Emergence of Consensus and Shared Vocabularies in Collaborative Tagging Systems

Valentin Robu
University of Southampton
Highfield Campus, Southampton, UK
Harry Halpin
University of Edinburgh and World Wide Web Consortium (W3C)
2 Buccleuch Place, Edinburgh, Scotland
and
Hana Shepherd
Princeton University
Wallace Hall, Princeton, NJ, USA
Vr2@ecs.soton.ac.uk, H.Halpin@ed.ac.uk, hshepher@princeton.edu

This paper uses data from the social bookmarking site del.icio.us to empirically examine the dynamics of collaborative tagging systems and to study how coherent categorization schemes emerge from unsupervised tagging by individual users.

First, we study the formation of stable distributions in tagging systems, seen as an implicit form of "consensus" reached by the users of the system around the tags that best describe a resource. We show that final tag frequencies for most resources converge to power law distributions and we propose an empirical method to examine the dynamics of the convergence process, based on the Kullback-Leibler divergence measure. The convergence analysis is performed both for the most utilized tags at the top of tag distributions and the so-called "long tail."

Second, we study the information structures that emerge from collaborative tagging, namely tag correlation (or folksonomy) graphs. We show how community-based network techniques can be used to extract simple tag vocabularies from the tag correlation graphs by partitioning them into subsets of related tags. Furthermore, we also show, for a specialized domain, that shared vocabularies produced by collaborative tagging are richer than the vocabularies which can be extracted from large-scale query logs provided by a major search engine.

Although the empirical analysis presented in this paper is based on a set of tagging data obtained from del.icio.us, the methods developed are general, and the conclusions should be applicable across all websites that employ tagging.

Categories and Subject Descriptors: H.5.3 [**Group and organizational interfaces**]: Collaborative computing; I.2.4 [**Artificial Intelligence**]: Knowledge Representation; H.1.1 [**Systems and information theory**]: Information theory

General Terms: Algorithms, Measurement, Human Factors

Additional Key Words and Phrases: Collaborative tagging, emergent semantics, complex systems, power laws, graphical models, knowledge extraction, community identification algorithms, search engines, del.icio.us

1. INTRODUCTION

1.1 Tagging versus Taxonomies on the Web

The issue of how knowledge engineering on the Web should proceed with the greatest efficiency and efficacy is a central concern as the amount of information on the Web grows. A small but increasingly influential set of web applications, including the social bookmarking site del.ici.ous, Flickr, Furl, Rojo, Connotea, Technorati, and Amazon allow users to *tag* objects with keywords to facilitate retrieval both for the acting user and for other users. Sets of categories derived based on the tags used to characterize some resource are commonly referred to as folksonomies. This approach to organizing online information is usually contrasted with taxonomies, including the approach some associate with the Semantic Web.

There are concrete benefits to the tagging approach. The flexibility of tagging systems is thought to be an asset; tagging is a categorization process, in contrast to a pre-optimized classification process such as expert-generated taxonomies. In defining this distinction, [Jacob 2004] believes that "categorization divides the world of experience into groups or categories whose members share some perceptible similarity within a given context. That this context may vary and with it the composition of the category is the very basis for both the flexibility and the power of cognitive categorization." Classification, on the other hand "involves the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and non-overlapping classes; it mandates consistent application of these principles within the framework of a prescribed ordering of reality" [Jacob 2004]. Other authors argue that tagging enables users to order and share data more efficiently than using classification schemes; the free-association process involved in tagging is cognitively much more simple than are decisions about finding and matching existing categories [Butterfield 2004]. Additionally, proponents of tagging systems show that users of tagging systems only need to agree on the general meaning of a tag in order to provide shared value instead of agreeing on a specific, detailed taxonomy [Mathes 2004].

However, a number of problems stem from organizing information through tagging systems, including ambiguity in the meaning of tags and the use of synonyms which creates informational redundancy. Additionally, an important open question concerning the use of collaborative tagging to organize metadata is whether the system becomes *stable* over time. By *stable*, we mean that users have collectively developed some implicit consensus about which tags best describe a site, and these tags do not vary much over time. We will assume that these tags that best describe a resource will be those that used most often, and new users mostly reinforce already-present tags with similar frequencies. Since users of a tagging system are not acting under a centralized controlling vocabulary, one might imagine that no coherent categorization schemes would emerge at all from collaborative tagging. In this case, tagging systems, especially those with an open-ended number of non-expert users like del.icio.us, would be inherently unstable such that the tags used and their frequency of use would be in a constant state of flux. If this were the case, identifying coherent, stable structures of collective categorization produced by users with respect to a site would be difficult or impossible.

Given the debate over the utility of collaborative tagging systems compared to other methods of knowledge engineering on the Web, it is increasingly important to understand whether a coherent and socially navigable method of categorization can emerge from collaborative tagging systems. This paper will empirically examine a crucial aspect of this

question: whether tag distributions stabilize over time and, if so, what type of distributions emerge. Since each tag for a given web resource (such as a web page) is repeated a number of times by different users, for any given tagged resource there is a distribution of tags and their associated frequencies. The collection of all tags and their frequencies ordered by rank frequency for a given resource is the *tag distribution* of that resource.

The hope among proponents of collaborative tagging systems is that stable tag distributions, and thus, possibly, stable categorization schemes, might arise from these systems. Again, by *stable* we do not mean that users stop tagging the resource, but instead that users collectively settle on a group of tags that describe the resource well and new users mostly reinforce already-present tags with the same frequency as they are represented in the existing distribution. Online tagging systems have a variety of features that are often associated with complex systems such as a large number of users and a lack of central coordination. These types of systems are known to produce a distribution known as a power law over time. A crucial feature of some power laws - and one that we also exploit in this work - is that they can be produced by scale-free networks. So regardless of how large the system grows, the shape of the distribution remains the same and thus *stable*. Researchers have observed, some casually and some more rigorously, that the distribution of tags applied to particular resources in tagging systems follows a power law distribution where there are a relatively small number of tags that are used with great frequency and a great number of tags that are used infrequently [Mathes 2004]. If this is the case, tag distributions may provide the stability necessary to draw out useful information structures.

This paper will empirically examine two important questions regarding the structure of tagging systems; first, whether tag distributions stabilize over time, and if so, what type of distribution emerges and second, whether the resulting structure of tags can be utilized to construct categorizations that provide meaningful information. This works seeks to make a contribution both to the theoretical understanding of the nature of tagging systems and to applied problems of information extraction from tagging systems.

1.2 Overview of Related Work

Existing research on tagging has explored a wide variety of problems, ranging from fundamental to more practical concerns. In this section, we provide a broad overview of the types of problems that interest researchers and practitioners in this area. We then focus on the research most relevant to the work presented here, in order to underscore our contribution.

A number of papers [Halvey and Keane 2007; Kuo et al. 2007; Hearst and Rosner 2008] examine which tag presentation techniques enable users to find information with greatest ease and speed. They often put a special emphasis on tag clouds, the most widely used presentation technique). [Halvey and Keane 2007] provide a systematic evaluation of the properties of tag interfaces which have the most effect on the accuracy and speed with which users find information. Using a set of 62 test subjects, they show that alphabetization, font size and position of the tags play an important role. They also conclude that users scan lists and clouds of tags, rather than reading them directly. [Kuo et al. 2007] perform a similar study, but focused on the field of biomedical information. They compare the results of user search based on the PubMed database with results from a search using tag clouds extracted from search summaries returned by PubMed. They conclude that a tag cloud interface is advantageous in presenting descriptive information, but it may be less effective in enabling users to discover relationships between concepts than full text summaries.

In more recent work, [Hearst and Rosner 2008] extend the study of tag clouds by also

examining the subjective reactions of the test users to different layouts. They also discuss the role that social signaling may play in motivating the use of tag clouds. Another paper concerned with visualization is [Kaser and Lemire 2007], who study the performance of different visualization algorithms for the 2-dimensional tag cloud drawing problem. The algorithms proposed are evaluated based on criteria such as minimization of the screen area required and computational speed. Compared to our work, this direction of research on tag visualization is different in scope, since we are more concerned with macro-level properties of tagging systems (e.g. convergence, emergence of shared vocabularies) that with visualization and usability aspects. However, as future work, comparing visualization methods using tag correlation graphs (as discussed in Sect. 4 of this paper) with existing approaches using tag clouds may prove insightful.

[Boydell and Smyth 2006] propose an approach for building a community-based snippet index that reflects the expertize and revolving interests of a group of searchers. They show how such an index could be used to re-rank the results produced by an underlying search engine, such as to give a higher rank to results that have been frequently selected by members of the same community in the past. [Boydell and Smyth 2007] build on the idea of using community knowledge, by proposing a social summarization technique which allows the generation of more community-focused and query-sensitive summaries than those returned by standard search engines. While this line of work does not focus explicitly on tagging, it uses the same underlying principle, that of capturing the expertize of a community of like-minded searchers to improve search results.

Other research examines the use of tagging for specific contexts and applications. [Hayes and Avesani 2007] provides a discussion of how tag clustering techniques could be used to retrieve information in blogs, while [Bateman et al. 2007] describe how using tagging in an e-learning system can supplement traditional metadata-gathering approaches. [Dubinko et al. 2006] consider the problem of visualizing the evolution of tags within the Flickr community. They develop several methods and algorithms for dynamically presenting tags to users given a sliding time window. [Rattenbury et al. 2007] present a method for the automatic extraction of event and place semantics from Flickr tags. [Chirita et al. 2007] develop a system for the automatic generation of personalized tags during browsing, based on the data residing on the surfer's desktop. All of these techniques would benefit from a method for determining whether a given set of tags has stabilized, such as the one proposed in this paper, in order to present the most stable tags to the user. If tags were presented before they stabilized, the information presented to the user might be less valuable.

In a direction of work that bears directly on the larger question of this research, [Mika 2005] addresses the problem of extracting taxonomic information from tagging systems in the form of Semantic Web ontologies. The paper extends the traditional model of taxonomies by incorporating a social dimension, thus establishing an essential connection between tagging and the techniques developed in the Semantic Web arena. However, unlike this work, Mika does not study the stabilization of the tag distributions themselves. Ideally, one would want to know if a tag distribution was stable before attempting to extract any taxonomic information from it.

There are several lines of research which take a perspective closely related to our work. Shen and Wu are interested in the structure of a tagging network for del.icio.us data as we are in Section 4. Unlike in our examples, their graph is unweighted [Shen and Wu 2005] and does not reflect the information in the tag distribution. They examine the degree

distribution (the distribution of the number of other nodes each node is connected to) and the clustering coefficient (based on a ratio of the total number of edges in a subgraph to the number of all possible edges) of this network and find that the network is indeed "scale-free" and has the features [Watts and Strogatz 1998] found to be characteristic of small world networks: small average path length and relatively high clustering coefficient.

An early line of research that has attempted to formalize and quantify the underlying dynamics of a collaborative tagging systems is [Golder and Huberman 2006], which also make use of del.ici.ous data. They show the majority of sites reach their peak popularity, the highest frequency of tagging in a given time period, within ten days of being saved on del.icio.us (67% in their data set), though some sites are "rediscovered" by users (about 17% in their data set), suggesting stability in most sites but some degree of "burstiness" in the dynamics that could lead to cyclical patterns of stability characteristic of chaotic systems. Importantly, Golder and Huberman find that the distribution of tags within a given site stabilizes over time, usually around one hundred tagging events. They do not, however, examine what type of distribution arises from a stabilized tagging process, nor do they present a method for determining the stability of the distribution which we see as central to understanding the possible utility of tagging systems.

