

The Social Structure of Tagging Internet Video on del.icio.us

John C. Paolillo and Shashikant Penumarthy

School of Informatics and School of Library and Information Science

Indiana University, Bloomington, IN 47401 USA

Email: {paolillo,sprao}@indiana.edu

Abstract—The ability to tag resources with uncontrolled metadata or “folksonomies” is often characterized as one of the central features of “Web 2.0” applications. Folksonomies are said to support emergent classification, where the semantic value of the tags and their relation to one another is worked out through a negotiated process of users applying their selected tags and seeing what others have tagged the same way. Few studies exist to show how folksonomic tagging is actually done, and to what extent users share each other’s tagging patterns. In this paper, we present the results of a social network analysis of two months worth of tagging Internet video on the social bookmarking system del.icio.us. The analysis reveals that specific videos are tagged in fairly coherent ways by a relatively tight group of users. However, contrary to our expectations, there does not appear to be much re-use of tags across different content, or even very many users tagging more than a few similar items. Overwhelmingly, specific clusters of tags and users are associated with individual video links. This result suggests that tagging bookmarks is highly local, and the overall collection of tags is unlikely to result in a coherent globally navigable classification system.

I. INTRODUCTION

The World-Wide Web, because of the many services offered using its protocols, raises in new ways the age-old problem of how to effectively organize and retrieve information. An early solution to this problem is the “bookmark”, i.e. an easy-to-locate link to a resource, generally with some kind of mnemonic label. Most modern web browsers have such a feature, with the associated problem that bookmarks remain on the client computer where the link was originally stored. Yet another solution to the information organization problem is “tagging”, or the use of unstructured, user-applied metadata. Numerous applications now exist in which some form of tagging is used to provide access to information, from blogging (e.g. WordPress) to social networking (LiveJournal, Friendster) to online photo-sharing (Flickr) to video sharing (YouTube), etc. All generally permit open-ended text input for assigning tags to content, where tags are separated by a simple delimiter (e.g. carriage return or space). This metadata is stored on the same system as the content which it tags, hence it is accessible from any client computer. Social bookmarking systems, of which del.icio.us is one example,¹ combine both methods of organization in a single application. Users save links to online resources, and label them with free-form

metadata. Furthermore, social bookmarking systems generally also provide some means of exposing different people’s bookmarks and tags. In this way, people can choose tags that converge with those of other users, or provide alternative tagging of other people’s tagged resources. A review of several social bookmarking systems can be found in [1].

While social bookmarking systems are growing in popularity, it is not clear to what extent or in what ways they improve information access. The deliberately designed social nature of collaborative tagging systems naturally raises questions about the social nature of the tagging process. The main motivation for using social bookmarking systems would appear to be the prospect of benefitting from the organizational efforts of other people. But are these systems really creating a coherent and easily navigable categorization system? Or is the organization of information only locally coherent? What is the extent of sharing actually observed between users? And to what extent are users actually collaborating with each other while tagging different resources?

In this paper, we attempt to shed light on these questions, using a social network analysis of video tagging on del.icio.us. Our findings reveal a highly local structure to individual tagging events, concentrated around individual links. Very little coherence across links is observed in the two months worth of video bookmarks that we have observed. These findings suggest that global coherence in the tagging of video, at the very least, does not emerge within this time scale, and that users require either longer time scales or better system supports to arrive at a globally coherent, navigable organization of resources usable by others. We interpret these findings in terms of the different purposes users may have when tagging various information resources.

II. BACKGROUND

A. Metadata and tagging

Collaborative tagging differs from traditional organization systems in certain important ways. At the same time, it shares certain important purposes with traditional information organization schemes, and must be understood in reference to them.

First, tags are a form of metadata, or data which label other data for the purpose of organization and access. Traditional forms of metadata, such as of MARC records or XML documents, are structured, i.e. different pieces of information

¹Other popular social bookmarking systems include CiteULike (www.citeulike.org) and Furl (www.furl.net).

are distinguished from one another as providing different sorts of information about the object they are attached to. Hence, author, title, year of publication, keywords, etc. can all be distinguished from one another. In a collaborative tagging system, tags are not systematically distinguished from one another this way. Rather, all tags are treated as equivalent for the purpose of retrieval. This “equanimity” toward metadata types (tags) has certain advantages. Users need not become trained specialists in an organization system in order to use tagging. Everything the system can do can be accomplished without detailed knowledge about how many different types of information are recognized, which types of information may be used with specific kinds of resources, etc., all things which are formally specified for traditional organization systems. The disadvantage of tagging is that neither user nor system can be sure in any specific instance if a particular tag refers to any specific type of information (e.g. author or subject of a page referred to by a link).

A second characteristic of tags as metadata is that they do not belong to a formal taxonomic hierarchy of concepts. This has two consequences. The first is that the organization accomplished by tagging is generally considered to be a form of categorization, rather than classification, which normally implies a hierarchical form of organization [2]. Each tag applied to an object ascribes some set of characteristics to the content it tags. The second consequence is that, because the tag vocabulary is not “controlled” or standardized, the categorization produced is informal, and not guaranteed to be the same from one person to the next.

Hence, tagging results in an organization system in which there is no explicit hierarchical relationship among different metadata labels, although there may be implicit associations among labels that are used on similar sets of objects. The exact nature of this similarity cannot always be deduced from the tags themselves on account of semantic issues such as synonymy, polysemy and variation in specificity [3], [4]. In other words, tags have to be interpreted in terms of their different contexts of use. These issues raise questions about the relative utility of collaborative categorization systems and more traditional forms of organization.

B. Folksonomies

Social metadata tagging systems like those described above are generally called “folksonomies” [5], a term formed by blending “folk” and “taxonomy”). The term is actually misleading because taxonomies are hierarchical classification systems whereas folksonomies are non-hierarchical categorizing systems.² Although the intent of [5] was not to describe these systems as taxonomies, the term has stuck. In fact, it might be questioned whether collaborative tagging systems are really categorizing systems or even collaborative. A more appropriate term for these systems might be social tagging systems [7].

