The Complex Dynamics of Collaborative Tagging

Harry Halpin
University of Edinburgh
2 Buccleuch Place
Edinburgh, Scotland
H.Halpin@ed.ac.uk

Valentin Robu
CWI, Center for Mathematics and Computer Science
Kruislaan 413
Amsterdam, Netherlands
robu@cwi.nl

Hana Shepherd
Princeton University
Wallace Hall
Princeton, NJ USA
hshepher@princeton.edu

ABSTRACT

The debate within the Web community over the optimal means by which to organize information often pits formalized classifications against distributed collaborative tagging systems. A number of questions remain unanswered, however, regarding the nature of collaborative tagging systems including whether coherent categorization schemes can emerge from unsupervised tagging by users. This paper uses data from the social bookmarking site del.icio.us to examine the dynamics of collaborative tagging systems. In particular, we examine whether the distribution of the frequency of use of tags for "popular" sites with a long history (many tags and many users) can be described by a power law distribution, often characteristic of what are considered complex systems. We produce a generative model of collaborative tagging in order to understand the basic dynamics behind tagging, including how a power law distribution of tags could arise. We empirically examine the tagging history of sites in order to determine how this distribution arises over time and to determine the patterns prior to a stable distribution. Lastly, by focusing on the high-frequency tags of a site where the distribution of tags is a stabilized power law, we show how tag co-occurrence networks for a sample domain of tags can be used to analyze the meaning of particular tags given their relationship to other tags.

Categories and Subject Descriptors

H.5.3 [**Group & Organisational Interfaces**]: Collaborative Computing; I.2.4 [**Artificial Intelligence**]: Knowledge Representation

General Terms

Algorithms, Experimentation

Keywords

tagging, Del.icio.us, power laws, complex systems, emergent semantics, collaborative filtering

1. INTRODUCTION

1.1 Folksonomies and Ontologies

The issue of how metadata for web resources should be generated with the greatest efficiency and efficacy continues to be 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 user and for other users. Sets of categories that are derived based on the tags that are used to characterize some resource are commonly referred to as "folksonomies." This approach to organizing online information is usually contrasted with formal ontologies that are imposed by experts, not by users [16].

There are both benefits and drawbacks to the tagging approach. Tagging is considered a categorization process, in contrast to a pre-optimized classification process as exemplified by expert-created Semantic Web ontologies. Jacob defines the distinction between categorization and classification in the following way: "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" while "classification 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"[9]. Tagging systems allow much greater malleability and adaptability in organizing information than do formal classification systems. Proponents of tagging systems argue that "groups of users do not have to agree on a hierarchy of tags or detailed taxonomy, they only need to agree, in a general sense, on the 'meaning' of a tag enough to label similar material with terms for there to be cooperation and shared value"[11]. Tagging is able retrieve the data and share data more efficiently than classifying: "Free typing loose associations is just a lot easier than making a decision about the degree of match to a pre-defined category (especially hierarchical ones). It's like 90% of the value of a proper taxonomy but 10 times simpler" [4].

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. The central concern with using collaborative tagging to organize metadata is whether or not the system becomes relatively "stable" with time and use. By "stable," we mean to indicate that users have developed some consensus about which tags best describe a site and those tags are used most often. The most problematic claim for tagging systems would be that because users are not under a centralized controlling vocabulary, no coherent categorization scheme *can emerge at all* from collaborative tagging. In this case, tagging systems would be inherently unstable, where the tags used and their frequency of use would be in a constant state of flux, especially those systems with an open-ended number of non-expert users like the social bookmarking site del.icio.us. It would

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

be difficult to identify or utilize any collective knowledge produced by users with respect to a site.

Given the debate over the utility of collaborative tagging systems compared to other methods of organizing information, it is increasingly important to understand whether a coherent and socially navigable way of organizing metadata can emerge from distributive 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 distribution emerges. Because each tag for a given resource 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.

There is hope among the proponents of collaborative tagging systems that a stable distribution might arise from these systems. Note that by ***stable*** we do not mean that users stop tagging the resource, but instead that the tagging eventually settles to a group of tags that describe the resource well and where new users mostly reinforce already present tags in the same frequency as in the stable distribution. There is reason to believe a stable distribution should arise. Online tagging systems have a variety of features that are often associated with complex systems such as a large number of users, a lack of central coordination, and non-linear dynamics, and these sort of systems are known to produce a type over time a distribution known as a power law. One important feature of power laws produced by complex systems is that they can often be "scale-free," such that regardless of how larger the system grows, the shape of the distribution remains the same, and thus "stable." Researchers have observed, some casually, some more rigorously, that the distribution of tags applied to particular URLs 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 [11]. We are concerned with a thorough demonstration, explanation, and empirical analysis of this phenomenon.

1.2 The Dynamics and Structure of Tagging

What are the underlying dynamics of a collaborative tagging system that could cause a tag distribution to reach some point of stability? Work by Golder and Huberman using del.ici.ous data has noted a number of patterns in tagging dynamics. The majority of sites reach their peak popularity, the highest frequency of tagging in a given time period, within 10 days of being saved on del.icio.us (67% in the data set of Golder and Huberman) 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 a cyclical relationship to stability characteristic of chaotic systems [8]. Importantly, Golder and Huberman find that the proportion of frequencies of tags within a given site stabilize over time; they find it occurs usually after around being bookmarked 100 times [8]. However, they do not measure what type of distribution arises from a stabilized tagging process, nor do they present a method for determining stability.

Golder and Huberman cite two important features of such collaborative tagging systems that might give rise to this type of stability: imitation of others and shared knowledge [8]. One of the specific features of del.icio.us is the inclusion of "most common tags" for a given site when a user saves that site, facilitating the use of the tags others have used with the greatest frequency. They explain that the stability of common tags, which are displayed for users when they save a site, is based on a shared background and set of assumptions among users. Given that the stability of tag frequencies presumably relies on both the interaction between users (imitation) and the shared cultural knowledge of users, the stability and patterns of tag frequency distributions might lend insight into the degree to which there is consensus within a community about how to characterize some site or into whether there are different groups of users with different sets of assumptions and who are tagging the same site. Or, as Golder and Huberman suggest, changes in the stability of such patterns might indicate that groups of users are migrating away from a particular consensus on how to characterize a site and its content or negotiating the changing meaning of that site. To the extent this consensus is stable, it is ripe for development into a classification system and perhaps even formalization into an ontology.

