The Dynamics and Semantics of Collaborative Tagging

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

Valentin Robu
Dutch 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 the dynamics of such systems and whether coherent classification schemes can emerge from undirected tagging by users. Currently millions of users are using collaborative tagging without centrally organizing principles, and many suspect this exhibits features considered to be indicative of a complex system. If this is the case, it remains to be seem whether collaborative tagging by users over time leads to emergent classification schemes that could be formalized into an ontology usable by the Semantic Web.

This paper uses data from "popular" tagged sites on the social bookmarking site del.icio.us to examine the dynamics of such collaborative tagging systems. In particular, we are trying to determine whether the distribution of tag frequencies stabilizes, which indicates a degree of cohesion or consensus among users about the optimal tags to describe particular sites. We use tag co-occurrence networks for a sample domain of tags to analyze the meaning of particular tags given their relationship to other tags and automatically create an ontology. We also produce a generative model of collaborative tagging in order to model and understand some of the basic dynamics behind the process.

1. INTRODUCTION

1.1 Folksonomies and Ontologies

The issue of how metadata on web resources should be generated for the greatest efficiency and efficacy continues to be a central debate. A small but increasingly influential set of websites, 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. Their categories are based on the set of tags that are used to characterize some resource, and these categories 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 these two processes 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 as process 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. Shirky explains the pitfalls of imposed classification: "If you've got a large, ill-defined corpus, if you've got naive users, if your cataloguers aren't expert, if there's no one to say authoritatively what's going on, then ontology is going to be a bad strategy"[16]. 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." [3]. However, a number of problems stem from organizing information through tagging systems including ambiguity in the meaning of tags, the use of synonyms which creates informational redundancy, and the possibility of idiosyncratic naming conventions where individuals string together many words or label items according to their personal utility, such as tagging a bookmarked site with "toread." These drawbacks are serious in that they have the ability to jeopardize the coherence of the informational content of the tagging system and render tagging systems less useful for groups of users.

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 and if so, how this might occur and whether particular features of sites using tagging facilitate or inhibit the emergence of coherence. This paper will empirically examine elements of two subsidiary issues of this larger project. In Section 4 we examine the dynamics of tag frequency in "popular" del.icio.us tags in order to detect whether the tag frequencies converge to a stable distribution and thus a categorization scheme. There is hope among the proponents of collaborative tagging systems that a stable sort of distribution might arise from these systems. Note that by "stable" we do not mean that users stop tagging the resource, but that the tagging eventually settles to a group of tags that describe the resource well and new users mostly reinforce already present tags in the same frequency as they are given in the stable distribution. 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 of distribution known as a "power-law." In Section 5 we examine how the information content of particular tags in relation to one another might be used to extract a classification scheme (ontology) from a categorization scheme (folksonomy). We present in detail some empirical work on the the first topic, and then more hypothetical work on the second.

1.2 Dynamics of Tagging

What are the underlying dynamics that could cause tagging to reach some point of stability such that the distribution of tags converge? 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]. 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 [7]. Most importantly for this paper, 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 [7].

To make inferences about the existence of some sort of structure in the distribution of tag frequencies, we need to understand the information inherent in the tags based on calculating the frequencies with which particular tags co-occur with other tags. Again, a number of critical questions remain regarding the **informational value** of tags used. By "informational value" we mean whatever information is conveyed by the natural language term used in the tag and how this makes the tag useful or not. Since the "meaning" of tags is elusive, one way to model their informational 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 informational 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 5, although unlike in our examples their graph is unweighted [15]. 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 scale free and has the features Watts and Strogatz found to characterize small world networks: small average path length and relatively high clustering coefficient [19]. A large amount of work exploring the structural properties of nature language networks finds similar results [5].

The dynamics of tagging systems are closely coupled to the informational value of tags. 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 [7]. 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 the stability of the less common tags, which are not displayed for users when they save a site, 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 frequencies 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 suggest 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 formalization into ontology.

1.3 Ontologies and Observed Patterns

Merholz uses the metaphor of "desire lines" for tagging systems; these "are the foot-worn paths that sometimes appear in a landscape over time" such that "a smart landscape designer will let wanderers create paths through use, and then pave the emerging walkways, ensuring optimal utility" [12]. This metaphor points towards a way of developing ontologies for the Semantic Web that maintains the advantages of both taxonomic classification and collaborative tagging, bridging the two sides of the debate about organizing metadata. After users have explored the space of possibilities and discovered some optimum categorization, an ontology could be formalized for classification purposes. Avoiding pre-optimization, a user-optimized ontology would take advantage of the often unexpected ways users categorize data, yet provide the amount of classificatory power provided by a smaller set of terms that can then be mapped to a Semantic Web ontology capable of expressing structured data facets, complex relationships, and scaling across the Web, which current collaborative tagging systems are incapable of doing. It is possible that in order to share data effectively users as a group, naturally and without external influence restricting their vocabulary, converge to tagging each URI with a fairly small set of semantically distinct tags.