Folksonomies or social tagging systems are governed by a theory of their workings that addresses the cognitive and social

functions of tagging, expected behaviors around tagging, and the design features of tagging systems to support the desired behaviors. For example, Golder and Huberman [3] identify a number of functions that tags perform and also suggest a model for collaborative tagging based on imitation and shared knowledge. In a similar vein, Marlow et al. [7] enumerates a number of design dimensions along which one can categorize existing social tagging systems and provides an overview of the user incentives for using these systems. Marlow et al. also suggest that users’ social networks might influence the vocabulary they adopt, so that users connected to each other have a higher probability of using similar vocabularies than users that are chosen at random. In other words, it is expected that different people’s tag usage will standardize around a set of tags (and meanings for those tags) that are shared by one’s local social sphere.

More recently, work has been undertaken from a stochastic modeling perspective, e.g. that of Cattuto et al [8] which used a modified version of the Yule-Simon stochastic model [9], [10] to incorporate effects of aging (tags used more recently are more likely to be used again than those used further in the past) and found that the temporal autocorrelation function of the tag sequence is matched quite well by the model. This essentially means that the process by which tags arrive in a collaborative tagging system can be simulated by the same stochastic process used to model the construction of free-form texts (but with long-term memory). Further, by addressing the correlation between high-frequency and lower-frequency tags, it suggests that high-frequency tags may effectively categorize the lower-frequency tags, possibly creating a latent hierarchical structure of tags. In this vein, [11] computes a few properties of the tag co-occurrence network by links and reports a power-law exponent of 1.4, and a clustering coefficient of 0.06 (higher than that for a random network), also suggesting that there may be interesting latent structure to be discovered among the network of tags.

The most useful perspective on the issues raised by social tagging would appear to be the Social Network perspective, in which relations among different entities are observed in order to explain the relationships among tags, users or links. In a social bookmarking system, each of these three types of entity participates in its own kind of network, and all three types of network can be studied together by the same means. Alternatively, the entire collection of users, tags and bookmarks may be considered as a tripartite network. For example, a pair of users may be considered to be related if they have tags or bookmarks in common. Similarly, a pair of tags can be related to each other through common users or common bookmarks. Finally, bookmarks can be related to each other through common users or common tags. [12] presents an interesting quantitative analysis of a tagging system by projecting the tri-partite network formed by users, tags and bookmarks onto successively simpler bipartite and uni-partite networks. The result of this is a uni-partite network where two resources or links are connected to each other if they are highly correlated with respect to users and tags. By thresholding, it

²Folksonomies must also be distinguished from “folk taxonomies” [6], which are non-formalized but relatively stable classification systems.

is possible to eliminate weaker correlations to reveal islands of strong correlation, possibly corresponding to the genres of content in the system.

C. Tagging internet video

The range of tags found in a social tagging system is as diverse as the set of its users and the content being tagged. The processes by which tags are applied to bookmarks also varies considerably to encompass a broad range of user intentions. Tagging is a two-step process. First, a URL gets bookmarked and second, the bookmark is tagged. Although tagging and bookmarking happen together, a significant percentage of bookmarks are not tagged (in our data, about 12% of bookmarks have no tags). Therefore, it seems reasonable to assume that the motivations behind tagging and bookmarking don't always overlap. Moreover, social tagging systems have a large number of secondary applications, created by developers who use the del.icio.us API to provide enhanced functionality³. Nonetheless, Marlow [7] was able to identify a small number of recognizable types of user incentives for bookmarking and tagging.

Although tagging in general has its uses with respect to organization and sharing of content, the nature of the medium can lead to differences in the way it is done. We will not attempt to present an exhaustive discussion of this issue here. However, it is worth looking at characteristics of video in terms of how it differs from other media bookmarked on del.icio.us in order to contextualize our results.

Videos on the Internet are heterogeneous in content and form – they fall into readily identified genres, such as music videos, movie and game trailers, humorous amateur productions, etc. Hence we had anticipated that these distinct types of video content might be tagged using different tags, or appeal to different user groups. The diversity of sources of videos also have possibly non-trivial effects. For instance, a website such as YouTube.com allows tagging as well as rating of videos and many bookmarks on del.icio.us simply point to a video on YouTube; hence the incentive to further tag such bookmarks on del.icio.us is low. On the other hand, videos put up by individual users or small websites are often mirrored on other websites so that heavy demand for the video does not overwhelm the original website. This means that the channels of distribution of content on the internet are varied and their effects on the diffusion of content are not understood.⁴ Whether the original video or its mirrored version is bookmarked can have significant impact on its continued availability. This implies that video links have a relatively high likelihood of being tagged by many different people, thus increasing the size of the potential sample we can collect.

Mirroring causes one additional issue that is quite unique to video: different URLs can point to copies of the exact same video, but they are treated as distinct bookmarks by del.icio.us.

³See <http://www.econsultant.com/delicious-by-function/index.html> for a list of such applications.

⁴Our initial intention was to study this diffusion, and this aspect of the research is still ongoing.

This means that with respect to such bookmarks, similarity of users' interests is determined to a great extent by their tag usage.

Since the system:media:video tag is automatically attached to bookmarks, we are able to access a stream of content whose characteristics are relatively independent from the users' tagging behavior. Otherwise it is very difficult to obtain a data sample that is not biased in some way toward particular users, tags or content. Consequently, we our focus is not on the behaviors of specific users. However, since we are interested describing large-scale effects we will not worry about this issue here.

We can now start looking at differences between video and other types of media with respect to bookmarking and tagging. We borrow the set of categories outlined in [7]:

Future retrieval Videos typically don't come with an abstract. Therefore, although the first few seconds of a video might give the user a general idea of its *genre*, it does not provide enough information for a user to decide whether or not to bookmark it. However, there are videos whose value, entertainment or otherwise, is easily understood by viewers without actually looking at the video. Videos which lend themselves to such immediate identification include ones heavily discussed in the blogosphere, a recent notable example being the video of Stephen Colbert's speech at the White House Correspondents' Dinner⁵. In cases like this, where the context for a video is provided by other means, a user may choose to tag the video using a tag such as *'TODO.'* as a way to remind oneself to view it later. Another tag of a similar nature that we encountered was *'blogthis'*, which seems to indicate tagging being used for fine-grained distinction between tasks. The mere act of bookmarking can sometimes be enough for task organization purposes, particularly when a user does not bookmark many URLs. In order to follow up on them, all one needs to do is filter bookmarks by date to find the most recent ones.