Assuming a stabilized distribution with a well-known shape and set of properties arises, in order to make inferences about the existence of some sort of meaning structure in the distribution, we need to understand the information inherent in the distribution of tags. This inherent structure can be traced to what we call the ***information value*** of a tag. By "information value" we mean the information conveyed by the natural language term used in the tag and how this makes the tag useful for retrieval of and distinction between resources or not. Since the "meaning" of tags is elusive, one way to model their information value is to look at their co-occurrence with other tags, and to try to answer questions about how these co-occurrence models reflect the information value of particular tags: Does the structure of tag networks based on co-occurrence make intuitive sense, doing justice to the common-sense ideas we have about the relationships between the concepts under scrutiny? Can tagging provide users with any new insight into the meaning of resources just by analyzing the structure of networks based on co-occurrence? Shen and Wu analyze the structure of a tagging network for del.icio.us data as we do in Section 6, although unlike in our examples their graph is unweighted [15] 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 so has the features Watts and Strogatz found to characterize small world networks: small average path length and relatively high clustering coefficient [18]. A large amount of work exploring the structural properties of nature language networks finds similar results [6].

In Section 3 we formalize a generative model for tagging in order to suggest how the patterns observed in tagging distributions might emerge. In Section 4 we empirically examine whether tagging distributions develop into stable power law distributions and in Section 5 we empirically analyze the trajectory of tagging distributions before they have stabilized. Establishing the convergence and stability of these distributions is essential to understanding whether coherent categorization schemes might emerge from distributed tagging systems. Finally, we use the importance of the information value of tags to demonstrate how the most frequent tags in a power law distribution can be used in inter-tag correlation graphs to chart their relation to one another in Section 4. It is conjectured that this method might be useful in extracting a classification scheme (ontology) from a categorization scheme (folksonomy).

2. THE TRIPARTITE STRUCTURE OF TAGGING

To begin, we need a conceptual model of generic collaborative tagging systems which is capable of being formalized in order to make predictions about collaborative tagging systems based on em-

pirical data and based on generative features of the model. A well-accepted tripartite model has already been theorized [10, 12], although we hope to clarify it below:

There are three main 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 each resource is normally denoted by a unique URI.[1] A tagging instance can be seen as the two edges that links together 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.

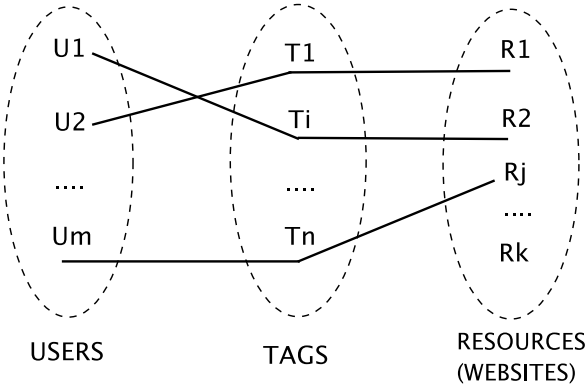

Figure 1: Tripartite graph structure of a tagging system. An edge linking a user, a tag and a resource (website) represents one tagging instance

From the above model and Fig.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 relevancy to the tags *pre-discovery*, displaying relevant resources to the user. In contrast, with collaborative tagging a user finds a resource, then adds one or more tags to the resource manually, with a 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 so relies on the collective intelligence of its human users to *pre-filter* the search results for relevancy. 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, giving further indication that a power law in the tagging distribution might emerge. Before going further we need to formalize these qualitative observations about collaborative tagging.

[1]A "Universal Resource Identifier" such as *http://www.example.com* that can return a webpage when accessed. Notice that 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. Regardless, our resource space is whatever is being tagged.

3. A GENERATIVE MODEL

Our model needs to combine the three-level model of tagging presented above with the manner in which feedback cycles and information value give rise to a stable distribution of tags over time. The notion of a feedback cycle is encapsulated in the simple idea that a tag that has already been used is likely to be repeated. This behavior is a clear example of ***preferential attachment***, known popularly as the "rich get richer" model. To model this phenomena, we need to have a baseline probability $P(a)$, or the probability of a user committing a "tagging action." This is the probability that for every time step t, a "tag" is added to a resource. There are very few empirical studies that estimate this parameter currently. Additionally, since users often tag more than once, there is $P(n)$ that determines the number (n) of tags a user is likely to add at once based on the distribution of the number of tags a given user employs in a single tagging action. As reported by other studies, this number varies between two and ten [8], although we will hold $n = 1$ in order to simplify our exposition. Once a tagging action ($P(a)$) has been done, a preferential attachment model can be formalized by use of a simple "shuffling theory" model [7]. This model holds that an "old tag" is reinforced with constant probability $P(o)$, so a "new tag" is added with probability $1 - P(o)$. If the old tag is added, it is added with a probability $\frac{R(x)}{\sum R(i)}$, where $R(x)$ is the number of times that particular previous tag x has been chosen in the past and $\sum R(i)$ is the sum of all previous tags. This leads to tags that have been heavily reinforced in the past being further reinforced in the future.

We illustrate this with a simple example, as given by Figure 2, where $P(tag)$ is $P(o)$ and assuming for simplification $P(a) = 1$. Also, we will have a user only add one new tag per time step. At time step 1 in our example, the user has no choice but to add a new tag, "piano", to the page. At the next stage, the user does not reinforce a new tag but chooses a new tag, "music", and so $P(piano) = \frac{1}{2}$ and $P(music) = \frac{1}{2}$. At $t = 3$, the user reinforces a previous "piano" tag and so $P(piano)$ increases to $\frac{2}{3}$, while $P(music)$ decreases to $\frac{1}{3}$. At $t = 4$, a new tag, "digital", is chosen and so $P(piano)$ goes up while $P(music)$ decreases to $\frac{1}{4}$ and $P(digital)$ is $\frac{1}{4}$. Taken to its conclusion, this process produces a power law distribution.

