"Supertagger" Behavior in Building Folksonomies

Jared Lorince[*†]
jlorince@indiana.edu

Sam Zorowitz[§]
szorowi1@jhu.edu

Jaimie Murdock[*‡]
jammurdo@indiana.edu

Peter M. Todd[*†‡]
pmtodd@indiana.edu

Indiana University:
[*]Cognitive Science,[†]Psychological & Brain Sciences,[‡]Informatics
John Hopkins University:
[§]Psychological & Brain Sciences

ABSTRACT

A folksonomy is ostensibly an information structure built up by the "wisdom of the crowds", but is the "crowd" really doing the work? Tagging is in fact a sharply skewed process in which a small minority of users generate an overwhelming majority of the annotations. Using data from the social music site Last.fm as a case study, this paper explores the implications of this tagging imbalance. Partitioning the folksonomy into two halves — one created by the prolific minority and the other by the non-prolific majority of taggers — we examine the large-scale differences in these two sub-folksonomies and the users generating them, and then explore several possible accounts of what might be driving these differences. We find that prolific taggers preferentially annotate content in the long-tail of less popular items, use tags with higher information content, and show greater tagging expertise. These results indicate that "supertaggers" not only tag *more* than their counterparts, but in quantifiably *different* ways.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—Collaborative computing, Web-based interaction

Keywords

Collaborative tagging, Folksonomy, Supertaggers

1. INTRODUCTION

Participation rates in a social tagging system vary widely. The semantic structure of a folksonomy — the collaboratively-generated classification scheme that emerges from many individual, assignments of free-form textual labels to content — is available and potentially useful to all users of a system with tagging features. But most users are precisely that: users. They may use tags to search for or gain information about resources, but only a minority of users actively contribute to the knowledge-generation process by assigning metadata to content. Even among those who do tag, only a small percentage do most of the work, with a small number of taggers contributing most of the annotations (i.e. tag assignments), and a comparatively large number only tagging a few times. The implications of these participation rates have deep consequences for the information architect wishing to implement a tagging system. Does the folksonomy represent the aggregated knowledge of its users, or only the few "supertaggers" among them? Would the behavior of prolific and non-prolific taggers actually create two distinct folksonomies?

We can partly attribute this lack of participation to the fact that tagging is most often a secondary feature of a given system. To tag is to make a deliberate choice with costs of time and effort outside the primary use of a service. Users may, for instance, use Flickr to find and share photos or Last.fm to listen to and learn about music, without making any substantive contribution to the folksonomies embedded in these systems. This fact is more pronounced in the latter case, where the principal activity on the site — listening to music — is a passive activity, while tagging requires active effort.

Underlying questions about folksonomy creation is the fundamental issue of motivation — why do users contribute to social tagging systems? A substantial literature has explored this topic in terms of why users tag in one manner rather than another [16, 1, 19], but there is little work addressing the question of why users choose to participate in the tagging process to begin with. By comparing the tagging patterns of the minority of prolific taggers to the majority of non-prolific taggers, here we contribute to an understanding of what differentiates the heavy contributors from their low-tagging counterparts in social tagging, what motivational factors distinguish these two groups, and whether their tags reflect different underlying folksonomies.

In summary, there are two high-level questions that interest us: First, how do the tagging patterns of the minority of prolific taggers differ from the majority of non-prolific taggers, and what does this suggest about motivations for tagging? Second, does the disproportionate contribution to the folksonomy by a small number of users compromise the

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

http://dx.doi.org/10.1145/2615569.2615686.

presumed crowdsourced nature of tagging? In other words, does the folksonomy truly represent the collective knowledge of its users, or just a subset who may or may not be representative of the general user base? Though we cannot offer complete answers to these questions, we present methods and results that help shed light on these relatively unexplored issues.

In this paper, we address these questions using a dataset of approximately 1.9 million users, with over 50 million annotations across 4.5 million items[1] crawled from the social music site Last.fm (Section 3.1). After presenting related work (Section 2) and an overview of the dataset (Section 3), we illustrate and formalize the the prolific- vs. non-prolific tagger dichotomy in Section 4. In Section 5 we present our main descriptive analyses showing differences in the tagging patterns and attributes of users in each of the two groups. Next, in Section 6, we explore possible causal accounts for the observed differences, focusing on expertise effects and differences in motivation. We conclude in Section 7 by synthesizing our results and considering their implications.

2. RELATED WORK

2.1 Folksonomies

A *folksonomy* is a collaborative organization scheme which uses tags (words or short phrases) to annotate objects for later retrieval. Thomas Vander Wal coined the word "folksonomy" in a 2004 listserv posting[20]. Folksonomies are most often social endeavors, with multiple users annotating the same objects with user-generated vocabulary.

Whereas many classification schemes are "top-down" hierarchies, a folksonomy is "bottom-up". In a taxonomy, a discrete set of pre-existing, often expert-generated, categories are assigned to resources. In a folksonomy, the vocabulary is unconstrained and comes from the users themselves, who may or may not be domain experts, bringing "power to the people" [17]. Many efforts have been made to infer taxonomies from folksonomies, synthesizing the advantages of controlled vocabulary and crowdsourced curation [10, 14].

The information retrieval advantages of folksonomies, combined with low economic cost of implementation and essentially free creation, provide a strong motivation for their use. Many folksonomies have been studied in diverse domains, including Flickr (photos, [15]), Delicious (web bookmarks, [3]), Last.fm (music, [11]), and BibSonomy (academic papers, [5]). A review of many early social tagging systems can be found in [12].

2.2 Tagging Motivation

One possible distinction between prolific and non-prolific taggers is tagging motivations. Though motivation in tagging behaviors has been operationalized in numerous ways, one prominent approach [9] characterizes users as either categorizers or describers. When tagging, categorizers use a limited vocabulary to construct a personal taxonomy conducive to later personal search. In contrast, describers do not constrain their vocabulary; instead they freely choose a variety of informative keywords to describe items. Strohmaier et al. [19] and Körner et al. [9] present several metrics with which to categorize users according to this dichotomy, discussed in Section 6.2.

Content produced by describers and categorizers has been shown to be useful for disparate tasks. Tags produced by describers, for example, are more useful in information retrieval [4] and knowledge acquisition [8]. Conversely, tags produced by categorizers are more useful for social classification tasks [25]. As such, it is important to determine whether prolific and non-prolific taggers differ in their tagging motivations along the lines of describers versus categorizers, to help understand how the folksonomy created by the top taggers may differ from that created by the rest.

