tagging, communities, vocabulary, evolution

Shilad Sen, Shyong K. (Tony) Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, John Riedl
GroupLens Research
University of Minnesota
4-192 EE/CS Building, 200 Union St. SE
Minneapolis, MN 55455 USA
{ssen,lam,arashid,cosley,dfrankow,oster,harper,riedl}@cs.umn.edu

ABSTRACT

A tagging community's vocabulary of tags forms the basis for social navigation and shared expression. We present a user-centric model of vocabulary evolution in tagging communities based on *community influence* and *personal tendency*. We evaluate our model in an emergent tagging system by introducing tagging features into the MovieLens recommender system. We explore four *tag selection algorithms* for displaying tags applied by other community members. We analyze the algorithms' effect on vocabulary evolution, tag utility, tag adoption, and user satisfaction.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*collaborative computing*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Design, Experimentation, Human Factors

Keywords

tagging, communities, vocabulary, evolution, social bookmarking

1. INTRODUCTION

Tagging sites have blossomed on the Internet since 2004 [11]. Tags are short free form labels used to describe items in a domain. They help people remember and organize information such as email (GMail), web sites (del.icio.us), photos (Flickr), blogs (Technorati), and research papers (CiteULike). Tags can also be a powerful tool for *social navigation* [14], helping people to share and discover new information contributed by other community members. In [13], Millen et al. suggest tags as a key reason current social bookmarking systems have enjoyed greater success than social bookmarking systems from the 1990s such as Fab [2], Knowledge Pump [8] and Pharos [3].

A critical characteristic of tagging systems that promote social navigation is their *vocabulary*, the set of tags used by members of the community. Instead of imposing controlled vocabularies or categories, tagging systems' vocabularies emerge organically from the tags chosen by individual members.

Although there is little peer-reviewed research on tagging, a number of bloggers and technology critics, such as Shirky[1], explore the value of tagging systems, including the relative merits of controlled versus evolved vocabularies. MacGregor and McCulloch collect and discuss these arguments in [12]. One valuable aspect of evolving vocabularies is that users invent personally meaningful tags, easing tasks such as organizing and re-finding items.

Individual invention, however, may not be best for the group as a whole. Social navigation may be more powerful in communities that share a common vocabulary. As an extreme example, people who speak different languages will find little value in each others' tags. User goals will also affect the value of others' tags. "Owned" is useful for remembering which books are in one's library, but not so helpful for others looking to discover new books to read. Even people trying to communicate the same idea often disagree how to describe it. Is your flavored carbonated drink a soda, a soft drink, a pop, a coke, or a tonic [7]?[2] The ESP Game [16] demonstrates how difficult it is for two people to agree on even simple descriptive words for a picture. In [10], Guy et al. suggest that correcting "sloppy tags" in a vocabulary can improve a tagging system's effectiveness.

In this paper we examine factors that influence both the way people choose tags, and ultimately, the degree to which community members share a vocabulary. Figure 1 shows three factors that are likely to influence how people apply tags: people's *personal tendency* to apply tags based on their past tagging behaviors, *community influence* of the tagging behavior of other members, and the *tag selection algorithm* that chooses which tags to display.

Personal tendency. People choose tags based on their *personal tendency*, their preferences and beliefs about the tags they apply. New users have an initial personal tendency based on their experiences with other tagging systems, their comfort with technology, their interests and knowledge [9],

CSCW'06, November 4–8, 2006, Banff, Alberta, Canada.

[1]http://shirky.com/writings/ontology_overrated.html
[2]See also http://www.popvssoda.com/.

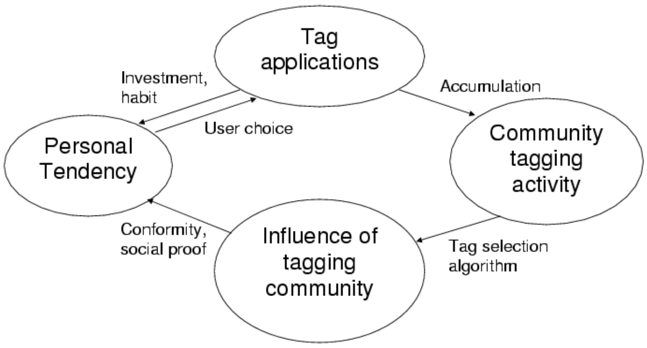

Figure 1: relationship between community influence and user tendency.

and so on. Personal tendency evolves as people interact with the tagging system.

Figure 1 indicates how users' own tagging behavior influences their future behavior through creating investment and forming habits. The tags one has applied are an investment in a personal ontology for organizing items. Changing ontologies midstream is costly. For someone who has labeled Pepsi, Coke, and Sprite as "pop", it would make little sense to label RC and Mountain Dew as "soda". Further, people are creatures of habit, prone to repeating behaviors they have performed frequently in the past [15]. Both habit and investment argue that people will tend to apply tags in the future much as they have applied them in the past.

There are also other factors that might influence a user's personal tendency to apply tags: they might lose or gain interest in the system, become more knowledgeable about tagged items, or become more or less favorably disposed to tagging as a way of organizing information. We do not model these factors in this paper.

Community influence. Figure 1 suggests that the community influences tag selection by changing a user's personal tendency. Golder and Huberman find that the relative proportions of tags applied to a given item in del.icio.us appears to stabilize over time [9]. They hypothesize that the set of people who bookmark an item stabilize on a set of terms in large part because people are influenced by the tagging behavior of other community members. Similarly, Cattuto examines whether the tags most recently applied to an item affect the user's tag application for the item [4].

The theory of social proof supports the idea that seeing tags influences behavior. Social proof states that people act in ways they observe others acting because they come to believe it is the correct way for people to act [5]. For example, Asch found that people conform to others' behavior even against the evidence of their own senses [1]. Cosley et al. found that a recommender system can induce conforming behavior, influencing people to rate movies in ways skewed toward a predicted rating the system displays, regardless of the prediction accuracy [6].