Preferential attachment models do not explain why a particular new tag is added to a resource; in practice, tags are not added at random because their information value is taken into account. For example, the oldest tags for a resource are not always the most popular tags. A new tag may be added that uncovers an informational dimension not captured by older tags. If this new dimension proves both relevant and useful then other users will reinforce the tag that represents the dimension, perhaps at the expense of older tags with less relevant informational dimensions. In this case, the

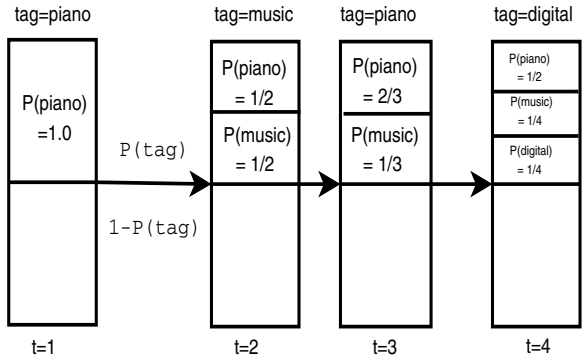

Figure 2: An example of how shuffling leads to preferential attachment

new relevant tag would experience a burst of reinforcement, perhaps surmounting the frequency with which older tags were used and eventually stabilizing towards the top of the tag distribution for a resource. The entire tagging process might be considered an "exploration" versus "exploitation" process where the exploration of possibly relevant dimensions of a resource is balanced with the exploitation of previously tagged dimensions of a resource. A stabilized distribution theoretically represents a state where the optimal number of dimensions has been tagged.

While it is impossible for a generic model to assign a priori the exact information value of a resource, it is possible to at least partially model the information value of a specific tag. A hypothetical tag applied to every relevant resource would, if used in a search by a user to discover resources, retrieve every document (imagine a tag such as "website," but used once by at least one user on every resource). This type of tag has an information value (I) of 0, and we assume that the information value of a tag that retrieves no resources is also 0. Another tag that hypothetically selects only the resource needed, would have an information value (I) of 1. This does not occur so precisely in practice, as users presumably want the optimal tag to return some cognitively appropriate (k) number of resources, such as the number of resources that fit on the screen or that allow users to effectively browse an area, and this may vary per user. However, for the purposes of our model we will assume that $k = 1$ when quantifying the information value to simplify our exposition. Notice also that a user may use multiple tags and these tag combinations may have different information values that are not additive. In our work with del.icio.us, we can empirically estimate the information value of a tag by retrieving the number of webpages a del.icio.us search with a tag (or combination of tags) returns and converting it into a probability, as done in Section 6.

In order to explain tight binding between information retrieval and value, we show an abstract example in Figure 3. In this example the act of "tagging" by a user (u_x) can be considered the assignment of a tag (t_y) to a given resource (r_z). Thus, a given search can be considered a transversal from u_x via a number of tags to a number of resources. The user wishes to minimize the number of tags needed to retrieve the relevant resources, which is unknown to both the system and the user. Following Zipf's famous "Principle of Least Effort," users presumably minimize the number of tags used [19]. In our example the user u_2 wishes to use a group of tags to discover a relevant resource, which an oracle would tell us is r_2. While tag t_1 and t_5 retrieve exactly one resource $I(t_1)$ and $I(t_5) = 1$, these tags do not identify r_2. $I(t_3) = 0$, since it retrieves all resources in the data-set. While $I(t_2)$ and $I(t_4) > I(t_3)$, the combination of both tags retrieve exactly the resource r_2 in our example so $I(t_3, t_2) = 1 > I(t_2)$ and $I(t_3)$. Notice that information value is not additive, since $I(t_1, t_5) = 0$ while both $I(t_1)$ and $I(t_5) = 1$.

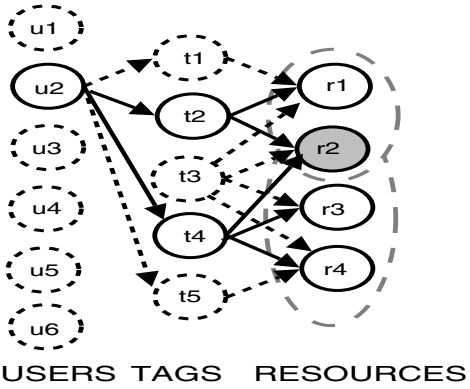

Figure 3: Tripartite tagging system graph used for search. The dotted edges represent options, while the dark edges represent a particular user engaging in a search for the shaded resource

If the user is satisfied with the search results and wishes to add a retrieved resource to their personal collection, they will reinforce one of the existing tags of the resource by repeating one of the pre-existing tags, and they might also add a new tag. If the user is not satisfied with the search results, they will likely add a new tag to a retrieved resource. This tag may allow them to use fewer tags in future searches to retrieve the same resource. Thus, if we linearly combine our two models of information value and preferential attachment, we can generate the probability of a tag x being reinforced or added as a linear interpolation of preferential attachment and information value, with λ being used to weigh the factors:

$$P(x) = \lambda * P(I(x)) + (1 - \lambda) * P(a) * P(o) * P(\frac{R(x)}{\sum R(i)}) \quad (1)$$

This formalizes a process that would give rise to a power law via preferential attachment, but one where the information value of a tag additionally figures into the dynamics of the tagging distribution. This model as it stands is heavily parameterized, where the values of the parameters no doubt vary from one tagging system to another. However, to see if this model stands, we need to determine whether a power law actually arises from empirical data.

4. DETECTING POWER LAWS IN TAGS

According to our model, there should be a connection between the stability of the distribution of tags and the general shape of the distribution. If our qualitative intuition about tagging systems as complex systems is correct, this distribution should follow a power law. Our complete data set includes 750 tagged sites from del.ici.ous, 500 of which were taken from the "Popular" section of del.icio.us and 250 of which were randomly selected from the "Recent" section of del.icio.us. Both sections are prominently displayed on the del.icio.us site. "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.