2.3 Expertise in tagging

Another possible distinction between tagger types is level of expertise. In other words, do prolific taggers demonstrate greater or lesser expertise than non-prolific taggers when annotating items? Detecting expert users in a folksonomy is motivated by an increasing need to distinguish users providing informative contributions from those producing unhelpful contributions (especially spammers) in large folksonomies [6, 22].

One noteworthy approach to expert detection is Spamming-Resistant Expertise Analysis and Ranking (SPEAR) [22, 23], a variant of the HITS Web page ranking algorithm [7], that identifies experts according to two principles: First, there should be mutual reinforcement between user expertise and the quality of the annotated items. In other words, an expert user is not only more adept at identifying high quality items, but is also defined by the quality of the items annotated. Second, expert users are more likely to "discover" quality items than less expert users.

Here, we utilize the SPEAR algorithm to quantify expertise among prolific and non-prolific taggers. The use of a spam-robust expertise measure is important, as Wetzker et al. [21] found an overwhelming majority of the most prolific taggers in a large taxonomy were spammers. SPEAR is particularly appropriate for detecting expertise in our dataset as users on Last.fm are provided tag recommendations when annotating items, and SPEAR reasonably assigns greater expertise to users who first annotate an item with a given tag than to users who tag later.

3. DATASET

To address our questions, we utilize a dataset crawled from the social music site Last.fm with data spanning July 2005 through December 2012. The data was first presented in [11], but has since been expanded to not only include tagging data, but friends, group memberships, items listened to, and loved/banned tracks[2] for an increased number of users.

3.1 Crawling Methodology

We crawled data with a combination of API queries and HTML scraping of users' publicly available profile pages. We did so on a user-by-user basis, such that we have the

[1] An "item" is a generic term referring to an atomic target of tagging activity on Last.fm, and can be an artist, album, or song. Although there is a hierarchical structure inherent to these item types (an artist has a set of albums, each made up of a set of songs), tag distributions exist on Last.fm at the item level, and we therefore perform our analyses at that level, as well.

[2] "Loving" a track is roughly equivalent to favoriting a tweet, or other similarly-defined activities, while "banning" allows a user to indicate disliked items and exclude them from any recommendations by Last.fm.

complete tagging history for every user in our data, but not necessarily the complete tagging history for any particular item. All temporal annotation data is at a monthly granularity, as users' profiles only list the month and year in which an item was tagged (no such data is available from the API).

Because users were crawled by traversing the site's social network, we necessarily only include those users with at least one friendship on the site, but we do not believe this is problematic for our analyses. See [11] for further discussion of our crawling methods and its limitations.

3.2 Data Summary

We crawled a total of nearly 1.9 million users, extracting the behavioral measures mentioned above, as well as self-reported demographic data. An "annotation" refers to a given instance of a user assigning a particular tag to a particular item at a particular time. It is best thought of as a four-element tuple in the form user-item-tag-time. For a subset of our users, we also have collected full scrobble histories[3]. Table 1 summarizes the data collected. All tagging analyses presented here reflect only those users with ≥ 1 annotation.

Total users	1,884,597
Friendship relations	24,320,919
Total annotations	50,372,895
Users with $\geq$1 annotation	521,780
Total unique tags	1,029,091
Unique items tagged	4,477,593
Total Scrobbles	1,181,674,857
Users with scrobbles recorded	73,251
Unique items scrobbled	32,864,795
Total loved tracks	162,788,213
Users with $\geq$1 loved track	1,355,859
Total banned tracks	23,321,347
Users with $\geq$1 banned track	502,758
Unique Groups	117,663
Users with $\geq$1 group membership	827,232

Table 1: Dataset summary.

The data show a long-tailed distribution for per-use annotation counts, with similar distributions[4] for other tagging (total uses of each tag, total annotations per item) and behavioral (number of groups, loved tracks, and banned tracks per user) measures of interest, as well as the total number of scrobbles per track. Figure 1 summarizes this data. The distribution of scrobbles differs from the others in lacking a long tail, showing that most users listen to a large number of items. While these scrobble counts come from a relatively small subsample of our users, the pattern is consistent with the distinction between passive listening and active tagging mentioned earlier.

[3]A "scrobble" is Last.fm's term for an instance of a user listening to a particular song at a particular time. The service tracks users' listening habits (either through the site directly, or via a plugin installed in a media player) providing recommendations and aggregated listening statistics, and each listen logged is a "scrobble".

[4]Though clearly long-tailed, we remain agnostic as to the precise mathematical form (e.g. power-law, lognormal) of these distributions, as it does not meaningfully affect our analyses.

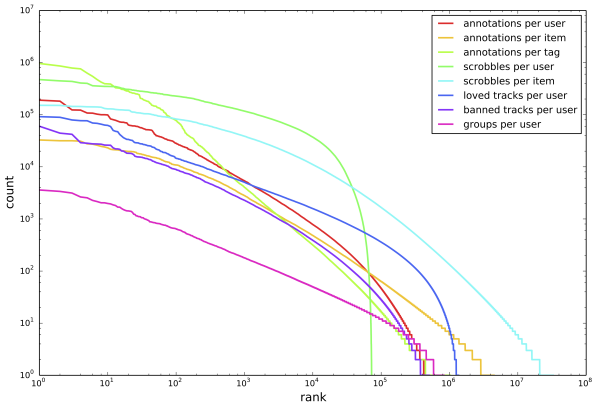

Figure 1: Rank-frequency plots for main measures from the dataset, on a log-log scale.

4. PROBLEM FORMALIZATION

The long-tail distribution of annotation counts in our data suggest the existence of two populations: a prolific-tagging minority and a non-prolific-tagging majority. To attempt to distinguish these two populations, we calculated the relative contributions of annotations across divisions between prolific and non-prolific taggers. We compared the proportion of taggers included in the prolific-tagger group to the proportion of annotations generated by that group (Figure 2). The top 20% of users generate over 90% of all tagging activity in our data, more skewed than the 20%/80% pattern commonly described by the Pareto Principle [13].

With this distribution in mind, we explored a variety of methods to seek a "natural" split between the prolific and non-prolific tagger populations, settling upon a 50-50 split in the number of annotations. This split at a threshold of 1,457 total annotations per user placed 5,086 users (0.97%) in the prolific-tagger group, and the remaining 516,694 users (99.03%) in the non-prolific group.