Contribution and sharing Tagging can be used to draw boundaries around groups of videos in terms of what they are about. For example, videos tagged with the tag *'rails'* are immediately recognized as being about the *'Ruby on Rails'* framework. Another example is using tags to identify the source of a video(e.g.: *'crooksandliars'*), which can then be used to make judgements about the trustworthiness, objectivity and neutrality of a video.

Attract Attention We encountered examples of URLs that were tagged with legitimate and popular tags that made them seem related to videos popular at the time, but were not. For example, the tag *'rails'* mentioned before was used to tag a bookmark pointing to a page that attempted to exploit security holes in the browser for malicious purposes.

Opinion Expression The overwhelming majority of bookmarks that are tagged with value judgements have positive evaluations. In our data, only about 5% of tags have negative connotations (*'stupid'*, *'(:'*). Among the positive ones, nearly

⁵See <http://video.google.com/videoplay?docid=-869183917758574879>

70% are associated with humor ('funny', 'humor'). The rest express approval in some form ('cool', 'amazing'). This seems to suggest that the most users use del.icio.us to keep track of videos that are either entertaining or extraordinary in some way.

There are two other categories that [7] mentions: *Play and Competition* and *Self-presentation*. In our data, we haven't found much evidence of tags being used for these purposes.

III. METHOD

A. Analytical method

The analytical method adopted here is most closely related to the approach of [13], [14]. Specifically, we employ the techniques of Social Network Analysis [15], [16] in order to illuminate the relations among users, tags and links in a social bookmarking system. Our analysis relies on the general technique of *social network reduction*, sometimes called "blockmodeling" [17], in which structurally equivalent nodes in a network are reduced into a small number of equivalence classes. Many techniques exist for social network reduction, but the most common of these uses hierarchical cluster analysis to identify a suitable partition of the nodes into "social positions", or groups of nodes whose linkage to other nodes is similar. Such classes of nodes are said to be "structurally equivalent". This approach has proven quite fruitful in identifying patterns of semantic coherence in analyzing free-text metadata such as the interests in personal profiles of LiveJournal users [13].

Social network analyses generally contend with data that are either uni-modal (nodes linking directly to other nodes over some relationship such as "friend of") or bi-modal (e.g. affiliation networks of people who are members of the same social club). Techniques of network analysis allow one to move between bi-modal and uni-modal forms of analysis. One way this can be done is to employ Principal Components Analysis (PCA) of a bi-modal affiliation network for identifying structural equivalence of nodes.

One result of a PCA is a projection of the nodes into a multi-dimensional space, the principal components scores. Each dimension captures a certain proportion of the variance in the original data. Therefore, by choosing a small number of dimensions that together explain most of the variance and projecting the original data into the space formed by these dimensions, PCA allows us to reduce the number of dimensions that need interpretation.

Distances between any pair of nodes in this reduced space can be treated as estimates of their relative network proximity, or inversely, in terms of the strength of the tie between them. Structurally similar nodes will thus be located close to one another in the principal components space, and Hierarchical Cluster Analysis can be performed over the principal components distances to partition the nodes into structurally equivalent classes.

An advantage of PCA is that a similar treatment can be done for rows as for columns, so in bi-partite networks, one can construct networks for both row entities (e.g. users) and

column entities (e.g. tags). In addition, one can take the equivalence classes for both rows and columns and examine the reduced bipartite network. All of the information available in the original network is retained and made available for interpretation.

For a three-mode network such as that among users, links and tags, some adaptation of the approach is necessary. From a common set of data, three distinct bimodal relations could be studied: user-link, user-tag and link-tag. Each potentially tells us about a different aspect of the data: the user-link relation tells us about how different kinds of users are oriented toward different kinds of content, the user-tag relation tells us how different types of users are oriented toward different types of semantic labels, and the tag-link relation tells us about association of different semantic labels with different types of content. While conceptually these are independent relations, they are actually mutually inter-dependent. A user chooses a specific set of tags when bookmarking a specific resource, in a single event. Hence, it is best to examine all three nodes in a network of bookmarks simultaneously. However it is not clear how to do this. If one were to treat individual tags as distinct relations among the entire set of users and links, one would obtain a three-dimensional data array, whose processing would be unwieldy.

The approach adopted here, after exploration of the three bipartite networks, was to focus on a common set of links, and to arrange a single matrix of users and tags (rows), and links (columns). The same matrix is easily partitioned into a user-link matrix and a tag-link matrix (it is actually the row-wise concatenation of these two matrices). When relations among the links are studied, we use a column-wise z-score normalization, primarily to account for the different numbers of users and tags associated with each link. Similarly, when users and tags are to be studied, we use a row-wise z-score normalization, to account for the popularity of different tags and the level of tagging activity of different users. Hierarchical cluster analysis was performed on the principal components scores of users and tags separately. We confirmed the analysis by comparing the PCAs of the bi-partite networks with those of this combined PCA that the relations among users or tags are common to both forms of analysis; the only difference is that of a small rotation by a few degrees of some of the principal component axes.

After performing the network reduction on the user/tag-link PCA, the user-tag relations were studied by partitioning the original matrix into separate user-link and tag-link matrices, and taking the inner product of the two. The cells in this matrix were then aggregated according to the structurally equivalent positions for users and tags already identified, and the resulting set of relationships was plotted as a two-mode network. By using color to represent the relations between the user or tag clusters and clusters of links, we are able to represent all three types of information in the final network plot.

B. Data

To investigate our questions, we chose to examine tagging of video on the social bookmarking site `del.icio.us`. This site is actually one of the first and best-known social bookmarking sites and also one of the first sites to explore collaborative tagging.⁶ `del.icio.us` provides a simple interface for saving and retrieving bookmarks and their associated metadata. It also automatically provides RSS feeds for any user, tag, or combination thereof, so it is relatively easy to monitor tagging activity on `del.icio.us` for a chosen sample period. Finally, `del.icio.us` automatically tags bookmarks with special “system” tags when they have certain recognized filetypes. For example, all videos bookmarked are tagged with the system tag “system:media:video”, and using the automatic RSS generating feature one can easily obtain a feed of video links and their associated metadata updated in real-time.