4.1 Power Law Distributions: Definition

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

$$y = cx^{\alpha} \tag{2}$$

Where α and c are constants characterizing the given power law. Without loss of generality, Eq. 2 can also be written as:

$$\log y = \alpha \log x + \log c \tag{3}$$

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. Recent work on the subject by Newman ([13]) suggests, however, that this may introduce a bias in the value of the exponent, and as the reliable alternative proposes the following formula to determine α:

$$\alpha = 1 + n * \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1} \tag{4}$$

where x_i, $i = 1..n$ are the measured values of x and x_{min} corresponds to the lowest value for which the power law behavior holds. This formula was also used in this work (the interested reader can consult the full derivation of the above formula in [13]).

In our tagging domain, the intuitive explanation of the above parameters 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 (for p=1 to 25) should be approximated by a function of the form (where $-\alpha > 0$):

$$Frequency(p) = \frac{Frequency(p = 1)}{p^{-\alpha}} \tag{5}$$

4.2 Empirical Results for Power Law Regression for Popular Sites

For this analysis, we used a subset of 500 "Popular" sites from del.icio.us that were tagged close to 2000 times. For each website, we considered the 25 most often used tags. Fig. 4 shows the observed data when plotted in the log-log scale. After the log-log transformation, we fit a linear regression to the resulting data points of each site individually. We computed the aggregate distribution for all sites by summing the frequency of tags that appear in each position across the sites and fitted a regression line to the data. The results are presented in Fig. 5. In all cases, logarithm of base 2 was used in the log-log transformation [2].

To summarize our results, we found that the data points can be fit with a linear regression line, with some error. With the aggregate function, the parameter for the slope of the power law, using the above equation (see Equation 4), had the value: $\alpha = -1.278$. For the individual sites (not shown graphically, for the sake of clarity of the picture), the slopes were in a similar range, i.e. with an average $\alpha = -1.22$, and standard deviation ± 0.03. Thus, it appears that the power law decay (i.e. slope) is relatively consistent, both in the cumulative case and across individual sites. Intuitively, this indicates a fundamental effect of the way tags are distributed in individual websites independent of the context and content of the specific website.

There is a caveat, however. We observed that tags in positions

[2]Note that the base of the logarithm does not actually appear in the power law equation (c.f. Eq. 2) but because we use empirical and thus possibly noisy data, this choice might introduce errors in the fitting of the regression line. However, we did not find significant differences when changing the base of the logarithm to e or 10.

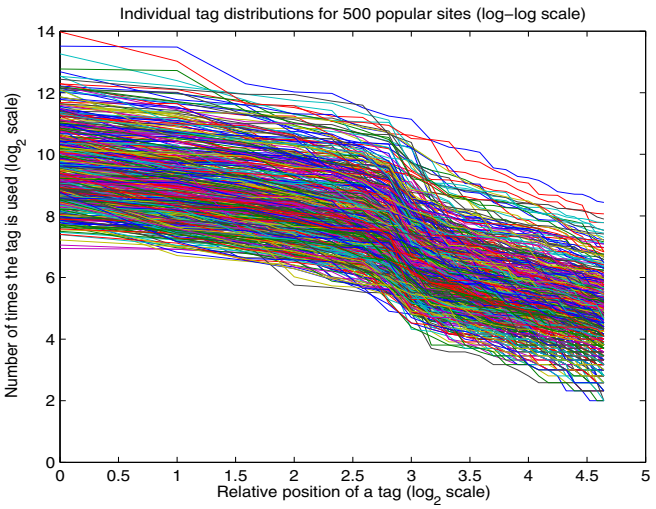

Figure 4: Frequency of tag usage, based on relative position. The dataset consists of 500 heavily tagged sites where for each, the 25 most frequently used tags were considered. The plot uses double logarithmic (log-log) scale: the horizontal scale gives the logarithm base 2 of the relative position (where the most used tag is in position 1, the second most used tag is in position 2 and so on), while the vertical scale gives the logarithm of the frequency of use

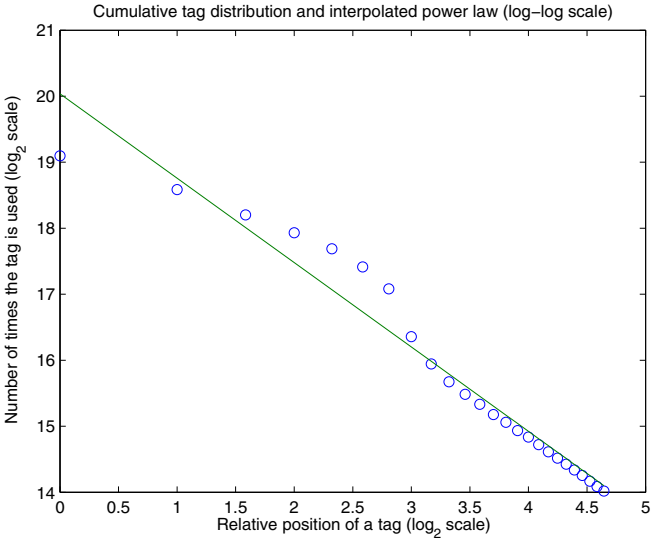

Figure 5: Cumulative frequency of tag use, based on relative position. The plot is on a log-log scale: the horizontal axis shows the logarithm of the relative position, while the vertical axis shows the logarithm of the cumulative frequency of tags in that relative position. The best fit linear regression (least-squares method) is shown.

seven to ten have a considerably sharper drop in frequency than the general trend line would predict. This means that if we were to do a piece-wise regression for the tags in the first seven positions and the tags in the last fifteen positions, we would get linear functions for both, though with different slopes. Furthermore, as Fig. 4 shows, this effect largely holds for almost all sites in the data set considered, so it is not attributable to noise alone, but a consistent effect of the way tagging is performed. We do not have yet a satisfactory explanation for this effect; it may have a cognitive explanation (i.e. it may be based on the number of tags the average user employs per website), or it may be an artifact of the user interface specific to del.icio.us (i.e. users see space for a particular number of tags or receive a particular number of suggestions for tags to use). This observation does not affect our basic result that tag distributions follow power laws.

