Can simple social copying heuristics explain tag popularity in a collaborative tagging system?

Jared Lorince
jlorince@indiana.edu

Peter M. Todd
pmtodd@indiana.edu

Cognitive Science Program
Department of Psychological & Brain Sciences
Indiana University
Bloomington, IN 47405

ABSTRACT

While research on collaborative tagging systems has largely been the purview of computer scientists, the behavior of these systems is driven by the psychology of their users. Here we explore how simple models of boundedly rational human decision making may partly account for the high-level properties of a collaborative tagging environment, in particular with respect to the distribution of tags used across the folksonomy. We discuss several plausible heuristics people might employ to decide on tags to use for a given item, and then describe methods for testing evidence of such strategies in real collaborative tagging data. Using a large dataset of annotations collected from users of the social music website Last.fm with a novel crawling methodology (approximately one millions total users), we extract the parameters for our decision-making models from the data. We then describe a set of simple multi-agent simulations that test our heuristic models, and compare their results to the extracted parameters from the tagging dataset. Results indicate that simple social copying mechanisms can generate surprisingly good fits to the empirical data, with implications for the design and study of tagging systems.

Author Keywords

Collaborative tagging, folksonomy, decision-making, ecological rationality, heuristics, cognitive science

ACM Classification Keywords

H.1.2 User/Machine Systems: Human information processing; H.3.5 Information Storage and Retrieval: Online Information Services—Web-based services; H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces—Collaborative computing, Web-based interaction; J.4 Social and Behavioral Sciences: Psychology

General Terms

Human Factors, Measurement, Theory

WebSci'13, May 1 – May 5, 2013, Paris, France.
ACM 978-1-4503-1889-1...$10.00.

INTRODUCTION

Motivation

Collaborative tagging systems have been the topic of serious research within the computer science community since the mid-2000s, gaining particular ground when Thomas Vander Wal coined the term "folksonomy" in 2004 [22] to refer to the bottom-up, user-generated organizational structure tagging systems enable. From early descriptive work [13] to computational models [5] and prediction of social links from tagging behavior [18], our understanding of the dynamics of tagging systems has increased greatly. Despite these advances, there has been little work exploring the decision-making processes of users in tagging environments. The high-level properties of collaborative tagging systems (e.g. their emergent vocabularies, distributions of tag frequencies, and so on) are the result of human behaviors in a dynamic environment where individuals' tagging decisions structure the information environment in which subsequent decisions are made. As such, the cognitive sciences can do much to inform our understanding of how and why people tag. This work is an early step in that direction, exploring the extent to which a simple multi-agent model of heuristic decision-making can account for the aggregate distribution of tag use in a collaborative tagging system. Though tagging is clearly a complex process, with both social influence (i.e. copying of other tags) and evaluation of the tagged content playing a role, the present study explores the strength of social copying in producing global patterns of tag use of the kind we see in Last.fm.

How do people tag?

Golder and Huberman [13] performed one of the first in-depth studies on the dynamics of a tagging system, exploring global patterns of tag use, changes in individuals' tagging vocabularies, and classification of different types of tags on the social bookmarking website Delicious. Subsequent work has elaborated all of these research threads. Al-Khalifa and Davis [2] and Marlow, et al. [15], for example, have developed taxonomies of tag types and motivations for tagging (e.g. tagging for future retrieval, for contribution and sharing, or for opinion expression), and studied their implications for the study and design of tagging systems. Farooq, et al. [6] reviews a number of other metrics for evaluating tagging systems, including tag-discrimination and tag reuse, using them to develop a set of design heuristics for collaborative tagging systems. More recently, Schifanella, et al. [18] studied the role of social relationships in tagging systems, discovering that social activity (i.e. friendships and group memberships) correlates with tagging activity, and furthermore that tagging habits were predictive of the existence of social ties between users.

This is only a sampling of the substantial body of work on tagging, but illustrates the expanding understanding of tagging systems. It remains a challenge, however, to connect high-level analyses of tagging patterns to a lower-level understanding of human decision processes in these environments. Making strong claims about cognitive processes based on Web data, whether the data is large- or small-scale, is a difficult problem [14], but one that has attracted increasing attention among cognitive scientists in recent years [12]. Though datasets like our own do provide ecologically valid data — that is, signals of behavior carried out in people's day-to-day lives without the artificiality of a lab context — they also suffer from a lack of experimental control. We can only make indirect inferences as to exactly what information was available to a user at the time of a tagging decision, and how this information did (or did not) factor into that decision. An understanding of why people tag to begin with is not even fully developed; theories and taxonomies of tagging motivations, as mentioned above and developed from survey research like [17] are helping, but it is still often unclear what motivated the decisions to generate the annotations we crawl from the Web, or how differences in motivations might imply different generating processes for the choice of which tag to apply. This of course does not make the development of models of tagging behavior a hopeless pursuit, and several notable efforts have been made. For example, Cattuto et al. [5] developed a modified Yule-Simon "rich-get-richer" model of tagging [19, 24], in which new tags are added to the folksonomy with a probability P, and with probability $1-P$ a tag is copied from from the existing distribution proportional to its frequency. This was then modified to include a memory kernel that resulted in more recently used tags being more likely to be copied. Other models have focused more directly on cognitive plausibility, such as Fu, et al.'s [8] semantic imitation model (the paper on which also provides a thorough review of other tagging models). In their model, users infer from the existing tags for a document the topics it contains, and assign semantically related tags based on that inference.

The approach we take in this paper is in some ways a hybrid of the two just described, in that we aim for the simplicity of a Yule-Simon model while simultaneously pursuing cognitive plausibility. We achieve this by exploring how simple decision-making heuristics, as developed within the ecological rationality research program, might account for patterns of tag use in a collaborative tagging system.