We collected bookmarks related to video by parsing the RSS feed corresponding to the `system:media:video` tag between Feb 7, 2006 and April 5, 2006. For each bookmark, we collected the username, bookmark URL, the date the bookmark was created and tags assigned by that user to the bookmark. The dataset contained 4247 bookmark events containing 2535 unique users, 1523 unique links and 2089 unique tags. From this data we extracted two matrices: user-link, and tag-link. Each of these matrices represents the cooccurrence of the entity represented by the row with the entity represented by the column. For example, the user-link matrix contains users as rows and links as columns and every element in the matrix indicates how many times a particular user used a particular tag. The sum of any row of this matrix is the total number of links bookmarked by the user corresponding to the row. Likewise, the sum of each column of the matrix is the number of times the link corresponding this column was bookmarked across all users. We then reduced these matrices by eliminating all links that were bookmarked less than twice, and removed all zero rows (tags or users). At the end of this process, we were left with 1394 users, 1013 tags, and 408 links. The two matrices were then row-wise concatenated for further analysis.

IV. ANALYSIS

In this section, we first consider relationships among the links, to ascertain if there are coherent clusters of content observable in the data.

A. Links are not strongly clustered

For the data we have observed, it turns out to be somewhat difficult to ascertain clusters of related content among the 408 links. We began by examining the user-link and the tag-link relations separately, but these proved to be unrevealing. We then conducted an analysis of the combined user/tag-link matrix, using column-wise normalization (centering and scaling), which analysis is presented below.

⁶`del.icio.us` was acquired in early 2006 by Yahoo!, as part of a series of Web 2.0 acquisitions.

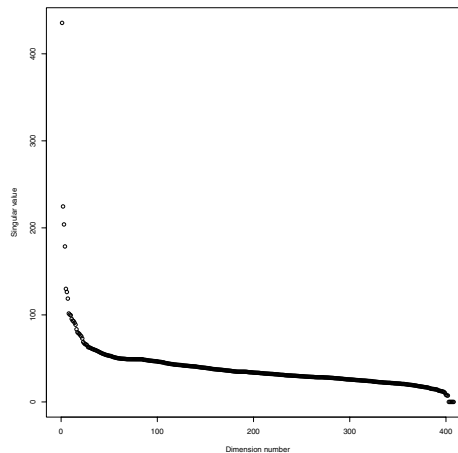


Fig. 1. Scree plot of link by user and tag matrix

Figure 1 shows the scree plot of this analysis, used to suggest a reasonable number of dimensions to investigate. The “elbow” of the scree plot is generally taken to be the point at which additional extracted dimensions result in no significant additional information. Here, the elbow is in a region between 20 and 30 dimensions (i.e. within the first 30 singular values). Hence we use no more than 20 dimensions in the subsequent analyses and interpretations. This result compares closely with that of [13].

We begin examining the associations among links with a hierarchical cluster analysis of the column principal component scores, using Euclidean distance as a measure of similarity, and Ward’s clustering method. Figure 2 presents a dendrogram plot of this cluster analysis. From this plot, we can see that there is a tendency toward “chaining” – small clusters are swept one-by-one into a single larger cluster. This occurs in spite of a general bias of Ward’s method to find spherical, rather than elongated clusters.⁷ Moreover, it is quite different from what was obtained using identical methods in analyzing LiveJournal profiles [13], [14], where large and clearly separable clusters of interests were obtained. Consequently we acknowledge that clusters of links in our data are generally small. If we examine the first cluster of links, we can see that it is composed of a relatively loose association of four clusters (i.e. they join together at a height between 30 and 40). Within each cluster, members are separated at a height of less than 10. Reading across the dendrogram, similar-sized clusters occur throughout the data set. A cut at this height reveals 47 distinct clusters, with a mean cluster size of 8.7 links and minimum and maximum cluster sizes of 2 and 28 respectively.

Since 47 clusters is too large a number to visualize successfully, we select instead a cut into six distinct clusters, which in addition to the first four clusters splits the second top-level cluster into one larger cluster and one smaller cluster, the smaller of which is about the size of the first four. This

⁷As are typically found by complete linkage and single linkage methods.

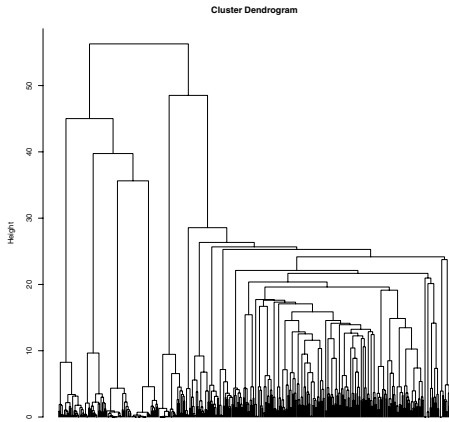


Fig. 2. Dendrogram of hierarchical cluster analysis of links, Euclidean distance and Ward's method.

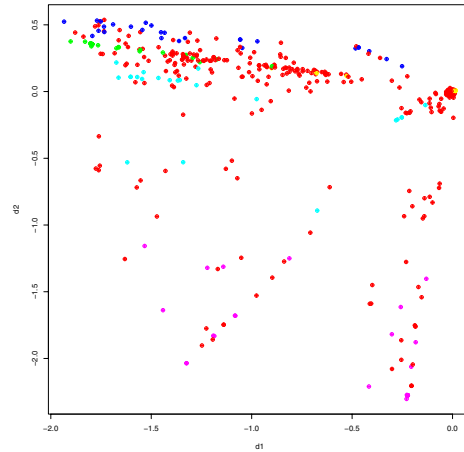


Fig. 3. Principal components plot of link clusters, dimensions 1 and 2.

cut can be accomplished by cutting at a height of 30. We then sorted the video links into their respective clusters, and listed them in Table II so that the clusters could be interpreted (Cluster 1 has been omitted from Table II because of its large size; also, the long URLs have been clipped down to just their final filenames, and delimiters such as `_` and `-` have been replaced with spaces to improve formatting and readability, but upper and lower case in the original file names have been retained). Following that, we visualized the locations of the clusters in the principal components dimensions, to assist in their interpretation. Two such principal components plots are presented as Figures 3 and 4.