While this partitioning is arbitrary, it yields two large folksonomic structures of equal size (in terms of total annotations) amenable to analysis, and also highlights the extreme

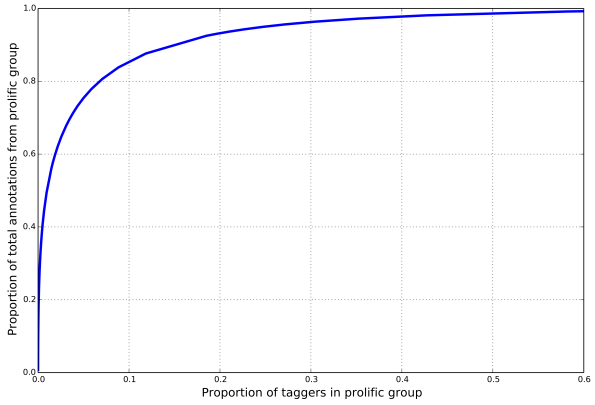

Figure 2: Proportion of total annotations created by the prolific taggers as a function of the proportion of top users included the prolific-tagging group.

skew in the behaviors of users on the site. Although other measures, such as the number of actual unique tags, users, and items vary between the two folkosnomies, this partitioning ensures that the total amount of tagging performed is equal in both.

5. DESCRIPTIVE ANALYSES & RESULTS

In this section we examine, at a descriptive level, how the *users* in each group defined by our partition differ, and how the two *folksonomies* generated by those groups differ.

5.1 User Attributes

Are the groups similar in terms of demographics and other attributes? There were few interesting demographic differences of note, but three points do warrant mention. First, prolific taggers are older on average than non-prolific taggers ($m = 31.1$ vs. $m = 26.4$). Second, they are more likely to be subscribers (users who pay a monthly fee for premium features): 7.3% of prolific taggers versus 1.2% of non-prolific taggers are subscribers. Finally, they are slightly more likely to report optional demographic data such as age (73.9 % versus 71.7%) and country (90.7% versus 84.5%)

Are the groups similar with respect to other behavioral measures? The behavioral measures we collected tend to show weak, but positive cross-correlations (with some exceptions, see Table 2), but our main interest is in how these measures covary with annotation volume. Following analyses in [18], we plot these measures for all users as a function of annotation count, binned logarithmically, in Figure 3. Users in the non-prolific tagging group appear on the left of the dashed line, and prolific taggers on the right.

Though the data is much noisier for the prolific taggers, the general trend is that of prolific taggers being more active than non-prolific taggers across all behavioral measures. This suggests that being a prolific tagger may, in part, be an artifact of being a heavy user of the site more generally (though not for all users; there are clear outliers in Figure 3).

	N_f	N_a	N_l	N_s	N_b	N_g
N_f		0.075	0.155	0.146	0.015	0.225
N_a	0.075		0.209	0.204	0.062	0.139
N_l	0.155	0.209		0.226	0.113	0.191
N_s	0.146	0.204	0.226		0.056	0.211
N_b	0.015	0.062	0.113	0.056		0.012
N_g	0.225	0.139	0.191	0.211	0.012	

Table 2: Cross-correlations (Pearson's r) between per-user counts of friends (N_f), annotations (N_a), loved tracks (N_l), scrobbles (N_s), banned tracks (N_b), and groups (N_g). In all cases $P \ll 0.0001$

5.2 Folksonomy Attributes

Table 3 presents several high-level measures of the two folksonomies. P denotes the prolific-tagger folksonomy, and NP denotes the non-prolific tagger folksonomy. With these global measures as our starting point, we can ask several concrete questions about the attributes of P and NP.

Do both groups use a similar global vocabulary? The non-prolific taggers clearly have a larger vocabulary overall, but note that both groups' vocabularies are largely shared:

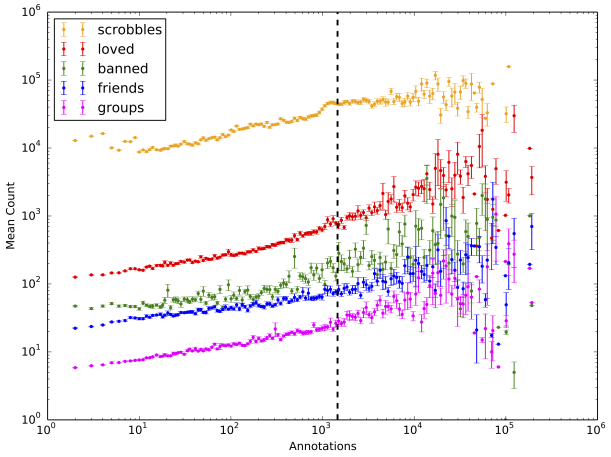

Figure 3: Users' mean number of scrobbles, loved tracks, banned tracks, friends and groups as a function of logarithmically binned annotation count. Error bars show +/- 1 SE, and the vertical line shows the prolific/non-prolific tagger threshold.

Though NP contains almost two times the tags of P, more than 90% of all annotations by both groups use one of the 168,245 tags the groups share (i.e. tags that occurred at least once in both folksonomies). This suggests the existence of many "singletons" — tags used only once, or a small number of times. This is verified in Figure 4A, which shows the distribution of annotation counts by tag for both groups.

The solid points in Figure 4A show, for a given number of annotations N, the proportion of unique tags in each folksonomy that are used N total times (i.e. having that many annotations). Clearly, more tags are used once overall than any other frequency for both P and NP. P does show, however, proportionally more tags with larger annotation counts (it follows that NP, which contains more unique tags than P, has a greater raw number of true singletons and other tags used a small number of times). This is an unsurprising result, given the very different number of users in each group. The crosses on the plot show the proportion of total annotations corresponding to a given N; that is, for a given N, the corresponding dot shows what proportion of tags were used N times, while the cross shows the combined propotion of annotations from all tags with N annotations. The most popular tags (far right of plot) represent the greatest overall contribution to the folksonomies, while the combined annotations of the many rarely used tags outweigh the contribution of the tags in between, creating a U-shaped relationship. P does have, however, more tags in this middle range (i.e. tags used $100 - 10000$ times).