In a very recent line of research, [Heymann et al. 2008] provide a large-scale comparison between social bookmarking and traditional web search, also using del.icio.us data. They find that tags used on del.icio.us are, on the whole accurate, while the class of users that use this system is broad, i.e. not restricted to a small subset of users. They also observe, however, that a large proportion of the tags assigned to a webpage (or resource) already appear in the title, forward and backward links to that page. Therefore, while tags assigned to resources are accurate, their distributions may not be suitable to make a significant impact on search performance. This is somewhat in line with our findings: while tags converge relatively fast to stable, power law distributions (c.f. Sect. 2), the top of these distributions may contain common (or obvious) tags. A solution to this problem (also suggested in [Heymann et al. 2008] may be a better mechanism for recommending tags. Conceivably, the local "vocabulary extraction" methods presented in Sect. 5 of this paper (and adaptations thereof) could be used to this end.

One important result is represented by [Cattuto et al. 2007], which discuss generative models to produce power law distributions for tag correlations. They also take a complex systems perspective to tagging and propose a generic model for the behavior of taggers, in the form of a Yule-Simon process with memory. However, [Cattuto et al. 2007] do not provide an analysis of how tag frequencies per website actually converge in time to stable distributions. [Dellschaft and Staab 2008] proposes a more-parametrized model that accounts for power law distributions in tag vocabulary growth and in tag distributions for websites. Overall, we see our work and that of [Cattuto et al. 2007] and [Dellschaft and Staab 2008] as complementary in scope. While they provide a theoretical model of a process which could give rise to power law distributions in tagging, we propose using an information-theoretic technique in Section 3 to analyze the convergence of power law distributions in already-existing tagging systems. Furthermore, we demonstrate its utility in several applications, such as extracting tag graphs and shared vocabularies.

Another important direction of work is represented by [Sen et al. 2006]. They present a user-centric model of tagging that distinguishes between personal tendency and community influence in the behavior of individual taggers. Furthermore, they propose a method to

select tags to be displayed to a user, such as to maximize tag utility, adoption and user satisfaction. In this work, tagging is introduced as a novel extension of an already existing recommender system, MovieLens. By contrast to [Sen et al. 2006], our work is more concerned with studying the tag distributions that emerge from the actions of a community of users and their mutual influence and choice. Of course, as shown by [Sen et al. 2006], some tags are simply personal bookmarks - and they are often very specific (even invented), as their scope is retrieval of a resource by an individual user. However, in the analysis performed in this paper, we focus on the aggregate tag distributions per resource which, in a large tagging system are highly unlikely to be personal bookmarks, but rather reflect the opinion or consensus of the user community. We also discuss the dynamics of the tags in the long tail, but as a macro-level convergence phenomenon.

In a recent position paper [Mikroyannidis 2007], using the empirical results presented in the conference version of this work [Halpin et al. 2007], argues that Semantic Web and Social Web approaches are essentially compatible and can co-exist. While we support the basic argument presented by [Mikroyannidis 2007], we should point out that convergence to stable tag distributions does not, by itself, imply that the converged distributions are directly usable for information retrieval. The process of constructing proper formal ontologies from folksonomies, while perhaps possible under certain conditions, is not a straightforward task.[1] The shared tag vocabularies (c.f. Sect. 5 of this paper) are not fully-fledged formal Semantic Web ontologies, but they can also be useful structures for many information retrieval applications, even without additional formalization.

Finally, in related work by some of the authors of this paper, [Robu et al. 2009] use complex systems techniques to study the dynamics of sponsored search markets, as well as the vocabularies which can be extracted from sponsored search query and click logs.

1.3 The Tripartite Structure of Tagging

To begin, we review the conceptual model of generic collaborative tagging systems theorized by [Marlow et al. 2006; Mika 2005] in order to make predictions about collaborative tagging systems based on empirical data and based on generative features of the model.

There are three main types of entities that compose any tagging system:

—The users of the system (people who actually do the tagging)

—The tags themselves

—The resources being tagged (in this case, the websites)

Each of these can be seen as forming separate spaces consisting of sets of nodes, which are linked together by edges (see Fig. 1). The first space, the ***user space***, consists of the set of all users of the tagging system, where each node is a user. The second space is the ***tag space***, the set of all tags, where a tag corresponds to a term (“music”) or neologism (“toread”) in natural language. The third space is the ***resource space***, the set of all resources, where usually each resource is a website denoted by a unique URI.[2] A tagging

[1] This may require for instance, some decision support in guiding the user, or a more structured design of the interface used to input the tags.

[2] A URI is a “Universal Resource Identifier” such as *http://www.example.com* that can return a webpage when accessed. Some tagging based systems such as Spurl (*http://www.spurl.net*) store the entire document, not the URI, but most systems such as del.icio.us store only the URI. The resource space, in this definition, represents whatever is being tagged, which may or may not be websites per se.

instance can be seen as the two edges that links a user to a tag and that tag to a given website or resource. Note that a tagging instance can associate a date with its tuple of user, tag(s), and resource.

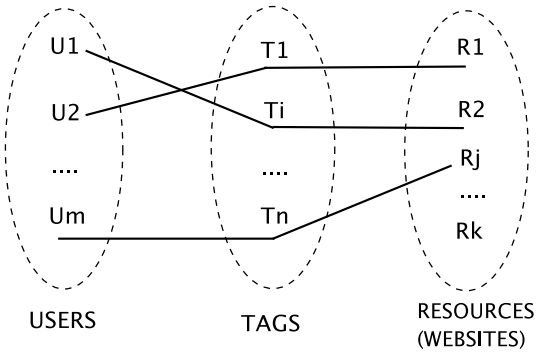

Fig. 1. Tripartite graph structure of a tagging system. An edge linking a user, a tag and a resource (website) represents one tagging instance

From Figure 1, we observe that tags provide the link between the users of the system and the resources or concepts they search for.

This analysis reveals a number of dimensions of tagging that are often under-emphasized. In particular, tagging is often *a methodology for information retrieval*, much like traditional search engines, but with a number of key differences. To simplify drastically, with a traditional search engine a user enters a number of tags and then an automatic algorithm labels the resources with some measure of relevance to the tags *pre-discovery*, displaying relevant resources to the user. In contrast, with collaborative tagging a user finds a resource and then adds one or more tags to the resource manually, with the system storing the resource and the tags *post-discovery*. When faced with a case of retrieval, an automatic algorithm does not have to assign tags to the resource automatically, but can follow the tags used by the user. The difference between this and traditional searching algorithms is two-fold: collaborative tagging relies on human knowledge, as opposed to an algorithm, to directly connect terms to documents before a search begins, and thus relies on the collective intelligence of its human users to *pre-filter* the search results for relevance. When a search is complete and a resource of interest is found, collaborative tagging often requires the user to tag the resource in order to store the result in his or her personal collection. This causes a *feedback cycle*. These characteristics motivate many systems like del.icio.us and it is well-known that feedback cycles are one ingredient of complex systems [Bar-Yam 2003], giving further indication that a power law in the tagging distribution might emerge.

1.4 Organization of the paper

This paper is organized as follows. In the first part of the paper, we examine how to detect the emergence of stable "consensus" distributions of tags assigned to individual resources. In Section 2 we demonstrate a method for empirically examining whether tagging distributions follows a power law distribution. In Section 3 we show how this convergence to a power law distribution can be detected over time by using the Kullback-Leibler divergence. We further empirically analyze the trajectory of tagging distributions before they

have stabilized, as well as the dynamics of the "long tail" of tag distributions. In the second part of the paper, we examine the applications of these stable power law distributions. In Section 4 we demonstrate how the most frequent tags in a distribution can be used in inter-tag correlation graphs (or folksonomy graphs) to chart their relation to one another. Section 5 shows how these folksonomy graphs can be (automatically) partitioned, using community-based methods, in order to extract shared tag vocabularies. Finally, Section 6 provides an independent benchmark to compare our empirical results from collaborative tagging, by solving the same problems using a completely different data set: search engine query logs. The paper concludes with a discussion of future work.

2. DETECTING POWER LAWS IN TAGS

This section uses data from del.icio.us to empirically examine whether intuitions regarding tagging systems as complex systems exhibiting power law distributions hold.

2.1 Power Law Distributions: Definition

A ***power law*** is a relationship between two scalar quantities x and y of the form:

$$y = cx^{\alpha} \tag{1}$$

where α and c are constants characterizing the given power law. Eq. 1 can also be written as:

$$\log y = \alpha \log x + \log c \tag{2}$$

When written in this form, a fundamental property of power laws becomes apparent; when plotted in log-log space, power laws are straight lines. Therefore, the most simple and widely used method to check whether a distribution follows a power law and to deduce its parameters is to apply a logarithmic transformation, and then perform linear regression in the resulting log-log space. In this paper we used a more powerful regression method to derive α that minimizes the bias in the value of the exponent (see [Newman 2005] for the technical details).

The intuitive explanation of power law parameters in the domain of tagging is as follows: c represents the number of times the most common tag for that website is used, while α gives the power law decay parameter for the frequency of tags at subsequent positions. Thus, the number of times the tag in position p is used (where $p = 1..25$, since we considered the tags in the top 25 positions) can be approximated by a function of the form:

$$Frequency(p) = \frac{c}{p^{-\alpha}} \tag{3}$$

where $-\alpha > 0$ and $c = Frequency(p = 1)$ is the frequency of the tag in the first position in the tag distribution (thus, it is a constant that is specific for each site/resource).

2.2 Empirical Results for Power Law Regression for Individual Sites

For this analysis, we used two different data sets. The first data set contained a subset of 500 "Popular" sites from del.icio.us that were tagged at least 2000 times (i.e. where we would expect a "converged" power law distribution to appear). The second data set considers a subset of another 500 sites selected randomly from the "Recent" section of del.icio.us. Both sections are prominently displayed on the del.icio.us site, though "Recent" sites are those tagged within the short time period immediately prior to viewing by the user and

"Popular" sites are those which are heavily tagged in general.[3] While the exact algorithms used by del.icio.us to determine these categories are unknown, they are currently the best available approximations for random sampling of del.icio.us, both of heavily tagged sites and of a wider set of sites that may not be heavily tagged.

The mean number of users who tagged resources in the "Popular" data set was 2074.8 with a standard deviation of 92.9, while the mean number of users of the "Recent" data set was 286.1 with a standard deviation of 18.2. In all cases, the tags in the top 25 positions in the distributions have been considered and thus all of our claims refer to these tags. Since the tags are rank-ordered by frequency and the top 25 is the subset of tags that are actually available to del.icio.us users to examine for each site, we argue that using the top 25 tags is adequate for this examination.

