al., 2005).

Pajek supports several export and import formats. For example, it can import network data from UCINET[1]. Furthermore, there exist tools which can be used for converting CSV and Excel files into the Pajek format (see A.3). The visualizations of networks can be amongst others exported to EPS, SVG, BMP and VRML. Furthermore, it offers interfaces to tools for statistical analysis like R and SPSS.

A.3 Text2Pajek

License: Free

OS: Windows

URL: `http://www.fas.at/science/downloads/apps/text2pajek.zip`

Objective: Creating input files for Pajek (see A.2).

This is a small tool for converting CSV files, e.g. created from the content of data bases, into input files for Pajek.

A.4 GNU Octave

License: GPL

OS: Windows, Linux, Mac OS X

URL: `http://www.gnu.org/software/octave/`

Objective: High-level language for numerical computations, Manipulation of matrix representations of network graphs

GNU Octave is a general, high-level language which is intended for doing numerical computations. It has a command line interface and may also be used as a batch-oriented language. Besides the numerical computation capabilities it also offers an interface to gnuplot[2] for displaying graphics. In TAGora Octave may be used for manipulating the matrix representations of network graphs.

[1] For a description of UCINET see (Huisman and van Duijn, 2005).
[2] http://www.gnuplot.info/

A.5 T-ORG

License: Free for non-commercial use

OS: Platform Independent

URL: `http://www.uni-koblenz.de/~abbasi/publications/T-ORG.pdf`

Objective: Classify resources present on a tagging system by tag classification.

T-ORG provides a mechanism to organize resources present on a tagging system by classifying the tags attached to them into predefined categories. Supervised classification in case of large tagging systems seems infeasible; therefore the classification algorithm of T-ORG called T-KNOW does not require any training. A more detailed description of the system and evaluation experiments are available in chapter 5. The source code and binaries of T-ORG are available at the above mentioned URL.

A.6 Boost – C++ libraries

License: Free

OS: Linux, Windows

Requirements: C++ Compiler

URL: `http://sourceforge.net/project/showfiles.php?group_id=7586&package_id=8041&release_id=504013`

Objective: Computation of graph characteristics.

For the computation of characteristic path length and the clustering coefficient of graphs, one can use the Boost Graph library. This library provides free peer-reviewed portable C++ source code and offers the functionality to apply standard analysis methods of graph characteristics. The generic graph components and algorithms were developed by Jeremy Siek and a team at the University of Notre Dame. To describe the characteristic path length of a graph, shortest paths were needed, which is computed by the breadth-first-search() function. According to our graph mode, the breadth-first-search() function includes performing a breadth-first traversal of an undirected graph. In more detail, 'a breadth-first traversal visits vertices that are closer to the root before visiting vertices that are further away. In this context, distance is defined as the number of edges on the shortest path from the source vertex'. The breadth-first-search() function can be used to compute the shortest paths from the root to all reachable vertices and the resulting shortest-path distances (B. Dawes, 2007). To measure the clustering coefficient, one can also use the implementation of the Graph Boost library and a clustering approximation according to (Schank and Wagner, 2005).

A.7 MCL

License: GNU General Public License

OS: Linux

URL: `http://micans.org/mcl/`

Objective: Community detection in networks

Notes: Part of official *Debian* releases

The MCL algorithm (van Dongen, 2000) is short for the **Markov Cluster Algorithm**, a fast and scalable unsupervised cluster algorithm for graphs based on simulation of (stochastic) flow in graphs. The algorithm was invented/discovered by Stijn van Dongen at the Centre for Mathematics and Computer Science (also known as CWI) in the Netherlands.

The MCL algorithm simulates flow using (alternating) two simple algebraic operations on matrices. The first operation used by MCL is expansion, which coincides with normal matrix multiplication. Expansion models the spreading out of flow, becoming more homogeneous. The second is inflation, which is mathematically speaking a Hadamard power followed by a diagonal scaling. Inflation models the contraction of flow, becoming thicker in regions of higher current and thinner in regions of lower current. The MCL process causes flow to spread out within natural clusters and evaporate in between different clusters. By varying parameters, clusterings on different scales of granularity can be found. The number of clusters can not and need not be specified in advance, but the algorithm can be adapted to different contexts.

The basic interface to the MCL algorithm is very simple: you need only one option (the **-I** flag) to get to the heart of it, and for large graphs you should also be aware of the the **-scheme** flag for regulating resources. The default approach is to vary the argument to **-I** over some interval (doing an MCL run for each value), and analyze the clustering output with the other utilities that come with the MCL package.

Bibliography

Rabeeh Abbasi, Steffen Staab, and Philipp Cimiano. Organizing resources on tagging systems using t-org. In *Proceedings of the Workshop "Bridging the Gap between Semantic Web and Web 2.0" at ESWC 2007*, June 2007. URL http://www.uni-koblenz.de/~abbasi/publications/T-ORG.pdf.

Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, 2003.

R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97, 2001. doi: 10.1103/RevModPhys.74.47. URL http://dx.doi.org/10.1103/RevModPhys.74.47.

Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer. Trend detection through temporal link analysis. *J. Am. Soc. Inf. Sci. Technol.*, 55(14):1270–1281, 2004. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.20082.