Is it possible that such a classification structure can be detected? While some have claimed that it is not possible since the "responsiveness and flexibility" of user categorization "effectively prohibit the establishment of meaningful relationships" because they are "fleeting and ephemeral," there are a number of other cases where complex structure emerges from simple behavior [9]. What are the types of local rules users might be employing which generate these observed aggregate patterns and can they be described mathematically? The paradigmatic case of local rules generating structure is natural language itself, where "one of the key questions to understand [is] how a communication system can arise...how distributed agents without a central authority and without prior specification can nevertheless arrive at a sufficiently shared language conventions to make communication possible" [18].

2. THE TRIPARTITE STRUCTURE OF TAGGING

To begin, we need a conceptual model to describe generic collaborative tagging systems which is capable of being formalized so that we can both make predictions about collaborative tagging systems based on empirical data and based on generative features of the model. A well-accepted tripartite model has already been theorized [10, 13], 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 vertices, 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 vertex 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 then that tag to a given website or resource. Note that a tagging instance can associate a date with its tuple of a user, a tag(s), and a resource.

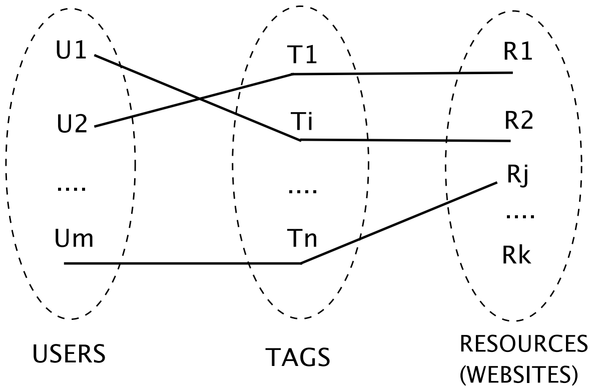

Figure 1: Tripartite graph structure of a tagging system. An edge linking a user, a tag and a resource (website) represents one tagging instance

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.

In particular, this analysis reveals a number of dimensions of tagging that are often under-emphasized. In particular, tagging is *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 in turn "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. However, before going further we need to formalize these qualitative observations about collaborative tagging.

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 informational 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 repeated. This behavior is a clear example of ***preferential attachment***, known popularly as a "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 [7], 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 [6]. 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 total 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 is chosen ("digital"), 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 informational 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, and 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 new relevant tag would experience a burst of reinforcement, perhaps surmounting the frequency with which older tags were used

[1] A "Universal Resource Identifier" such as *http://www.example.com* that can return a web-page 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.

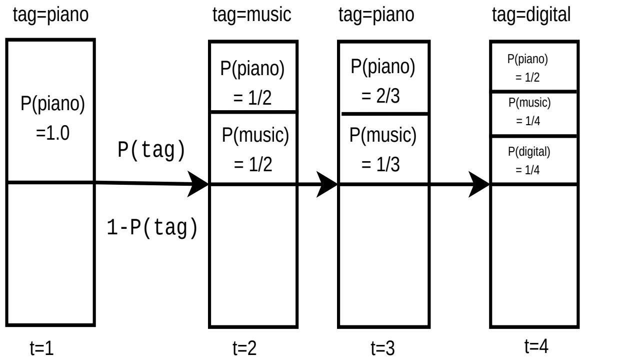

Figure 2: An example of how shuffling leads to preferential attachment

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 have been tagged.

While it is impossible for a generic model to assign a priori the exact informational value of a resource, it is possible to at least partially model the informational 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 informational value (I) of 0, and we assume that the informational value of a tag that retrieves no resources is also 0. Another tag that hypothetically selects only the resource needed, would have have an informational 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 informational value to simplify our exposition. Notice also that a user may use multiple tags and these tag combinations may have different informational values that are not additive. In our work with del.icio.us, we can empirically estimate the informational value of a tag by retrieving the number of web-pages a del.icio.us search with a tag (or combination of tags) returns and converting it into a probability, as done in Section 5.

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. [20]. 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 informational value is not additive, since $I(t_1, t_5) = 0$ while both $I(t_1)$ and $I(t_5) = 1$.

