Music Tagging and Listening:
Testing the Memory Cue Hypothesis in a Collaborative Tagging System

Jared Lorince and Peter M. Todd

Dept. of Psychological & Brain Sciences, Cognitive Science Program

Indiana University

Abstract

As an example of exploring human memory cue use in an ecologically valid context, we present ongoing work to examine the "memory cue hypothesis" in collaborative tagging. In collaborative tagging systems, which allow users to assign freeform textual labels to digital resources, it is generally assumed that tags function as memory cues that facilitate future retrieval of the resources to which they are assigned. There is, however, little emprirical evidence demonstrating that this is in fact the case. Employing large-scale music listening and tagging data from the social music website Last.fm as a case study, we present a set of time series and information theoretic analytic methods we are using to explore how patterns of content tagging and interaction support or refute the hypothesis that tags function as retreival cues. Early results are, on average, consistent with the hypothesis. There is an immediate practical application of this work to those working with collaborative tagging systems (are user motivations what we think they are?), but our work also comprises contributions of interest to the cognitive science community: First, we are expanding our understanding of how people generate and use memory cues "in the wild". Second, we are enriching the "toolbox" available to cognitive scientists for studying cognition using large-scale, ecologically valid data that is latent in the logged activity of web users.

Music Tagging and Listening:
Testing the Memory Cue Hypothesis in a Collaborative Tagging System

Introduction

Humans possess a unique capacity to manipulate the environment in the pursuit of goals. These goals can be physical (building shelter, creating tools, etc.), but also informational, such as when we create markers to point the way along a path or leave a note to ourselves as a reminder to pick up eggs from the market. In the informational case, the creations of reminders or pointers in the environment functions as a kind of cognitive offloading, enriching our modes of interaction with the environment while requiring reduced internal management of information.

The proliferation of web-based technologies has massively increased the number of opportunities we have for such offloading, the variety of ways we can go about it, and the need to do so (if we are to keep up with the ever expanding mass of information available online). This is particularly true with respect to the various "Web 2.0" technologies that have recently gained popularity. As jargony and imprecise a term it may be, "Web 2.0" entails a variety of technologies of interest to cognitive scientists, including the sort of informational environment manipulations that interest us here. More than anything else, the "upgrade" from Web 1.0 that has occurred over the past 10–15 years has seen the evolution of the average web user from passive information consumer to active information producer, using web tools as a means of interacting with digital content and other individuals. The active web user generates a variety of data of interest to our field, facilitating the study of cognitive processes like memory and categorization, as well as a wealth of applied problems that methods and theory from the cognitive sciences can help address. The systematic recording of user data by Web systems means there is a wealth of "big data" capturing such behavior available to cognitive scientists.

Collaborative tagging is one of the core technologies of Web 2.0, and entails the assignment of freeform textual labels (tags) to online resources (photos, music, documents, etc.) by users. These tag assignments are then aggregated into a socially generated semantic structure known as a “folksonomy.” The commonly-assumed purpose of tagging is for personal information management: Users tag resources to facilitate their own retrieval of tagged items at a later time. In effect, then, such tags serve as a memory cues, signals offloaded to the (virtual) environment that allow users to find resources in the future. If this assumption holds, tagging behavior can serve as a useful window on the psychological processes described above. However, while the “tags as memory cues” hypothesis is assumed across an majority of tagging research, there is little in the way of empirical evidence supporting this interpretation of tagging behavior. Our current research thus serves to test this hypothesis, examining big data from social tagging systems to determine whether users are in fact using tags as memory cues. Using a unique dataset from the social music website Last.fm that includes records of both what music users have tagged and how they have interacted with that music over time (in the form of music listening histories), we examine if and how patterns of content interaction support or contradict the memory cue interpretation. There is an immediate practical application of this work to those working with collaborative tagging systems (are user motivations what we think they are?), but our work also comprises contributions of interest to the cognitive science community: First, we are expanding our understanding of how people generate and use memory cues “in the wild”. Second, we are enriching the “toolbox” available to cognitive scientists for studying cognition using large-scale, ecologically valid data that is latent in the logged activity of web users.

We begin the chapter by providing background on precisely what collaborative tagging entails, describing the existing theories of tagging motivation, and briefly summarizing the relevant work in psychology and cognitive science on memory cue generation and use, relating it to the case of online tagging. We then formalize our research objectives, outlining the difficulties in making claims about why people are tagging based on histories of what they have tagged, presenting the details of our dataset and how it offers a partial solution to those difficulties, and delineating our concrete hypotheses. Finally, we present the novel analysis methodologies we are employing and some of the results they have generated.

Background

What is Collaborative Tagging?