Research questions. Our work differs from Golder, Huberman, and Cattuto in an important way. Their analyses focus on how vocabulary emerges around items, i.e., how tags applied to an item affect future tags applied to that item. In contrast, we focus on factors affecting the way individual *users* apply tags across the domain of tagged items. Our first two research questions address the strength of the two factors we believe most affect the evolution of individuals' vocabularies:

RQ1: How strongly do investment and habit affect personal tagging behavior?

RQ2: How strongly does community influence affect personal tagging behavior?

To the extent that the community influences individual taggers, system designers have the power to shape the way the community's vocabulary evolves by choosing which tags to display. In the extreme case, a system might never show others' tags, thus eliminating community influence entirely. Even systems that do make others' tags visible will often have too many tags to practically display. Figure 1 shows the tag selection algorithm acts as a filter on the influence of the community. We ask two research questions about the effect of choosing tags to present:

RQ3: How does the tag selection algorithm influence the evolution of the community's vocabulary?

RQ4: How does the tag selection algorithm affect users' satisfaction with the system?

Finally, we examine whether communities converge on the *classes* of tags they use (e.g., factual versus subjective), rather than on individual tags. We explore whether these different classes of tags are more or less valuable to users of tagging systems:

RQ5: Do people find certain tag classes more or less useful for particular user tasks?

Our work differs from prior tag-related research in a number of ways. First, we focus on people rather than items. Second, we study a new tagging system rather than a relatively mature one. Third, we compare behavior across several variations of the same system rather than looking at a single example. Fourth, we study tagging as a secondary feature, rather than as the community's primary focus.

We believe that our perspective and questions will give fresh insight into the mechanisms that affect the evolution and utility of tagging communities. We use this insight to provide designers with tools and guidelines they can use to shape the behavior of their own systems.

The rest of this paper is organized as follows. In section 2 we discuss the design space of tagging systems and present the tagging system we built for users of the MovieLens recommender system. Section 3 presents our experimental manipulations and metrics within this tagging system. Sections 4, 5, and 6 address our first three research questions related to personal tendency, community influence, and tag selection algorithm. Section 7 covers research questions four and five, which explore the value of a vocabulary to the community. We conclude in section 8 with a discussion of our findings, limitations, design recommendations, and ideas for future research in tagging systems.

2. DESIGN OF TAGGING SYSTEMS

In this section, we briefly outline a design space of tagging systems and then describe the choices we made for the MovieLens tagging system.

2.1 Tagging Design Space

Numerous collaborative tagging systems have been deployed on the web in recent years. While all of them follow the same high-level principle of allowing people to apply free form textual labels (tags) to items in the system (i.e., web log entries, bookmarks, or pictures), there are several important choices that define a design space for tagging systems. We discuss each of these in this section.

Tag sharing. This dimension describes the extent to which a user's tags are shown to other users of the system. At one extreme, there are fully private systems such as Gmail, where a tag application is only visible to the person who applied it. These systems are not very interesting from the standpoint of enabling social navigation. At the other extreme are fully shared systems, where all tag applications are visible to all users. Systems in the middle of this scale offer a balance between privacy and openness, often allowing people to control who is able to view their tags. Similarly, some designers may consider it undesirable to make others' tag applications readily visible. The information may still be public, but require additional effort to view.

Tag selection. Systems that allow tag sharing may not be able to display every tag applied to every item because of the sheer number of tag applications that may exist (for example, del.icio.us has over 2 million applications [11]). It may be necessary to choose a small number of tags to display. Systems may opt to display meta-information related to a tag, such as the number of times a tag has been applied or the number of users who have applied it. The method by which a system selects and displays tags serves as a user's lens into the tagging community, and thus directly impacts the community's influence on a user's choice of tags.

Item ownership. Who owns an item likely affects tagging behavior. In some systems, people apply tags to items they created. On Flickr, for example, people most often apply tags to pictures they posted themselves. On the other hand, books, movies, and albums in systems such as Listal and Amazon, are not created by individual members of the community. As a result, people apply tags to works that were created by others. This distinction has implications on the value and benefits of a collaborative tagging system that, while interesting, are beyond the scope of this paper.

Tag scope. This dimension describes whether tag applications belong to individual users or shared by the entire community. This characterization of collaborative tagging systems was originally proposed by Vander Wal[3]. There are two alternatives:

- Broad. A broad system has the property that any number of people may apply their own personal set of tags to an item (e.g. individual users own tag applications). Each tag application is as a $< user, item, tag >$ triple. An example of this type of system is del.icio.us.
- Narrow. A narrow system represents a single shared set of tags for each item (e.g. the community owns tagging application). The tags are stored as $< item, tag >$ pairs and can be collectively maintained by the community. In practice, narrow systems are often also systems in which users create and own the items they tag. Examples of such systems are Flickr and Technorati.

Other Dimensions. Tagging systems pose a number of other design decisions that can impact a system's usability.

[3]http://personalinfocloud.com/2005/02/explaining_and_.html

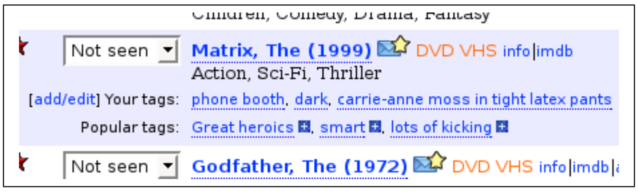

Figure 2: MovieLens movie list with tags.

These include the choice of tag delimiter, whether a tag can be one or multiple words, and how to normalize tags across factors like punctuation, letter case, and white space. The tagging systems in use today vary widely in how they approach these design choices, and there have been no studies regarding which choices are better than others. Although we do not study them here, each of these has the power to affect a community's vocabulary.