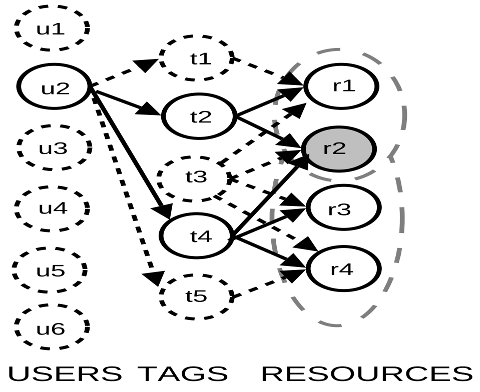

Figure 3: Tripartite tagging system graph used for search. The dotted edges represent options, while the dark edges represent a particular user engaging in a search for the shaded resource

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 preexisting 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 informational 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:

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

This formalizes a process that would give rise to a power-law via preferential attachment, but one where the informational 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. We are in process of collecting enough empirical data from del.icio.us to provide estimations of the model parameters such that we could compare model-generated results to empirical distributions. However, first we need to determine whether a power law actually arises from empirical data.

4. THE EMERGENCE OF POWER LAWS FROM TAG DISTRIBUTIONS

According to our model, there should be a connection between the relative rank of the tag for a given resource (defined by the ordering of the number of users who used that tag to mark the website), and the frequency of use of the tag for that particular resource. If our qualitative intuition about tagging systems as complex systems is correct, we hypothesize that this distribution should follow **power laws**. We consider the distribution of data from a subset of 100 heavily tagged sites (defined as those that were tagged over 1000 times) and we present the results of power law interpolation on this data. Finally, we discuss some of the reasons the observed distributions could emerge, based on the tagging behavior of individual users.

4.1 Power Law Distributions: Definition

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

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

Where α and c are constants characterizing the given power law. Without loss of generality, 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 represent straight lines. Therefore, the easiest way to check whether a distribution follows a power law is to apply a logarithmic transformation, and then use linear interpolation of the data points to determine the parameters α and c.

In our tagging domain, the intuitive explanation of the above parameters is that c represents the number of times the most chosen 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 $-\alpha > 0$):

$$Frequency(p) = \frac{Frequency(p = 1)}{p^{-\alpha}} \tag{3}$$

4.2 Results from Considered Data Set

As discussed above, in our experiments we considered a test set of 100 “popular” sites, defined as those that were tagged at least 1,000 times (the most heavily tagged sites in the considered data set were tagged over 30,000 times). For each website, we considered in our analysis the most heavily used 25 tags.

For this distribution, we first applied a transformation to a log-log scale and then we linearly interpolated the resulting data points (this was done individually for each site, though in Fig. 4 we show only the actual data, not the interpolated functions in order to preserve the clarity of the image). We computed the aggregate (cumulative) distribution for all sites (by summing up the frequency of tags that appear in each position) and interpolated the resulting points. The results are presented in Fig. 4 and Fig. 5, respectively.

In all of the cases, logarithm of base 2 was used in changing to a log-log representation. Note that the base of the logarithm does not actually appear in the power law equation (c.f. Eq. 1), but because we are interpolating empirical and thus possibly noisy data, this choice can influence errors recorded in the interpolation phase. However, we did not find significant differences from changing the base of the logarithm to e or 10.

To summarize our results, we found that the data points can (with some error) be linearly interpolated. The constants of the power law we found (see Equations 1 and 2), for the cumulative case, had the values: $\alpha = -1.28$ and $c = 18.3$. The results from linear interpolations for each of the individual sites (not shown graphically, due to lack of space) all had slopes in the same range, i.e. $\alpha \in [1, 1.5]$, with the average close to 1.28. Thus is can be said that the power law decay (i.e. slope) measured for the cumulative case is relatively consistent across individual sites (though the constant factor c varies, of course). Intuitively, this means there is a fundamental effect of the way tags are distributed in individual websites that is independent of the context and content of the specific website.

There is an important caveat though. We observed that somewhere between tags in positions 7 and 10 there is a considerably sharper drop in frequency than the general trendline. This means that for example, if we do a piece-wise interpolation for the tags in the first 7 positions and the last 15 we would get, in both cases, linear functions, but with slightly different slopes (while for the first positions α is closer to -1, it decreases slightly to -1.28 afterwards). Furthermore, as Fig. 4 shows, this effect largely holds for almost all sites in the data set considered, so it is not just attributable to noise, but a consistent effect of the way tagging is performed. We do not have yet a satisfactory explanation for this effect, but it might have a cognitive explanation, based on the number of tags the aver-