Ecological rationality

Research on ecological rationality, spearheaded by the Adaptive Behavior and Cognition Group [11, 20], approaches decision making from a different perspective than traditional judgment and decision-making theory. Rather than developing models that start with the assumption that people are rational, optimal decision makers, and then modifying them to account for humans' observed deviation from rational principles (a strategy especially common in behavioral economics [3]), the ecological rationality approach takes a decidedly different view of what it means for an agent to be rational. To be ecologically rational is not to achieve optimality, or even optimality under constraints, but rather to utilize simple decision-making heuristics — rules of thumb — derived from evolutionary adaptations to problems humans faced in ancestral environments. These adaptive strategies allow humans to make quick decisions under the simultaneous constraints of time, available information, and cognitive processing capacity. For decades, this kind of decision making was lumped under the category "heuristics and biases" (as in the work of, e.g., Kahneman and Tversky [21]), with the negative connotations of "bias" overshadowing the fact that heuristic strategies are often effective under the right circumstances: "Heuristics are efficient cognitive processes that ignore information. In contrast to the widely held view that less processing reduces accuracy, the study of heuristics shows that less information, computation, and time can in fact improve accuracy" [9].

There exists a large variety of cognitive heuristics, collectively making up the mind's so-called "adaptive toolbox", that vary in their appropriate domain of application. A thorough review is beyond the scope of this paper (see [10, 11, 20]), and we specifically wish to focus on social heuristics. These strategies, such as imitate-the-successful or imitate-the-majority [4], leverage available social information to make quick decisions. Such heuristics of course can be problematic when misapplied (e.g. by allowing the spread of misinformation within a population), but have immense adaptive value as tools for making decisions in uncertain environments. When first-hand information is scarce or costly to acquire, emulating the behavior of successful individuals or of large groups can be an efficacious strategy.

While evolution of course did not equip humans with a "tagging heuristic", it likely did give us innate strategies for appraising and building on the cultural productions of others around us, and there are reasons to believe that a tagging environment is one such cultural setting conducive to social heuristic strategies. When tagging, a user can apply any of an effectively infinite set of possible labels to an item, and has access to social informations in the form of the existing distribution of tags for an item (often displayed as tag clouds, or partly communicated via tag recommendations). In an uncertain environment with accessible social data, it is reasonable to hypothesize that users will employ social imitation strategies. In the tagging context, this presumably takes the form of users copying the tagging decisions of others. In what follows we formalize the form such copying might take in the tagging system of Last.fm, and describe the formal multi-agent model we developed to test our hypotheses. Clearly, people do not use copying alone to decide what tags to use in a given situation — content will account for much of how the choices made. But *how* much? By exploring the extent to which global tagging patterns can be explained by the use of copying heuristics, we also help to identify the extent to which other factors, including semantics, also play a role.

Last.fm

Our analysis is of data from Last.fm (http://www.last.fm), a social music service built around the concept of "scrobbling". Using plugins for a wide variety of music software, users upload their listening history to their profiles on the site (each instance of listening to a track is termed a "scrobble"), which Last.fm then uses to generate intelligent music recommendations. This functionality is coupled with social networking capabilities, and, most relevant to our work, the ability for users to annotate artists, albums, and songs with textual tags. This is supported directly through Last.fm's web interface and via API functionality. Tagging is completely freeform, such that users may apply as many tags of arbitrary length (including multi-word tags with spaces) as they want for any given item. The same user can assign multiple tags, but cannot assign the same tag more than once to a given item. The systems fits Vander Waal's definition of a "broad" folksonomy, in which many individuals tag the same content [23]. The site maintains a robust API that allows for easy access to many of their user and content metrics.

The choice of Last.fm as our object of study was, to a certain degree, arbitrary, as our questions of interest could be asked of any broad folksonomy. In a narrow folksonomy, like the photo-sharing website Flickr, users predominantly only tag content they have uploaded themselves, making questions around social copying minimally applicable. However, even among broad folksnomies, the music tagging domain is especially appropriate for our investigation. Classification of musical content is notoriously challenging, so much so that work has been published in the music informatics community explicitly asking if musical genre classification is even a problem worth pursuing [16]. We thus propose that in such an uncertain domain, social copying may play a particularly important role.

Terminological notes

Throughout this paper, an "annotation" refers to a given instance of a user assigning a tag to an item at a particular time, and can be thought of as a unique four-element tuple in the form user-item-tag-time. 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. Last.fm maintains distinct tag distributions for each unique item (even if one is a subcategory of another, e.g. an album by a particular artist).

DATASET

Crawling the data

Previous work using Last.fm tagging data (e.g. [18, 7]) has crawled content primarily using the site's API, but this proved insufficient for our purposes. Though the API does provide accurate summary information of a user's tagging habits (i.e which tags they have used, and which items they have annotated with each tag), its utility is limited because it provides no temporal tagging information (that is, when a user tagged a particular item with a particular tag). User profile pages, however, contain a timestamped history of a user's tagging activity, with a temporal resolution of one month for all annotations made more than one month prior to when the page is loaded. Though a finer-grained resolution would have been ideal, this limitation could not be avoided. Thus all temporal information presented hereinafter is binned into monthly increments.

To crawl the data, we developed a hybrid crawler combining API methods and direct parsing of the HTML content of user profile pages. All data was stored in a MySQL database. We began with a set of arbitrarily selected seed users, whose usernames were used to initialize a crawl queue. The program then proceeded by (1) selecting the next username from the crawl queue; (2) querying the Last.fm API for the user's numeric ID and list of friends on the network, adding each to the crawl queue and recording the friendship relations; (3) extracting the user's list of unique tags utilized from his/her profile; and finally, (4) for each tag, extracting annotation information for each instance of that tag. Each annotation was stored in a MySQL table as a four-element tuple consisting of a user ID, an item URL (the relative URL on Last.fm for the artist, album, or song tagged, e.g. "/music/Mogwai/_/Auto+Rock"), the tag assigned, and the month and year in which the item was tagged. The process then repeated, drawing usernames from the crawl queue until a sufficiently large sample of content had been crawled. The number of usernames increased at a faster rate than we could crawl the content of users' profiles, so we did not run into the problem of exhausting the queue. The crawling process is schematized in Figure 1.