It is difficult to characterize the clusters in Table II in terms of genre or content. Any cluster may contain a number of different kinds of videos: music videos, trailers, amateur videos, etc., so it does not seem that genres are being separated out by the clusters. Although Cluster 2 contains a number of technically focused videos (First Subversion Setup, Folksonomy Cast, knowledge navigator, etc.), it also contains humor videos (japanese tradition sushi) trailers (U2 trailer), political videos (generally copied from TV sources, such as rush limbaugh vs pro choice) and animated shorts (gopher broke). Similarly, although identical videos may show up in the same cluster, meaning that they are tagged similarly or bookmarked by a similar set of users (e.g. Mighty McPilgrim brokemac [mountain] in Cluster 4; natalie portman rap in Cluster 4), sometimes different copies of the same video appear in different clusters (e.g. japanese tradition sushi in clusters 2 and 3). Hence, it appears that bookmarking and tagging don't guarantee that similar (even identical) videos will be treated the same by different users.

With this caveat in mind, we offer the following observations about the cluster contents. Cluster 2 is one of the more clear in that it appears to contain almost all of the technical videos (web programming, podcasting, etc.). Cluster 5 contains almost exclusively music videos from profes-

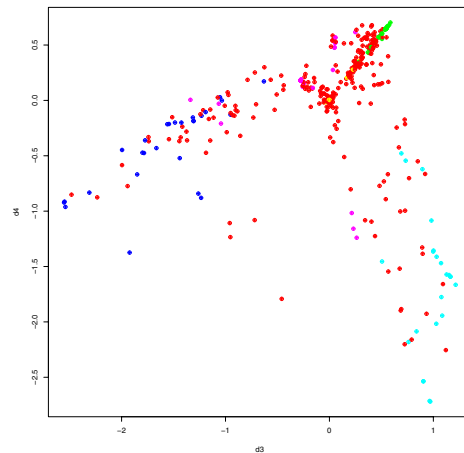


Fig. 4. Principal components plot of link clusters, dimensions 3 and 4.

sional musicians and video artists. Clusters 3, 4 and 6 are more heterogeneous, although both 3 and 4 have links with pornographic or sexual references (arnold [Schwarzenegger] in brazil and new pornos in Cluster 3; dildo song and farm sluts in Cluster 4), something apparently absent in Cluster 6. Clusters 3 and 6 also contain some political content that is apparently absent from Cluster 4 (e.g. 911 made in the USA with spare change in Cluster 3; Rumsfeld civil w[ar] and NSA Alberto in Cluster 6). Apart from Clusters 2 and 5, however, these characterizations are somewhat tenuous.

An alternative possibility is that members of a given cluster are linked on `del.icio.us` at around the same time. This would also explain the tendency toward chaining in the cluster analysis; links are gathered into a single large cluster by sweeping up smaller, temporally coherent clusters, one at a time. This suggestion requires further verification, however.

Plotting the six clusters in the first four principal components provides some indication of how the clusters are separated. In both figures, the clusters are colored as follows:

1, red; 2, yellow; 3 green; 4, cyan; 5 blue and 6 violet. Figure 3 shows the first two principal components, in which Clusters 1, 3, 4 and 5 fall along the x-axis, beginning from the origin and rising slightly toward the left. The bulk of their variation falls along dimension 1. Cluster 6 (violet) is somewhat looser than the other clusters, being distributed over much of the plot but concentrated below the x axis. Hence, its primary dimension of variation is dimension 2. Cluster 2 is concentrated at the origin in this plot.

In Figure 4, the points are distributed in a sharp right angle pattern, which suggests strong patterns of inter-correlation among the links. However, few of the links participate in the spokes of the distribution, where the correlations are strongest. This observation amplifies the point made regarding the cluster analysis, namely that there are a number of fairly tight clusters of small size, but few clusters are spherical, and few are both large and tight. The clusters characterizing the third and fourth dimensions are Clusters 5 (blue) and 4 (cyan) respectively. Thus, dimension 3 is possibly a “music video” dimension, whereas dimensions 1, 2 and 4 represent more heterogeneous content. Also on dimensions 3 and 4 but tending in the opposite (positive) direction, though not as strongly, is the fairly tight green cluster. The Yellow cluster, which was located exclusively at the origin in Figure 3, has also begun to separate in this figure.

Further principal components plots were made to explore additional structure among the links. All are similar to Figure 4 in that they generally have two or three spokes. These spokes are made up of different members of the red cluster (Cluster 1), which indicates that these additional dimensions serve to distinguish sub-clusters of this largest cluster from one another. These could be explored in much the same way as we have explored the other clusters so far, but to keep this part of the analysis from getting unwieldy, we do not pursue this further here.

The patterns we have observed show only weak tendencies toward clusters of links distinguished by semantics or genre. The music video cluster and its associated dimension are perhaps the clearest semantic/genre category that emerges from the data observed. Other potential categories (Clusters 1, 2, 3, 4 and 6) are weaker yet, with many non-canonical members among them, however they might be characterized. Moreover, there are even some copies of the same video that occur in different clusters, which suggests that some other principle must be at work in grouping the different video links. The suggestion that seems the most reasonable is that there is a temporal relation binding the links together.

B. Users and tags are clustered

We turn now to the relationships among users and tags, as they pertain to the linking of video on *del.icio.us*. For this analysis, we took the same matrix used in the previous analysis and normalized row-wise using z-scores. We computed the principal components of the resulting matrix, and clustered separately rows corresponding to users and rows corresponding to tags. These analyses were compared with

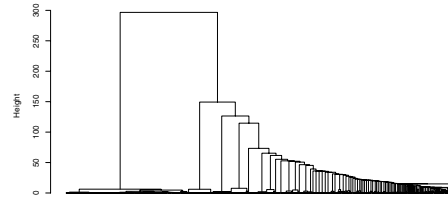


Fig. 5. Dendrogram of user cluster analysis.

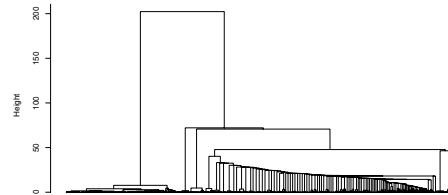


Fig. 6. Dendrogram of tag cluster analysis.

the previous analysis and with each other in order to identify likely clusters among users and tags. Figures 5 and 6 present the cluster analyses for users and tags, respectively.