4.3 Regression Results for Less Popular Sites

The analysis presented in the above section refers to heavily tagged sites (tagged close to 2000 times) and considers the 25 most used tags for each site. In order to further illustrate and verify our results, we considered an additional sample of 500 sites selected randomly from the "Recent" section of del.icio.us and plotted their distribution on a log-log scale. This set of sites is much less heavily tagged: the mean number of users of the "Recent" distribution is 286.1 with a standard deviation of 18.2, as opposed to the previously examined "Popular" distribution which has a mean of 2074.8 users and a standard deviation of 92.9 users.

Results are shown in Fig. 6. Our analysis shows that for the less heavily tagged individual sites, the slopes differed from each other to a much greater extent than with the heavily tagged data, with an average $\alpha = -3.9$ and standard deviation ± 4.63. 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 data. For sites in the "Popular" category, the standard deviation of the power law decay slope from the average slope is only 0.03, while the set of less heavily tagged sites has a standard deviation of 4.63. In fact, for random sites with relatively few instances of tagging, the results reveal little other than noise, though even for some of these less popular sites, a power law is beginning to emerge.

5. THE DYNAMICS OF TAG DISTRIBUTIONS

In Sect. 4, we have shown that tag distributions converge to power law distributions. Again, because power laws are scale-free, the emergence of this type of distribution suggests the emergence of a stable distribution. In this section, we study another aspect of the problem, namely how the shape of these distributions forms in time from the tagging actions of individual users. In practice, this involves measuring the distance between the distributions of tags of a given site at different time points (in our case, each time point approximately corresponds to a calendar month, which is discussed below). We take a novel approach to this problem by employing a method derived from information theory, namely the Kullback-Leibler divergence.

5.1 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:

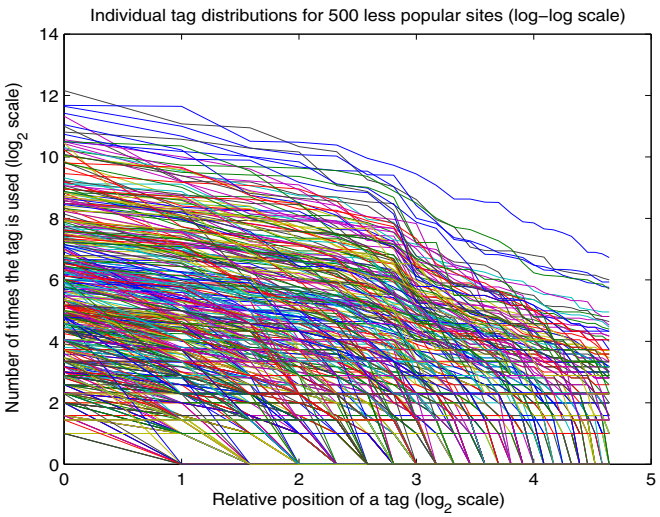

Figure 6: Frequency of tag usage based on relative position for a dataset consisting of 500 less-heavily tagged sites. The 25 most frequently used tags are considered for each.

$$D_{KL}(P||Q) = \sum_x P(x) log(\frac{P(x)}{Q(x)}) \quad (6)$$

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 $D_{KL}(P,Q) \neq D_{KL}(Q,P)$.

5.2 Application to Tag Dynamics

There are two complementary ways to detect whether or not a distribution has converged to a steady state using the Kullback-Leibler divergence (relative entropy):

- 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 the rank-ordered tag frequencies for the top 25 highest ranked unique tags for any one website. Each point of time was a given month where the tag distribution had changed. Months where there was no change in tagging 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.

- The second method involves taking the relative entropy of the tag distribution for each time step with respect to the final tag distribution for that site (where "final" indicates the distribution at the time the measurement was taken, or the last observation in the data). This method is most useful for heavily tagged sites, for which (as shown in Sect 4) the final distribution has already converged to a power law.

The two methods are complementary because the first methodology would converge to zero if the two consecutive distributions are the same, and thus could illustrate when distributions converged if even temporarily. One could imagine a cyclical pattern of stabilization and destabilization being detected using this first method.

The second method assumes that the final time point is the stable distribution and thus illustrates convergence only towards the final distribution. If both of these methods produce relative entropies that approach zero, then we can be certain the distributions have converged over time to a single distribution, which is the distribution at the final time point. Since we have already shown that final distributions converge to power laws, what is examined here are the dynamics of convergence to a power law.

5.3 Empirical Results for Tag Dynamics

The analysis of the dynamics of tagging is considerably more involved than the analysis of the 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 Sect 4, so our analysis has to consider each resource individually. In Fig. 7 (A and B), we plot the results for the convergence of the 500 "Popular" sites, selected as to simultaneously satisfy several requirements. First, their final distribution must have converged to a power law. Second, their complete tagging history must have been available from the first tagging instances and this history must have had a substantial length. In the data set considered, up to 35 time points are available for some sites (which roughly corresponds to 3 years of data, since one time point represents one month).

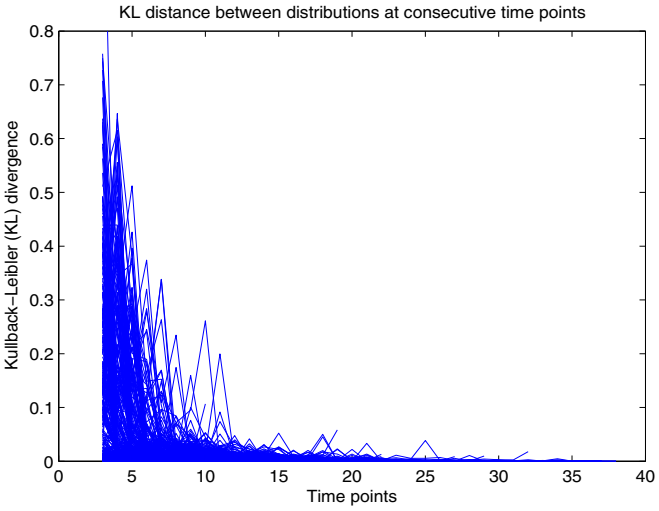

Figure 7: Relative entropy (Kullback-Leibler divergence) between tag frequency distributions at consecutive time-steps