In Figure 4B we show how this same data is distributed over users: For a given N, what proportion of tags (within each folksonomy) are used by N users? Consistent with the first plot, more tags are used by a single user than by any other number of users for both folksonomies. We again plot the corresponding annotation proptions, which show a similar U-shaped pattern. This indicates that annotations are concentrated among the few most popular tags (in this case defined in terms of number of users instead of total annotations) and the many tags used by the fewest users.

	P	NP
Total Users	5,086	516,694
Total Tags	399,552	797,784
Unique Tags	231,307	629,539
Shared Tags	168,245	
Total Items	2,992,046	2,515,070
Unique Items	1,962,523	1,485,547
Shared Items	1,029,523	

Table 3: Summary measures of prolific- and non-prolific tagger folksonomies.

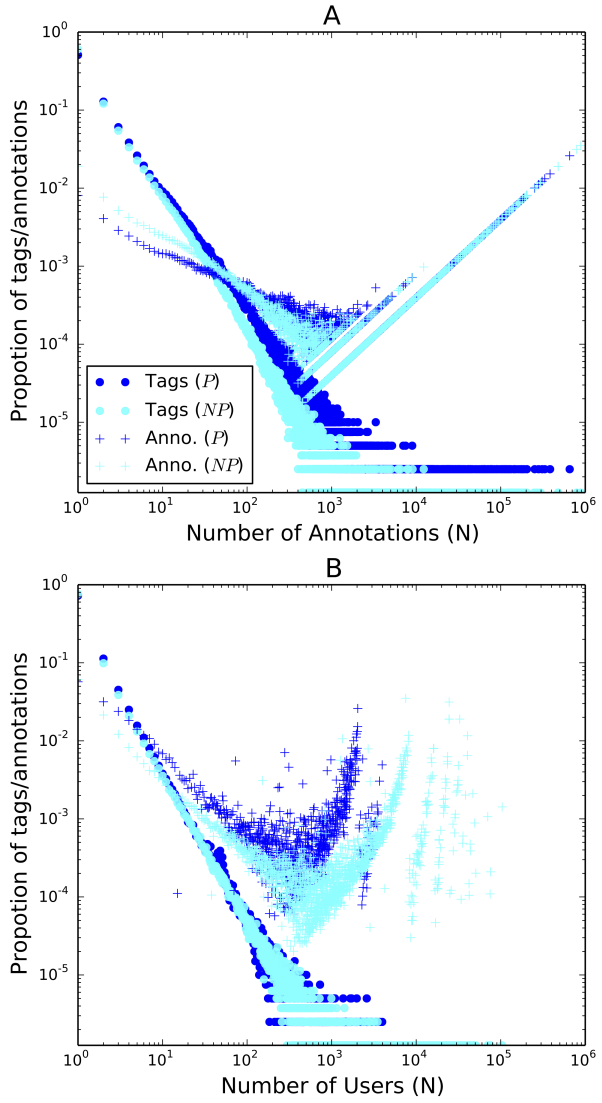

Figure 4: Log-log probability distributions of number of unique *tags* with N total annotations (A) and N total users (B), marked with dots. Crosses indicate the corresponding proportions of total *annotations* from tags with the corresponding annotation/user count.

Two simple summary measures of the similarity between P and NP are the rank correlation, Spearman's ρ, of tags for each folksonomy (i.e. is the rank order of overall tag popularity the same in both distributions) and the cosine similarity between the two global tag vocabularies (i.e. calculated across vectors of the frequency of each tag in each of the two folksonomies). Considering all tags, we find a rank correlation $\rho = -0.219$ and a cosine similarity of 0.8719 between P and NP. These give rather opposing impressions of the distribution similarities, so it is informative to consider these measures for smaller subsets of the data. We calculated both measures for the top N tags in both folksonomies, and in Figure 5 plot the results as a function of increasing N[5]. We find that the rank correlation coefficient is maximized by only considering the top 225 tags from each folksonomy, yielding $R = 0.836$. Considering more tags leads to monotonic decreases in ρ. The cosine similarity does increase as we consider more tags, but only marginally (for the top 225 tags the cosine similarity is .8713). These results indicate that there are substantial differences in the use of the many, rare tags in the tail of the distribution (hence the decreasing ρ past the top 225 tags), but that these do little to affect the overall similarity of the two vocabularies. The lower ρ and decreasing cosine similarity when considering fewer than the top 225 tags shows, however, that there are non-negligible differences in the most popular tags used by the two groups.

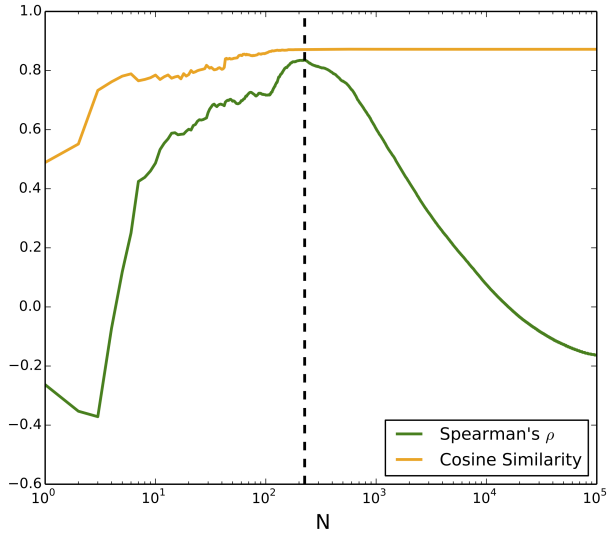

Figure 5: Spearman's ρ and cosine similarity between P and NP as a function of N, considering only the top N most popular tags overall from each folksonomy. The dashed line shows $N = 225$.

Do both groups tag the same content? P clearly covers a larger number of items than does NP, but the overlap is substantial, with 72.6 and 83.7 percent, respectively, of the annotations in P and NP allocated to items tagged by both groups. The higher percentage for NP suggests that they are concentrating their tagging on popular items more so than P. To better understand these patterns, we repli-

[5]As an example for clarification, if $N = 100$, we consider the top 100 most frequent tags in each folksonomy. Tags that appear in P but not NP (and vice versa) are assumed to have rank $N + 1$ for the purposes of calculating the rank correlation. This was repeated for N from 1 to 100000.

cate the analyses shown in in Figures 4 and 5 for items as opposed to tags.