Results are presented in Figure 2. In all cases, logarithm of base 2 was used in the log-log transformation. [4]

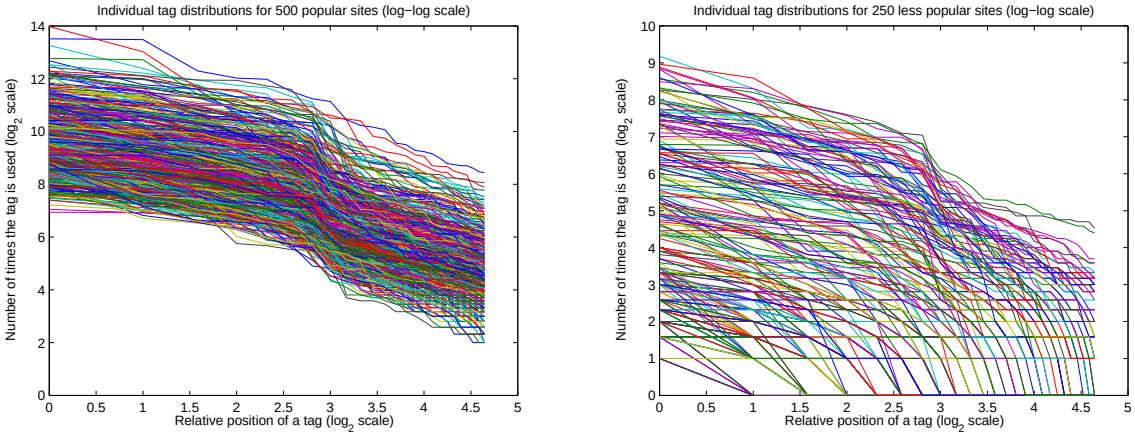

Fig. 2. Frequency of tag usage relative to tag position. For each site, the 25 most frequently used tags were considered. The plot uses a double logarithmic (log-log) scale. The data is shown for a set of 500 randomly-selected, heavily tagged sites (left) and for a set of 500 randomly-selected, less-heavily tagged sites (right).

As shown by [Newman 2005] and others, the main characteristic of a power law is its slope parameter α. On a log-log scale, the constant parameter c only gives the "vertical shift" of the distribution with respect to the y-axis. For each of the sites in the data set, the corresponding power law function was derived and the slopes of each (α parameters) were compared. The slopes indicate the fundamental characteristic of the power laws, as vertical shifts can and do vary significantly between different sites.

Our analysis shows that for the subset of heavily tagged sites, the slope parameters are very similar to one another, with an average of $\alpha = -1.22$ and a standard deviation ± 0.03. Thus, it appears that the power law decay slope is relatively consistent across

[3] All data used in the convergence analysis was collected in the week immediately prior to 19 Nov 2006.

[4] Note that the base of the logarithm does not actually appear in the power law equation (c.f. Eq. 1), but because we use empirical and thus possibly noisy data, this choice might introduce errors in the fitting of the regression phase. However, we did not find significant differences from changing the base of the logarithm to e or 10.

all sites. This is quite remarkable, given that these sites were chosen randomly with the only criteria being that they were heavily tagged. This pattern where the top tags are considerably more popular than the rest of the tags seems to indicate a fundamental effect of the way tags are distributed in individual websites which is independent of the content of individual websites. The specific content of the tags themselves can be very different from one website to the other and this obviously depends on the content of the tagged site.

For the set of less-heavily tagged sites, we found the slopes differed from each other to a much greater extent than with the heavily tagged data, with an average $\alpha = -5.06$ and standard deviation ± 6.10. Clearly, the power law effect is much less pronounced for the less-heavily tagged sites as opposed to the heavily tagged sites, as the standard deviation reveals a much poorer fit of the regression line to the log-log plotted aggregate data. For sites with relatively few instances of tagging, the results reveal mostly noise.

2.3 Empirical Results for Power Law Regression Using Relative Frequencies

In the previous section, we applied power law regression techniques to individual sites, using the number of hits for a tag in a given position in the distribution. In this section, we examine the aggregate case where we no longer use the raw number of tags (because these are not directly comparable across sites), and instead use the relative frequencies of tags. The relative frequency is defined as the ratio between the number of times a tag in a particular position is used for a resource and the total number of times that resource is tagged [5]. Thus, relative frequencies for a given site always sum to one. These relative frequencies based on data from all 500 sites of the "Popular" data set were then averaged. Results are presented in Figure 3.

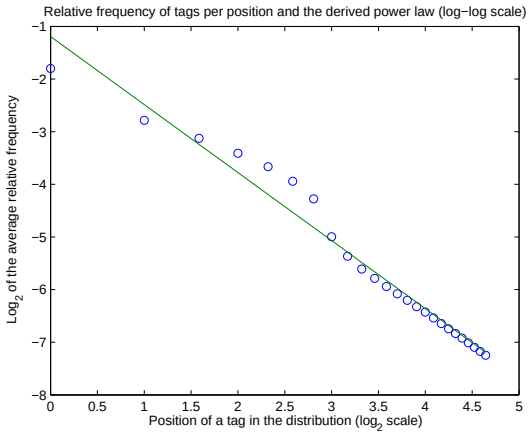

Fig. 3. Average relative frequency of tag usage, for the set of 500 "Popular" sites from above. On the y-axis, the logarithm of the relative frequency (probability) is given. (The plot uses a double logarithmic (log-log) scale, thus on the y-axis values are negative since relative frequencies are less than one.)

As before, a power law was derived in the log-log space using least-means squares

[5] To be more precise, the denominator is taken as the total number of times the resource is tagged with a tag from the top 25 positions, given available data.

(LMS) regression. This power law was found to have the slope $\alpha = -1.25$. The regression error, computed through the LMS method in the normal, not logarithmic space, was found to be 0.038. Note that the LMS regression error computation only makes sense when converted back in the normal space, since in the log-log space exponents are negative and, furthermore, deviations on the y-axis only denote actual error only after the exp_2 function is applied. This corresponds to a LMS error rate in the power law regression of 3.8% over the total number of tags in the distribution, which is low enough to allow us to conclude that tag distributions do follow power laws.

We note, however, that there is a deviation from a perfect power law in the del.icio.us data in the sense that there is a change of slope after the top seven or eight positions in the distribution. This effect is also relatively consistent across the sites in the data set. This may be due to the cognitive constraints of the users themselves or an artifact of the way the del.icio.us interface is constructed, since that number of tags are offered to the users as a suggestion to guide their search process. Nevertheless, given that the LMS regression error is rather low, we argue the effect is not strong enough to change the overall conclusion that tag distributions follow power laws.

3. THE DYNAMICS OF TAG DISTRIBUTIONS

In Section 2, we provide a method for detecting a power law distribution in the tags of a site or collection of sites. In this section, we study another aspect of the problem, namely how the shape of these distributions develops in time from the tagging actions of the individual users. First, we examine the how power law distributions form at the top (the first 25 positions) of tag distributions for each site. For this, we employ a method from information theory, namely the Kullback-Leibler divergence. Second, we study the dynamics of the entire tag distributions, including all tags used for a site, and we show that the relative weights of the top and tail of tag distributions converge to stable ratios in the data sets.

3.1 Kullback-Leibler Divergence: Definition

In probability and information theory, the Kullback-Leibler divergence (also known "relative entropy" or "information divergence") represents a natural distance measure between two probability distributions P and Q (in our case, P and Q are two vectors representing discrete probability distributions). Formally, the Kullback-Leibler divergence between P and Q is defined as:

$$D_{KL}(P||Q) = \sum_x P(x) log(\frac{P(x)}{Q(x)}) \tag{4}$$

The Kullback-Leibler distance is a non-negative, convex function, i.e. $D_{KL}(P,Q) \geq 0, \forall P, Q$ (note that $D_{KL}(P,Q) = 0$ iff. P and Q coincide). Also, unlike other distance measures it is not symmetric, i.e. in general $D_{KL}(P,Q) \neq D_{KL}(Q,P)$.

3.2 Application to Tag Dynamics

We use two complementary ways to detect whether a distribution has converged to a steady state using the Kullback-Leibler divergence:

—The first is to take the relative entropy between every two consecutive points in time of the distribution, where each point in time represents some change in the distribution.

Again, in our data, tag distributions are based on the rank-ordered tag frequencies for the top 25 highest-ranked unique tags for any one website. Each point in time was a given month where the tag distribution had changed; months where there was no tagging change were not counted as time points. Using this methodology, a tag distribution that was "stable" would show the relative entropy converging to and remaining at zero over time. If the Kullback-Leibler divergence between two consecutive time points becomes zero (or close to zero), it suggests that the shape of the distribution has stopped evolving. This technique may be most useful when it is completely unknown whether or not the tagging of a particular site has stabilized at all.

—The second method involves taking the relative entropy of the tag distribution for each time step with respect to the final tag distribution, the distribution at the time the measurement was taken or the last observation in the data, for that site. This method is most useful for heavily tagged sites where it is already known or suspected that the final distribution has already converged to a power law.

The two methods are complementary; the first methodology would converge to zero if the two consecutive distributions are the same, and thus one could detect whether distributions converged if even temporarily. Cyclical patterns of stabilization and destabilization may be detected using this first method. The second method assumes that the final time point is the stable distribution so this method detects convergence only towards the final distribution. If both of these methods produce relative entropies that approach zero, then one can claim that the distributions have converged over time to a single distribution, the distribution at the final time point. Given our interest in distributions that have converged to power laws, we are actually examining the dynamics of convergence to a power law.

3.3 Empirical Results for Tag Dynamics

The analysis of the intermediate dynamics of tagging is considerably more involved than the analysis of final tag distributions. Because the length of the histories varies widely, there is no meaningful way to compute a cumulative measure across all sites as in Section 2, so our analysis has to consider each resource individually. In Figure 4 (A and B), we plot the results for the convergence of the 500 "Popular" sites, on the basis that their final distribution must have converged to a power law, that their complete tagging history was available from the first tagging instances, and that this history was of substantial length. In the data set considered, up to 35 time points are available for some sites (which roughly corresponds to three years of data, since one time point represents one month).

There is a clear effect in the dynamics of the above distributions.[6] At the beginning of the process when the distributions contain only a few tags, there is a high degree of randomness, indicated by early data points. However, in most cases this converges relatively quickly to a very small value, and then in the final ten steps, to a Kullback-Leibler distance which is graphically indistinguishable from zero (with only a few outliers). If the Kullback-Leibler divergence between two consecutive time points (in Figure 4A) or between each step and the final one (Figure 4B) becomes zero or close to zero, it indicates that the shape of the distribution has stopped changing. The results here suggest that the power law may form relatively early on in the process for most sites and persist throughout. Even if the

[6]Note that in Figure 4, the first two time points were omitted because their distribution involved few tags and were thus very highly random.

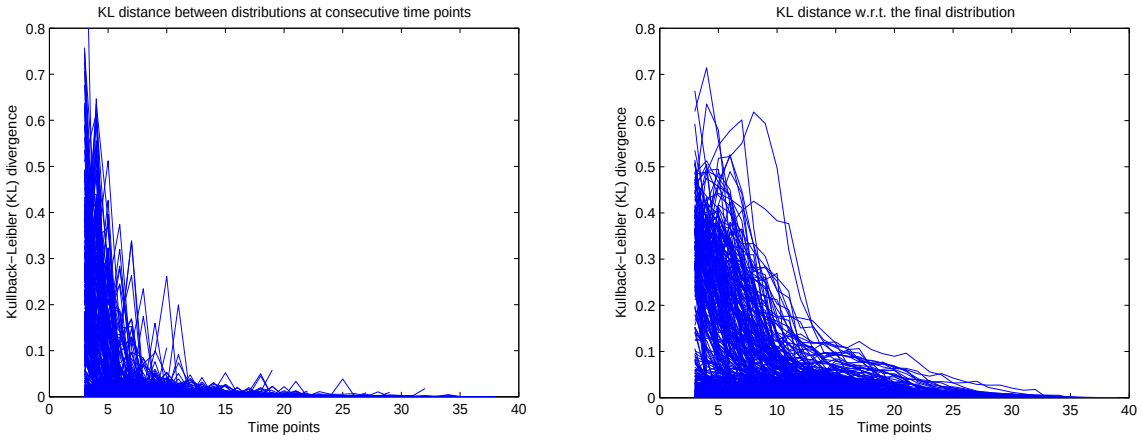

Fig. 4. A (left). Kullback-Leibler divergence between tag frequency distributions at consecutive time steps for 500 ”Popular” sites. B (right). Kullback-Leibler divergence of tag frequency distribution at each time step with respect to the final distribution.

number of tags added by the users increases many-fold, the new tags reinforce the already-formed power law. Interestingly, there is a substantial amount of variation in the initial values of the Kullback-Leibler distance prior to the convergence. Future work might explore the factors underlying this variation and whether it is a function of the content of the sites or of the mechanism behind the tagging of the site. Additionally, convergence to zero occurs at approximately the same time period (often within a few months) for these sites.