There is a clear effect which can be observed from the dynamics of the above distributions. (Note that in Figs. 7 and 8, the first two time points were omitted because their distribution involved few tags and were thus highly random.) 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 all cases this converges relatively quickly to a very small value, and then (in the final ten steps) to a Kullback-Leibler distance which is so low that is graphically indistinguishable from zero (with a few outliers). If the Kullback-Leibler divergence between two consecutive time points (in Fig. 7) or between each step and the final one (Fig. 8) becomes zero (or close to zero), it indicates that the shape of the distribution has stopped changing. This result suggests that the power law may form relatively early in the process for most sites and persist with remarkable consistency throughout. Even if the number of tags added by the users increases many

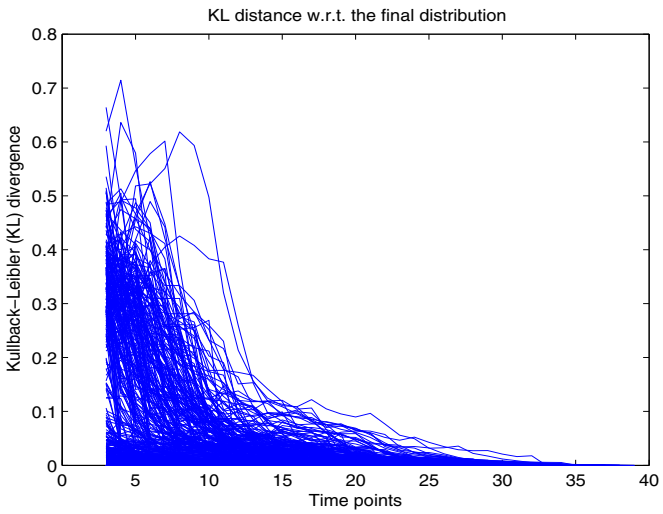

Figure 8: Relative entropy (Kullback-Leibler divergence) of tag frequency distribution at each time step, with respect to the final distribution

fold afterwards, the new tags reinforce the already-formed power law. Convergence to zero occurs at approximately the same time period, often within a few months, for these sites. 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.

6. CONSTRUCTING INTER-TAG CORRELATION GRAPHS

In addition to the role of processes of social influence between users, the information value of tags is a central aspect governing the evolution of tag distributions. We examine one of the most simple information structures that can be derived through collaborative tagging: inter-tag correlation graphs. First, we discuss the methodology used for deriving such graphs. Next we illustrate our approach through an example, using tags from a limited domain. Finally, we discuss the importance of tag-tag graphs and how they could be used to shed light on the underlying dynamics of the tagging process.

6.1 Methodology

The act of tagging resources by different users induces, at the tag level, a simple distance measure between any pair of tags. In our case, define the distance between two tags T_i, T_j through a cosine distance measure:

$$Dist(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}} \quad (7)$$

Where we denote by $N(T_i)$, respectively $N(T_j)$, the number of times each of the tags was used individually to tag all pages, and by $N(T_i, T_j)$ the number of times two tags are used to tag the same page, summed over all pages. The distance measure captures a degree of co-occurrence, which we interpret as a similarity metric, between the concepts represented by the two tags. The distance measure which is used can play a large role in the actual structure

retrieved and we note that there are more sophisticated distance measures proposed in existing item-item collaborative filtering (see e.g. [14]). For this paper, cosine distance was adequate.

Next, from these similarities we can construct a tag-tag correlation graph or network, where the nodes represent the tags themselves, weighted by their absolute frequencies, while the edges are weighted by the cosine distance measure. We build a visualization of this weighted tag-tag correlation by using a "spring-embedder" of algorithm - in this case we used the well-known Kawada-Kawai algorithm [1]. An analysis of the structural properties of such tag graphs may provide important insights into how people tag and how semantic structure emerges in distributed folksonomies (we return to this issue in Section 6.3, where we discuss the relation between this approach and the structures derived in the literature on language evolution). While it would be difficult if not impossible for independent researchers to collect enough data to construct and analyze the entire space of tags used in del.icio.us, we did collect enough data to provide an illustration of the approach for a restricted sub-domain.

6.2 Constructing tag-tag correlation networks

In order to exemplify our approach, we collected the data and constructed visualizations for a restricted class of 15 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.[3] The visualizations were made using Pajek [1]. The purpose of the visualization was to study whether our method retrieves a connection between a central tag "complexity" and related disciplines. We consider two cases:

- Only direct dependencies between the tag "complexity" and all other tags are taken into account in building the graph (Fig. 9).
- 30 other edges (i.e. 45 edges in total for 15 tags) are considered (Fig. 10). Those taken have the highest expected correlations, though future work will consider more sophisticated methods for determining the cut-off based on examining deviation from the mean.

In both figures, the size of the nodes is proportional to the absolute frequencies of each tag, while the distances are, roughly, inversely related to the distance measure (as returned by the "spring-embedder" algorithm).[4] We tested two energy measures for the "springs" attached to the edges in the visualization: Kamada-Kawai and Fruchterman-Reingold [1]. For lack of space, only the visualization returned by Kamada-Kawai is presented here, as it is more faithful to the proportions present in the data.

The results from the visualization algorithm match well with what one would intuitively expect to see in this domain. Some nodes are much larger than others, which, again shows that taggers prefer to use 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 (some of them were actually developed in the context of complex systems), come 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 the mathematics/complexity aspects in art. In fact, there are quite a few of such sites - but they represent only a small proportion of the total sites tagged with "art", leading to a large distance measure. There are, upon inspection, some problems in the structure retrieved: the tag "ecology" would be expected to appear much closer to "complexity," since much research on complexity in biological systems has focused on applications in ecology.