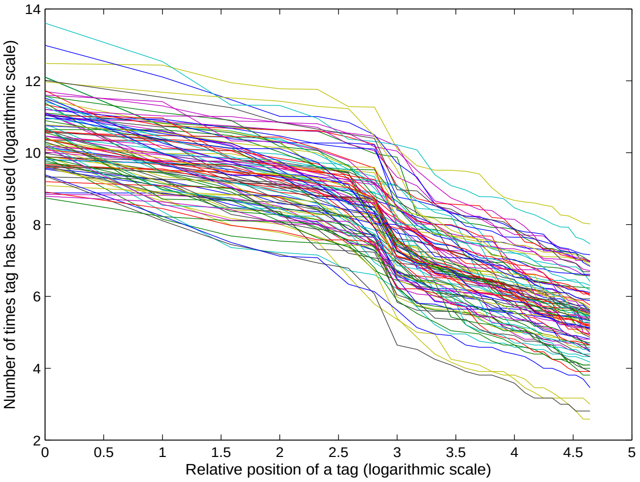

Figure 4: Frequency of tag usage, based on relative position. The dataset consists of 100 heavily tagged sites where for each, the most frequently used 25 tags were considered. The plot uses double logarithmic (log-log) scale: the horizontal scale gives the natural logarithm of the relative position, while the vertical scale gives the logarithm of the frequency of use

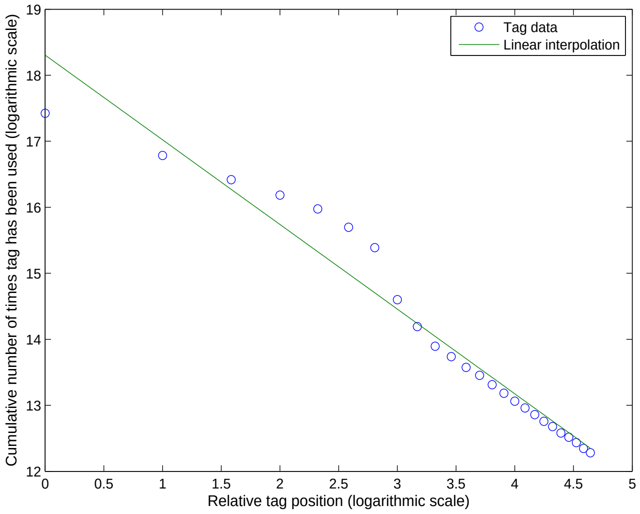

Figure 5: Cumulative number of tag usage frequency, based on their relative position. The plot is on a log-log scale: the horizontal axis shows the logarithm of the relative position, while the vertical axis shows the logarithm of the cumulative frequency of tags in that relative position. The best fit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

age user employs per website. However, this observation does not affect our basic result that tag distributions follow power laws.

We note, however that the above analysis refers to heavily tagged sites (tagged more than 1000 times), and considers the most used 25 tags for each site. We have also looked at a set of less popular sites, for which the power law interpolation produces somewhat less clear results - although some of these can be expected to become more heavily tagged and eventually evolve clear power law distributions. Furthermore, for each of the sites, below the first 25 highest-ranking tags there are a lot of unique tags that are used more scarcely (some only by a few people). This forms the "long tail" of the distribution, which does not usually follow the same power law decay pattern as the head.

5. CONSTRUCTING INTER-TAG CORRELATION GRAPHS

So while we have shown that power laws evolve on popular sites, is there any way to model the informational value that partially drives the process? We look at one of the simplest information structures that can be derived through collaborative tagging: inter-tag correlation graphs. First, we discuss the methodology used for getting such graphs. Next we illustrate our approach through an example, with 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.

5.1 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:

$$Dist(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}} \qquad (4)$$

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 up 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 can play a big role in actual structure retrieved and we note that there are more sophisticated distance measures proposed both in item-item collaborative filtering (see [14]), and from text mining literature. For this paper, cosine distance seemed to work well enough.

Next, 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 this weighed tag-tag correlation, by using a "spring-embedder" type of algorithm - in our case we preferred 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 5.3, 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.

5.2 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 does the user community of del.icio.us see as most related to "complexity" science (a problem which has traditionally elicited some discussion).[2] The visualizations were made on Pajek [1]. 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).
- 30 other edges (i.e. 45 edges in total for 15 tags) are considered (Fig. 7). These taken as the ones with the highest expected correlations, though in future work we'll consider more sophisticated methods for determining the cut-off, based on examining the 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 speaking, inversely related to the distance measure (as returned by the "spring-embedder" algorithm).[3] 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, since we feel it is more faithful to the proportions present in the data.

The results from the visualization algorithm do match well what one would intuitively expect to see in this domain. Some nodes are much larger than others, which, again shows the 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 (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, however, 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.

We should mention that in similar work, Mika [13] concluded (for another domain than the one in this paper), that filtering based on users produces more useful results than filtering based on items. Although we cannot preciscely assess whether the same measures were used for building the graph, for this particular domain, our approach produced reasonably useful results. Before reaching any definite conclusions, however, further work examining the results with different similarity criteria, with different methods of applying

[2] 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.