The results of the Kullback-Leibler analysis provide a powerful tool for analyzing the dynamics of tagging distributions. This very well might be the result of the “scale-free” property of tagging networks, so that once the tagging of users have reached a certain threshold, regardless of how many tags are added, the distribution remains stable [Shen and Wu 2005]. This method can be immensely useful in analyzing real-world tagging systems where the stability of the categorization scheme produced by the tagging needs to be confirmed.

3.4 Examining the dynamics of the entire tag distribution

In the previous sections, we focused on the distributions of the tags in the top 25 positions. However, heavily tagged or popular resources, such as those considered in our analysis, can be tagged several tens of thousands of times each, producing hundreds or even thousands of distinct tags. It is true that many of these distinct tags are simply personal bookmarks which have no meaning for the other users in the system. However, it is still crucial to understand their dynamics and the role they play in tagging, especially with respect to the top of the tag distribution. Some sources (e.g. Anderson [Anderson 2006]), have argued that the dynamics of long tails are a fundamental feature of Internet-scale systems. Here we were particularly interested in two questions. First, how does the number of times a site is tagged (including the long tail) evolve in time? Second, how does the relative importance of the head (top 25 tags) to the long tail change as tags are added to a resource?

Results for the same set of 500 “Popular” sites described above are shown in Figure 5. Note that the tag distributions were reconstructed through viewing the tagging history of

the individual site as available through del.icio.us and collecting the growth of this tagging distribution over time, thus allowing us to record the growth of tags outside the 25 most popular.

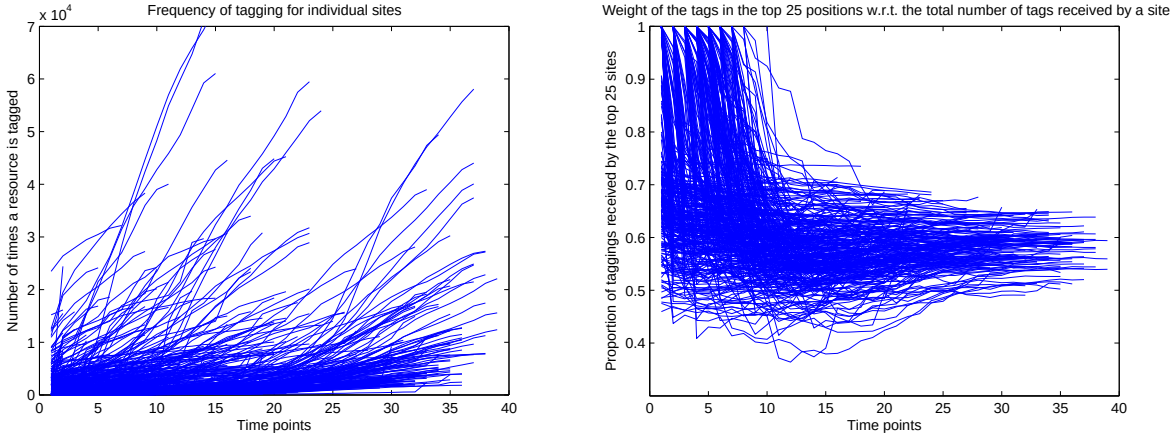

Fig. 5. A (left). Cumulative number of times a resource is tagged for each time point. B (right). Proportion of times a tag in the top 25 spots of the distribution has been used to tag a resource to the total number of times the resource has been tagged with any tag.

As seen in Figure 5, the total number of times a site is tagged grows continuously at a rate that is specific to each site and this probably depends on its domain and particular context. Though the results are not shown here due to space constraints, a similar conclusion can be formulated for the number of distinct tags, given that the number of distinct tags varies considerably per site and does not seem to stabilize in time. However for virtually all of the sites in the data set considered, the proportion of times a tag from the top 25 positions is used relative to the total number of times that a resource is tagged did stabilize over time. So, while the total number of tags per resource grows continuously, the relative weight of the tags in the head of the tag distribution compared to the those in the long tail does stabilize to a constant ratio. This is an important effect and it represents a significant addition to our analysis of the stability analysis of the top 25 positions, since it shows the relative importance of the long tail with respect to the head of the distribution does eventually stabilize regardless of the growth of tags in the long tail.

4. CONSTRUCTING TAG CORRELATION GRAPHS

The previous section examines the type of frequency distributions that emerge from the collective tagging actions of individual users, as well as the dynamics of this process. This section examines the type of information structures that form from these actions, given the hypothesized importance of the information value of tags in understanding tagging systems. We look at one of the most simple information structures that can be derived through collaborative tagging: inter-tag correlation graphs (or, perhaps more simply, “folksonomy graphs”) . We discuss the methodology used for obtaining such graphs and then illustrate our approach through an example domain study.

4.1 Methodology

The act of tagging resources by different users induces, at the tag level, a simple distance measure between any pair of tags. This distance measure captures a degree of co-occurrence which we interpret as a similarity metric, between the concepts represented by the two tags.

The collaborative filtering [Sarwar et al. 2001; Robu and Poutré 2006] and natural language processing [Manning and Schutze 2002] literature proposes several distance or similarity measures that can be employed for such problems. The metric we found most useful for this problem is *cosine distance*.[7]

Formally, let T_i, T_j represent two random tags. We denote by $N(T_i)$ and $N(T_j)$ respectively the number of times each of the tags was used individually to tag all resources, and by $N(T_i, T_j)$ the number of times two tags are used to tag the same resource. Then the similarity between any pair of tags i and j is defined as:

$$similarity(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}} \tag{5}$$

In the rest of the paper, we use the shorthand: sim_{ij} to denote $similarity(T_i, T_j)$.

From these similarities we can construct a tag-tag correlation graph or network, where the nodes represent the tags themselves weighed by their absolute frequencies, while the edges are weighed with the cosine distance measure. We build a visualization of this weighed tag-tag correlation, by using a "spring-embedder" or "spring relaxation" type of algorithm. We tested two such algorithms: Kawada-Kawai and Fruchterman-Reingold [Batagelj and Mrvar 1998]; the two graphs included in this paper are based on the latter. An analysis of the structural properties of such tag graphs may provide important insights into both how people tag and how structure emerges in collaborative tagging.

4.2 Constructing the tag correlation (folksonomy) graphs

In order to exemplify our approach, we collected the data and constructed visualizations for a restricted class of 50 tags, all related to the tag "complexity." Our goal in this example was to examine which sciences the user community of del.icio.us sees as most related to "complexity" science, a problem which has traditionally elicited some discussion. The visualizations were made on Pajek [Batagelj and Mrvar 1998]. The purpose of the visualization was to study whether the proposed method retrieves connection between a central tag "complexity" and related disciplines. We considered two cases:

—Only the dependencies between the tag "complexity" and all other tags in the subset are taken into account when building the graph (Fig. 6).

—The weights of all the 1175 possible edges between the 50 tags are considered (Fig. 7).

In both figures, the size of the nodes is proportional to the absolute frequencies of each tag, while the distances are, roughly speaking, inversely related to the distance measure as returned by the "spring-embedder" algorithm.[8] We tested two energy measures for the

[7]This should not be interpreted as a conclusion on our part that cosine distance is always an optimal choice for this problem. This issue probably requires further research and even larger data sets.

[8]For two of the tags, namely "algorithms" and "networks," morphological stemming was employed. So both absolute frequencies and co-dependencies were summed over the singular form tag, i.e. "network" and the plural "networks," since both forms occur with relatively high frequency.

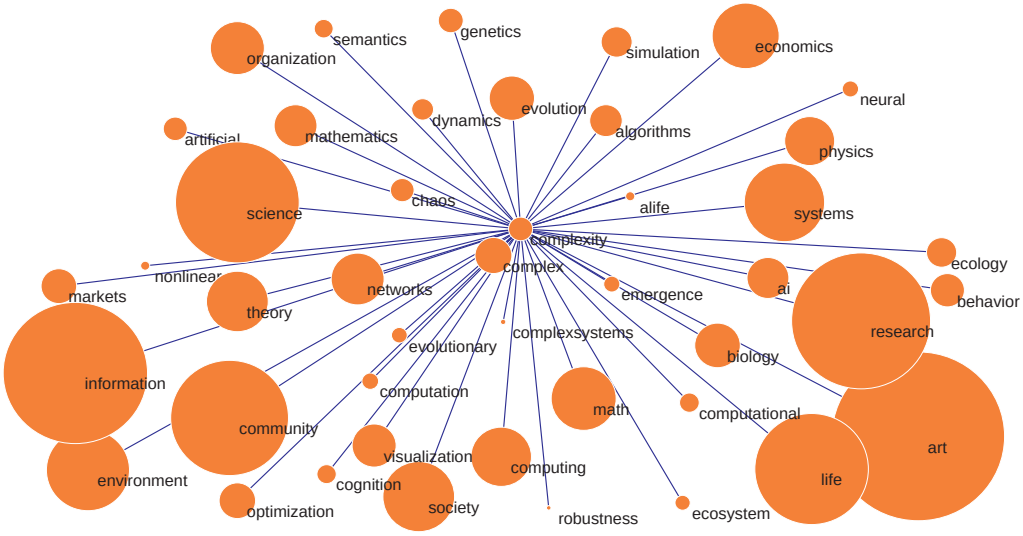

Fig. 6. Folksonomy graph, considering only correlations corresponding to central tag “complexity”

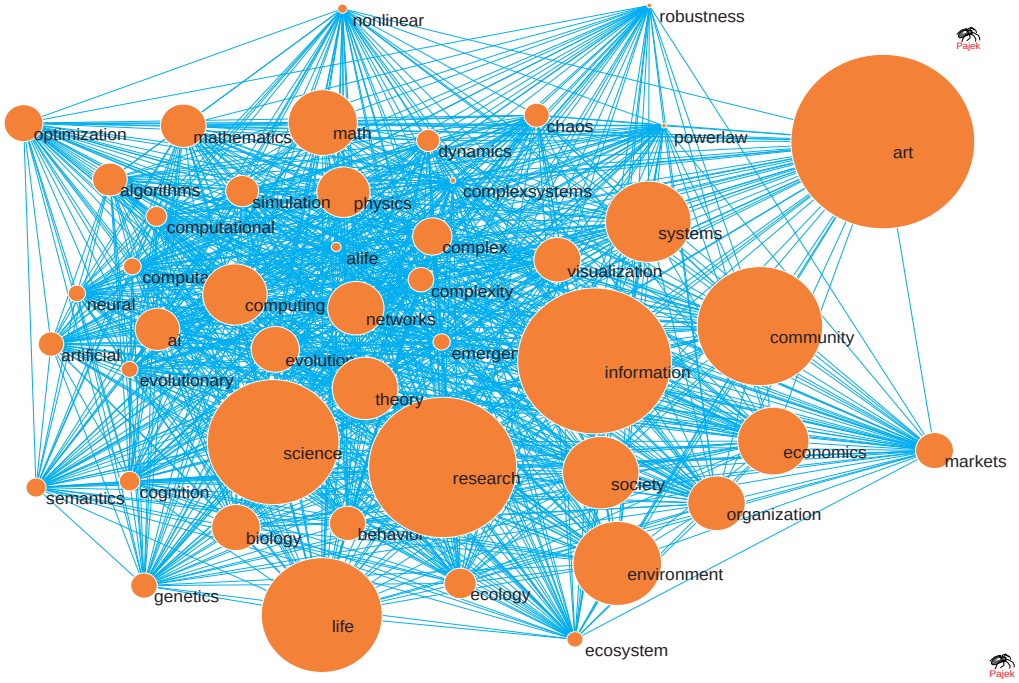

Fig. 7. Folksonomy graph, considering all relevant inter-tag correlations

“springs” attached to the edges in the visualization: Kamada-Kawai and Fruchterman-Reingold [Batagelj and Mrvar 1998]. For lack of space, only the visualization returned by

Kamada-Kawai is presented here, since we found it more faithful to the proportions in the data.