2.2 The MovieLens Tagging System

As a platform for our experiments, we incorporated tagging features into the MovieLens movie recommendation system[4]. The choice of movies as an item domain constrained some design decisions. The item creator is almost always *not* the tagger, so we opted to design a broad system where people apply their own set of tags to each movie. In our experiments, we manipulate the tag sharing and tag selection dimensions, creating several parallel but independent tagging communities. We describe our experimental manipulation in more detail in section 3.

We implemented tagging features using *AJAX* based functionality provided by the script.aculo.us[5] library to allow for lightweight interaction between users and the tagging system. Below we describe how the system displayed tags, allowed users to apply tags, and helped users navigate using tags.

Because the tagging system was new and we wanted to encourage usage and foster awareness, we chose to display tags throughout the interface. All MovieLens users who logged in during the experiment saw a tagging introduction page that explained how to tag movies and invited them to tag three movies. The MovieLens home page displayed a list of the ten most recently applied tags by members of the user's experimental group. On lists of movies returned by search results and recommendations requests, the system displayed up to three tags applied by the community and three tags applied by the user. A full list of tags applied to a movie was available on a movie's details page. Figures 2 and Figure 3 show how tags were displayed in movie lists and on the details page, respectively.

Figure 4 shows the interface for applying tags, which was available on both movie lists and movie details pages. To apply a tag, a user clicks the "add tag" link, which opens a text box. MovieLens dynamically generates an *auto-completion* list of tags matching what has been typed thus far. The user may select a tag from this list, or she may ignore it and continue typing her tags. Users could also quickly "steal" tags applied by other community members by clicking the plus icon next to a tag.

Users could navigate through tag space in three primary ways. First, all tags that appear in the interface are clickable

[4]http://www.movielens.org

[5]http://script.aculo.us/

Figure 3: Movie details page tag display.

Figure 4: Adding tags with auto-complete.

links that display a list of movies that have been tagged with the clicked tag. Second, a tag search box with the auto-completion feature is provided to facilitate quick access to lists of movies that have been tagged with a particular tag. Finally, we added a "Your Tags" page that lists all the tags that a user has applied along with a sampling of movies that each tag was applied to.

3. EXPERIMENTAL SETUP

Each user was provided with the common tagging elements described in section 2.2. We now describe the experimental manipulations we performed to gain insight into our research questions.

We randomly assigned users who logged in to MovieLens during the experiment to one of four experimental groups. Each group's tags were maintained independently (i.e. members of one group could not see another group's tags).

Each group used a different tag selection algorithm that chose which tags to display, if any, that had been applied by other members of their group. We used these algorithms to manipulate the dimensions of tag sharing and tag visibility.

The **unshared** group was not shown any community tags, corresponding to a private system where no tags are shared between members.

The **shared** group saw tags applied by other members of their group to a given movie. If there were more tags available than a widget supported (i.e. three tags on the movie list, seven tags on the auto-complete list), the system *randomly selected* which tags to display.

The **shared-pop** group interface was similar to that of the shared group. However, when there were more tags available than a widget supported, the system displayed the *most popular* tags, i.e., those applied by the greatest number of people. Both the details page and the auto-complete drop-down displayed the number of times a tag was applied in parentheses. We expected this group to exhibit increased community influence compared to the shared group because, since everyone would see the most popular items, people would tend to share the same view of the community's behavior.

Table 1: Overall tag usage statistics by experimental group. Note that the tags column overall total is smaller than the sum of the groups, because two groups might independently use the same tag.

group	users	taggers	tags	tag applications
unshared	830	108	601	1,546
shared	832	162	809	1,685
shared-pop	877	154	1,697	4,535
shared-rec	827	211	1,007	3,677
overall	3,366	635	3,263	11,443

The **shared-rec group** interface used a *recommendation algorithm* to choose which tags to display for particular movies. When displaying tags for a target movie, the system selected the tags most commonly applied to both the target movie and to the most similar movies to the target movie. Similarity between a pair of movies was defined as the cosine similarity of the ratings provided by MovieLens users. Note that this means that a tag that was never actually applied to a movie could appear as being associated with that movie–and further, that tags could be displayed for a movie that had never had a tag applied to it.

We collected usage data from January 12, 2006 through February 13, 2006. Table 1 lists basic usage statistics overall and by experimental group. During the experiment, 3,366 users logged into MovieLens, 635 of whom applied at least one tag. A total of 3,263 tags were used across 11,443 tag applications. (A *tag* is a particular word or phrase used in a tagging system. A *tag application* is when a user applies a particular tag to a given item.)

3.1 Metrics

As shown in Table1, basic usage metrics differed widely between experimental groups. However, these differences are not statistically significant due to effects from "power taggers." Most tag applications are generated by relatively few users, approximating a power law distribution ($y = 15547x^{-1.4491}$, $R^2 = 0.9706$). The mean number of tag applications per user was about 18, but the median was three. The most prolific user applied 1,521 tags, while 25 users applied 100 or more. Because of these skewed distributions, differences such as the number of tags applied per group, are not statistically significant.

Further, most of our research questions are not about differences in quantity, but rather, about how the tags people apply and view influence their future decisions on which tags to apply. In most cases, we study this influence at the level of categories of tags, which we call *tag classes*. Golder et al. present seven detailed classes of tags[9]. We collapse Golder's seven classes into three more general classes that are related to specific user tasks that tags could support in the MovieLens community. We list short descriptions of Golder's tag classes that were folded into each of our tag classes in parentheses.

1. **Factual tags** identify "facts" about a movie such as

Table 2: Ten most popular tags in each tag class, and how often each tag was applied.

Factual	Subjective	Personal
action (134)	classic (235)	bibliothek (253)
drama (104)	chick flick (61)	in netflix queue (177)
disney (86)	funny (60)	settled (148)
comedy (86)	overrated (54)	dvd (122)
teen (64)	girlie movie (51)	my dvds (110)
james bond (62)	quirky (39)	netflixq (87)
super-hero (57)	special (29)	get (58)
japan (56)	funny as hell (25)	ohsoso (48)
true story (55)	funniest movies (23)	buy (35)
crime (54)	must see! (22)	(s)vcd (32)

people, places, or concepts. We operationally define factual tags as tags that most people would agree apply to a given movie. Factual tags help to describe movies and also help to find related movies (Golder's classes: item topics, kinds of item, category refinements).