[3] 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.

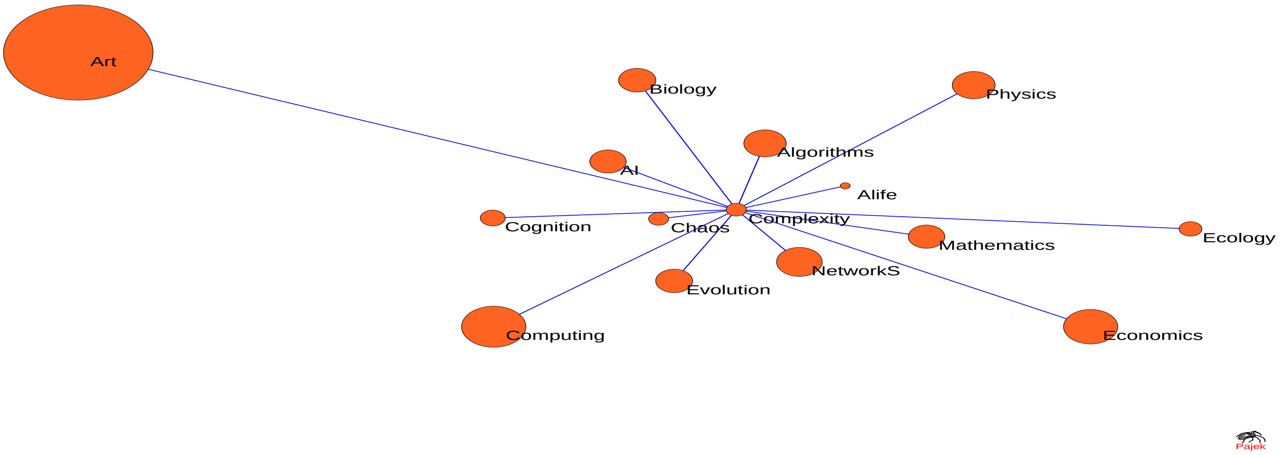

Figure 6: **Visualization of a tag correlation network, considering only the correlations corresponding to one central node "complexity"**

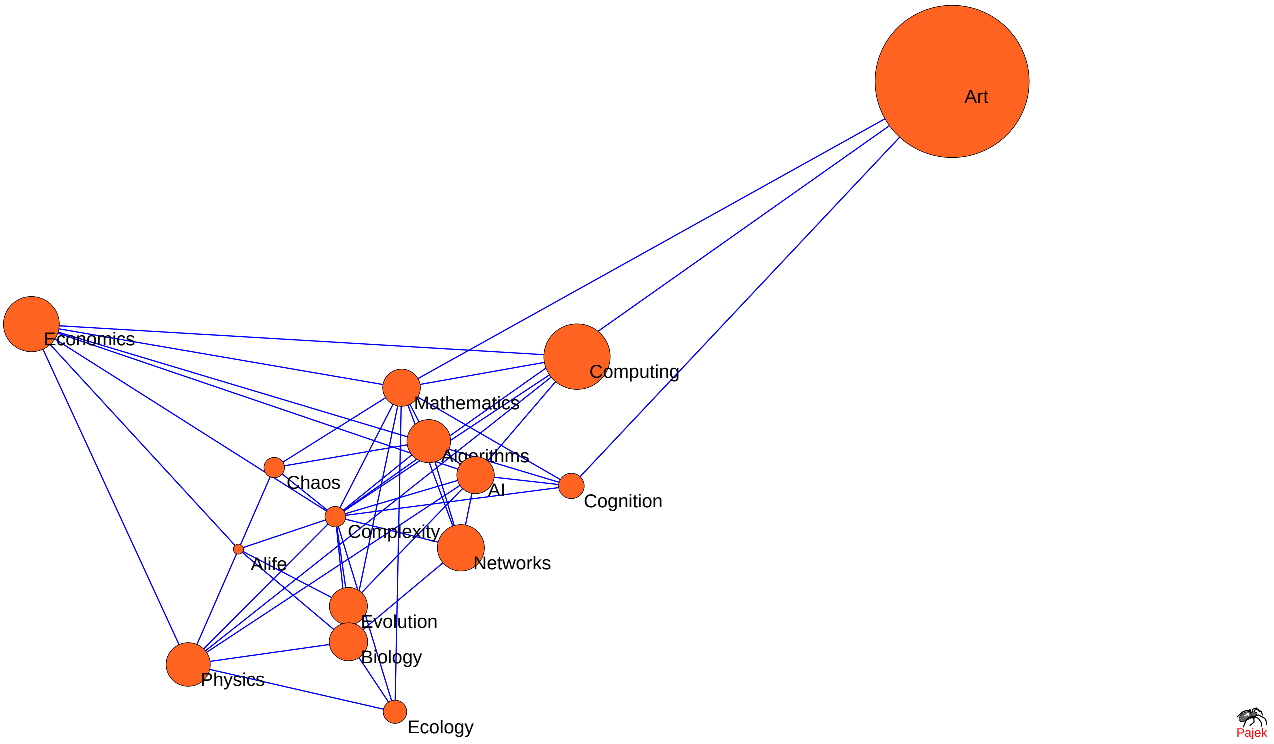

Figure 7: Visualization of a tag correlation network, considering all relevant correlations