The results from the visualization algorithm match relatively well with the intuitions of an expert in this field. Some nodes are much larger than others which again shows that taggers prefer to use to general, heavily used tags (e.g. the tag “art” was used 25 times more than “chaos”). Tags such as “chaos”, “alife”, “evolution” or “networks” which correspond to topics generally seen as close to complexity science are close to it. At the other end, the tag “art” is a large, distant node from “complexity.” This is not so much due to the absence of sites discussing aspects of complexity in art as there are quite a few of such sites, but instead due to the fact that they represent only a small proportion of the total sites tagged with “art,” leading to a large distance measure.

In Figure 7, the distances to “complexity” change significantly, due to the addition of the correlations to all other tags. However, one can observe several clusters emerging which match reasonably well with intuitions regarding the way these disciplines should be clustered. Thus, in the upper-left corner one can find tags such as “mathematics”, “algorithmics”, “optimization”, “computation”, while immediately below are the disciplines related to AI (“neural” [networks], “evolutionary” [algorithms] and the like). The bottom left is occupied by tags with biology-related subjects, such as “biology”, “life”, “genetics”, “ecology” etc, while the right-hand side consists of tags with more “social” disciplines (“markets”, “economics”, “organization”, “society” etc.). Finally, some tags are both large and central, pertaining to all topics (“research”, “science”, “information”).

We also observed some tags that are non-standard English words, although we filtered most out as not relevant to this analysis. One example is “complexsystems” (spelled as one word), which was kept as such, although the tags ””complex” and “system” taken individually are also present in the set. Perhaps unsurprisingly, the similarity computed between the tags “complexsystems” and “complex” is one of the strongest between any tag pair in this set. One implication of this finding is that tag distances could be used to find tags that have minor syntactic variance with more well-known tags, such as “complesystems,” but which cannot simply detected by morphological stemming.

5. IDENTIFYING TAG VOCABULARIES IN FOLKSONOMIES USING COMMUNITY DETECTION ALGORITHMS

The previous sections analyzed the temporal dynamics of distribution convergence and stabilization in collaborative tagging as well as some information structures, like tag correlation (or folksonomy) graphs, that can be created from these tag distributions. In this Section, we look at how these folksonomy graphs could be used to solve an important problem in collaborative tagging: identifying shared tag vocabularies.

The problem considered in this section can be summarized as: given a heterogeneous set of tags (which can be represented as a folksonomy graph), how can we partition this set into subsets of related tags? In this paper, we call this problem a “vocabulary identification” problem. It is important to note that we use the term “vocabulary” only in a restricted sense, i.e. as a collection of related terms, relevant to a specific domain. For instance, a list of tropical diseases is a “vocabulary”, a list of electronic components in a given electronic device is a vocabulary, and a list of specialized terms connected to a given scientific subfield would all be “vocabularies” in our definition.

We acknowledge that this is a restricted definition: in some applications, especially

Semantic Web approaches, we would also like to know precisely how these terms are related. This type of structural information is difficult to extract only from tags, given the simple structure of folksonomies. Nevertheless, our approach could still prove useful in such applications: for example, one could construct the set of related terms as a first rough step and then a human expert (or, perhaps, another [semi]-automated method) could be used to add more more detail to the extracted vocabulary set.

However, there are many settings in which the fully automated technique presented in this paper could prove very useful. For example, drawing of tag clouds has received significant attention, but how to select the subset of related tags that will be presented in a cloud is an open problem. Another potential application is in selecting terms for sponsored search auctions. Some keywords (tags) bring a high value to advertisers, and knowing all the related keywords in a category that people can potentially use in search for can be very useful information for an advertiser. Conversely, the information regarding subsets of related tags could also be useful for the search engine in pricing searches using these tags.

Note that the complexity-related disciplines data set (already introduced in Sect. 4) is a useful tool to examine this question, since the initial set of tags are heterogeneous (complexity science is, by its very nature, an interdisciplinary field), but there are natural divisions into sub-fields, based on different criteria. This allows easier intuitive interpretation of the obtained results (besides the mathematical modularity criteria described below).

The technique we will use in our approach is based on the so-called "community detection" algorithms, developed in the context of complex systems and network analysis theory [Newman and Girvan 2004; Newman 2004]. Such techniques have been well studied at a formal level and have been used to study large-scale networks in a variety of fields from social analysis (e.g. analysis of co-citation networks), analysis of biological nets (e.g. food chains) to gene interaction networks. [Newman and Girvan 2004] provide an overview of existing applications of this theory, while [Newman 2004] presents a formal analysis of the algorithm class used in this paper. To the best of the authors' knowledge, however, this is the first paper that studies the application of these techniques to tagging systems and folksonomies. In a somewhat related direction of work, [Jin et al. 2007] study the application of community detection techniques to aggregate bidder preferences in Ebay auctions.

5.1 Using community detection algorithms to partition tag graphs

In network analysis theory, a community is defined as a subset of nodes that are connected more strongly to each other than to the rest of the network. In this interpretation, a community is related to clusters in the network. If the network analyzed is a social network (i.e. vertexes represent people), then "community" has an intuitive interpretation. For example, in a social network where people who know each other are connected by edges, a group of friends are likely to be identified as a community, or people attending the same school may form a community. We should stress, however, that the network-theoretic notion of community is much broader, and is not exclusively applied to people. Some examples [Newman and Girvan 2004; Jin et al. 2007] are networks of items on Ebay, physics publications on arXiv, or even food webs in biology. We will use a community detection algorithm to identify "vocabularies" within a folksonomy graph, identifying "communities" as "vocabularies."

5.1.1 *Community detection: a formal discussion.* Let the network considered be represented a graph $G = (V, E)$, when $|V| = n$ and $|E| = m$. The community detection

problem can be formalized as a partitioning problem, subject to a constraint. The partitioning algorithm will result in a finite number of explicit partitions, based on clusters in the network, that will considered "communities."

Each $v \in V$ must be assigned to exactly one cluster $C_1, C_2, ...C_{n_C}$, where all clusters are disjoint, i.e. $\forall v \in V, v \in C_i, v \in C_j \Rightarrow i = j$.

Generally speaking, determining the optimal partition with respect to a given metric is intractable, as the number of possible ways to partition a graph G is very large. [Newman 2004] shows there are more than 2^{n-1} ways to form a partition, thus the problem is at least exponential in n. Furthermore, in many real life applications (including tagging), the optimal number of disjoint clusters n_C is generally not known in advance.

In order to compare which partition is "optimal", the global metric used is *modularity*, henceforth denoted by Q. Intuitively, any edge that in a given partition has both ends in the same cluster contributes to increasing modularity, while any edge that "cuts across" clusters has a negative effect on modularity. Formally, let $e_{ij}, i, j = 1..n_C$ be the fraction of all edges in the graph that connect clusters i and j and let $a_i = \frac{1}{2}\sum_j e_{ij}$ be the fraction of the ends of edges in the graph that fall within cluster i (thus, we have $\sum_i a_i = \sum_{i,j} e_{ij} = m$).

The modularity Q of a graph $|G|$ with respect to a partition C is defined as:

$$Q(G, C) = \sum_i (e_{i,i} - a_i^2) \tag{6}$$

Informally, so Q is defined as the fraction of edges in the network that fall within a partition, minus the expected value of the fraction of edges that would fall within the same partition if all edges would be assigned using a uniform, random distribution. These partitions are identified as communities by [Newman and Girvan 2004]. In tagging, each of these partitions is identified as a vocabulary.

As shown in [Newman 2004], if $Q = 0$, then the chosen partition c shows the same modularity as a random division.[9] A value of Q closer to 1 is an indicator of stronger community structure - in real networks, however, the highest reported value is $Q = 0.75$. In practice, [Newman 2004] found (based on a wide range of empirical studies) that values of Q above around 0.3 indicate a strong community structure for the given network.

We will return shortly to define the algorithm by which this optimal partition can actually be computed, but first some additional steps are needed to link this formal definition to our tagging domain.

5.2 Edge filtering step

As shown in tag graph construction step above, for our data set the initial inter-tag graph contains $\binom{50}{2} = 1225$ pairwise similarities (edges), one for each potential tag pair. Most of these dependencies are, however, spurious as they represent just noise in the data, and our analysis benefits from using only the top fraction, corresponding to the strongest dependencies.

In this paper, we make the choice to filter and use in further analysis only the top $m = k_d * n$ edges, corresponding to the strongest pairwise similarities. Here, k_d is a parameter that controls the density of the given graph (i.e. how many edges are there to be considered

[9] Note that Q can also take values smaller than 0, which would indicate that the chosen partition is worse than expected at random.

Algorithm 1 *GreedyQ Determination*: Given a graph $G = (V, E), |V| = n, |E| = m$ returns partition $< C_1, ...C_{n_C} >$

1. $C_i = \{v_i\}, \forall i = \overline{1, n}$
2. $n_C = n$
3. $\forall i, j$, e_{ij} initialized as in Eq. 7
4. repeat
5. $\quad < C_i, C_j >= \text{argmax}_{c_i, c_j}(e_{ij} + e_{ji} - 2a_i a_j)$
6. $\quad \Delta Q = \text{max}_{c_i, c_j}(e_{ij} + e_{ji} - 2a_i a_j)$
7. $\quad C_i = C_i \bigcup C_j$, $C_j = \emptyset$ *//merge* C_i *and* C_j
8. $\quad n_C = n_C - 1$
9. until $\Delta Q \leq 0$
10. $maxQ = Q(C_1, ..C_{n_C})$

vs. the number of vertexes in the graph). In practice, we take values of $k_d = 1..10$, which for the tag graph we consider means a number of edges from 500 down to 50.

5.3 Normalized vs. non-normalized edge weights

The graph community identification literature [Newman and Girvan 2004] generally considers considers graphs consisting of discrete edges (for example, in a social network graph, people either know or do not know each other, edges do not usually encode a "degree" of friendship). In our graph, however, edges represent similarities between pairs of tags (c.f. Eq. 5). There are two ways to specify edge weights.

The non-normalized case assigns each edge that is retained in the graph, after filtering, a weight of 1. Edges filtered out are implicitly assigned a weight of zero.