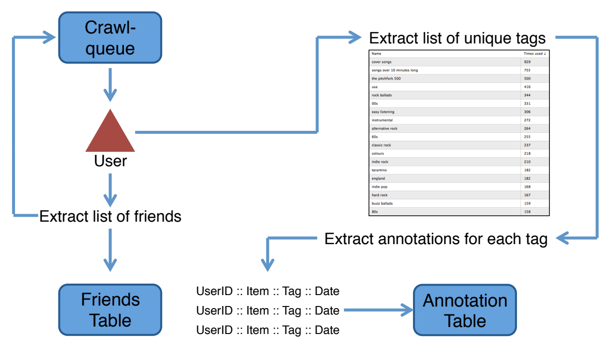


Figure 1. Schematic of the crawling process. Friendship relations were recorded, though not used in the analyses presented here.

Caveats

Several potential pitfalls with our crawling methods must be addressed. First, we were forced to crawl on a user-by-user (as opposed to item-by-item) basis, because temporal annotation data was only available on user profile pages. This allows for a complete picture of any particular user's tagging habits, but means we cannot guarantee that the data for any particular item is complete. There is no way around this problem, as Last.fm only makes available normalized, summary information for the tags assigned to any particular item. We purposefully collected as large a sample of data as possible, however, making the probability quite low that the tag distribution for any particular item is not representative of its true distribution, especially for items with many annotations. A second possible problem was our choice to crawl users via friendship relations. With a small sample this could present problems of covering a biased subsample of the social network, but with over one million users crawled, we have no reason to believe our network coverage is biased.

A more notable problem with this decision is that it necessarily limits our data to Last.fm users who have at least one friend on the network. Previous work, however, indicates that tagging behavior is correlated with having more friends on the social network [18], and this was confirmed within our own data. To further verify this, we crawled a small truly random sample of users by generating random numbers and crawling the annotation data for users whose numeric IDs matched those values. As we would expect, this resulted in far less annotation data (approximately 479,000 total annotations from 200,000 users, an average of 2.4 annotations per user) than we collected using the friend-based crawler (approximately 33 annotations per user, see Table 1). Given our goal of analyzing tagging behavior, we thus concluded that selectively collecting data from users with friends was more likely to give us greater amounts of analyzable tagging data.

Data summary

We crawled a total of slightly over one million users, extracting a total of approximately 33 million annotations. Last.fm does not regularly report the size of their user base, but a 2009 announcement [1] claimed approximately 30 million users. Considering that our analyses suggest that user activity has been declining since 2009, we estimate that our sample covers between two and three percent of Last.fm users. Table 1 presents a detailed summary of our data collected, after performing database maintenance to clean the data and remove any erroneous or duplicate entries. Consistent with other large-scale tagging datasets, we find long-tailed distributions[1] for several key summary metrics of the dataset, visualized in Figure 2

Users	1,053,163
Active Taggers	318,415
Total Annotations	33,140,605
Total Unique Items	3,262,724
Total Unique Tags	747,275
Friendship Relations	12,408,953

Table 1. Dataset summary. Active taggers are those users with at least one annotation.

SIMPLE TAGGING HEURISTICS

Exploring the interface

In considering heuristic decision-making strategies people may use in a tagging environment, we seek possibilities that are both simple (from both the cognitive processing and computational modeling perspectives) and psychologically plausible. In the tagging domain this can be challenging, because the environment is quite complex, and the available data tell us little about precisely how users were interacting with the system when they made their tagging decisions. While on the Last.fm website, there are various ways to arrive at the tagging interface, including via the radio feature (Figure 3A), an item information page (Figure 3B), or the tag cloud for an item (Figure 3C). Furthermore, the site's API supports methods for assigning tags via external applications and services, and we can only speculate as to how and to what extent users tag content with such services.

[1]We remain agnostic here as to the precise mathematical form of these distributions. The curves are suggestive of power laws with exponential decay, but our analyses are not dependent on the distributions taking any particular form, and we thus refer to them only as "long-tailed distributions".

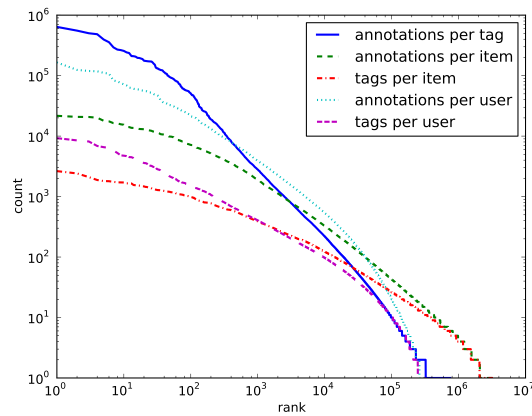


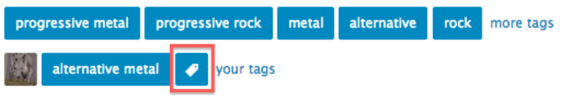
Figure 2. Frequency-rank plots of key metrics for the dataset, on a log-log scale. Displayed are the total number of times each tag was used across the dataset ("annotations per tag"), the total number of times each item was tagged ("annotations per item"), the total number of unique tags assigned to each item ("tags per item"), the total number annotations made by each user ("annotations per user"), and the total number of unique tags used by each user ("tags per user"). In all cases, values are ranked from highest to lowest ("rank" on x-axis), while the count for each metric is shown on the y-axis, and one point is plotted per user/item.