spring-embedder algorithms and based on larger and more precise data sets is needed.

Overall, we expect that when applying fully automated retrieval methods on larger data sets (for example, not choosing the subset of tags considered "by hand"), the same problems we identified for the tag distributions on individual sites would appear. This means, some very common, general-purpose tags could have a very large weight and centrality, although for a particular domain or application they do not carry too much information.

5.3 Tag Graphs and Human Language Networks

In the previous section, we have shown that tag networks can be easily constructed and visualized and that they could prove useful in simple information retrieval. However, exploring the properties of these tag graphs (e.g. node centrality, degree distribution, etc.) - 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 in our further modelling is the 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, 4] 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, in fact a "small world" structure.[4] Furthermore, this protostructure is remarkably similar across different languages.

Graphs which exhibit a small world network effect have the distribution of the mean degree of the edges follow Zipf's law.[5] 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 [5], they simulate a version of Zipf's classic generative model of human language: speakers prefer to use ambiguous, 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 content.

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. The resources (websites) could correspond to the objects in the real world - that need to be described by the language, the users to the speakers of the language, and the tags to the tokens of the language (i.e. the words). Tags also likely have a Zipf law distribution of node degrees, and while the massive data harvesting needed to show this is difficult, even our provisional results do point in this direction. In such a case, generative models proposed by Sole et al. [5] may be useful to explain the online behavior of taggers as regards informational value. 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 (i.e. "hearers") who need to use this tags to find as precise as possible information resources and so use tags with the highest informational value. In our generative model shown in Section 3, the results of this "least effort principle" would be the parameter λ. However, the next question we have to encounter, one that is particularly relevant to the Semantic Web community, is given that a stable power law distribution emerges from highly tagged sites, is there any methodology to formalize that distribution into an ontology?

6. ONTOLOGY EXTRACTION FROM TAGGING

Ontology extraction from tagging is a difficult research question given the ambiguity in formally quantifying what constitutes an ontology. In this section we demonstrate the first brush strokes of a speculative ontology extraction methodology using a tagging example. One should only apply any sort of ontology extraction from tagging after the tag distribution of the site has reached a stable power law. In order to avoid the higher variance of the tail of the power law distribution, we look only at what del.icio.us measures to be the most "Common Tags," which is the upper end of the power-law distribution. Since RDF (Resource Description Framework) is the most basic component of the Semantic Web, we will use it as our extraction format instead of a more complex OWL (Web Ontology Language).

Since we can expect to extract only the most basic structure from tag space, the very simplicity of the RDF "triple" structure is actually a boon. RDF differs from traditional knowledge representation systems in two major ways. First, every statement is composed on three atoms arranged into a "triple" in the following manner: *Subject, Predicate, Object*. Second, each atomic subject, predicate, or object is denoted by a unique URI. There are a number of special properties given to us by RDF, although we will focus only on sub-classing. The predicate *rdfs:subClassOf* is defined by RDF Schema to "to state that all the instances of one class are instances of another," while *rdf:type* states that "a resource is an instance of a class" [8]. Notice that *rdfs:subClassOf* is how RDF deals with the idea of hierarchies, something that folksonomies are supposedly incapable of providing.

One finds in many folksonomies that a hierarchy is not eliminated; instead the user in some instances "types the hierarchy out" or "mixes a single name among multiple tags," such as repeating the two tags "digital piano." A comprehensive ontology-based system would already formalize "digital piano" as a single class that is the sub-class of "piano." There are also ***facets***, or structured relationships that are not strictly hierarchical relationships. Traditional facets are given by relationships like that each "Person" has a "name" or that photos are taken on a "date." These types of facets clearly fits within the "triple" structure of RDF, such as *foaf:Person foaf:name xsd:string 'Harry Halpin"* or *flickr:ryansking/160217283/ dc:date xsd:date '2006-05-25"*. However, the real problem with any automatic mining of facets is that the property is almost always implicit. For example, many people just add the date to Flickr photos without adding a "date" tag, much less a "name" tag for a person when mentioning their name. Worse, the object is usually variable, as the dates and names of people change. However, while facets may be beyond our grasp, we can attempt to capture the sorts of structure exemplified by our first example: When two tags are used to describe a single class or instance, and when one class is the subclass of another class. In our example, we inspect an article on the Web that reviews "digital pianos." In this context, the term"digital" is closely related to "piano" since the term "digital" is never men-

[4] A small-world graph is a graph in which any two nodes are connected by a path of small maximum length - usually 2-4.