While formal ontologies are of great utility in highly structured domains such as biology, in other domains, like webpages, collaborative tagging may be a better way of organizing information. Even in an open-ended domain such as webpages, our data seem to show some consensus about how to categorize the information. These tag-correlation graphs could provide knowledge engineers guidance in how people naturally categorize data, although any knowledge engineer should check for stabilization of the distribution as outlined here. Some of these tag correlation networks might, with the help of a careful engineer, serve as the base of a taxonomy using a Semantic Web languages such as RDF Schema [3]. Though, a fully automatic process for deducing any type of taxonomy based only on tag co-occurrence is unlikely to work, since "there is no one-to-one correspondence between concepts and keywords. It is not always possible for the users to express a complex concept with a single keyword and thus they may use more than one tag to express the concept association that the item brings up in them" [12].

6.3 Tag Graphs and Human Language Networks

In the previous section, we demonstrated that tag networks can be easily constructed and visualized and that they might prove useful in simple information retrieval. However, exploring the properties of these tag graphs (e.g. node centrality, degree distribution, and so on) - and their evolution - can provide us with much deeper insights into how folksonomies develop from the aggregate behavior of individual users. They could additionally provide insight into how more complex semantic structures evolve.

A starting point for further modeling is work that seeks to explain the emergence of structure and syntax in human language. In recent high-profile work, Ferrer i Cancho and Sole [17, 5] study the evolution of several human languages by constructing their graphical protostructure. They do this by taking large corpuses of natural language texts and constructing inter-correlation graphs between all pairs of words in the language, based on the distance they appear from each other in these texts.

Next, they analyze the resulting graphical structure for each of the considered languages. Following the seminal work of Zipf, they show that the retrieved networks, far from having the structure predicted by random graph theory for such large networks [2], have a "small world" structure.[5] Furthermore, this protostructure is remarkably similar across different languages.

Graphs which exhibit small world network properties have a mean degree distribution that follows Zipf's law.[6] Sole et al.[17] argue that, far from being a mere coincidence, this is an essential underlying property of human languages, and furthermore, syntax and structure in human languages emerges "for free" from these simpler structures. In [6], they simulate a version of Zipf's classic generative model of human language: speakers prefer to use ambiguous,

[3]The choice of terms considered in the subset is loosely based on the topics covered at the 2006 summer course on complexity offered by the Santa Fe Institute.

[4]For two of the tags, namely "algorithms" and "networks," 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.

[5]A small world graph is a graph with a high clustering coefficient and a low average path length.

[6]The degree of a node is the number of edges connected to that node. The distribution of the degrees across all nodes is an important property of a graph

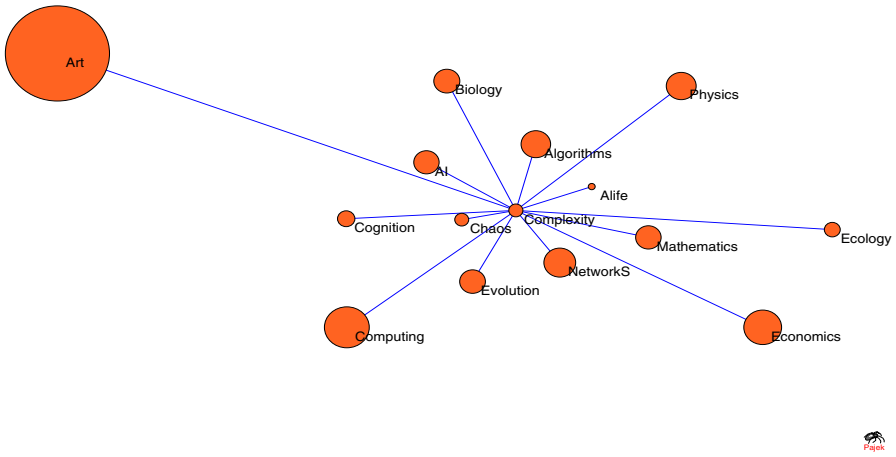

Figure 9: Visualization of a tag correlation network, considering only the correlations corresponding to one central node "complexity"

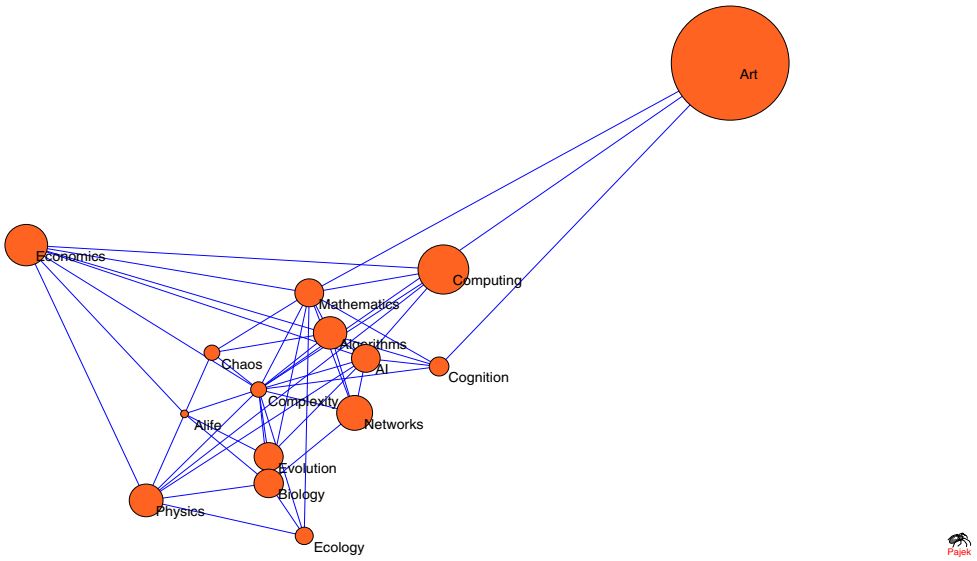

Figure 10: Visualization of a tag correlation network, considering all relevant correlations

general words which have minimum entropy (and minimize their effort for choosing the word), while hearers prefer words with high entropy, and thus high information value.