In Figure 6A and B we see that, just as many *tags* are only used once, many *items* are tagged only once. Unlike the tags, however, we do not find highly-annotated items with combined totals of annotations rivaling those of the singleton items. In other words, for neither P nor NP is the item distribution as skewed as the overall tag distribution.

The notable differences between the folksonomies is that annotations in P are skewed towards items with proportionally fewer total items and taggers, suggesting that prolific taggers not only tag more items, but preferentially tag less popular, more obscure music. We confirm this in Figure 6C, plotting the mean number of annotations for items with a given *global* number of scrobbles (i.e. across all users). Though there is a general trend of items with more scrobbles attracting more annotations, there is clear pattern of users in P allocating more annotations to items with low scrobble counts.

We also repeat the cosine similarity and rank correlation measurements at the item level. That is, we calculate the rank correlation and cosine similarity over the distributions of items tagged (as opposed to tag vocabularies) for both folksonomies. Calculated over the entireties of P and NP, we find a rank correlation of $R = 0.216$, and a cosine similarity of 0.768, but considering smaller subsets of the data is again informative. Figure 7 shows cosine similarities and rank correlations between the folksonomies when only considering the top N most tagged items overall. There is clearly greater disagreement between P and NP when it comes to which items are tagged than which tags are used. The rank correlation peaks at 0.540 when considering the top 944 items, while the corresponding cosine similarity is 0.760. These results indicate that the overall differences between P and NP with respect to items tagged are more extreme than differences with respect to the tags used. Furthermore, there is relatively greater deviance in the top tagged items within each group as compared to the top used tags. In other words, prolific and non-prolific taggers agree more regarding what the most popular tags are than regarding what the most popular (or, at least, tag-deserving) items are.

5.3 Information Theoretic Measures

In addition to traditional statistical methods, we examined several information theoretic differences between P and NP. These metrics enable us to express differences not just in *what* is being tagged, but in *how* they are being tagged. Combined with a time-dependent analysis of cumulative tagging behaviors, we can also see if the prolific and non-prolific populations diverge as the folksonomies grow.

Are tags generated by prolific taggers more informative than those generated by non-prolific taggers? To answer this question, we calculated naive Shannon entropy for items and tags, as defined below:

$$H(T) = -\sum_i p(i) \log p(i)$$

where $p(i)$ is the ratio of appearances of that entity to the total number of annotations. The results are shown in Table 4. We see that the tags provided by the prolific and non-prolific taggers have roughly equivalent uncertainty with slightly higher uncertainty for the prolific taggers. Given that there are 399,552 tags used by the prolific taggers and 797,784 tags used by the non-prolific taggers (see Table 3), the roughly equivalent entropy shows that each tag contains roughly equivalent information, no matter how often it is used. Similarly, the item entropy is higher for the prolific taggers (with 2,992,046 items tagged by the prolific group and 2,515,070 items tagged by the non-prolific group), consistent with the greater diversity of items observed in Section 5.2.

	P	NP
tag	11.7548	11.2922
item	19.2823	17.8425

Table 4: Entropy of each annotation component for the prolific and non-prolific folksonomies.

Does the behavior of each population change over time? To answer this question, we calculated the monthly Kullback-Leibler (KL) divergence for the cumulative folksonomies for both the items tagged and the tags used. KL divergence is also known as the "relative entropy" and can be interpreted as the amount of information gained by using the distribution A instead of B. It is formally defined as:

$$D_{KL}(A||B) = \sum_i \ln\left(\frac{A(i)}{B(i)}\right) A(i)$$

KL divergence is asymmetric, which allows us to tell if one distribution is mimicking the other. If a distribution has a low divergence relative to another, it requires little information to transcode into the other distribution. A higher divergence indicated that more bits are required to store the same amount of information in the second distribution. If these divergence scores differ widely between distributions, the direction with a lower divergence indicates that the other set has a better fit to the underlying information. KL divergence is often used in a modeling context, in which the second distribution is a model, and the first distribution is the observed data. As opposed to a correlative measure, it is able to show changes in *how* items are tagged as opposed to *what* items are tagged.

Thus, we calculated KL in both directions ($P \rightarrow NP$ and $NP \rightarrow P$) and only over elements (items or tags) contained in both populations at that point in time. For tagging divergence, we calculated the cumulative divergence at each time step by creating a new folksonomy P_m and NP_m which consisted of all annotations containing tags present in both folksonomies up to that point of time. Thus, annotations which were previously excluded may be included once a given time step is reached. Figure 8 shows the results. Similarly, monthly folksonomies were generated based on the intersection of item annotations. Item folksonomies may have non-intersecting tags, and tag folksonomies may have non-intersecting items.

We found that for tags (solid lines) the KL divergence grew over time, indicating that the ability of each population to fully capture the other's annotations decreased. Furthermore, as the $P \rightarrow NP$ non-prolific divergence was larger than the $NP \rightarrow P$ divergence, highlighting that the prolific taggers were generating a schema that more closely matched the communal tag usage than the non-prolific taggers. This emphasizes some of the expertise effects noted