Two things are striking in these two figures. First, both figures show strong evidence of chaining, as in the link dendrogram of Figure 2. Again this is unexpected given the nature of the algorithm used (Euclidean distance with Ward’s method), which favors similarly-sized spherical clusters linked together at a fairly loose level. Consequently we must consider the chaining to be a genuine property of the data, rather than an artifact of our analysis. Second, the shape of the user and tag dendrograms is almost identical: both have two large clusters, joined at the top level, with one of those being very tight. The size of the two top-level clusters in each dendrogram is also very similar, as is the chained structure in the second cluster. This suggests a close association between users and tags, which we need to test.

The chaining in the two dendrograms has the additional consequence that it makes it hard to visualize the relations among the clusters in principal components plots using color to distinguish clusters, since there is no satisfactory cut that defines a small set of comparably-sized clusters. In the plots below, we use color to distinguish users (blue) from tags (red), to illustrate the types of relationships that emerge in representative principal components plots. Figure 7 plots the first and second principal components, while Figure 8 plots the third and fourth.

These two figures show a striking contrast to the principal components of the links. Instead of being spread throughout the first two principal components, and showing sharp spokes on other dimensions, both users and tags are tightly clustered

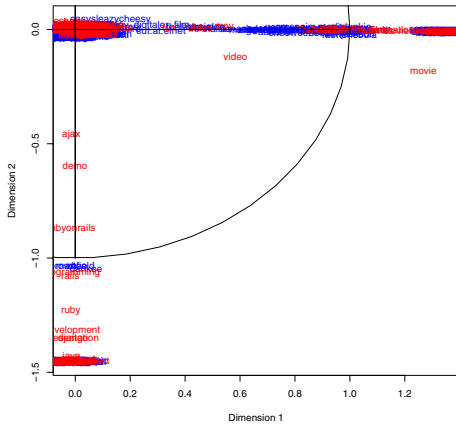


Fig. 7. Principal components 1 and 2 of users and tags.

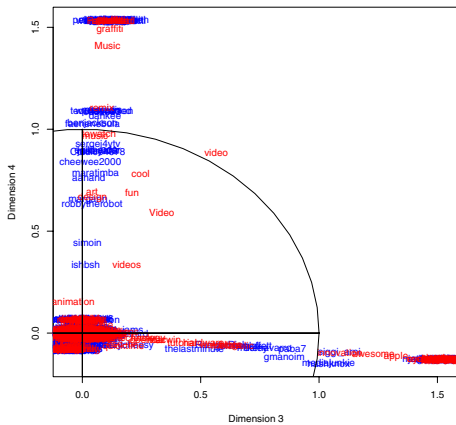


Fig. 8. Principal components 3 and 4 of users and tags.

around the coordinate axes on these and a large number of other principal components (greater than 20, when we explored them). Moreover each axis is associated with a distinct set of users and/or tags, that are both few in number and readily interpretable. For example, the first dimension in Figure 7 corresponds to the fairly general tags such as trailer, animation, video and movie, whereas the second dimension corresponds quite specifically to a set of screencasts for the web programming environment Ruby on Rails. Likewise in Figure 8, the third dimension corresponds to tags primarily used on apple product announcements and commercials (ipod, apple, awesome, etc.), while the fourth dimension corresponds to terms used for music videos (music, graffiti, remix, video).

A further observation that can be made in these plots is that there is often a very tight association between the locations of users in the space and that of some set of tags. The association is so close that they often land at exactly the same locations. This reinforces the impression lent by the dendrograms in Figures 5 and 6, that there is a strong association between

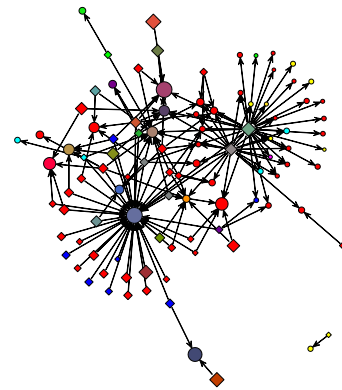


Fig. 9. User-tag association network.

users and tags.

C. The relationship between user clusters and tag clusters

To investigate the nature of the user-tag relationship, we chose to view user and tag clusters as a bipartite network. So that we would be able to capture the apparently quite specific nature of at least some user and tag clusters, we chose a cut at an arbitrary height for each cluster analysis that gave 50 clusters. We then re-tabulated the user-tag co-occurrences according to these clusters and visualized the resulting network as a two-mode network using the sna (social network analysis) package of R (www.r-project.org). Our expectation was that we would find a cloud of dyadic relations, where individual user clusters link to a specific tag cluster, but which are otherwise unconnected. The user and tag clusters so joined should be the same color, indicating that they are associated with the same type of content.

Since we wanted to be able to interpret the network in terms of our link clusters from the analysis above, we decided to color the nodes in the plot to reflect this association. However, since the association is typically not “pure”, we needed to mix RGB colors in appropriate proportions to reflect the relative contribution of each type of link to a user or tag cluster. This was done in two steps. First, using the color assignments for each of the six clusters above, we cross-tabulated user and link clusters, and tag and link clusters, in separate matrices. These were scaled as column-wise proportions, to take account of the larger size of certain link clusters (e.g. Cluster 1), and again as row-wise proportions, to adjust the gray value of the colors upward, toward white. We then computed the RGB values as weighted averages of the columns with an appropriate component, e.g. for red, we averaged the red, yellow and violet columns, etc. The resulting values were plotted as the node colors. For clarity, user clusters are indicated in the plot as squares and tags are indicated as circles. In addition, links are indicated as arrows from the user to the tag clusters. This network visualization is presented as Figure 9.