2. **Subjective tags** express user opinions related to a movie. They can be used to help evaluate a movie recommendation (item qualities).

3. **Personal tags** have an intended audience of the tag applier themselves. They are most often used to organize a user's movies (item ownership, self-reference, task organization).

In order to assign tags to classes, we manually coded the 3,263 distinct tags into one of the three classes. If tags were incomprehensible, or did not fit in a class, the tag was coded as class **other**. Each tag was coded by two people. Coders agreed on 87% of tags. When coders differed, the coders discussed the tag and reached a consensus.

The final distribution of tags across tag classes was 63% factual, 29% subjective, 3% personal, and 5% other. Unless mentioned otherwise, we ignore the class "other" when performing tag-class-based analyses. For each tag class, Table 2 shows the ten tags of that class applied most often, across groups.

We often define influence in terms of the cosine similarity between tag class distributions. By tag class distribution we mean the proportion across these three tag classes of a group of tags, tag applications, or tag views. Cosine similarity is useful because it normalizes for the size of the distributions.

For example, suppose we wish to talk about the community influence on a specific tag application by a user. We can treat the tags the user saw before applying that tag as a distribution across the three tag classes. Suppose that 62% of the tag views were of factual tags, 25% were subjective, and 13% were personal. Likewise, we can look at the class of the tag applied and think of it as a tag class distribution. If the tag is subjective, the distribution would be 0% factual, 100% subjective, and 0% personal. We can encode these as vectors: $x = [0, 1, 0]$ $y = [0.62, 0.35, 0.13]$. We then compute cosine similarity of x and y as $\frac{x \cdot y}{\|x\|\|y\|}$, or $\frac{0*0.62+1*0.25+0*0.13}{\sqrt{0^2+1^2+0^2}\sqrt{0.62^2+0.25^2+0.13^2}} \approx 0.37$. If the tag applied had been a factual tag, then the similarity would have been about 0.91.

One disadvantage of using cosine similarity it that it can be hard to understand how to interpret differences between two similarity values. As a frame of reference, the similarity between the uniform tag class distribution $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ and any

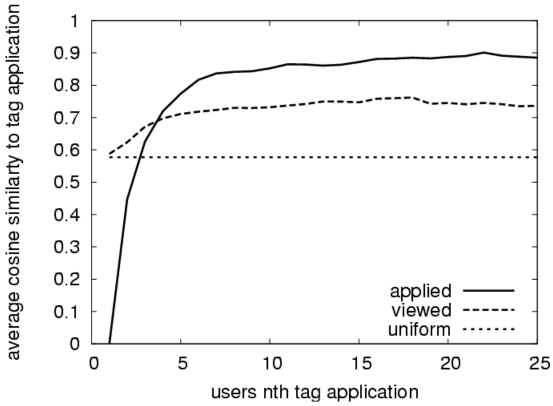

Figure 5: Similarity of tag class of the nth tag applied by a user to tag class distributions of other tags applied by the user before the nth tag (applied), of tags viewed by the user (viewed), and of the uniform tag class distribution (uniform). Both habit/investment and tags viewed appear to influence the class of applied tags.

tag application is $1/\sqrt{3} \approx 0.58$.

Finally, upon completion of the tagging experiment we conducted a survey of all MovieLens users. A detailed description of the survey is presented in section 7. We include results from the survey as they are relevant.

4. PERSONAL TENDENCY

We are now ready to explore our first research question:

RQ1: How strongly do investment and habit affect personal tagging behavior?

In the model in Figure 1, a user's personal tendency determines the types of tags they apply. In this section, we examine how strongly investment and habit affect user choices. We measure the strength of this association by comparing the tags a user has applied in the past to the tags they apply in the future.

The solid line in Figure 5 shows the average cosine similarity between the tag class distribution across the users first $n - 1$ tags and the tag class of their nth tag. We smoothed lines exponentially with weight 0.7. The horizontal line graphically displays the similarity of any tag application to the uniform tag class distribution. We will discuss the third line in section 5.

Once a user has applied three or more tags, the average cosine similarity for the nth tag application is more than 0.83. Moreover, similarity of a tag application to the user's past tags continues to rise as users add more tags.

As well as reusing tag classes, users also reuse individual tags from their vocabulary. Figure 6 shows that as users apply more tags, the chance that an applied tag will be new for them drops. In total, 51% of all tag applications are tags that the tagger has previously applied (experimental groups are grouped together to increase statistical power). As a baseline, we determined through simulation that users randomly selecting tags without any tendency to repeat tags would have about 27% tag reuse.

Clearly habit and investment influence tagging behavior. We wanted to determine if these factors are entirely respon-

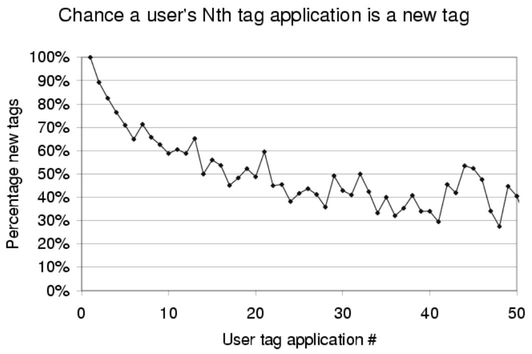

Figure 6: Chance a user's Nth tag application is a new tag.

sible for user behavior. If habit and investment were the only factors determining personal tendency, the law of large numbers implies that tag class distributions for the four experimental groups should converge as more users enter the system. Table 3 shows that this is not the case; at the end of the experiment, the tag class distribution for the four groups was very different. In Section 6, we showed that the ending tag class distribution in groups vary significantly, both from a user-centric, and tag-centric viewpoint.