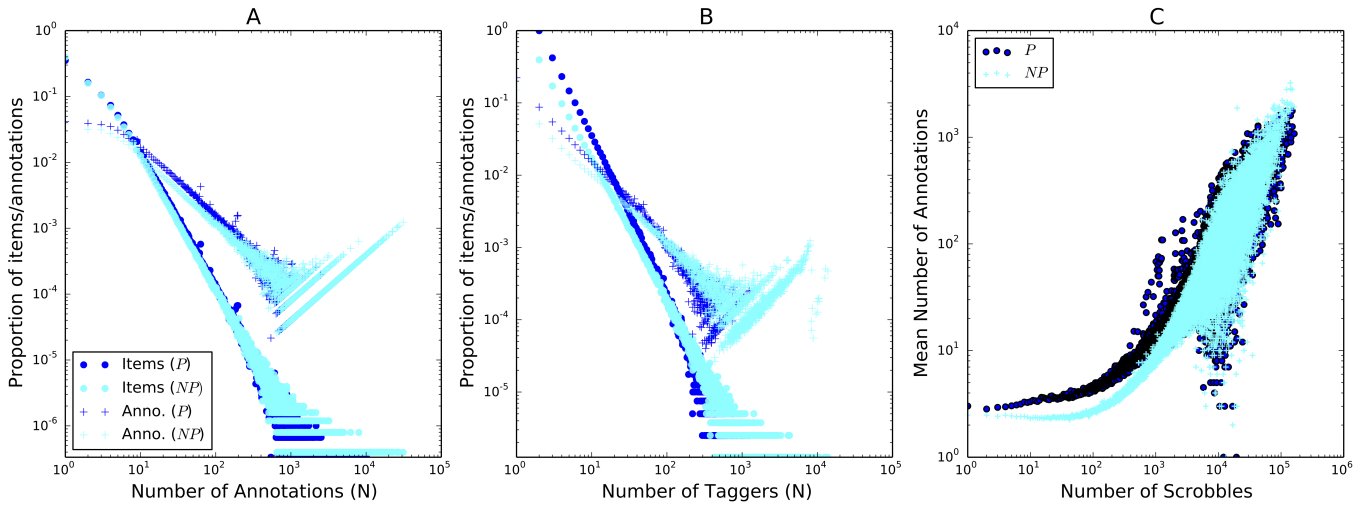

Figure 6: Log-log probability distributions of number of *items* with N total annotations (A) and N total taggers (B), marked with dots. Crosses indicate the corresponding proportions of total *annotations* assigned to items with the corresponding annotation/user count. C shows the mean number of annotations for items with a given global scrobble count.

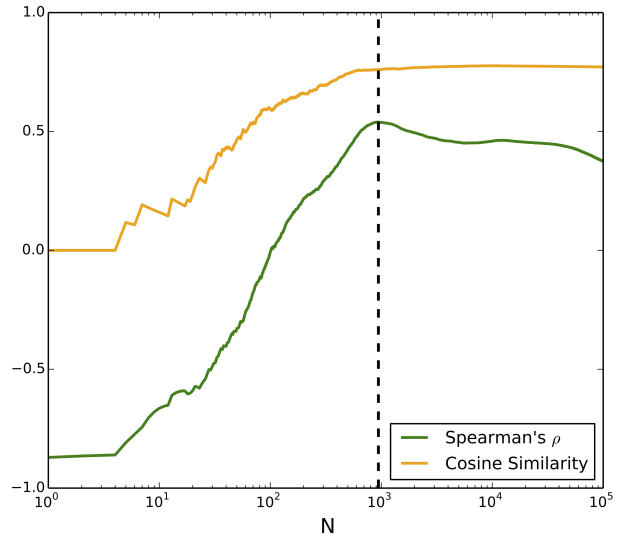

Figure 7: Spearman's ρ and cosine similarity between P and NP as a function of N, considering only the top N items from each folksonomy. The dashed line shows $N = 943$.

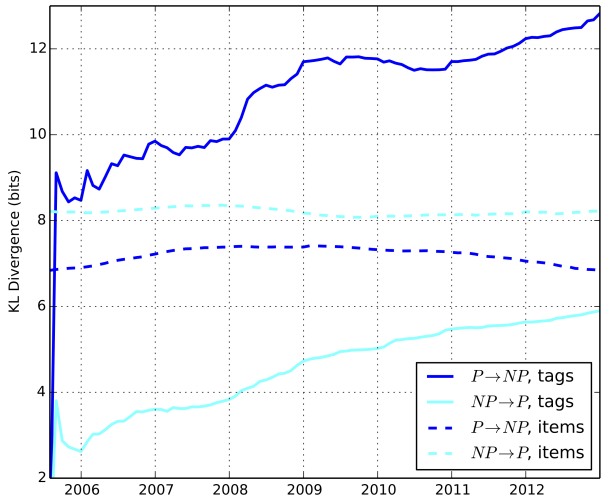

Figure 8: Kullback-Leibler (KL) divergence between P and NP, calculated over tags (solid lines) and items (dashed lines).

below in Section 6.1. However, for items (dashed lines), the KL divergence stayed fairly consistent, indicating that the types of objects annotated were equally accessible to either population.

6. POSSIBLE CAUSAL FACTORS

6.1 Expertise Effects

To measure expertise, we implemented the SPEAR algorithm using its associated package in Python. Briefly, SPEAR works as follows. For every tag t, there are two corresponding vectors: E, a vector of expertise scores of users annotating with t, i.e. $E = (e1; e2; \ldots; e_M)$, and Q, a vector of quality scores for items annotated with tag t, i.e. $Q = (q1; q2; \ldots; q_N)$, where M and N are the total number of users and items associated with t, respectively. From this, an adjacency matrix A of size $M \times N$ is constructed, where $A_{m,n} = 1 + k$ if user m had assigned a tag to item n, and k users had assigned tags to item n after user m, and $A_{m,n} = 0$ otherwise. Thus, if user m was the first to tag item n, $A_{m,n}$ would be set to the total number of users who tag resource n; but if user m was the last one, then $A_{m,n}$ would be set to 1. Following recommendations by [22, 23], the value of $A_{m,n}$ was adjusted by the square root function, such that $A_{m,n} = \sqrt{A_{m,n}}$. Then, user expertise scores per tag are derived by $E = Q \times A^T$.

In computing user expertise scores, we included the first 5,000 most popular tags of the entire folksonomy. In doing so, we obtain a set of 908,494 total across 5,086 users in P, and a set of 5,060,983 expertise scores across 516,694 users in NP. A majority of these expertise scores exhibit an apparent floor effect, with nearly all values less than 0.1. In fact, only 4317 (0.475%) and 587 (0.011%) expertise scores from 1358 users in P and 561 users in NP, respectively, are above 0.1. Thus, a much larger proportion of users from P (27%) have non-negligIbile expertise scores as compared to users in NP (0.1%). In order to get a clearer picture of the distribution of expertise across the two folksonomies of users, we only show scores above this threshold. In addition, we do not compute average expertise scores by user, reflecting the intuition that being an expert in one or several tags does not necessitate being an expert in most or all tags.

Figure 9 presents the distribution of expertise values across P and NP. The differences are most striking in the extremes of expertise scores. A greater proportion of expertise scores in NP are clustered towards the lower end of the range of expertise scores. In contrast, a greater proportion of expertise scores in P are clustered towards the higher end of the range of expertise scores. To reiterate, the SPEAR algorithm assigns higher expertise scores per tag to users annotating quality items (i.e. items more associated with a given tag) more often and earlier than other users. Therefore, the results suggest that users in P are more adept than users in NP at identifying and annotating quality items associated with the 5000 most popular tags.