In collaborative tagging, many individuals assign freeform metadata in the form of arbitrary strings (tags) to resources in a shared information space. These resources can, in principal, be any digital object, and web services across a wide variety of domains implement tagging features. Examples include web bookmarks (Delicious.com, Pinboard.in), music (Last.fm), photos (Flickr.com, 500px.com), academic papers (academia.edu, mendeley.com), books (LibraryThing.com) and many others. When many users engage in tagging of a shared corpus of content, the emergent semantic structure is known as a folksonomy, a term defined by Thomas Vander Wal as a "user-created bottom-up categorical structure…with an emergent thesaurus" (Vander Wal, 2007). Under his terminology, a folksonomy can either be broad, meaning many users tag the same, shared resources, or narrow, in which any given resource tends to be tagged by only one user (usually the content creator or uploader). Last.fm, on which we are performing our analyses, is a canonical example of the former, and Flickr, where users upload and tag their own photos, is a good example of the latter.

Folksonomies have been lauded as a radical new approach to content classification (Weinberger, 2008; Sterling, 2005; Shirky, 2005; Heckner et al., 2008). In principle, they leverage the "wisdom of the crowds" to generate metadata both more flexibly (multiple classification of content is built in to the system) and at lower economic cost (individual users are, generally, self-motivated and uncompensated) than in traditional, expert- or computer-generated taxonomies, as one might find in a library. The approach is not uncontroversial, however, with critics from library science in particular (Macgregor &

McCulloch, 2006) pointing out the difficulties that the wholly uncontrolled vocabularies of folkonomies can introduce (especially poor handling of homonyms and hierarchical relationships between tags). In broad folksonomies, the existence of social imitation effects (Lorince & Todd, 2013; Floeck et al., 2011) can also cast doubt on whether agreement as to how an item ought to be tagged reflects true consensus, or instead bandwagon effects that don't "correctly" categorize the item. Given our current focus on individuals' tagging motivations, the level of efficacy of tagging systems for collective classification is not something we address here.

Hotho et al. (2006) formally define a folksonomy as a tuple $\mathbb{F} := (U, T, R, Y)$[1] where U, T, and R are finite sets representing, respectively, the set of all unique users, tags, and resources in the tagging system. Y is a ternary relation between them ($Y \subseteq U \times T \times R$), representing the set of tag assignments (or, equivalently, annotations) in the folksonomy (i.e. instances of a particular user assigning a particular tag to a particular resource). They also define the personomy of a particular user, $\mathbb{P} := (T_u, R_u, Y_u)$, which is simply the subset of $\mathbb{F}$ corresponding to the tagging activity of a single user.

Collaborative tagging systems began to be developed in the early 2000s, with the launch of the social bookmarking tool Delicious in 2003 marking the first to gain widespread popularity. Three years later, Golder and Huberman's (2006) seminal paper on the stabilization of tag distributions on Delicious sparked interest in tagging as an object of academic interest. In the years since, a substantial literature on the dynamics of tagging behavior has developed. Research has covered topics as diverse as the relationship between social ties and tagging habits (Schifanella et al., 2010), vocabulary evolution (Cattuto, Baldassarri, et al., 2007), mathematical and multi-agent modeling of tagging behaviors (Lorince & Todd, 2013; Cattuto, Loreto, & Pietronero, 2007), identification of expert

[1] The original definition contains a fourth element, such that $\mathbb{F} := (U, T, R, Y, \prec)$. The last term, $\prec$, represents a user-specific subtag/supertag relation that folksonomy researchers (including the authors who define it) do not typically examine, and we do not discuss it here.

taggers (Noll et al., 2009; Yeung et al., 2011), emergence of consensus among taggers (Halpin et al., 2007; Robu et al., 2009), and tag recommendation (Jäschke et al., 2007; Seitlinger et al., 2013), among others.

This small sample of representative work is indicative of the fact that, at least at the aggregate level, researchers have a fairly good idea of *how* people tag. What is comparatively poorly understood (and relevant to our purposes here) is exactly *why* users tag.

Why People Tag

The prevailing assumption about tagging behavior is that tags serve as retrieval or organizational aids. Take the original definition of "folksonomy" as a canonical example: "Folksonomy is the result of personal free tagging of information and objects (anything with a URL) *for one's own retrieval*" (Vander Wal, 2007, emphasis added). Couched in psychological terms, this is to say that tags function as memory cues of some form, facilitating future retrieval of the items to which they are assigned. There are various manifestations of this perspective (see, among many examples, Glushko et al. 2008, Halpin et al. 2007, and Golder & Huberman 2006), and it is one generally in line with the design goals of tagging systems. Though tagged content can be used in various ways beyond retrieval, such as resource discovery and sharing, the immediate motivation for a user to tag a given item is most often assumed (not illogically) to achieve an information organization and retrieval goal. This is not to imply that other tagging objectives, such as social sharing, are necessarily *illogical*, only that they are less often considered primary motivators of tagging choices. Such retrieval goals are implemented in tagging systems by allowing users to use tags as search keywords (returning items labeled with a particular tag from among a user's own tagged content, or the global folksonomy) and by allowing them to directly browse the tags they or others have generated. On Last.fm, for example, a user can click on the tag "rock" on the tag listing accessible from her profile page, and view all the music to which she has assigned the tag “rock”.