TABLE I
SET OF TAGS IN THE TAG CLUSTER IN FIGURE 9

.mov, 2006-03, ad, ads, advertisement, advertising, ai, ajax, amen, animation, art, art_technology, audio, baby, beat, benassi, benny, bizarre, blogs, blogthis, bluescreen, break, bush, carpaviar, cartoons, cg, channel4, checkout, cheney, clip, clothing, clowes, cm, comedy, comic, commercial, cool, copyright, cover, crazy, crossgenre, crowd, culture, cwreck, cyberstudies, dailyshow, dance, darwin, database, delta, demented, design, design:motion, dessin, development, digitalvision, disturbing, disturbing, dnb, documentary, dog, dogs, dolls, drums, drle, english, enurefeeds, esc-tv, event, evolution, executive, experiment, feedback, fire, flickr, fold, folding, fonts, fox, frohsinn, fun, fun, fundsachen, funny, geek, gondry, graffiti, graffiti, grenouille, guinness, guitar, gun, ha, hacks, heh, hilarious, hilarious, hiphop, history, house, howto, humor, humor, humour, idiot, ilustracin, in, incredible, innovation, inspiration, interface, interviews, it, japan, jazz, joke, jon_stewart, jonstewart, josefmullerbrockmann, joshspicks, jungle, just, katrina, keitai, kids, kindafunny, language, larry_king, laundry, law, lawnmowers, lego, letterpress, lifehacks, mac, make, making_of, makingof, man, media, metaholic, metal, microsoft, microsoft, migrations, mit, mixedculture, mixer, mmedia, montypython, motion, motiongraphics, motorhead, mov, ms, multi-touch, music, music, music_video, msica, national, noise, noitulove, parody, performance, pierce-bush, political, political.commentary, politics, prank, prez08, printing, programming, puppets, quicktime, radiohead, radiohead, rails, rails, rain, reference, reinhardt, remix, resources, rock, ruby, rubyonrails, satire, school, shirt, silly, simpsons, site, skit, sleaze, software, sound, stevejobs, street.fighter, streetart, streetfighter, sublime, subversion, system:imported, t-shirt, taste, tech, technology, ted, television, termite, the, thesimpsons, tips, toads, toblog, tosee, touch, towatch, turntables, tutorial, tutorials, tv, typography, urban, verdi-favorite, via:lessig.org/blog, video, video, video.games, videoselections, vidoe, vids, vido, vikings, violin, visualization, vrvwc, vw, watchthis, website, weird, werbung, windows, winnebago, winnebago_man, wittlin, wtf, xp, zg, zwigoff

Figure 9 shows a remarkable and unexpected characteristic, namely the network has two star patterns, one being a user cluster with links to many different tag clusters, the other being a tag cluster with links from many different user clusters. Both clusters, apart from having different central roles in the graph, are somewhat larger than the other clusters. Hence, the user cluster should be interpreted as an active group of users (they create more bookmarks to video links than other users), while the tag cluster represents a set of popular tags. While this interpretation appears to hold in the case of the tag cluster, (the full set of tags is shown in Table I) closer examination of the set of users reveals a disappointment. What is shared in common among the seven users is that all of them bookmarked a link (for some this is their only link in our data) for an amateur humorous video hosted at the MIT media lab depicting a “voice activated” blender called “blendie”. The tags they employ merely happen to fall into a range of semantic categories, presumably defined by others. Hence both stars observed here indicate a relatively weak semantic structure among the users, tags and links employed in the sample of video links we have observed.

V. DISCUSSION AND CONCLUSIONS

The investigation into tagging video on `del.icio.us` presented here reveals a number of unexpected findings. These findings suggest that, unless measures are taken to address them, folksonomic tagging systems are likely to encounter certain problems.

First, the associations among the links, users and tags are relatively weaker than we might have expected, especially given the findings of [13], [14]. In part this can be attributed to the well-known problems of uncontrolled vocabulary being applied by non-specialists. Many alternative versions of the same terms are seen in the tagging data (movie, Movie, video, vid, vids, etc.), and while such alternatives often correlate, this is not always the case. A similar problem is observed when different copies of the same video are linked by different users. At times they are found to be closely associated, but sometimes they are not, and there is no obvious way to guarantee that they will be. Given that mirroring is a common way of dealing with high demand for video, ideally the bookmarking system would provide a way to help identify videos that are the same but hosted at different locations. Without such support, a system like `del.icio.us` might be summarizing the network of video sources, rather than users’ semantic overlaps.

A characteristic common to all of the cluster analyses above is the chaining pattern in which smaller clusters are accreted one-by-one into a single large cluster. This pattern suggests that links, users and groups most naturally have only small clusters of related elements, at least within the time depth we have observed here, and that the semantics of video links is highly particularistic. Where users and links are concerned, this is perhaps not surprising. At the same time, it represents a challenge for the use of tags for collaborative access to information. Given the small number of tags actually applied to most links, it would be difficult to guarantee that users would be able to identify tags that lead to useful content, or to navigate among sets of links using tags to discover useful information. Thesauri, tag suggestion utilities and other similar means may help, but it is likely that they will need to be built using more information than two months worth of tagging to be optimally useful.

Although the tagging of videos appears to be highly particularized, this does not mean that the process of tagging lacks a social structure. Rather, it appears that such social structure as exists (e.g. the two stars in our network analysis) is not particularly helpful to the purpose of organizing or accessing information. Moreover, we find little evidence to suggest that specific users are minding specific content. Instead, we find instead that even a small group users with unusual activity (linking a common video in different and unexpected ways) can have a disproportionate effect on the structure observed. This is a situation with some important consequences.

First, a large share of the semantic range available among the tagged videos on `del.icio.us` is influenced by a single group of users. It is not clear how people would come into contact with this group of users, or how one might come to understand or share their tagging conventions, but such influence could either be a resource for information organization through tagging or an impediment to it, depending on whether those tagging relations are found to be useful.

Second, a different situation is posed by the set of popular tags. Large number of otherwise distinct groups of users all employ this same set of tags, hence it is natural to ask what

this set of tags means. Is it a useful set? Or are they tags whose meaning is relatively limited in use, perhaps too general or too specific? It is important to answer this question, since some of these tags are likely to appear in the “tag cloud ” that comprises the most commonly implemented tag-based navigation affordance among folksonomy applications.

System designers who wish to implement social metadata systems would do well to consider these two aspects of collaborative tagging. On the one hand, there may well be an active group of users whose tagging behavior would lead to useful semantic differentiation. Systems should ideally provide ways in which this kind of information could be exposed. On the other hand there are the popular tags which do little to contribute to semantically differentiating content, but which many users are likely to come up with. Unfortunately, such tags are likely to be exposed in tag clouds. What is needed is a way to measure the semantic work performed by a tag, perhaps like the use of inverse document frequency to weight terms in Information Retrieval. If high-value tags are exposed, rather than high-use ones, then the goal of semantics emerging through collaborative process is more likely to be realized.

ACKNOWLEDGMENT

The authors would like to thank Susan Herring, Elijah Wright, other members of BROG and three anonymous reviewers for their comments on this research.