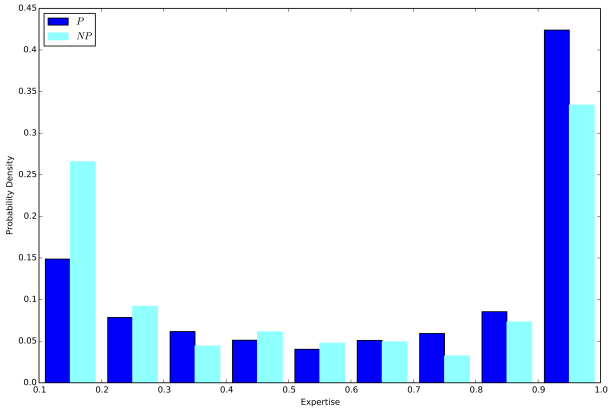

Figure 9: Histogram of all expertise scores exceeding 0.1 for P and NP

6.2 Motivational Effects

To quantify user motivations along the describer-categorizer spectrum, we employed three common metrics: tags per post (TPP), tag/resource ratio (TRR), and the orphan ratio (OR). TPP measures a user's number of total annotations to the total number of annotated items. We expect describers to, on average, annotate items with a greater number of tags and thus score higher on this measure. TRR compares the size of a user's tag vocabulary to the total number of annotated items. We expect categorizers to maintain a constricted vocabulary, and thus score lower on this measure. OR compares a user's number of seldom used tags to the tag vocabulary. We expect describers to be minimally motivated to reuse tags, and thus score higher on this measure. Though there exist other measures, we limit our analyses to these three in light of previous research reporting high correlations between TPP, TRR, OR, and other measures [25]. For full details on the calculations of each measure, see [9].

Figure 10 present, as function of total annotations N, the TPP, TRR, and OR scores for P and NP. As is evident in Figure 10A, user TPP scores increase as total annotations increase. This suggests that P are not simply annotating more items than NP; rather, P are, on average, annotating any given item with more tags than are NP. Similarly, Figure 10C presents a trend of increasing OR scores as total annotations increase. As such, P appear to have far more orphaned tags in their vocabulary than do NP. These two results indicate that P is populated with a far greater number of describers than is NP.

Figure 10B presents as an anomaly then to the above interpretation. Indeed, it presents an unclear, if not negative, relationship between total user annotations and TRR. If greater TRR scores are representative of describers, then the TRR scores are contrary to the TPP and OR scores described above. We believe the discrepancy can be resolved, however, by looking at the relation between total items annotated and the size of user tag vocabularies. For NP, there is a strong correlation of 0.522 between these two values across users; this correlation decreases dramatically to 0.143 for P. This is in line with the results of Cattuto and Baldassarri [2] who report sub-linear growth of user tag vocabularies as compared to total annotated items, perhaps reflecting a saturation point in user tag vocabularies. In sum then, the results suggest that P and NP differ in tagging motivations, where P is populated with more describers than NP.

7. DISCUSSION AND CONCLUSIONS

The principal contributions of this work are the following:

- A formalization of the disproportionate contribution by "supertaggers" to a folksonomy;
- an analysis of the differences between these prolific taggers and their non-prolific counterparts, at the levels of the users themselves and the folksonomic structures they generate; and
- analysis of how these two groups of taggers differ in terms of established measures of expertise and tagging motivation.

Our results demonstrate that, while it is the case that they are more active across a variety of behavioral measures, the most prolific taggers are not simply generating a greater volume of annotations in a manner consistent with "the crowd". Instead, their tagging patterns quantifiably different from the non-prolific taggers. With respect to tag vocabulary, we find that both groups use many of the same most popular tags, but disagree on the long-tail of less common tags, with prolific taggers using fewer true singletons and more moderate popularity (100-10000 total annotations). With respect to items tagged, prolific taggers allocate proportionally more annotations to less popular items, while non-prolific taggers are more likely to tag more popular items. This suggests that the tagging of users in P is more exploratory, favoring items further down the long-tail of popularity instead of tagging the most popular items.

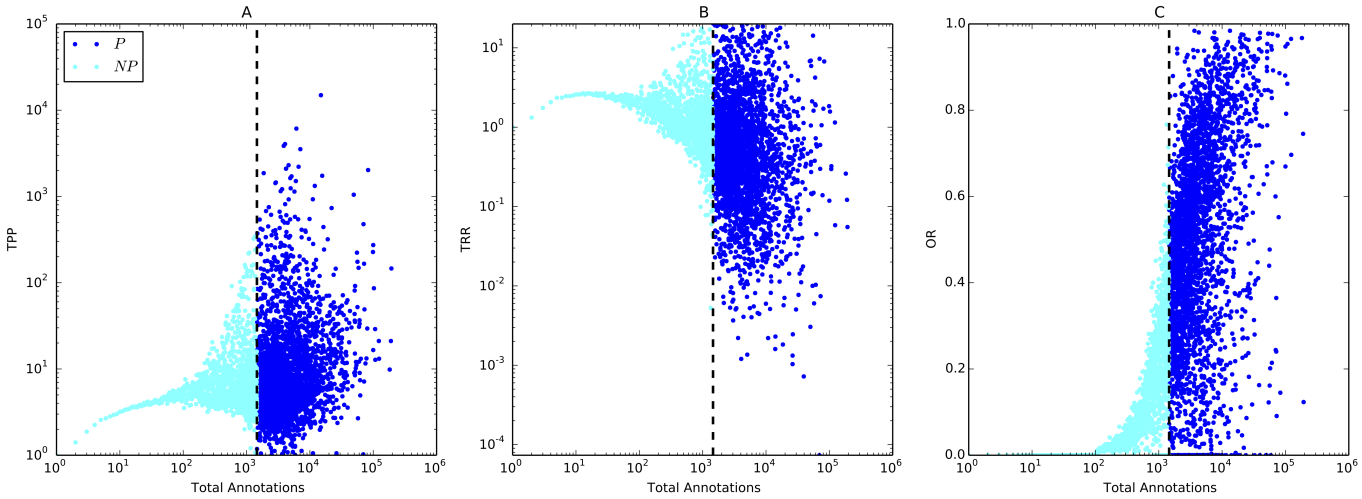

Figure 10: Scatterplots of principal categorizer/describer measures from [25], averaged over users with a given number of annotations. Shown are Tags Per Post (TPP, A), Tag-Resource Ratio (TRR, B), and Orphan Ratio (OR, C).