Nonetheless, we can still develop plausible hypotheses as to people's tagging strategies based on knowledge of the Last.fm interface. Though there is variance in the presentation of tag information on Last.fm, the web-based interface via which users can actually assign tags to items has a standard format (Figure 4) that prominently displays the top five most popular tags for the item being tagged, ordered by frequency, and describes them as "suggested tags". Thus when a user decides to tag an item, (s)he is presented with aggregated social data on other individuals' tagging decisions in a way that facilitates copying those decisions, explicitly presenting them as suggestions, and making their assignment to an item a simple matter of clicking the name of the tag, rather than typing it in. Re-use of one's own commonly-applied tags ("your tags" in Figure 4) is facilitated in a similar manner, with a user's most commonly used tags listed in overall frequency order (i.e. "español" is the most commonly-used tag across all items for the user in Figure 4).

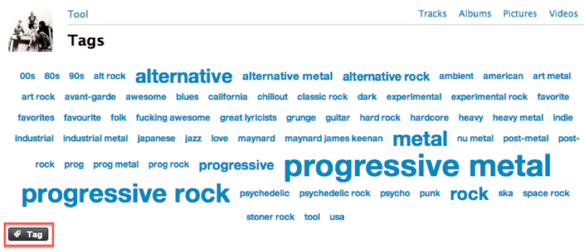
Description of possible heuristics

If people do in fact utilize simple social copying heuristics in deciding how to tag, what form might these heuristics take? Our goal here is not to develop a complex cognitive model, but rather a simple strategy that, if embodied by agents in a tagging system, would approximate at least some of the be-

(a) Tag information available when listening to Last.fm radio.



(b) Tag information available on an item's main information page.



(c) Tag cloud for an item on Last.fm showing more than 50 unique tags, with font size proportional to their relative frequency.

Figure 3. Different presentations of tagging information on Last.fm. In each image, the button triggering the display of the tagging interface is highlighted in red.

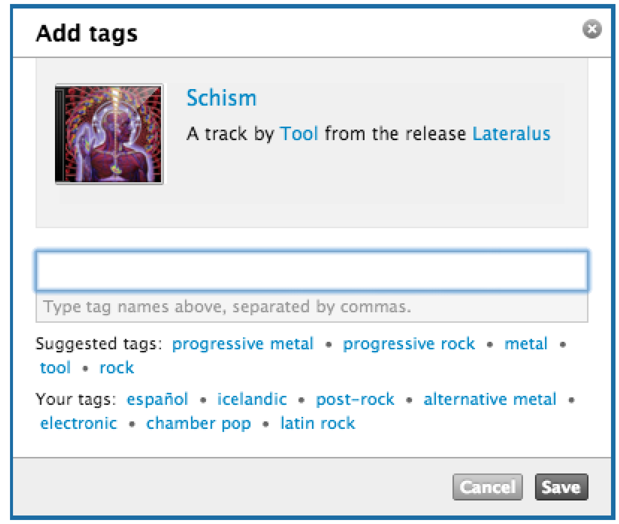


Figure 4. Web interface for tagging an item on Last.fm

havior we observe in our dataset. Let us formalize the decision strategy as a stochastic process in which a user, upon deciding to tag an item (we will leave aside, for the present model, how a user decides *which* items to tag) engages in copying behavior with probability P, and novel tagging behavior (that is, generating a new tag rather than copying) with probability $1 - P$. A user copies at time T by re-using a tag from the cumulative tag distribution for an item existing at time $T - 1$. We begin our discussion with three simple strategies:

1. Uniform heuristic: The simplest possible copying strategy, from an algorithmic perspective, would be to randomly pick a tag from the existing tag distribution at time $T - 1$ for the item being tagged. Though simple, this is not likely behaviorally, as there is no way for a user to review the complete tag distribution for a given item.

2. Normalized heuristic: Instead of sampling randomly, we could assume that the probability of copying any particular tag is proportional to its relative frequency for that item at time $T - 1$. Though unrealistic for the same reason as the uniform heuristic, this would at least capture the idea that more frequent tags are more likely to be copied (as they are displayed more prominently in an item's tag cloud). This model is akin to a preferential attachment on the level of items.

3. Top-5 heuristic: Rather than make the unrealistic assumption that users have access to the entire distribution of tags for a particular item (or that they would be able to effectively use that information, were it available), the top-5 heuristic assumes that, when copying, a user selects randomly among the top 5 most popular tags (the "suggested tags"in Figure 4) for an item. This remains agnostic as to how precisely the user selects which of the five top tags to copy, but provides a plausible tagging heuristic that is based on social information that is (a) directly available and (b) of manageable size. We thus hypothesized that this heuristic would be the top performer when modeled.

Estimating copying parameters

We have described three simple tagging heuristics, each dependent on a single parameter P. Before exploring modeling strategies, we were curious if these parameters could be extracted directly from our dataset. Keeping in mind that our data is at the temporal resolution of one month, we formalized the problem as follows: Among all annotations of an item i in month M, what proportion of these new annotations were "copies" under the definitions in described above? This provided, on an item-by-item basis, a summary value describing the extent to which people copied tags from the existing tag distribution. In the case of the uniform heuristic, this amounted to simply calculating the proportion of annotations for item i during month M that used tags that already existed (i.e. had a frequency $>= 1$) in the cumulative tag distribution in month $M - 1$. For the normalized heuristic, we used the equation:

$$P = \frac{\sum_{i=1}^{n} \left(freq(t_{i,M}) \times \frac{freq(t_{i,M-1})}{t_{max,M-1}} \right)}{\sum_{i=1}^{n} freq(t_{i,M})}$$

So, for each unique new tag t used in month M, we multiplied its frequency of use in that month by its normalized frequency in the preceding month $M - 1$. This normalized frequency was simply its frequency in the tag distribution divided by the frequency of the most common tag in the distribution. This ensured the metric would be on a 0-1 scale, where a value of one indicated copying of the most popular existing tag. The product of the two values described was then divided by the total number of annotations in month M ($\sum_{i=1}^{n} freq(t_{i,M})$), providing the final normalized value of P for that item and month.