The normalized case assigns each edge a weight proportional to the similarity between the tags corresponding to the ends. Formally, using the notations from Eq. 5 and 6 from above, we initialize the values e_{ij} as:

$$e_{ij} = \frac{m}{\sum_{ij} sim_{ij}} sim_{ij} \tag{7}$$

Where $\frac{m}{\sum_{ij} sim_{ij}}$ is simply a normalization factor, which assures that $\sum_{ij} ei_{ij} = m$.

5.4 The graph partitioning algorithm

Since we have established our framework, we can now formally define the graph partitioning algorithm. As already shown, the number of possible partitions for this problem is at least 2^{n-1} (e.g. for our 50 tag setting $2^{50} > 10^{15}$). Therefore, to explore all these partitions exhaustively would be clearly unfeasible. The algorithm we use to determine the optimal partition (Alg. 1) is based on [Newman 2004], and it falls into the category of "greedy" clustering heuristics.

Informally described, the algorithm runs as follows. Initially, each of the vertexes (in our case, the tags) are assigned to their own individual cluster. Then, at each iteration of the algorithm, two clusters are selected which, if merged, lead to the highest increase in the modularity Q of the partition. As can be seen from lines 5-6 of Alg. 1, because exactly two clusters are merged at each step, it is easy to compute this increase in Q as: $\Delta Q = (e_{ij} + e_{ji} - 2a_i a_j)$ or $\Delta Q = 2 * (e_{ij} - a_i a_j)$ (the value of e_{ij} being symmetric).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
computation	markets	semantics	powerlaw	genetics	robustness	art
optimization	economics	cognition	nonlinear	biology		
visualization	society	neural	complexsystems	evolution		
physics	community	ai	dynamics	evolutionary		
mathematics	organization	alife	chaos	science		
math	ecology	artificial	emergence			
computational	ecosystem	life	networks			
algorithms	environment	behavior	systems			
information		simulation	complex			
computing		research	complexity			
theory						
Tags that increase modularity the most, if eliminated: theory, science, research, simulation, networks.						

Fig. 8. Optimal partition in tag clusters (i.e. "communities") of the folksonomy graph, when the top 200 edges are considered. This partition has a Q=0.34. After eliminating the 5 tags mentioned at the bottom, Q can increase to 0.43.

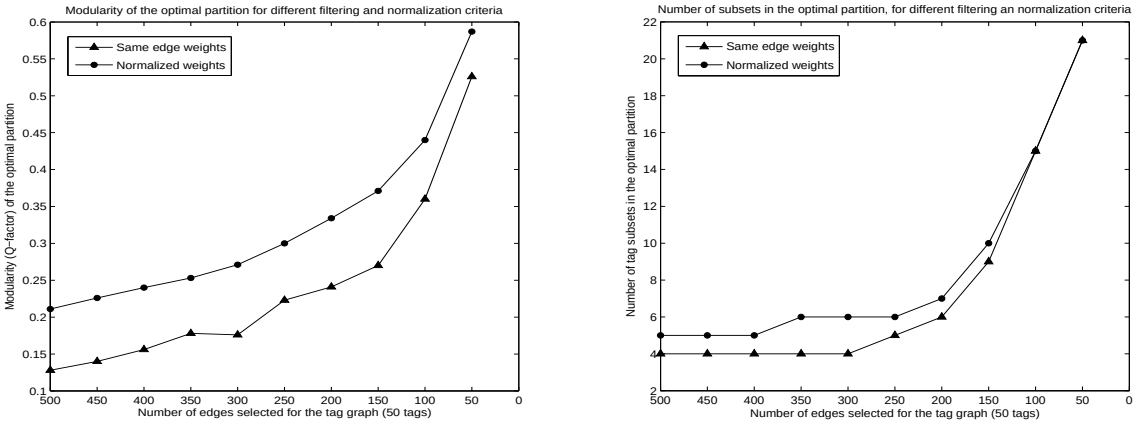

Fig. 9. Modularity (Q-factor) and number of partitions obtained from applying community detection algorithms to the scientific disciplines data set

The algorithm stops when no further increase in Q is possible by further merging.

Note that it is possible to specify another stopping criteria in Alg. 1, line 9, e.g. it is possible to ask the algorithm to return a minimum number of clusters (subsets), by letting the algorithm run until n_C reaches this minimum value.

5.5 Graph partitioning: experimental results

The experimental results from applying Alg. 1 to our data set are shown in Fig. 9. In Fig. 8 we present a detailed "snapshot" of the partition obtained for one of the experimental configurations. There are several interesting features of the results.

First, it becomes clear that using normalized edge weights produces partitions with higher modularity than assigning all the top edges the same weight of 1. This was intuitively hypothesized by us, since edge weights represent additional information we can use, but it was confirmed experimentally. Second, we are clearly able to identify partitions with a modularity higher than around 0.3, which exhibit a strong community structure ac-

cording to [Newman and Girvan 2004]. Yet perhaps the most noteworthy feature of the partitions is the rapid increase both in the modularity factor Q and in the number of partitions, as the number of edges filtered decreases (from left to right, in our figure).

The filtering decision represents, in fact, a trade-off. Having too many edges in the graph may stop us from finding a partition with a reasonable modularity, due to the high volume of "noise" represented by weaker edges. However, keeping only a small proportion of the strongest edges (e.g. 100 or 50 for a 50-tag graph, in our example), may also have disadvantages, since we risk throwing away useful information. While a high modularity partition can be obtained this way, the graph may become too "fragmented": arguably, dividing 50 tags into 10 or 15 vocabularies may not be a very useful.

Note that it is difficult to establish a general rule for what a "good" or universally "correct" partition should be in this setting. For example, even the trivial partition that assigns each tag to its own individual cluster cannot be rejected as "wrong" but such a trivial partition would not be considered a useful result for most purposes. In this paper we generally report the partitions found to have the highest modularity for the setting. However, for many applications, having a partition with a certain number of clusters, or some average cluster size, may be more desirable. The clustering algorithm propose here (Alg. 1) can be easily modified to account for such desiderata, by changing the stop criteria in line 9.

Fig. 8 shows the solution with the highest modularity Q for a graph with 200 edges, in which 7 clusters are identified. This partition assigns tags related to mathematics and computer science to Cluster 1, tags related to social science and phenomena to Cluster 2, complexity-related topics to Cluster 4 etc., while "art" is assigned to its own individual cluster. This matches quite well our intuition, and its modularity $Q = 0.34$ is above (albeit close) to the theoretical relevance threshold of 0.3. In Section 6 we will compare this partition (as well as the entire tag graphs constructed in Section 4) against an independent benchmark that addresses the same problem, but based on a completely different data set: search engine query logs. However, first we briefly present a method that can further improve the modularity of the retrieved tag graphs.

5.6 Eliminating tags from resulting partitions to improve modularity

The analysis in the previous section shows that community detection algorithms were able to produce useful partitions, with above-relevance modularity. Still, there are a few general-meaning tags that would fit well into any of the subsets resulting after the partition. These tags generally reduce the Q modularity measure significantly, since they increase the inter-cluster edges. Therefore, we hypothesized that the modularity of the resulting partitions could be greatly improved by removing just a few tags from the set under consideration. In order to test this hypothesis, we tested another greedy tag elimination algorithm, formally defined as Alg. 2. Result graphs are shown in Fig 10, while in Fig. 8 we show the top 5 tags that, if eliminated, would increase modularity Q from 0.34 to 0.43.

As seen in Fig. 2, for this data set only 5-6 tags need to be eliminated as eliminating more does not lead to a further increases in Q. In the example in Fig. 8, we see which these are, in order of elimination: theory, science, research, simulation, networks. In fact, these tags, that are marked for elimination automatically by Alg. 2, are exactly those that are the most general in meaning and would fit well into any of the subsets.

Regarding scalability, it is relatively straightforward to show that both Alg. 1 and 2 have linear running time the number of vertexes n, i.e. in this case, number of tags considered in the initial set. In the case of Alg.1, exactly two clusters of tags are merged at each step,

Algorithm 2 *GreedyQ Elimination*: Given a partition $C_1, ...C_{n_C}$ of graph $G = (V, E)$ removes all vertexes $v_i \in V$ that increase Q

1. repeat
2. $v_i = \text{argmax}_{v_i}[Q(.., C_k \setminus \{v_i\}, ..) - Q(.., C_k, ..)]$
3. $\Delta Q = \max_{v_i}[Q(.., C_k \setminus \{v_i\}, ..) - Q(.., C_k, ..)]$
   where $v_i \in C_k$ *//C_k is the partition of vertex i*
4. until $\Delta Q \leq 0$

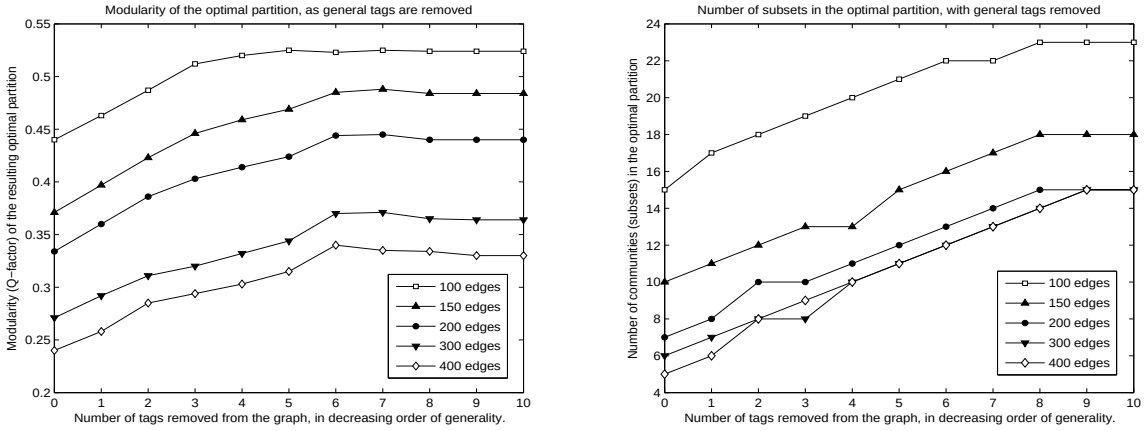

Fig. 10. Modularity (Q-factors) and number of partitions obtained after gradually eliminating tags from the data set, such as to increase the modularity. At each step, the tag that produced the highest increase in modularity between the initial and resulting partition was selected. In these results, all edge weights are normalized.

so one cluster increases in size by a minimum of one, until the algorithm terminates. In case of Alg. 2, one tag is eliminated per step, until termination. In practice, this scalability property means they are easily applicable to analyze much larger folksonomy systems.

To our knowledge, this is the first paper to investigate the applicability of this type of algorithms to tagging, and we can conclude that results are very encouraging. We leave some aspects open to further work. For instance, in the current approach, similarity distances between pairs of tags are computed using all the tagging instances in the data set. In some applications, it might be useful to first partition the set of users that do the tagging, and then consider only the tags assigned by a certain class of users. For example, for tags related to a given scientific field, expert taggers may come up with a different vocabulary partition than novice users. This may require a two-fold application of this algorithm: first to partition and select the set of users, and then the set of tags based on the most promising category of users.

While these applications of tagging distributions have shown promise, one question that can be reasonably asked is how well these applications of tagging compare to some benchmark that does not use tagging distributions. In the next section we will compare the results obtained here from collaborative tagging data against a benchmark case, which uses "classic" search engine query data.