Though in the aggregate expertise scores from the SPEAR algorithm are low, the "supertaggers" make up a disproportionate number of those users with higher expertise scores. Furthermore, the divergence metrics presented in Section 5.3 are consistent with the SPEAR algorithm, which favors "discoverers" who tag content earlier. Finally, with respect to tagging motivation as formalized in [25], we find the prolific taggers show tagging habits more consistent with describers than categorizers.

The implications of these findings are significant. They reveal that the most prolific taggers on Last.fm exhibit behavior systematically distinct from that of the majority of the tagging population. Our results suggest that the minority of prolific taggers annotate more obscure items using more describer-like vocabularies, and the non-prolific taggers annotate more popular content with categorizer-like vocabularies. This, combined with our information theoretic analyses, challenges the notion that collaborative tagging truly captures the "wisdom of the crowd" in the traditional sense of the term. Whether or not this is "good" for the folksonomy is an entirely different question, however. It may be the case that such a "division of labor" between prolific and non-prolific taggers serves to generate a more usable semantic structure than would be created by users with more homogenized tagging strategies. Addressing this question (e.g. via multi-agent modeling) is a promising direction for future research that is beyond the scope of this paper.

There are, of course, unaddressed complexities at play here. It could be the case, for instance, that the measured differences in motivation of "supertaggers" are partly a function of their tagging more obscure items. This might occur if more obscure items do not fit canonical musical categories and demand multiple classifications such that users tagging them appear more like describers than categorizers, even when this does not reflect a fundamental motivational difference. Relatedly, the motivations of "supertaggers" may not reflect internal, stable user traits but may instead result from interacting with the folksonomy over time. By virtue of discovering more obscure items through increasing use, users' motivations may transition from resembling categorizer to describer behavior for the reasons described above. There also is the question of spam tagging, which we did not address here, other than to use the SPEAR expertise assessment methods to avoid spam tagging problems. Effective identification and elimination of prolific spam taggers might shift the dominance in annotation counts away from the most prolific taggers. A final issue is that the two folksonomies we analyzed here are not independent; a user we have classified as "prolific" could certainly see and be influenced by tags assigned by a non-prolific user, and vice versa. A possible approach to address this would be to only consider items uniquely tagged by users in one folksonomy or the other (i.e. for P, limit analysis to those items tagged *only* by users in P, and vice versa), but more work is needed to determine if and how this might alter our conclusions.

Other future work will need to examine a number of issues, including methods for for identifying more formally what constitutes a "supertagger" and determining other relevant metrics along which these users may differ from the general tagging population. It will also be important to replicate these analyses on more datasets from different tagging systems, to help determine if the patterns observed are idiosyncratic aspects of Last.fm or common across tagging systems in general. At a minimum, we would expect different dynamics in tagging systems with differing (or nonexistent) tag recommendation functionality [24], a feature prominent on Last.fm

Nevertheless, our work presents compelling evidence that the bulk of tagging activity comes from a minority of users. Moreover, the tagging patterns of this unique minority are quantifiably distinct from other users. Thus, it is important for both researchers and designers of collaborative tagging systems to identify and differentially interpret the metadata generated by these "supertaggers".

8. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *SIGCHI conference on Human factors in computing systems*, pages 971–980. ACM, 2007.

[2] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. *ArXiv e-prints*, Apr. 2007.

[3] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, Apr. 2006.

[4] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining - WSDM '08*, pages 195–206, New York, New York, USA, 2008. ACM Press.

[5] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, 2006.

[6] I. Ivanov, P. Vajda, and T. Ebrahimi. In tags we trust: Trust modeling in social tagging of multimedia content. *Signal Processing Magazine, IEEE*, 29(2):98–107, 2012.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, Sept. 1999.

[8] C. Körner, D. Benz, A. Hotho, and M. Strohmaier. Stop Thinking , Start Tagging : Tag Semantics Emerge from Collaborative Verbosity. In *In Proceedings of the 19th international conference on World wide web*, pages 521–530, 2010.

[9] C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 157–166. ACM, 2010.

[10] M. Kubek, J. Nützel, and F. Zimmerman. Automatic Taxonomy Extraction through Mining Social Networks. In *Proc. of the 8th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods*, pages 109–114, Namur, Belgium, 2010.

[11] J. Lorince and P. M. Todd. Can simple social copying heuristics explain tag popularity in a collaborative tagging system? In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 215–224, Paris, France, 2013. ACM.

[12] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40. ACM, 2006.

[13] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[14] M. Niepert, C. Buckner, and C. Allen. A dynamic ontology for a dynamic reference work. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 288–297, Vancouver, British Columbia, 2007. ACM Press.

[15] O. Nov, M. Naaman, and C. Ye. What drives content tagging: the case of photos on Flickr. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1097–1100. ACM, 2008.

[16] O. Nov and C. Ye. Why do people tag? *Communications of the ACM*, 53(7):128–131, July 2010.

[17] E. Quintarelli. Folksonomies: Power to the People. In *Proceedings of the 1st International Society for Knowledge Organization (Italy) (ISKOI), UniMIB Meeting*, Milan, Italy, June 2005.

[18] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 271–280. ACM, 2010.

[19] M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *ICWSM*, 2010.

[20] T. Vander Wal. Folksonomy: Coinage and Definition, 2007.

[21] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing Social Bookmarking Systems : A del.icio.us Cookbook. In *Proceedings of the ECAI 2008 Mining Social Data Workshop.*, pages 3–7, 2008.

[22] C.-m. A. Yeung, M. G. Noll, N. Gibbins, C. Meinel, and N. Shadbolt. SPEAR: Spamming-Resistant Expertise Analysis amd Ranking in Collaborative Tagging Systems. *Computational Intelligence*, 27(3):458–488, 2009.

[23] C.-m. A. Yeung, M. G. Noll, C. Meinel, N. Gibbins, and N. Shadbolt. Measuring Expertise in Online Communities. *IEEE Intelligent Systems*, 26(1):26–32, Jan. 2011.

[24] A. Zubiaga, V. Fresno, R. Martinez, and A. P. Garcia-Plaza. Harnessing Folksonomies to Produce a Social Classification of Resources. *IEEE transactions on knowledge and data engineering*, 25(8):1801–1813, 2013.

[25] A. Zubiaga, C. Körner, and M. Strohmaier. Tags vs shelves: from social tagging to social classification. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 93–102. ACM, 2011.