Comparing this setting with the considered tripartite model of tagging systems (presented above and in Fig. 1), we observe some important similarities to models of language evolution. Resources (websites) would correspond to the objects in the real world to be described by language, the taggers correspond to the speakers of the language, and the tags correspond to the tokens of the language, i.e. the words. Tags likely have a Zipf's law distribution of node degrees and while the massive data harvesting needed to show this is difficult, our provisional results do point in this direction. In such a case, generative models proposed by Sole et al. [6] is useful to explain the online behavior of taggers with respect to the information value of tags. Thus, folksonomy structure could also be seen as emerging at the intersection between the efforts of taggers who try to minimize their effort and thus prefer to choose more common tags with less information value, and retrievers or "hearers" who need to use these tags to find as precise resources as possible and thus use tags with the highest information value. In our generative model shown in Section 3, the results of this "least effort principle" would be the parameter λ.

7. CONCLUSION AND FUTURE WORK

This work has explored a number of issues highly relevant to the question of whether a coherent way of organizing metadata can emerge from distributive tagging systems. We began by outlining a principled generative model of tagging. Our model is based on Mika's formalization of tagging, but additionally incorporates the information value of tags which we believe allows for a more complete account of tagging [12]. Our model formalizes many of the common-sense observations made by people who are informally studying folksonomies.

Using a larger set of empirical data than previous studies have used, we have shown that tagging distributions tend to stabilize into power law distributions. This is important in that a stable distribution is an essential aspect of what might be user consensus around the categorization of information driven by tagging behaviors. Furthermore, as shown by our empirical study of the tagging history of these items, this behavior depends on the number of users and to some extent on the temporal duration of the tagging process. Therefore, given sufficient active users, over time a stable distribution with a limited number of stable tags and a much larger "long-tail" of more idiosyncratic tags develops. One might consider this stabilized distribution to be an emergent categorization scheme. This stable categorization scheme is described by a scale-free power law, such that in the future, unless a new tag with a high information value is discovered, further tagging will only reinforce the pre-existing categorization scheme given by the limited number of stable tags. One might claim that the users have collectively discovered a collective categorization scheme. The optimality of such a scheme merits further attention.

Using an example domain, we explored one of the most empirically challenging aspects of the generative model: the information value of a tag as a function of the number of pages the tag retrieves when searched. We examined how this information can be used with multiple tags to visualize correlation graphs that lend insight into the categorization process and into existing intuitions about how concepts are related.

It seems quite plausible that folksonomies and ontologies, which are merely new incarnations of the age-old distinction between categorization and classification respectively, are not mortal enemies, but fundamentally compatible, as tagging-based categorization in our data exhibits emergent consensus. By focusing on the tags which are most common in the tagging distribution, one should be able to understand the essence of the collective categorization scheme. One could then safely ignore the "long-tail" of idiosyncratic and low frequency tags that are used by users to tweak their own results for personal benefit, or alternatively, treat the "long-tail" as an object of examination for other reasons. As shown by our visualization graphs, insightful categorization and classification schemes can be gained by focusing on the high frequency "short head" (as opposed to the "long tail") of a stabilized tag distribution.

Using the methodology outlined above, in particular the detection of power law distributions and measures of relative entropy, any tagging application should be able to detect whether and at what point a tagged resource has stabilized to a power law. Using the Kullback-Leibler divergence, interested parties can test their tag distributions for stabilization. This is of practical importance since the results of data mining or knowledge engineering from stabilized tag distributions will be more mature and less likely to change with time, and therefore can take advantage properties of the power law distribution. Future work will elaborate on the results presented here regarding categorization schemes based on tag co-occurrence and information value and will examine whether these results hold among many different tagging applications. Lastly, these results suggest that treating collaborative tagging systems as complex systems yields important insights into the dynamics and processes of these systems. Insights gained by taking collaborative tagging systems seriously as an empirical object of study could provide insight into the complexity of the one of the world's most complex systems, the World Wide Web.

8. ACKNOWLEDGEMENTS

This work was performed during the authors' visit at the Santa Fe Institute, Santa Fe, NM, USA. The authors wish to thanks the SFI for its support in the initial stages of this research.

9. REFERENCES

[1] V. Batagelj and A. Mrvar. Pajek - A program for large network analysis. *Connections*, 21:47–57, 1998.

[2] B. Bollobas. *Random Graphs*. Academic Press, London, England, 1985.

[3] D. Brickley and R. Guha. RDF Vocabulary Description Language 1.0: RDF Schema,W3C Recomendation, 2004. http://www.w3.org/TR/rdf-schema.

[4] S. Butterfield. Folksonomy, 2004. http://www.sylloge.com/personal/2004/08/folksonomy-social-classification-great.html.

[5] R. F. Cancho and R. V. Sole. The small world of human language. *Proc. Roy. Soc. London*, B 268:2261–2266, 2001.

[6] R. F. Cancho and R. V. Sole. Least effort and the origins of scaling in human language. *Procs. Natl. Acad. Sci. USA*, 100:788–791, 2003.

[7] P. Diaconis, M. McGrath, and J. Pitman. Riffle shuffles, cycles and descents. *Combinatorica*, 15:11–29, 1995.

[8] S. Golder and B. Huberman. The structure of collaborative tagging systems, 2006. HP Labs Technical Report http://www.hpl.hp.com/research/idl/papers/tags/.

[9] E. Jacob. Classification and categorization: A difference that makes a difference. *Library Trends*, 52(3):515–540, 2004.

[10] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Collaborative Web Tagging Workshop at WWW'06, Edinburgh, UK*, 2006.

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

[12] P. Mika. 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, 2005.

[13] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[14] V. Robu and J. A. L. Poutré. Retrieving utility graphs used in multi-item negotiation through collaborative filtering. In *Proc. of RRS'06, Hakodate, Japan*, 2006.

[15] K. Shen and L. Wu. Folksonomy as a complex network, 2005. http://arxiv.org/abs/cs.IR/0509072.

[16] C. Shirky. Ontology is over-rated, 2005. http://www.shirky.com/writings/ontology-overrated.html.

[17] R. V. Sole. Syntax for free? *Nature*, 434:289, 2005.

[18] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[19] G. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusets, 1949.