Finally, the top-5 heuristic was calculated in much the same way as the uniform heuristic, except in this case we calculated only the proportion of new annotations in month M using tags that were among the cumulative top five most popular tags for a given item in month $M-1$. All calculations were based on our sample of annotation data, so we cannot guarantee that these estimates perfectly represent what users observed when tagging. However, given the large size of our sample, we believe our estimates to be reasonable.

We calculated the copy index P for all possible items and months. It was of course impossible to calculate values for (a) the first month in which we had recorded annotation data for a particular item, and (b) for those items for which we only had annotation data from a single month. We were thus limited to considering 1,119,345 of the items in our dataset. There was enough variance across items and months that simple summary statistics were uninformative, so we instead considered how the three copying metrics varied as a function both of time and increases in annotations. Figure 5 shows the three copy indexes as a function of total annotations, with each point in the scatterplot indicating the the copy index (y-axis) for all items with the corresponding total number of annotations shown on the x-axis, averaged across all months for which the index was calculated. Thus each point is the average value of the copy index across all time for all items with a particular number of annotations. The lines show moving averages of the points in the scatter plot, and clearly indicate a convergence of copy index as items acquire progressively more annotations. The data is noisy for items with few annotations, but the convergence of copy index values as annotations increase allows for an estimate of the underlying value of the copy index.

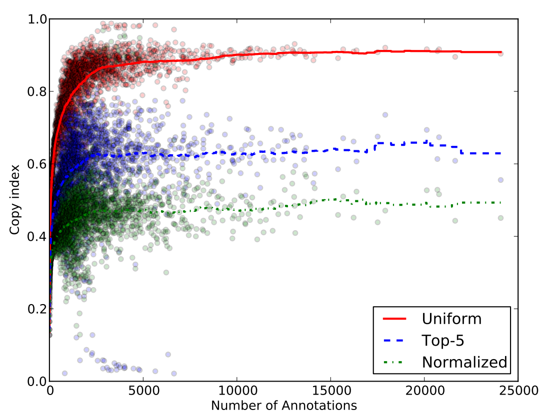


Figure 5. Copy index as a function of total annotations. Each point indicates the mean copy index across all months for items with the corresponding number of annotations. Solid lines are moving averages of increasing window size equal to $a^{0.9}$, where a is the total number of annotations shown on the x-axis.

The data as presented in Figure 5 may overrepresent those items with the most annotations, so we also calculated the copy index metrics as a function of time. Figure 6 shows the same three copy indexes over items' lifetimes. We calculated, for all items, the average copy index each month after the first month in which we had tagging data for that item. This was then averaged across items for each month index. If an item had no annotations in a given month, it was excluded from the calculation for that month. For example, the copy index where item age equals 30 is the average copy index of all all items in the thirtieth month after their first annotation. Thus we have a picture of the average progression of the copy indexes over the life of items. Copy indexes for the largest values of item age show greater error because relatively few items have existed across the full 89 months for which we have data.

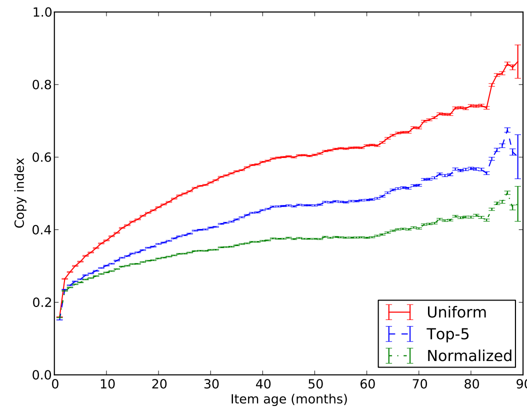


Figure 6. Copy index as a function of time. The x-axis shows the "age" of an item (months since first annotation), while the y-axis shows the mean copy index for all items x months after their first annotation. The average for a particular month does not incorporate those items with 0 annotations in that month. Error bars show +/- one standard error.

Making a precise estimate of the underlying value of P from these data is a challenge, as it shows a clear temporal dependence; all three metrics increase as items accumulate more annotations over time. This analysis does, however, provide us with a range of plausible values to focus on in the development of our multiagent models (in the range of approximately 0.6 to 0.9 for the uniform heuristic, 0.3 to 0.5 for the normalized heuristic, and 0.4 to 0.6 for the top-5 heuristic). Even with these estimates, it remains to be seen if models using the described heuristics as generating mechanisms can reproduce the patterns of tag popularity observed in our dataset. In the following section, we address precisely that.

MODELING

Core model description

With a set of candidate heuristic models in hand, we developed a basic multi-agent simulation framework in which to test them. The core model required a single parameter P and the metric by which we evaluated the models' performance was the overall distribution of tag use it generated (c.f. the solid blue line in Figure 2). In a complex environment consisting of one million or more agents, difficult-to-predict, emergent effects are to be expected, so we constrained the model to emulate the environmental structure of the empirical dataset as much as possible, but not so much so that we trivially ensured that it would mirror the true distribution of tags. We thus ran all simulations with the same number of users, items, and possible unique tags as in our crawled data (see Table 1). Furthermore we constrained it such that the distributions of total annotations per item and total annotations per user were the same as in the real data. We accomplished this by generating random user-item-timestep triplets that mirrored the distributions just mentioned. This amounts to the assumption that taggers' activity levels (i.e. how many items they tag), item popularity (i.e. how many times a particular item is tagged), and taggers' decisions of which items to tag are all processes independent of the heuristic process we modeled of deciding how to tag an item. This is of course a gross simplification, but permits a focused analysis of the effects of the simple heuristics described above.