We should note that our study examines only a portion of the feedback loop between personal tendency and tag applications. It may be possible to directly measure personal tendency through more invasive iterative surveying techniques. We leave such studies for future work.

We conclude by restating our findings related to a user's personal tendency:

1. Habit and investment influence user's tag applications.
2. Habit and investment influence grows stronger as users apply more tags.
3. Habit and investment cannot be the only factors that contribute to vocabulary evolution.

5. INFLUENCE OF TAG VIEWS

We now turn our attention to our second research question:

RQ2: How does the tagging community influence personal vocabulary?

Our model from Figure 1 shows that viewing community tags indirectly impacts a user's tag applications by changing a user's personal tendency. We measure this by comparing the tags a users saw before each of their tag applications. Note that while [4] analyzes community influence in del.icio.us from an item-centric (i.e. web-page-centric) point of view, we consider a user-centric analysis. Because we control our tagging system, we can record the tags each user is shown while using the system (this is not possible with external analyses of del.icio.us).

During our experiments, we recorded all tags a user had viewed on the home page, search results, and "your tags" pages. Additionally, we recorded all tags displayed in autocomplete lists.[6]

[6]Due to an implementation error, we failed to log tag views on movie details pages. However, we estimate that movie details pages account for less than 5% of total tag views.

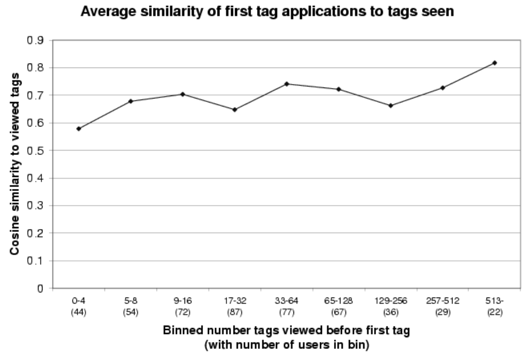

Figure 7: Average cosine similarity of the class of a user's first tag application to class distribution of tags viewed before the user applied the first tag. Results are grouped by number of tags viewed before the first application. Bins are logarithmic in an effort to put roughly an equal number of people in each bin.

As in our analysis of personal tendency, we measure the cosine similarity between a tag application's tag class and the tag class distribution the user *has seen* up until that point. This average similarity over user's nth tag applications is shown by the dotted curved line in Figure 5. Although the similarity between tag views and tag applications is weaker than the similarity between a user's personal tendency and their tag applications, it is stronger than the uniform tag distribution baseline.

We also examine how the number of tags viewed before a user's first tag application influences the choice of tags to apply. Figure 7 shows the average cosine similarity between a user's first tag class and the class distribution they saw before applying their first tag. A gentle upward trend is apparent; users who view more tags before their first tag application are more likely to have their first tag influenced by the community.

Based on our analysis, community influence plays an important role in vocabulary. In particular:

1. Community influence affects a user's personal vocabulary.
2. Community influence on a user's first tag is stronger for users who have seen more tags.

6. CHOOSING TAGS TO DISPLAY

We have shown that users are influenced by the community tags that they see. In our tagging model (Figure 1), the algorithm for choosing tags to display serves as the user's lens into community tagging activity. We explore this relationship in our third research question:

RQ3: How does the tag selection algorithm affect a user's personal vocabulary?

We examine algorithm influence using two approaches. First we explore the relationship between tag selection algorithms and resulting tag class distributions. Second, we examine the distribution of the actual tag phrases themselves.

Table 3: Final tag application class distribution by experimental group. The dominant tag class for each group is bolded. (Each row sums to 100%.)

Group	Subjective	Factual	Personal
Unshared	24%	38%	**39%**
Shared	**60%**	37%	3%
Shared-pop	9%	**82%**	9%
Shared-rec	20%	**67%**	12%

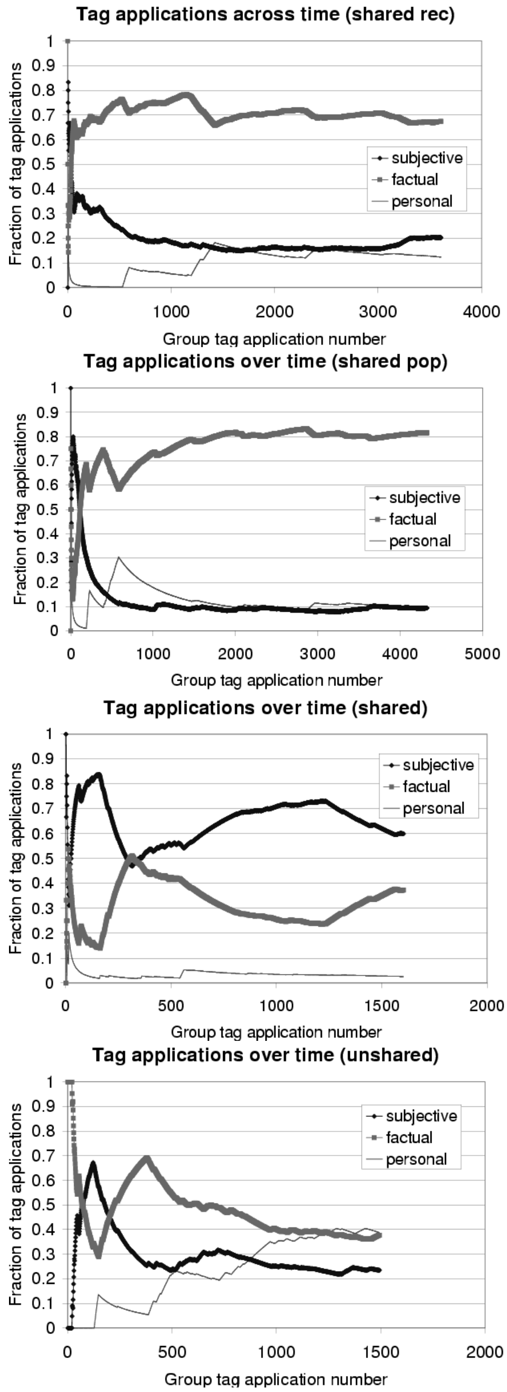

Figure 8: Tag class distribution over time for each experimental group: shared rec, shared pop, shared, and unshared, respectively. The first two appear to converge more strongly, to factual tags.