REFERENCES

[1] T. Hammond, T. Hannay, B. Lund, and J. Scott, “Social bookmarking tools (i): A general review,” *D-Lib Magazine*, vol. 11, no. 4, April 2005. [Online]. Available: <http://www.dlib.org/dlib/april05/hammond/04hammond.html>

[2] J. Elin K, “Library trends: Classification and categorization: a difference that makes a difference,” *Library Trends*, 2004. [Online]. Available: http://www.findarticles.com/p/articles/mi_m1387/is_v3_n52/ai_n6080402

[3] S. A. Golder and B. A. Huberman, “Usage patterns of collaborative tagging systems,” *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006. [Online]. Available: <http://jis.sagepub.com.offcampus.lib.washington.edu/cgi/content/abstract/32/2/198>

[4] J. W. Tanaka and M. Taylor, “Object categories and expertise: Is the basic level in the eye of the beholder?” *Cognitive Psychology*, vol. 23, pp. 457–482, 1991. [Online]. Available: <http://citeseer.ist.psu.edu/context/559418/0>

[5] T. Vander Wal, “Folksonomy definition and wikipedia,” <http://www.vanderwal.net/random/entrysel.php?blog=1750>, November 2005. [Online]. Available: <http://www.vanderwal.net/random/entrysel.php?blog=1750>

[6] B. Berlin, D. E. Breedlove, and P. H. Raven, “Folk taxonomies and biological classification,” *Science*, 1966. [Online]. Available: <http://links.jstor.org/sici?sici=0036-8075%2819661014%293%3A154%3A3746%3C273%3AFTABC%3E2.0.CO%3B2-U>

[7] C. Marlow, M. Naaman, D. Boyd, and M. Davis, “Position paper, tagging, taxonomy, flickr, article, toread.” in *Collaborative Web Tagging Workshop, 15th International World Wide Web Conference*, May 2006.

[8] C. Cattuto, V. Loreto, and L. Pietronero, “Collaborative tagging and semiotic dynamics,” May 2006. [Online]. Available: <http://arxiv.org/abs/cs.CY/0605015>

[9] H. A. Simon, “On a class of skew distribution functions,” *Biometrika*, vol. 42, pp. 425–440, 1955. [Online]. Available: <http://dx.doi.org/doi:10.1093/biomet/42.3-4.425>

[10] G. Udney Yule, “A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s.” *Royal Society of London Philosophical Transactions Series B*, vol. 213, pp. 21–87, 1925.

[11] K. Shen and L. Wu, “Folksonomy as a complex network,” Sep 2005. [Online]. Available: <http://arxiv.org/abs/cs.IR/0509072>

[12] R. Lambiotte and M. Ausloos, “Collaborative tagging as a tripartite network,” Dec 2005. [Online]. Available: <http://arxiv.org/abs/cs.DS/0512090>

[13] J. Paolillo and E. Wright, *Visualizing the Semantic Web. XML-Based Internet and Information Visualization*, 2nd ed. Springer-Verlag London Ltd, 2005, ch. Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF.

[14] J. Paolillo, S. Mercure, and E. Wright, “The social semantics of livejournal foaf: Structure and change from 2004 to 2005,” in *ISWC 2005 Workshop on Semantic Network Analysis*, ser. CEUR Workshop Proceedings, vol. 171, 2005.

[15] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY, USA: Cambridge University Press, 1994.

[16] A. Degenne and M. Forse, *Introducing Social Networks*. Thousand Oaks, CA, USA: Sage Publications, 1999.

[17] P. Doreian, V. Batagelj, and A. Ferligoj, *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge University Press, 2004.

TABLE II

CLUSTERS OF VIDEO LINKS ON DEL.ICIO.US

| Cluster | Links |
|---------|--|
| 2 | 011606 ovechkin p3g6, ALA Midwinter-2006Jan20, Chopper Pirates, First Subversion Setup, Folksonomy Cast, Pat Duffy, Phuwadon 070306 H, SNME 18.03.066, T Override Help 1, U2 trailer, bee.39.espn, dance challenge, facet 24, gopher broke 480, how close, japanese tradition sushi, joi predavanje, knowledge navigator, led touch, levitation-trick-revealed, lisp van, bigpezones, rendu quicktime leger, rush limbaugh vs prochoice 060302a, systm 0004 podcasting large h264, tamori, ushpezin movie, video conference nintendo ds1 |
| 3 | 031106, 2658805 200, 911 Made In The USA with Loose Change 1of4, 911 Made In The USA with Loose Change 2of4, ALWAYS READY, Ekai SXSW 2006 Now That Its 1997 698, Envelopes Free Jazz High, FINAL OMALOUIS MED 060209, IE7 big, Orson, Say, arnold in brazil, bathroom, crazy dance, cross fader, japanese tradition sushi, life day, new pornos spantech, rd, red-web, rendezvous20 04, sos recap, stop p showreel xlarge, toystory 2 requiem, toystory 2 requiem, zelda ds large |
| 4 | 03 el choque, Flickr, Frank Caliend on, Jelly D, Mighty McPilgrim brokemac large, Mighty McPilgrim broke mac large, Pi, TDS FOX Serial Ki, TH reel truth 3a, air machine, berlitz junior 40 mpeg 2, berlitz tv commercial 2006, dildo song large, farm sluts ref, fuck the shit, german coast guard, hadoken, heston of the apes 320x240, iBrator, jardin, karate chimp, portman rap, snl natalie portman rap md |
| 5 | Bjork Hi, Boney M Rasputin, Pleix vitalic birds L, Radiohead just, Video Export Lo, ace of spades 15 mb, amen web, cracker redux small, criminal 02, criminal 02, django, django, dosh, hot club, last flight sonic, lets pretend we dont exist, licata, marimba ponies, modern world, party dream, pink stupid girls web rip, slow dogs, system D128 muppets, the doves, the sun romantic death ref, touch the sky, urban medium preview |
| 6 | 2005-12-5 GMU, Block Jam 0, Ginger, ITTD trailer, Multi Touch, Rumsfeld civil w, TCR Long War 3-9, TDS Nsa Alberto, The Tornados Robot, UFO, We Will Become Sihouettes (Hi Res Quicktime) 289, ariel pink video, cas devo, cow bell, dc great wall med, dm bob country jem lou, doggy poo features, house of cosbys 1 high, jimmy kimmel, leaf house70, little birds, living proof 300, parfum teaser, rb 06 mar 14, rondo, shower, tea, victory auto |