For each user-item-timestep triplet, the model simulated one of the heuristics previously described. In pseudocode:

for *each user-item-timestep* **do**
    rand = random real number in range [0,1];
    if *rand* $< P$ **then**
        select tag T with copying heuristic;
    else
        generate novel tag T;
    end
    assign tag T to item;
end

The generation of a novel tag, in this simple model, was simply the assignment of a random tag from the set of possible tags (i.e. the number of unique tags in our crawled dataset)[2]. The program maintained a data structure for each item containing the frequency of each item assigned to it, and tagging decisions at any given time step were based upon the distribution of tags existing for the item at that time. Tags lacked any semantic content or relationship to one another, being represented simply as integers. The same held for items and users.

Our model, especially when using the normalized heuristic, is akin to a multi-agent version of the Yule-Simon model upon which Cattuto, et al. [5] based their model. The key difference is that our model, in an effort to represent the information available to real users when tagging, operates on the level of item tag distributions (complete distributions for the normalized heuristic, or only the most popular tags in those distributions for the top-5 heuristic). A true Yule-Simon model, on the other hand, operates on the level of the global tag distribution, but we cannot assume that users have access to that overall distribution.

Results of basic models

We ran the first version of the model a total of 27 times, 9 times across a sampling of P values (from 0.1 to 0.9, in increments of 0.1) for each of the three basic tagging heuristics. For each run we calculated and plotted the distribution of total tag use as in Figure 2. Figure 7 shows the results. No quantitative analysis is required to see that none of the three models — with any possible parameter value — generate reasonable fits to the data. With all three heuristic strategies, and across the range of possible P values, we see an underrepresentation of the most popular tags and overrepresentation of rare tags, as compared to the empirical distribution. However, it is clear, as we expected, that the top-5 heuristic comes closest to matching the empirical data among these three basic models. Although the normalized heuristic, for high values of P, shows a similar pattern as the top-5 heuristic, the latter model is more robust over a wider variety of parameter values, including those previously estimated. The other two models, for the range of parameter values estimated, show much poorer fits.

Model refinement

Leaving aside the somewhat unrealistic uniform and normalized heuristic models, we explored a variety of extensions and modifications of the top-5 model. With such a simple model, there are countless possible changes that could have led to better fits to our empirical data, but we only explored those options that retained our focus on simple copying heuristics (without making assumptions about, e.g., individuals assigning semantically appropriate tags to content) and did not require that users have any knowledge about the global distribution of tags. We experimented with several model modifications, such as:

- Varying the number of suggested tags displayed to users (effectively making the "top-5" heuristic, for example, a "top-10" or "top-15" heuristic). Though not in line with our interface-based predictions, if users often explore the tag clouds (Figure 4C) of items, it is possible that the effective set of suggested tags is in fact larger than the five we hypothesized. This modification resulted in minimal change to the overall tag popularity distribution.

- Implementing a fixed set of common tags that were more likely to be used than a random tag when users did not copy. This amounted to there existing a set of popular or well-known tags that all users have in mind, and are more likely to use. This modification was not successful either, as it resulted in those top tags dominating the distribution while being used with roughly equal frequency to one another.

After extensive experimentation, we found that implementing a secondary form of copying within the tagging decision preocess led to substantially better fits to our data. Leaving the principal copying mechanic of the top-5 heuristic as it was, we added the assumption that users keep track of the tags they have been exposed to every time they tag any item (i.e. the five top tags suggested for any item seen). This information was used to modify the novel tag generation process such that, rather than simply assigning a random tag, users would engage in one of two behaviors. With probability Q, they would re-use (i.e., copy from a different item) a tag they had previously seen with probability proportional to the number of times they had encountered it, and with probability $1 - Q$ they would assign a random tag. This new, two-parameter version of the model embodies the simple assumption that

[2] It was possible for an agent to assign, by chance, an existing tag to an item even when engaging in novel tagging behavior.

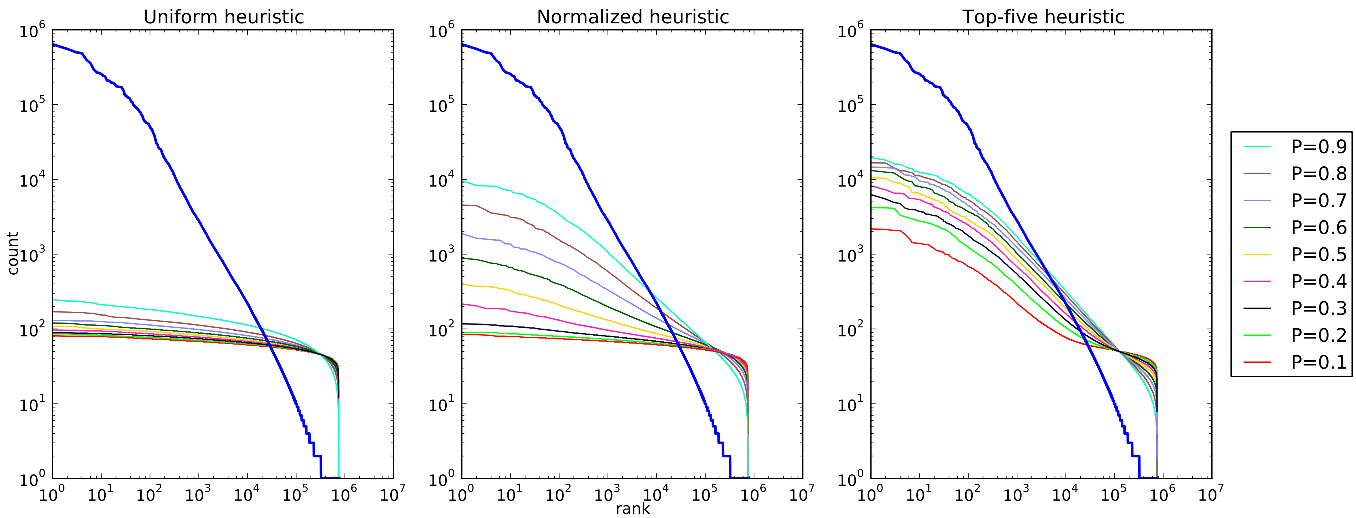


Figure 7. Frequency-rank plots of overall tag use for each of the three basic models, across a sampling of P values. The bold blue line shows the empirical tag distribution.