6.1 Tag Class Distributions

We begin by looking at how tag display algorithms influence the distribution of tag classes (subjective, factual, and personal). We measure this influence by comparing the tag classes distributions between experimental groups, each of which had a different display algorithm. We consider both the final distribution, and the distribution as it varied across time during the experiment.

The final distributions have very large differences across our experimental groups. Table 3 shows the shared-rec and shared-pop groups are dominated by factual tags, the shared group by subjective tags, and the unshared group is divided more evenly.

We compared the proportion of tag applications in each tag class across groups, and found the difference between groups to be significant ($\chi^2(6, N = 11073) = 2730$, $p < .001$). We wanted to know if particular power users were skewing tag-class distribution, so we also compared user-centric distributions. We compared the dominant tag class (i.e. the class with the most tag applications by the user) for all users with five or more tags across the experimental groups. The proportions of dominant tag classes between groups is again significant ($\chi^2(6, N = 168) = 123$, $p < .001$).

Differences in tag class convergence may be due to our experimental manipulation, community influence, or evolved personal tendency (though probably not initial personal tendency, since we randomly assigned users to groups). There are multiple possible explanations.

In addition to final tag class distributions, we looked at whether tag class distributions converged quickly, slowly, or not at all by examining per-group plots of tag class distribution over time (Figure 8). In all graphs, the X axis is the tag application number for the group, and the Y axis is the fraction of tag applications of the given class. The tag application number represents the number of tag applications by users in the experimental group since the beginning of the experiment (this can be roughly thought of as time).

Visually, it looks as if the shared-pop and shared-rec groups converged. In each, factual tags rapidly became and remained the dominant class by a large margin. By contrast, the shared and unshared groups have less visual evidence of convergence. In the shared group, subjective tags were often the dominant class, although there were more factual tags during tag applications 300-331, and there is more drifting in general. In the unshared group personal tags rise continuously until they become the dominant class at the end.

Likely the shared-rec and shared-pop display algorithms both favor tags applied by many different people, and those tend to be factual in nature (80% of tags applied by 5 or more people are factual). Perhaps this contributes to the greater number of factual tags in these groups. It is also perhaps interesting that these groups were the more convergent ones, and also had greater numbers of tags (see Table 1). This suggests that perhaps the interface strengthened convergence.

Finally, we note that these graphs represent a tag-centric view of tag class distribution, and power taggers may disproportionately influence these graphs. We also considered graphs that use a user-centric view of tag class distribution, where every user gets equal weight. We do not include these graphs due to lack of space. Personal tag class proportions in the user-centric graphs are tempered due to the fact that, on average, personal tags are applied many more times (14.9) than factual (3.5) or subjective (2.6) tags. For example, while personal tags are most common in tag-centric view of the unshared group, they are least common in the user-centric view. Other graphs show similar behaviors.

In summary, experimental groups exhibit different final tag class distributions and rates of tag class convergence. While we cannot definitively attribute these differences to tag selection algorithms, we hypothesize that the shared-rec and shared-pop algorithms may encourage vocabularies to converge on factual tags, while the unshared selection algorithm encourages personal tag use by eliminating any motivation to create tags that are good for the community.

6.2 Tag Reuse

In additional to looking at tag class convergence, we would like to know if the convergence of actual tag phrases differs across groups. As a measure of tag convergence, we look at the average number of users who apply a tag. We chose this metric because it is more robust to power taggers than, for example, average applications per tag. Since every tag is applied by at least one user, the minimum value for this metric is 1.0. As a baseline, the unshared group averages 1.10 users per tag. The shared group follows with 1.27 users per tag. Next, the shared-pop group averages 1.31 users per tag. Finally, the shared-rec group, which exposed users to the largest number of tags during their use of MovieLens, yields 1.73 users per tag. Clearly the user interface has some effect on tag convergence.

Figure 9 breaks down origination of user tags (the first application of a tag by a particular user) based on the original creator of the tag. If the user is the first person to use a tag, we say they *invented* the tag. If somebody else in the experimental group used the tag, but the user has not seen the tag, we say that the user *reinvented* the tag. Finally, if the user saw the tag before applying it for the first time, we say the user *borrowed* the tag. For example, because the unshared group doesn't see other users it has no borrowed tags, but does have invented and reinvented tags. The origination results match our tag reuse metric: the shared-rec group uses more borrowed tags while the unshared group invents and reinvents more tags.

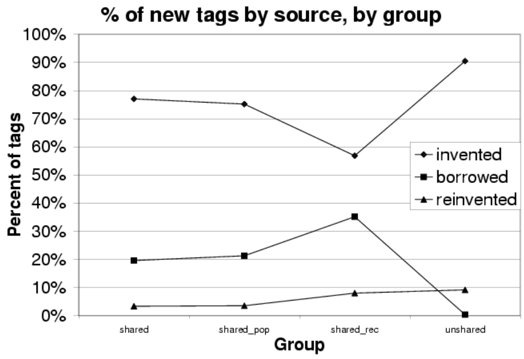

Figure 9: Origination of user tags, per experimental group. Re-invented tags are tags that are not seen by the user prior to creation, but were previously applied by other group members. The shared_rec group has more borrowed tags, and the unshared group has more invented tags.

7. VALUE OF TAGS TO THE COMMUNITY

In the previous three sections, we analyzed factors that contribute to vocabulary evolution in tagging communities. We now turn our attention towards exploring the value of a vocabulary to the community. We frame our exploration using our last two research questions:

RQ4: How does the tag selection algorithm affect users' satisfaction with the tagging system?