D. Abrahams B. Dawes. Boost c++ libaries, 2007. URL http://www.boost.org/.

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.

Vladimir Batagelj and Andrej Mrvar. Pajek – program for large network analysis. *Connections*, 21: 47–57, 1998.

B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

K. Borner, Soma Sanyal, and A. Vespignani. Network science: a theoretical and practical framework. *Annual Review of Information Science and Technology*, 41:537–607, 2007.

Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998. doi: 10.1016/S0169-7552(98)00110-X.

A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori. Detecting communities in large networks. *Physica A: Statistical and Theoretical Physics*, 352(2-4):669–676, July 2005. doi: 10.1016/j.physa.2004.12.050. URL http://www.sciencedirect.com/science/article/B6TVG-4FB91CD-8/2/508224d0d5e1fc0635dbaf18ee058541.

Andrea Capocci and Francesca Colaiori. Mixing properties of growing networks and the simpson's paradox, 2005. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0506509.

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, 2007. URL http://www.pnas.org/cgi/content/short/104/5/1461.

P. Cimiano, G. Ladwig, and S. Staab. Gimme' the context: Context-driven automatic semantic annotation with C-PANKOW. In *Proceedings of the 14th World Wide Web Conference*, pages 332–341, 2005. URL http://www.aifb.uni-karlsruhe.de/Publikationen/showPublikation?publ_id=889.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

Connotea. Connotea Mailing List. https://lists.sourceforge.net/lists/listinfo/connotea-discuss.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.

Li Ding, Tim Finin, Anupam Joshi, Rong Pan, Scott R. Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management*, pages 652–659, New York, NY, USA, 2004. ACM Press. URL http://portal.acm.org/citation.cfm?id=1031289.

Li Ding, Rong Pan, Timothy W. Finin, Anupam Joshi, Yun Peng, and Pranam Kolari. Finding and ranking knowledge on the semantic web. In *International Semantic Web Conference*, pages 156–170, 2005. URL http://ebiquity.umbc.edu/get/a/publication/197.pdf.

S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW (Physics)*. Oxford University Press, Inc., New York, NY, USA, 2003. ISBN 0198515901.

M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. 15th Int. WWW Conference*, May 2006.

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

Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems, Aug 2005. URL http://arxiv.org/abs/cs.DL/0508082.

T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

M. Hearts. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France*, pages 539–545, 1992.

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

Bettina Hoser, Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Semantic network analysis of ontologies. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 514–529, Heidelberg, June 2006. Springer. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hoser2006semantic.pdf.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006a. Springer. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006information.pdf.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006b. Springer. ISBN 3-540-49335-2. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006trend.pdf.

Mark Huisman and Marijtje van Duijn. Software for social network analysis. In Peter Carrington, John Scott, and Stanley Wasserman, editors, *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.

J. Kleinberg. Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke, and R. Rastogi, editors, *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2006. ISBN 3540286071. URL http://www.cs.cornell.edu/home/kleinber/stream-survey04.pdf.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632, 1999. URL citeseer.ist.psu.edu/article/kleinberg98authoritative.html.

Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981. ISBN 0-201-03822-6.

F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *Lecture Notes in Computer Science*. Springer, 1995. ISBN 3-540-60161-9.

Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.

A. Mathes. Folksonomies-Cooperative Classification and Communication Through Shared Metadata. *Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December*, 2004a.

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

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

S. Milgram. The small world problem. *Psychology Today*, 1:60–67, 1967. URL ftp://cs.ucl.ac.uk/genetic/papers/Milgram1967Small.pdf.

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.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0303516.

M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006. doi: 10.1103/PhysRevE.74.036104.

Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521826985.

G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

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

Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Őiberna, 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.

Christoph Schmitz, Miranda Grahl, Andreas Hotho, Gerd Stumme, Ciro Cattuto, Andrea Baldassarri, Vittorio Loreto, and Vito D. P. Servedio. Network properties of folksonomies. In *Proceedings of the Tagging and Metadata for Social Information Organization workshop held in conjunction with WWW2007*, 2007.

Gerd Stumme. A finite state model for on-line analytical processing in triadic contexts. In *ICFCA*, pages 315–328, 2005.

TAGora. Theoretical tools for modeling and analyzing collaborative social tagging systems. Deliverable D4.1, TAGora project, 2007.

Stijn van Dongen. A cluster algorithm for graphs, 2000.

A. Vazquez, J. Gama Oliveira, Z. Dezso, K. I. Goh, I. Ko ndor, and A. L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006. URL http://www.citebase.org/abstract?id=oai:arXiv.org:physics/0510117.

Alexei Vazquez. Exact results for the barabasi model of human dynamics. *Physical Review Letters*, 95:248701, 2005. URL http://www.citebase.org/abstract?id=oai:arXiv.org:physics/0506126.

Stanley Wasserman and Katherine Faust. *Social Network Analysis. Methods and Applications.* Cambridge University Press, 1995.

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.

R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470. Reidel, Dordrecht-Boston, 1982.

W. Xi, B. Zhang, Y. Lu, Z. Chen, S. Yan, H. Zeng, W. Ma, and E. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proc. 13th International World Wide Web Conference*, New York, 2004.