users, when selecting a new tag to assign to an item, are more likely to use tags they are more familiar with. In pseudocode:

```
for each user-item-timestep do
    rand1 = random real number in range [0,1];
    if rand1 < P then
        randomly select one of top five tags;
    else
        rand2 = random real number in range [0,1];
        if rand2 < Q then
            select tag T from distribution of encountered tags
            with probability proportional to frequency;
        else
            select random tag T from set of all possible tags
        end
    end
    for each tag in item's top five tags do
        add tag to frequency distribution of encountered tags
    end
    assign tag T to item;
end
```

This of course makes the unrealistic assumption that users have perfect memory of tags encountered, and does not account for effects of users preferentially re-using tags that they have previously used (rather than observed), but for the purposes of our simple model captures a secondary form of social copying that is plausible to hypothesize in a tagging environment. This secondary copying is distributed over time (rather than being based on the five top-tags observed at the time of copying), but the weighting of tag selection by relative frequency presents a realistic assumption that the tags a user has most often seen are those (s)he is most likely to use. Furthermore it generates more realistic fits to the empirical tag popularity distribution without assuming any knowledge of the global tag distribution on the part of of users.

Fitting and model comparison

Due to processing constraints, we were unable to cover as broad a range of parameter values in our second set of simulations. Our initial parameter estimations, however, constrained the range of values for P that we were interested in exploring. We thus ran simulations across P values of 0.4 to 0.6, and varied Q from 0.1 to 0.9 (both in increments of 0.1). Within these constraints, our simulation best fit the data with the parameters $P = 0.6$ and $Q = 0.9$. These parameters were simple to arrive upon, as fit to the empirical data increased monotonically in both parameters. Further increases in P may have generated better fits, but such parameterization would have been inconsistent with our earlier analysis, so we selected the highest value of P consistent with the range estimated above.

The tag rank-frequency plot generated by this distribution is shown in Figure 8. We also show, for comparison, the best fit of the standard top-5 model within the estimated range ($P = 0.6$), the empirical distribution, and a null model distribution. The null model, which is equivalent to users tagging randomly across all 33 million simulated annotations, has an equal frequency across all possible tags (approximately 44 annotations per unique tag).

To state the form of our model in simple terms, users engage in simple copying of one of the top five tags for an item roughly 60% of the time. In the 40% of cases where they do not copy, they are far more likely to utilize a tag they have seen at some point before, with the most seen tags being the most likely to be re-used. Only in a small proportion of cases (10%) do they generate “truly” novel tags.

Though the relative fits of the of the basic top-5 model versus the modified version can be directly observed in Figure 8, we can quantify them by calculating the root mean squared error (RMSE) of each as compared to the null model. The basic model has an RMSE of 2224.72 as compared to the empirical data, versus the null model's RMSE of 2296.46 as compared to the empirical data. This represent an improvement of only

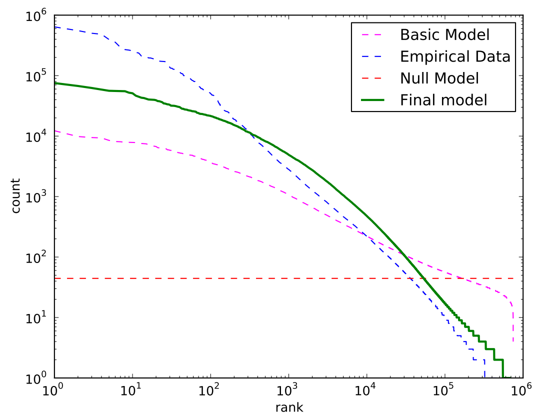


Figure 8. Frequency-rank plots of overall tag use for the fitted version of the final, two-parameter model ($P = 0.6$, $Q = 0.9$, solid line), the original top-5 model ($P = 0.6$, lower curved line), the null model (flat distribution), and the empirical data (topmost line).

about 3%, but if we bear in mind the scale of the empirical data (the most popular tag is used close to one million times), a relatively small percent improvement in RMSE corresponds to the substantial improvement in model fit that is evident in the plot. The modified model performs much better, with an RMSE of 1898.88, a 17% improvement over the null model. The final model is still far from closely matching the empirical data, in particular underestimating the prevalence of the most popular tags. The fit across the middle portion and tail of the distribution is quite good, however. Given that we had no expectations that our simple model would fully explain the data, the modest fit we find is in fact meaningful, showing what copying alone can achieve, and what is left to be explained by other mechanisms. As we have argued from the beginning, copying is of course not the only process driving tagging decisions, and the deviance between the model and empirical data at the head of the distribution is likely due to their being content-based non-copying mechanisms that lead to consensus as to how the most popular tags are applied.

CONCLUSIONS

In this paper we have presented the following:

- A novel methodology for crawling Last.fm that allows us to explore the temporal evolution of tags used within our dataset;
- A set of possible social copying heuristic models in a tagging environment;
- Methods for estimating from empirical data plausible parameter values for our heuristic models; and,
- A set of multi-agent models employing social copying heuristics that demonstrate the extent to which the patterns of tag popularity we see in our crawled data can be explained through copying behavior.

The final model we describe is of course a simplification of people's behavior when tagging for a variety of reasons, most notably in that it assumes that tagging decisions are in no way based upon the content of the item being tagged (which ostensibly should be the primary determinant of tagging decisions), instead being driven by simple copying and stochasticity. Tagging decisions result from the interplay of background knowledge, evaluation of the content, and social influence, and our model is a proof of concept of the importance of the last factor. We have demonstrated that extremely simple, psychologically plausible mechanisms are capable of generating data that account for a surprising amount of the empirical data as compared to a null model.