6. COMPARISON BENCHMARK: AUTOMATIC CONSTRUCTION OF KEYWORD VOCABULARIES FROM SEARCH ENGINE QUERY DATA

The previous sections of this paper provide a compelling argument that show that a stable categorization scheme can arise from collaborative tagging, and these stable tagging distributions can produce vocabularies that can be harnessed in a wide range of applications. However, in order to truly establish the case for tagging, we need a benchmark to compare the results extracted from collaborative tagging data to results that can be obtained by means of other web search methods.

The obvious candidate for finding such a comparison benchmark is to use of large-scale query data produced by a search engine. The idea of approximating semantics by using search engine data has, in fact, been proposed before, and is usually found in existing literature under the name of "Google distance." [Cilibrasi and Vitanyi 2007] were the first to introduce the concept of "Google distance" from an information-theoretic standpoint, while other researchers [Gligorov et al. 2008] have recently proposed using it for tasks such as approximate ontology matching. It is fair to assume (although we have no way of knowing this with certainty), that current search engines and related applications, such as Google Sets [http://labs.google.com/sets 2008], also use text or query log mining techniques (as opposed to collaborative tagging) to solve similar problems.

There are two ways of comparing terms (in this case, keywords) using a search engine. One method would be to compare the number of resources that are retrieved using each of the keywords and their combinations. Another method is to use the query log data itself, where the co-occurrence of the terms in the same queries vs. their individual frequency is the indicator of semantic distance. We employ this latter method as it is more amendable to comparison with our work on tagging. In the latter method, the query terms are comparable to tags, where instead of basing our folksonomy graphs and vocabulary extraction on tags, we used query terms. In general, query log data is considered proprietary and much more difficult to obtain than tagging data. We were fortunate to have access to a large-scale data set of query log data, from two separate proposals awarded through Microsoft's "Beyond Search" awards.[10] In the following we describe our methodology and empirical results.

6.1 Data set and methodology employed

The data set we used consists of 101,000,000 organic search queries, produced from Microsoft search engine Live.com, during a 3-month interval in 2006. Based on this set of queries, we computed the bilateral correlation between all pairs from the set of of complexity related terms considered in Sect. 4 and 5 above. The set of terms are, however, no longer treated as tags, but as search keywords.[11] The correlation between any two keywords T_i and T_j is computed using the cosine distance formula in Equation 5 from Section 4 above. However, here $N(T_i, T_j)$ represents the number of queries in which the keywords T_i and T_j appear in together, while $N(T_i)$ and $N(T_j)$ are the numbers of queries in which T_i, respectively T_j appear in total (irrespective of other terms in the query), from the 100 million queries in the data set.

[10] The authors wish to thank Microsoft Research for their kind support in providing this data.

[11] We acknowledge this method has some drawbacks, as a few terms in the complexity-related set, such as "powerlaw" and "complexsystems" (spelled as one word) or "alife" (for "artificial life") are natural to use as tags, but not very natural as search keywords. However, since there are only 3 such non-word tags, they do not significantly affect our analysis.

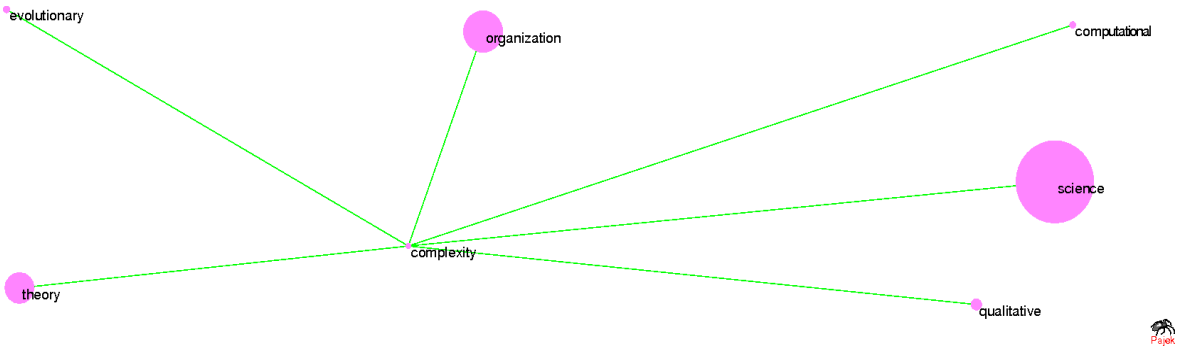

Fig. 11. Correlation graph from Microsoft queries, showing only correlations to the term “complexity”.

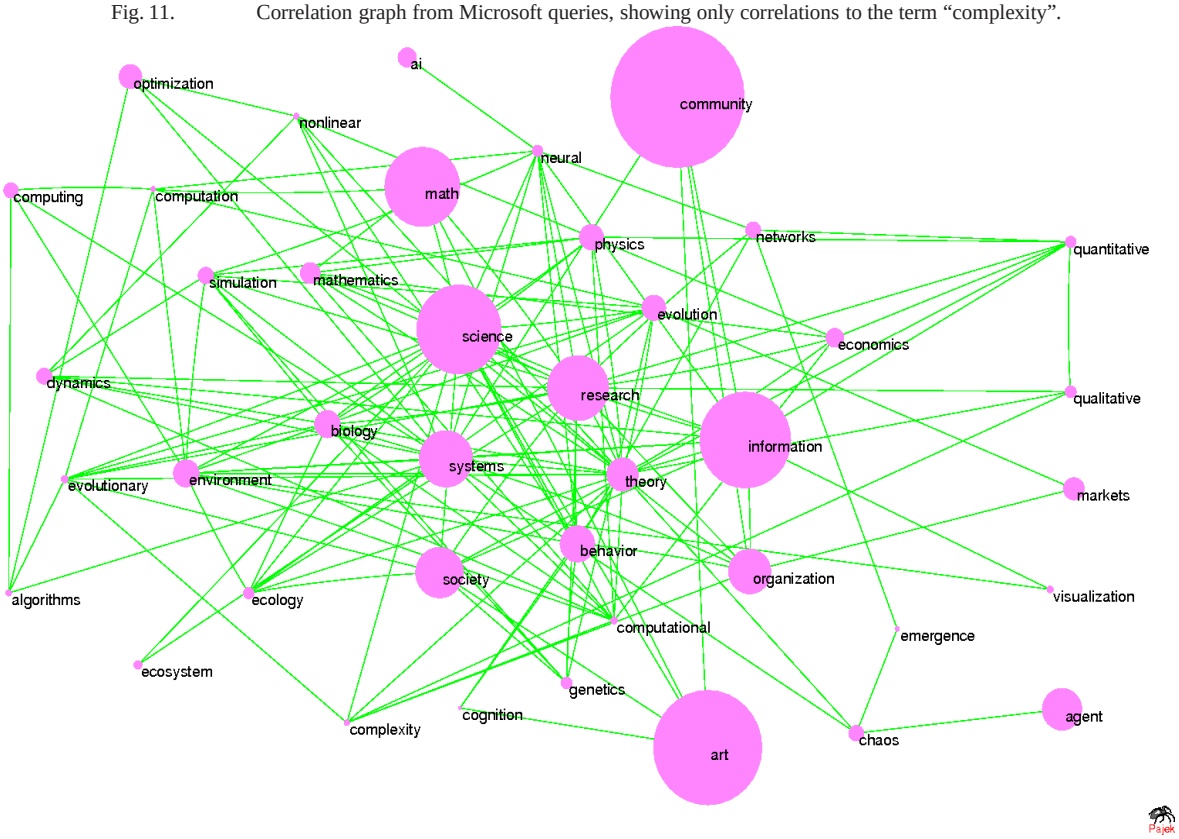

Fig. 12. Correlation graph obtained from Microsoft query logs, considering all relevant search terms.

The rest of the analysis mirrors closely the steps described in Sections 4 and 5, but optimizing the learning parameters which best fit this data set, in order to give both methods a fair chance in the comparison. More specifically, the Pajek visualization of the keyword graphs in Figs.11 and 12 were also built by using a spring-embedder algorithm based on the Kamada-Kawai distance, while Fig. 13 shows the keyword vocabulary partition that

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
complexity	systems	networks	algorithms	mathematics	research
evolution	visualization	ai	ecology	physics	quantitative
evolutionary	organization	emergence	math	economics	qualitative
chaos	information	neural	computing	art	society
cognition	community		optimization	science	
biology			computation	simulation	
theory			environment	dynamics	
behavior				nonlinear	
markets				computational	
genetics				ecosystem	
agent					
Terms left unclassified (i.e. one word clusters): complex, complexsystems, robustness, multi-agent, life, artificial, semantics, powerlaw, alife.					

Fig. 13. Optimal partition into clusters, obtained from the Microsoft query data, when the top 200 edges are considered. The resulting partition has a Q=0.536. However, 9 terms were assigned to their own cluster, thus basically left unclassified.

maximizes the modularity coefficient Q in the new setting, considering the top 200 edges. For clarity, the graph pictures are depicted in a different color scheme, to clearly show they result from entirely different data sets: Figures 6 and 7 from del.icio.us collaborative tagging data, and Figures 11 and 12 from Microsoft's Live.com query logs.

6.2 Discussion of the results from the query log data and comparison

When comparing the graphs in Figures 6 and 11 (i.e. the ones which only depict the relations to the central term "complexity") an important difference can be observed. While the graph in Fig. 6, based on collaborative tagging data, shows 48 terms related to complexity, the one is Fig. 11, based on query log data, shows just 6. The basic reason is that no relationship between the term "complexity" and the other 40+ terms can be inferred from the query log data. These relationships either do not appear in the query logs or are statistically too weak (only based on a few instances).

It is important to emphasize here that this result is not an artifact of the cosine similarity measure we use. Even if we use another, more complex distance measure between keywords, such as some suggested in the previous literature [Cilibrasi and Vitanyi 2007], we get very similar results. The fundamental reason for the sparseness of the resulting graph is that the query log data itself does not contain enough relevant information about complexity-related disciplines. For example, among the 101,000,000 queries, the term complexity appears exactly 138 times, a term such as "networks" 1074 times. Important terms such as "cognition" or "semantics" are even less common, featuring only 47 and 26 times respectively among more than 100 million queries. Therefore, it is fair to conclude that the query log data, while very large in size, is quite poor in useful information about the complexity-related sciences domain. As a caveat, we do note that more common terms, such as "community" (78,862 times), "information" (36,520 times), "art" (over 52.000), or even "agent" (about 7,000) do appear more frequently, but these words have a more general language usage and are not restricted to the scientific domain. Therefore, these higher frequencies do not actually prove very useful for identifying the relationship of these terms to complexity science, which was our initial target question.

Turning our attention to the second graph in Fig. 12 and the partition in Fig. 13, we can

see that query logs can also produce good results in comparison with tagging, although they are somewhat different from the ones obtained from tagging. For example, if we compare the partitions obtained in Fig. 8 (resulting from tagging data) and the one in Fig. 13 (from query log data), we see that tagging produces a more precise partition of the disciplines into scientific sub-fields. For instance, it is clear from Fig. 8 that cluster 1 corresponds to mathematics, optimization and computation, cluster 2 to markets and economics, cluster 5 to biology and genetics, cluster 4 to disciplines very related to complexity science and so forth. The partition obtained from query log data in Fig. 13, while is still very reasonable, reflects perhaps how a general user would classify the disciplines, rather than a specialist: organization is related to both information, systems and community (cluster 2), research is either qualitative or quantitative (cluster 6), and the like. There are also some counter-intuitive associations, such as putting biology and markets in the same cluster (number 1). Note that the clustering (or modularity) coefficient Q is higher in Fig. 13 than 8, but this is only because there are less inter-connections between terms in general in the query log data, thus there are less edges to "cut" in the clustering algorithm.