[5] The degree of a vertex is the number of edges connected to that vertex. The distribution of the degrees across all vertexes is an important property of a graph

tioned without "piano," so we can safely say that the two terms "digital" and "piano" can be considered in some contexts just "digital piano." However, "piano" is a more abstract class than "digital piano" since "piano" often appears in the context of other tagging instances without the tag "digital."

Since we do not have access to the entire del.icio.us tagging database, we have to focus on creating ontologies from a given resource, looking at what knowledge can be extracted from a webpage about a digital piano rather than the set of all pages about digital pianos tagged by users. The network is a directed and weighted graph whose weights are given by how many times two given tags ($t_x \| t_y$) co-occur in a tagging instance (i.e. when a user tags the site) and dividing this by their total number of occurrences of the tags, so $C(y, x) = \frac{t_x \| t_y}{t_y}$. Because a given tag may co-occur with multiple terms in a single tag and also a tag may co-occur with non-common tags $\sum(c_x \| c_y) \neq c_x$. In our example, the most common tags are "piano" (8), "music"(4), "digital"(2), and "review"(2)." The single tagging instance "piano music digital review select" has "select" eliminated since it is not a common tag, and the tag "piano," while it appears 8 times in total, appears in this single instance with three other tags. An example graph of what this "common tag" graph looks like is shown here in Figure 8. From this graph, one can tell that the word "piano" co-occurs most frequently, even with a number of words that are uncommon. Lastly, one can then tell that for common tags, the tag "digital" always occurs with "piano" while the word "music" usually occurs with "piano." The tag correlation graph is shown in Figure 9, plotted using the same methodology as in Section 5. This visually estimates the information values.

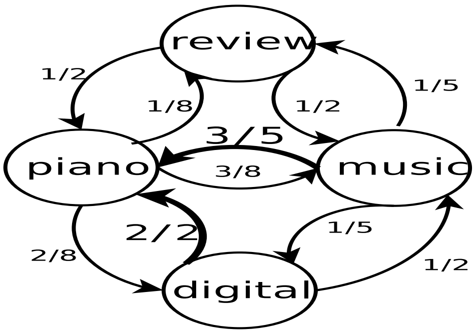

Figure 8: Example "common tag" network for a resource in del.icio.us

The two heuristics we use to extract ontologies are, given two tags in the common-tag graph x and y whose "common-graph" values are given as $C(y, x)$ for the directed edge between y and x:

- If the $I(y) < I(x)$ and if $1 > C(y, x) > \epsilon$ then *y rdfs:subClassOf x.*
- If $I(y, x) \leq I(y) < I(x)$, and if $C(y, x) = 1$ then *"y x" rdfs:subClassOf x.*

The first rule merely states that for a given tag that may belong to a more abstract class than another tag, the more abstract tag should have a lower information value (i.e. retrieve more pages) than the less abstract tag. Also, the rule states that the more abstract tag should also be used in combination with other tags besides the less abstract tag, but used with the less abstract class often (as quantified by ϵ). In this manner, we can guess that the more abstract class is not equivalent to the less abstract class. The second rule is similar, that if two tags are to be considered a single "compound tag resource," then those two tags should appear together all the time in tagging instances for a given resource. Furthermore, if the informational value of one of the tags by itself is less than the rest of the tags of the newly created "compound tag resource" then this "compound tag resource" is a subclass of the tag with the lower informational value.

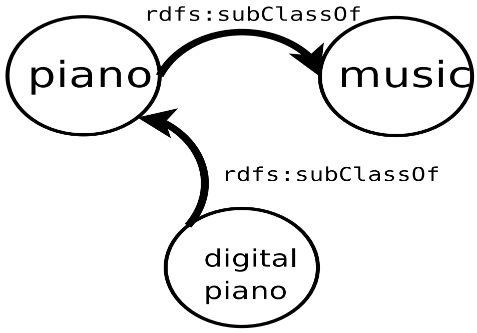

Figure 10: An example extracted ontology as a RDF graph.