RQ5: Do people find certain tag classes more or less useful for particular user tasks?

We base our answers to these questions on a survey that we administered to MovieLens users. Before turning to the two research questions, we describe this survey in detail.

7.1 Survey Description

At the conclusion of our tagging study, we emailed 1,596 MovieLens users and asked them to complete a survey about their tagging experiences. The selected users comprised all taggers and non-taggers who had seen at least one tag, opted in to receive MovieLens emails, and had not received another MovieLens email in the past three months.

We divided the tagging survey into two main sections. In the first section, we asked general questions about a user's tagging experience, such as why they created tags and how much they liked the MovieLens tagging features. 365 users (23% of emailed users) completed this section of the survey.

In the second section, we asked users about specific tag applications. For each tag application, the user was presented with the tag, the movie it was applied to, and a series of questions about the application. All users were asked about five tag applications created by other people:

- One tag application from each of the unshared, shared, and shared-pop groups.
- One actual tag application from the shared-rec group, where a user actually applied the tag to the movie.
- One inferred tag application from the shared-rec group where the shared-rec algorithm inferred a tag that had not ever been applied to the movie.

In addition to selecting one tag from each of the five groups, we ensured that the five applications spanned the three tag classes. Finally, for those users who were taggers, we asked questions about up to four of their own tag applications, again including at least one tag from each of the three tag classes.

327 users answered questions about at least five tags. After users answered their first set of questions about tag applications, they were given the option to continue answering questions about tag applications. 173 users answered ten or more sets of questions. In total, users answered questions about 3960 tag applications.

7.2 Mapping Tag Classes to User Tasks

We now skip forward to our fifth research question, which relates to some of our high-level survey results:

RQ5: Do people find that different tag classes are more and less useful for supporting various user tasks?

task	factual	subjective	personal
self-expression	39%/NA	**80%/NA**	20%/NA
organizing	62%/NA	61%/NA	**87%/NA**
learning	**60%/49%**	46%/36%	10%/7%
finding	**59%/48%**	35%/27%	12%/8%
decision support	41%/33%	**45%/35%**	13%/8%
overall	**56%/44%**	44%/31%	13%/9%

Figure 10: Usefulness of tag classes for user tasks. Percentages list agreement of responses by (only users who applied a at least one tag / all users). For each user task, the most highly rated tag class is bolded. The bottom row lists lists overall user satisfaction for each tag class.

Hammond et al. suggest that reasons for tagging are generally application-specific [11]. Based on prior experience with MovieLens users, we selected five user tasks related to tagging. In the tagging survey, we asked users whether they agree that tags are helpful for each of the user tasks. Below we list the user tasks as they were described in our survey, and show the percentage of taggers and the percentage of overall users that agreed that tags were helpful for the task.

1. *Self-expression* - I think the tagging features help me express my opinions. (50% of taggers agree / 30% of all users agree)
2. *Organizing* - I think the tagging features help me organize my movies. (44% / 23%)
3. *Learning* - I think the tagging features help me know more about the movies to which they are applied. (37% / 27%)
4. *Finding* - I think the tagging features help me find movies I am looking for. (27% / 19%)
5. *Decision support* - I think the tagging features help me decide whether or not to watch the movie to which they are applied. (21% / 14%)

In addition to asking whether tagging supports the five user tasks in general, we asked whether each tag application supported the five tasks. The questions about learning, finding, decision support, and overall usefulness were asked about both tags the user applied and tags the users did not apply. We only asked questions about self-expression and organizing for tags a user had actually applied, since these tasks are most relevant to the tagger herself. Figure 10 details our results per tag class.

Figure 10 indicates that different tag classes are useful for different tasks. Factual tags are useful for learning about and finding movies. Subjective tags are useful for self-expression. Personal tags are useful for organization. Both factual and subjective tags are moderately useful in decision support.

The final row in Figure 10 gives results per tag class for overall user satisfaction with the tag. Users generally prefer factual tags and dislike personal tags. Additionally, users said they would prefer *not* to see 67% of personal tags they were asked about (compare this to 27% for factual tags and 37% for subjective tags).

7.3 Differences by Choice of Tag Display

In section 6 we demonstrated that different tag display algorithms appear to lead to different tag class distributions for a community's vocabulary. We return to our fourth research question, which examines user satisfaction resulting from different tag display algorithms:

RQ4: How does the algorithm for choosing tags to display affect user satisfaction with the tagging system?

Users complained that the shared-rec tag selection algorithm resulted in an overly invasive tagging interface. Our choice of user interface may be partly to blame. In order to encourage tagging, we designed the tag input box to automatically pop open on movies users had rated. Furthermore, the auto-completion list automatically appeared, suggesting tags inferred by the algorithm. MovieLens users did not like these design decisions, perhaps because they interfered with other common user tasks such as rating movies.

Secondly, users did not like the tag inference algorithm itself. While users said they would like to see 36.5% of the actual tag applications in the shared-rec group, they only wanted to see 18.0% of the tag applications that were inferred using our algorithm. Users were confused by some of the inferred tags, and understandably so, because they were not informed that the displayed tags may not actually have been applied to the movie. For example, one user comments about the tag "small town" which was inferred for the movie "Swiss Family Robinson":

> *I'm confused - I thought it was about people on a deserted island???*

In addition, the algorithm led to a far higher number of tags being displayed in the interface. It generated 5,855,393 tag views, compared to 710,313 for shared-pop, 379,313 for shared, and 12,495 for unshared. 64% of surveyed users in the shared-rec said they would like to be able to hide tagging features - more than any other experimental group. However, it appears that the pervasive presence of tags had some effect in converting users to taggers. 25% of the users in the shared-rec group applied at least one tag, compared to 19%, 17%, and 13% of users in the shared, shared-pop, and unshared groups respectively. A chi square analysis indicates that this difference is significant $\chi^2(3, N = 3,357) = 43.7, p < 0.001$.