Tagging systems, whether for music on Last.fm, bookmarks on Delicious, or elsewhere, are thought to generate effective, crowd-sourced classifications of content. This work, however, suggests that a surprising proportion of tagging activity may in fact be driven by heuristic decision-making that manifests as replication of existing popular tags. This is not necessarily a criticism of tagging systems; their utility is ultimately determined by the extent which user-generated tags can be put to good use for search applications, content recommendation, and so on, and the scope of this work does not permit us to address the implications of copying for those applications. We do contend, however, that our results should inform subsequent research on tagging systems, as well as how such systems are designed in the future. Sensitivity to humans' propensity to utilize simple social copying heuristics is necessary for the design and study of tagging systems. Our analysis suggests, for instance, that Last.fm's choice to selectively suggest the top five most popular tags for an item may in fact be a driving force behind the pattern of tags we see on the site. More broadly, these results should arouse some skepticism around folksonomy practices, as they raise the possibility that the terminology that develops within such systems may be driven by simple copying mechanisms (as opposed to meaningful labeling of content) more so than is typically assumed.

This work has several limitations, a number of which we have already discussed in the course of developing the model. Further unrealistic assumptions of the model could of course be mentioned (e.g. assuming homogeneity of P and Q across agents, lack of variance in agents' vocabulary sizes), but these kinds of assumptions were necessary for the development of the simple model that interested us. One relatively small change that we would like to implement in a subsequent version of the model is a memory mechanism similar to that used in [5]. Having users "forget" tags that they have not seen recently may very well result in reinforcement of the popular tags that our model currently underestimates. Another possibility will be to incorporate preferential re-use of previously used tags on a user-by-user basis (assuming that the suggestion of "your tags", as seen in Figure 4, elicits a form of self-copying).

But beyond details of the model, it is important to note that the very process of making inferences about psychological processes from large-scale data such as that used here is an inherently difficult problem, with no guarantee that the proposed generative processes are in fact what drives the patterns in the empirical data. Further work must be done to gain a greater understanding of the individual psychological mechanisms driving decisions in tagging environments, the results of which would do much to inform future models of the kind we have described. These may take the form of laboratory studies, surveys, and other methods that can provide greater insight into individual behavior. We must also consider how heuristic use may differ in different tagging systems, either as a result of differences in the content being tagged or the information structure of the system itself. Would we, for example, find equally strong evidence for copying in a domain where there is greater agreement on the properties of the items being tagged? How do users' goals (e.g. attempting to accurately classify content, versus tagging so as to more easily retrieve it at a later time) in the system affect their decisions and likelihood of copying? Answering these and other questions will do much to further our understanding of tagging dynamics, and of the realism of the types of copying mechanisms we have proposed.

ACKNOWLEDGMENTS

We wish to thank Saurabh Malviya for extensive technical help is designing and implementing the the crawler and database. The first author was supported by an NSF IGERT fellowship while performing this research.

REFERENCES

1. Last.fm Radio Announcement, 2009.
2. Al-Khalifa, H. S., and Davis, H. C. Towards better understanding of folksonomic patterns. *Proceedings of the eighteenth conference on Hypertext and hypermedia* (2007), 163–166.
3. Berg, N., and Gigerenzer, G. As-if behavioral economics: Neoclassical economics in disguise. *History of Economic Ideas 18*, 1 (2010), 133–166.
4. Boyd, R., and Richerson, P. *The origin and evolution of cultures*. Oxford University Press, USA, 2005.
5. Cattuto, C., Loreto, V., and Pietronero, L. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences of the United States of America 104*, 5 (2007), 1461–1464.
6. Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., and Giles, L. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proceedings of the 2007 international ACM conference on supporting group work*, ACM Request Permissions (Nov. 2007).
7. Firan, C. S., Nejdl, W., and Paiu, R. The benefit of using tag-based profiles. In *Latin American Web Conference*, IEEE (2007), 32–41.
8. Fu, W., Kannampallil, T., and Kang, R. A semantic imitation model of social tag choices. In *International Conference on Computational Science and Engineering*, vol. 4, IEEE (2009), 66–73.
9. Gigerenzer, G., and Brighton, H. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science 1*, 1 (2009), 107–143.
10. Gigerenzer, G., and Gaissmaier, W. Heuristic decision making. *Annual review of psychology 62* (2011), 451–482.
11. Gigerenzer, G., and Todd, P. M. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.
12. Glushko, R. J., Maglio, P. P., Matlock, T., and Barsalou, L. W. Categorization in the wild. *Trends in Cognitive Sciences 12*, 4 (Apr. 2008), 129–135.
13. Golder, S. A., and Huberman, B. A. Usage patterns of collaborative tagging systems. *Journal of Information Science 32*, 2 (Apr. 2006), 198–208.
14. Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. Psychological research online: report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist 59*, 2 (2004), 105.
15. Marlow, C., Naaman, M., Boyd, D., and Davis, M. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, ACM (2006), 31–40.
16. McKay, C., and Fujinaga, I. Musical genre classification: Is it worth pursuing and how can it be improved. *Proc. of the 7th Int. Conf. on Music Information Retrieval* (2006), 101–106.
17. Nov, O., and Ye, C. Why do people tag? *Communications of the ACM 53*, 7 (July 2010), 128–131.
18. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. Folks in folksonomies: social link prediction from shared metadata. *Proceedings of the third ACM international conference on Web search and data mining* (2010), 271–280.
19. Simon, H. A. On a class of skew distribution functions. *Biometrika 42*, 3-4 (1955), 425–440.
20. Todd, P. M., and Gigerenzer, G. *Ecological Rationality: Intelligence in the World*. Oxford Univ Pr, Dec. 2011.
21. Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science 185*, 4157 (1974), 1124–1131.
22. Vander Wal, T. Folksonomy, 2007.
23. Vander Wal, T. Explaining and Showing Broad and Narrow Folksonomies, 2012.
24. Yule, G. U. A mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213*, 402-410 (1925), 21–87.