To conclude, while both methods produce reasonable results, collaborative tagging does better, at least for this domain. Tagging data appears to be more rich in information about interconnections between the terms that can be exploited by the filtering algorithms proposed in this paper. This can probably be explained by the fact the del.icio.us users have more expertise and interest in complexity-related topics than general web searchers. Furthermore, they are probably more careful in selecting resources to tag and in selecting labels for them that would be useful to other users as well (general web searchers are known to be "lazy" in typing queries). As a caveat, we note that this target domain (i.e. complexity-related disciplines) is scientific and very specialized. If the target would be more general (for example, if we selected a set of terms related to pop-culture), the comparison might lead to different results.

In future work, it may be interesting to study the formation of such vocabularies, but taking into account only the opinion (expressed in terms of bookmarks or queries) of a group or sub-community of users rather than all users, for example the community of users expert in a particular field. While this should be theoretically possible for both approaches, in practice, it may be easier to trace identities of users with collaborative tagging, not least due to privacy concerns. People who sign up to use a collaborative tagging system are implicitly more willing to share their knowledge and expertise with a community of other users. By contrast, web search is implicitly a private activity, where tracing users' actual identity, hence his/her expertise level may be undesirable.[12]

7. CONCLUSIONS AND FUTURE WORK

This work has explored the important question of whether a coherent, stable way of characterizing information can emerge from collaborative tagging systems and has presented several novel methods for analyzing data from such systems.

First, we show that tagging distributions of heavily tagged resources tend to stabilize into power law distributions and present a method for detecting power law distributions in tagging data. We see the emergence of stable power law distributions as an aspect of what may be seen as collective consensus around the categorization of information driven by

[12] Although, from anecdotal evidence, this probably happens to some degree in current practice.

tagging behaviours. We have additionally presented a method for examining the dynamics and convergence of stable tag distributions over time by the use of Kullback-Leibler divergence measures between distributions at different time steps. Also included is an empirical study of the importance of the "long tail" of the tag distributions in the convergence process.

In the second part of the paper, we propose a method for constructing and visualizing correlation graphs from tags, and show how they can lend important insights into how a community of users sees the relations between a set of terms. We also use a method from network theory for partitioning tag correlation graphs that can be used to identify vocabularies shared by a community of users. Finally, we show that vocabularies that from collaborative tagging data can be significantly richer, at least for some domains, than the ones that can be extracted from general search engine query logs. While these methods were empirically tested using del.icio.us data, the proposed methods are general enough to be applicable to most existing tagging systems.

This work suggests a number of exciting problems, both theoretical and applied, that merit further research. These include examining whether aspects of tagging distributions and dynamics are subject to the influence of particular features of tagging sites, to human cognitive limits, or some mixture of the two. A thorough examination of this aspect would represent a significant contribution to work in this area and would be important to many practical tagging applications.

Another important direction of work would be examining the effects of using specialized sub-communities of users in the study of convergence of tag distributions and resulting information structures, rather than the entire user population as in this paper. As shown by [Heymann et al. 2008], del.icio.us is not dominated by a small number of core users, but other tagging sites may be. We know relatively little about how user concentration might influence the types of information structures that can be derived from tags. Furthermore, the shared vocabulary used by a specialized sub-community of users may differ considerably to that of a larger user base.

Based on these results, it seems quite plausible that folksonomies can be fruitfully utilized for a wide category of applications related to organization of information on the web. Insights gained by taking collaborative tagging systems seriously as an empirical object of study could result in insight into the complexity of the one of the world's most complex systems, the World Wide Web.

8. ACKNOWLEDGMENTS

The initial stage of this work was performed during the authors' stay at the Santa Fe Complex Systems Institute, Santa Fe, NM, USA. The authors wish to thank the Sante Fe Institute for its support. We also thank Microsoft Research for their support in providing the query log data used in the analysis in Sect. 6 of the paper.

We also thank the anonymous reviewers for their careful reading of the manuscript and their insightful comments, that greatly improved both the contents and the readability of this paper. Finally, we thank the participants of the Dagstuhl 2008 Seminar on Social Web Communities for sharing their advice and interesting insights.

REFERENCES

ANDERSON, C. 2006. *The Long Tail*. Random House Business Books.

BAR-YAM, Y. 2003. *Dynamics of Complex Systems (Studies in Nonlinearity)*. Westview Press.

BATAGELJ, V. AND MRVAR, A. 1998. Pajek - A program for large network analysis. *Connections 21*, 47–57.

BATEMAN, S., BROOKS, C., MCCALLA, G., AND BRUSILOVSKY, P. 2007. Applying collaborative tagging to e-learning. In *WWW'07 Workshop on Tagging and Metadata for Social Information Organization*.

BOYDELL, O. AND SMYTH, B. 2006. Capturing community search expertise for personalized web search using snippet-indexes. In *Proc. of the Int. Conference on Information and Knowledge Management (CIKM'06)*. ACM Press, 1313–1314.

BOYDELL, O. AND SMYTH, B. 2007. From social bookmarking to social summarization: an experiment in community-based summary generation. *Proc. of the 2007 Int. Conf. on Intelligent User Interfaces*, 42–51.

BUTTERFIELD, S. 2004. Folksonomy. http://www.sylloge.com/personal/2004/08/folksonomy-social-classification-great.html.

CATTUTO, C., LORETO, V., AND PIETRONERO, L. 2007. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences (PNAS) 104*, 5, 1461–1464.

CHIRITA, P., COSTACHE, S., HANDSCHUH, S., AND NEJDL, W. 2007. P-tag: Large scale automated generation of personalised annotation tags for the web. In *Proceeding of the 16th Int. World Wide Web Conference (WWW 2007)*. ACM Press, 845–854.

CILIBRASI, R. AND VITANYI, P. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering 19(3)*, 370–382.

DELLSCHAFT, K. AND STAAB, S. 2008. An epistemic dynamic model for tagging systems. In *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia (HYPERTEXT'08*. ACM Press, 71–80.

DUBINKO, M., KUMAR, R., MAGNANI, J., NOVAK, J., RAGHVAN, P., AND TOMKINS, A. 2006. P-tag: Large scale automated generation of personalised annotation tags for the web. In *Proceeding of the 15th Int. World Wide Web Conference (WWW 2006)*. ACM Press, 193–202.

GLIGOROV, R., ALEKSOVSKI, Z., TEN CATE, W., AND VAN HARMELEN, F. 2008. Using google distance to weight approximate ontology matches. In *Proc. of 16th Int. World Wide Web Conference (WWW'07)*. ACM Press, 767–775.

GOLDER, S. AND HUBERMAN, B. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science 32(2)*, 198–208.

HALPIN, H., ROBU, V., AND SHEPHERD, H. 2007. The complex dynamics of collaborative tagging. In *Proc. of the 16th Int. World Wide Web Conference (WWW'07)*. ACM Press, 211–220.

HALVEY, M. AND KEANE, M. T. 2007. An assesment of tag presentation techniques. In *Proceedings of the 16th Int. World Wide Web Conference (WWW 2007)*. ACM Press, 1313–1314.

HAYES, C. AND AVESANI, P. 2007. Using tags and clustering to identify topic-relevant blogs. In *Proceedings of the 1st International Conference on Weblogs and Social Media*, N. Nicolov, N. Glance, E. Adar, M. Hurst, M. Liberman, J. H. Martin, and F. Salvetti, Eds. Boulder, CO, U.S.A. http://www.icwsm.org.

HEARST, M. A. AND ROSNER, D. 2008. Tag clouds: Data analysis tools or social signaller? In *Proc. of 41st Hawaii Int. Conference on System Sciences*. IEEE.

HEYMANN, P., KOUTRIKA, G., AND GARCIA-MOLINA, H. 2008. Can social bookmarking improve search? In *Proc. of Int. Conference on Web Search and Data Mining (WSDM'08)*. ACM Press, 195–205.

HTTP://LABS.GOOGLE.COM/SETS. 2008. Google sets. Retrieved: 1st Sept.

JACOB, E. 2004. Classification and categorization: A difference that makes a difference. *Library Trends 52*, 3, 515–540.

JIN, R. K.-X., PARKES, D. C., AND WOLFE, P. J. 2007. Analysis of bidding networks in eBay: Aggregate preference identification through community detection. In *Proc. AAAI Workshop on Plan, Activity and Intent Recognition (PAIR)*.

KASER, O. AND LEMIRE, D. 2007. Tag-cloud drawing: Algorithms for cloud visualization. In *WWW'07 Workshop on Tagging and Metadata for Social Information Organization*.

KUO, B. Y.-L., HENTRICH, T., GOOD, B. M., AND WILKINSON, M. D. 2007. Tag clouds for summarizing web search results. In *Proceedings of the 16th Int. World Wide Web Conference (WWW 2007)*. ACM Press, 1203–1204.

MANNING, C. AND SCHUTZE, H. 2002. *Foundations of statistical natural language processing*. MIT Press, London.

MARLOW, C., NAAMAN, M., BOYD, D., AND DAVIS, M. 2006. Position paper, tagging, taxonomy, flickr, article, toread. In *Collaborative Web Tagging Workshop at WWW'06, Edinburgh, UK*.

MATHES, A. 2004. Folksonomies: Cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

MIKA, P. 2005. Ontologies are us: A unified model of social networks and semantics. In *Proc. of the 4th Int. Semantic Web Conference (ISWC'05)*. Springer LNCS vol. 3729.

MIKROYANNIDIS, A. 2007. Towards a social semantic web. *IEEE Computer Magazine Nov.'07*, 113–115.

NEWMAN, M. 2005. Power laws, pareto distributions and zipf's law. *Contemporary Physics 46*, 323–351.

NEWMAN, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Phys. Rev. E 69, 066133*.

NEWMAN, M. E. J. AND GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E 69, 026113*.

RATTENBURY, T., GOOD, N., AND NAAMAN, M. 2007. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of SIGIR'07, Amsterdam, The Netherlands*, A. Press, Ed. 103–110.

ROBU, V., POUTRÉ, H. L., AND BOHTE, S. 2009. The complex dynamics of sponsored search markets. *Agents and Data Mining Interaction Springer LNCS vol. 5680*.

ROBU, V. AND POUTRÉ, J. A. L. 2006. Retrieving utility graphs used in multi-item negotiation through collaborative filtering. In *Proc. of RRS'06, Hakodate, Japan (Springer LNCI, to appear)*.

SARWAR, B., KARYPIS, G., KONSTAN, J., AND RIEDL, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Tenth International WWW Conference (WWW10), Hong Kong*.

SEN, S., LAM, S. K., RASHID, A. M., COSLEY, D., FRANKOWSKI, D., OSTERHOUSE, J., HARPER, F. M., AND RIEDL, J. 2006. tagging, communities, vocabulary, evolution. In *CSCW '06: Proc. of the 2006 20th Conference on Computer supported cooperative work*. ACM Press, 181–190.

SHEN, K. AND WU, L. 2005. Folksonomy as a complex network. http://arxiv.org/abs/cs.IR/0509072.

WATTS, D. AND STROGATZ, S. 1998. Collective dynamics of 'small-world' networks. *Nature 393*, 6684, 440–442.