These two rules are illustrated by their application to Figure 8. First, the only tags that are always applied together are "digital" and "piano," and since "digital" has a higher informational value than "piano," and both have a higher informational value than "digital piano" (as should be trivially expected), then "digital piano" is a subclass of "piano" (*ex:digital-piano rdfs:subClassOf ex:piano*). When the first rule is applied with an $\epsilon = 0.5$, only the connection between "music" and "piano" is strong enough to (0.6) to qualify without being absolutely a perfect correlation (1.0), and since "music" has a higher information value than "piano," therefore "piano" can be considered a subclass of "music" (*ex:piano rdfs:subClassOf ex:music*). In this particular example, the tag "music" is being used as shorthand for "musical instruments." While this technique is in need of refinement and widespread testing, we do think simple rules like this might be useful within the context of a stable tagging distribution of a single resource.

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 with outlining a principled generative model of tagging. Unlike other proposed models, our model is based on Mika's formalization of tagging and incorporates the informational value of tags which we believe allows for a more complete account of tagging. We also have shown that our model formalizes many of the common-sense observations made by people who are informally studying folksonomies. Because we currently lack the empirical data (since data is not made easily accessible) to currently estimate the model parameters using a training set in order to compare the results of the generative model to an empirical test set, this will be our next goal.

Using empirical data, we have shown that tagging distributions tend to stabilize into power law distributions. This suggests that consensus around the categorization of information driven by tagging behaviors occurs. Using example domains, we have explored the most empirically challenging aspects of the generative model: the informational value of a tag as a function of how many pages a given tag can retrieve using a search. We examined how this information can be used with multiple tags to visualize correlation

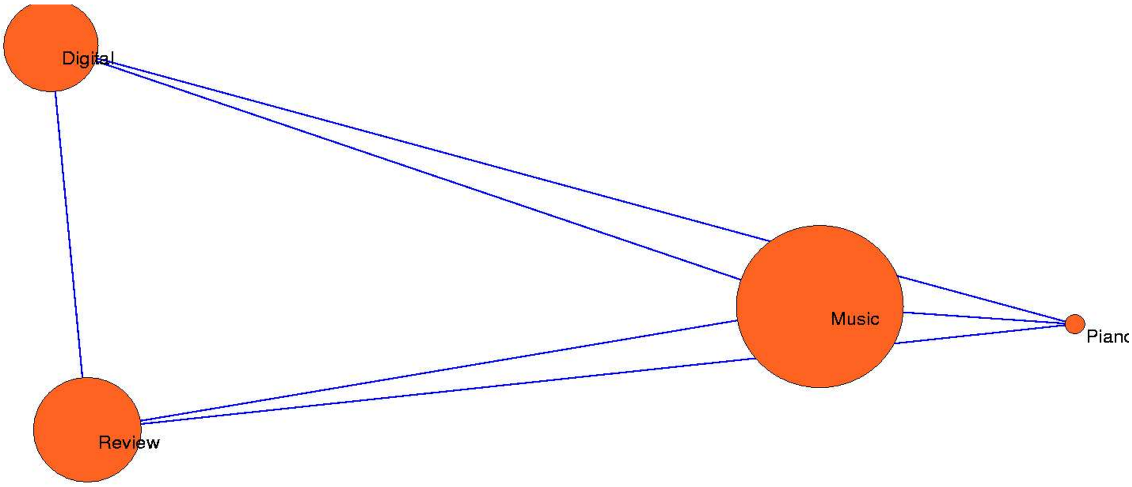

Figure 9: Tag-tag correlation graph for the individual tags of a resource about digital pianos

graphs that lend insight into the categorization process and into existing intuitions about how concepts are related. This provides preliminary evidence that some type of latent classification scheme and taxonomic structure may lie behind tagging.

Finally, we have shown a simple methodology for extracting Semantic Web ontologies (in particular RDF and RDF Schema) that can be used on tagged resources whose tagging distribution has stabilized into a power law. Again, we need more empirical data to validate these ontologies and produce them en masse, and currently we are gathering this data and as such will likely refine our heuristics in time. From these results, it seems quite plausible that folksonomies and ontologies, which are merely new incarnations of categorization and classification respectively, are not mortal enemies, but fundamentally compatible, as tagging-based categorization can evolve into stable classification schemes that can be formalized as ontologies. Further work will contribute more rigorous analyses to these observations.

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] Bela Bollobas. *Random Graphs*. Academic Press, London, England, 1985.

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

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

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

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

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

[8] Pat Hayes. RDF Semantics, W3C Recomendation, 2004. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

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

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

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

[12] Peter Merholz. Metadata for the masses, 2004. http://adaptivepath.com/publications/essays/archives/000361.php.

[13] Peter 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.

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

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

[16] Clay 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] Luc Steels. The evolution of communication systems by adaptive agents. In *Adaptive agents and multi-agent systems*, pages 125–140. Springer LNAI, 2004.

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

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