In contrast, the unshared group had a relatively unobtrusive tag display. While 36.3% of users in the shared-rec group disliked tags overall, only 13% of users in the unshared group disliked the tagging features (along with 25% of users in the shared and shared-pop groups).

8. DISCUSSION

Our results point towards several guidelines for designers of tagging systems. First, some popular systems such as flickr do not support the notion of private tags. Hammond et al. argue in support of solely public tags [11]:

> *Social bookmarking tools, as with the Web at large, usually pay users back many times over in utility for whatever privacy they may have surrendered.*

While users cite organizing their movies as one of the most important reasons for creating tags, they overwhelmingly dislike seeing others' personal tags:

> *There should potentially be a private/public tag option. I don't really need to see how many people have a movie on their NetFlix list.*

Therefore, we suggest that designers create affordances for hiding a user's personal tags from other users. In some tagging systems, other design dimensions may reduce the need for explicit personal tags. For instance, in del.icio.us, tags are public, but the most popular tags are chosen for display, so personal tags are unlikely to appear. Moreover, in common uses of del.icio.us, such as viewing one's own saved pages, or viewing an acquaintance's saved pages, tags from the rest of the community do not appear at all.

Recall that our user feedback suggests that tagging features should not be overly intrusive. In MovieLens, there are a subset of users who do not value tags, and would prefer to hide them entirely:

> *Tagging is very heavy on the movielens user interface, and it would be good to be able to hide it. I can see their use when searching for movies, but most of the time I just look up a known movie to see its expected score...*

One key difference between MovieLens and other tagging systems is that MovieLens is not primarily a tagging system. MovieLens exists to make recommendations, and users sometimes found the tagging features interfered with their primary goals. MovieLens has existed for over eight years, and adding new highly-visible features such as tagging was not welcomed by some long-standing users. Indeed, our survey results show that new MovieLens users were significantly less likely to want to hide the tagging features than users who existed before the features were introduced ($\chi^2(1, N = 248) = 7.6$, $p < .01$). As tagging is increasingly added to existing systems, designers should consider the full range of use cases of their system.

One reason some users did not tag is because they could not think of any tags. This problem was cited by fully 68% of non-taggers in the unshared group, but only 40% of non-taggers in other groups. Offering tag suggestions is one way for designers to encourage more people to use tags.

It may be desirable to "steer" a user community toward certain types of tags that are beneficial for the system or its users in some way. To this end, designers may wish to take advantage of our finding that pre-existing tags affect future tagging behavior. For instance, a new tagging system might be seeded by its designers with a large set of tags of the preferred type. Our results suggest that users would tend to follow the pre-seeded tag distribution. At the extreme, a site owner could seed a tag system with a nearly complete ontology of useful tags.

Finally we point out several areas for further research. First, differences in users' attitudes towards different tag classes suggest that it may be valuable for tagging systems to classify tags. Researchers should investigate both automatic techniques to infer tag classes, and user interface designs that support manual classification of tags by the community.

Second, deriving relationships and structure from the tags that are applied may provide additional guidance in how to display tags in ways that aid search and navigation. Perhaps automated tools can be developed to help guide the emerging ontology. For instance, users could be steered to prefer the tag "soda" rather than "pop", if soda is being used heavily by other users. Perhaps the system could conflate the terms transparently, so that users could use either term effectively.

Third, studying the intrinsic information value of a particular tag may be useful to inform a tag display algorithm about which tags to choose in maximizing the value to users. For instance, the density of tag applications across objects may provide information about their value to users. A tag that is applied to a very large proportion of items may be too general to be useful, while a tag that is applied very few times may be useless due to its obscurity.

9. ACKNOWLEDGEMENTS

We would like to thank Sara Drenner for her help in implementing the MovieLens tagging system, the rest of Grouplens for their discussion and input, and most importantly our MovieLens users for their passionate tagging participation and feedback.

This work is funded in part by National Science Foundation, grants IIS 03-24851 and IIS 05-34420.

10. REFERENCES

[1] S. E. Asch. Effects of group pressure upon the modification and distortion of judgements. *Groups, Leadership, and Men*, pages 177–190, 1951.

[2] M. Balabanovic and Y. Shoham. Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[3] V. Bouthors and O. Dedieu. Pharos, a collaborative infrastructure for web knowledge sharing. In *ECDL*, pages 215–233, London, UK, 1999. Springer-Verlag.

[4] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics in online social communities. In *The European Physical Journal C (accepted)*. Springer-Verlag, 2006.

[5] R. B. Cialdini. *Influence Science and Practice*. Allyn and Bacon, MA, USA, 2001.

[6] D. Cosley, S. K. Lam, I. Albert, J. Konstan, and J. Riedl. Is seeing believing? How recommender system interfaces affect users' opinions. In *CHI*, 2003.

[7] D. Coye. A linguistic survey of college freshmen: Keeping up with standard American english. *American Speech*, 69(3):260–284, 1994.

[8] N. Glance, D. Arregui, and M. Dardenne. Knowledge pump: Supporting the flow and use of knowledge. In *Information Technology for Knowledge Management*. Springer-Verlag, 1998.

[9] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science (accepted)*, 2006.

[10] M. Guy and E. Tonkin. Tidying up Tags? *D-Lib Magazine*, 12(1):1082–9873, 2006.

[11] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools : A general review. *D-Lib Magazine*, 11(4), April 2005.

[12] G. MacGregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library View (accepted)*, 55(5), 2006.

[13] D. Millen, J. Feinberg, and B. Kerr. Social bookmarking in the enterprise. *Queue*, 3(9):28–35, 2005.

[14] A. J. Munro, K.Höök, and D. Benyon. *Social navigation of information space*. Springer, 1999.

[15] J. A. Ouellette and W. Wood. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Review*, 127:54–74, 1998.

[16] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of SIGCHI*, pages 319–326, 2004.