While our current goal is test whether this assumption holds when considering users’ histories of item tagging and interaction, it is important to recognize that alternative motivations for tagging can exist. Gupta et al. (2010), for instance, posit no less than nine possible reasons, beyond future retrieval, for which a user might tag: Contribution and sharing, attracting attention to one’s own resources, play and competition, self-presentation, opinion expression, task organization, social signaling, earning money, and “technological ease” (i.e. when software greatly reduces the effort required to tag content). We will not analyze each of these motivational factors in depth, but present the list in its entirety to make clear that tagging motivation can extend well beyond a pure retrieval function. We do, however, briefly review the most well-developed theories of tag motivation in the literature.

What is likely the most critical distinction in a user’s decision to tag a resource is the intended audience of the tag, namely whether it is self- or other-directed. This distinction maps onto what Heckner et al. (2008) refer to as PIM (personal information management) and resource sharing. The sort of self-generated retrieval cues that interest us here are fall under the umbrella of PIM, while tags generated for the purposes of resource sharing are intended to help other people find tagged content. For example, a user may apply tags to her Flickr photos that serve no personal organizational purpose, but are intended to make it easier for others to discover her photos. Social motivations can be more varied, however. Zollers (2007), for instance, argues that opinion expression, performance, and activism are all possible motivations for tagging. Some systems also implement game-like features to encourage tagging (Weng & Menczer, 2010; Weng et al., 2011; Von Ahn & Dabbish, 2004)that can invoke socially-directed motivations.

Ames & Naaman (2007) present a two-dimensional taxonomy of tagging motivation, dividing motivation not only along dimensions of sociality (like Heckner et al. 2008), but also a second, functional dimension. Under their terminology, tags can be either organizational or communicative. When self-directed, organizational tags are those used for future retrieval, while communicative tags provide contextual information about a tagged resource, but are not intended to aid in retrieval. Analogously, social tags can either be intended to help other users find a resource (organizational) or communicate information about a resource once it is found (communicative).

While all of these theories of tagging motivation appear reasonable (to varying degrees), there is little in the way of empirically rigorous work demonstrating that user tagging patterns actually align with them. The most common methods for arriving at such taxonomies are examining the interface and features of tagging systems to infer how and why users might tag (e.g. in a system where a user can only see her own tags, social factors are likely not at play, see Marlow et al., 2006), semantic analysis and categorization of tags (e.g. "to read" is likely a self-directed organizational tag, while tagging a photo with one's own username is likely a socially-directed tag suggesting a variety of self-advertisement, see Zollers, 2007; Sen et al., 2006), and qualitative studies in which researchers explicitly ask users why they tag (e.g. Ames & Naaman, 2007; Nov et al., 2008.) All of these methods certainly provide useful insights into why people tag, but none directly measure quantitative signals of any proposed motivational factor. One notable exception to this trend is the work of Körner and colleagues (Körner, Benz, et al., 2010; Körner, Kern, et al., 2010; Zubiaga et al., 2011), who propose that taggers can be classified as either categorizers (who use constrained tag vocabularies to facilitate later browsing of resources) or describers (who use broad, varied vocabularies to facilitate later keyword-based search over resources). They then develop and test quantitative measures that, they hypothesize, should indicate that a user is either a categorizer or describer. Though Körner and colleagues are able to classify users along the dimensions they define, they cannot know if describers actually use their tags for search, or that categorizers use them for browsing. This is a problem pervasive in work on tagging motivation (for lack of the necessary data, as we will discuss below); there is typically no way to verify that users actually use the tags they have applied in a manner consistent with a given motivational framework.

Connections to Psychological Research on Memory Cues

We now turn to work from the psychological literature on how humans generate and employ the kinds of externalized memory cues that tags may represent. There is little work directly addressing the function of tags in web-based tagging systems as memory cues, but some literature has explored self-generated, external memory cues. This research finds its roots more broadly in work on mnemonics and other memory aids that gained popularity in the 1970s (Higbee, 1979). Although most work has focused on internal memory aids (e.g. rhyming, rehearsal strategies, and other mnemonics), some researchers have explored the use of external aids, which are typically defined as "physical, tangible memory prompts external to the person, such as writing lists, writing on one's hand, and putting notes on a calendar" (Block & Morwitz, 1999, p. 346). We of course take the position that digital objects, too, can serve as memory cues, and some early work (Harris, 1980; Hunter, 1979; Intons-Peterson & Fournier, 1986) was sensitive to this possibility long before tagging and related technologies were developed.

The work summarized above, though relevant, provides little in the way of testable hypotheses with respect to how people use tags. Classic research on human memory – specifically on so-called cued recall – can offer such concrete hypotheses. If the conceptualization of tags as memory cues is a valid one, we would expect users' interaction with them to conform, to at least some degree, with established findings on cued retrieval of memories. The literature on cued recall is too expansive and varied to succinctly summarize here (see Kausler & Kausler 1974 for a review of classic work), but broadly speaking describes scenarios in which an individual is presented with target items (most typically words presented on a screen) and associated cues (also words, generally speaking), and is later tested on her ability to remember the target items when presented with the previously-learned cues. The analog to tagging is that tags themselves function as cues, and are associated with particular resources that the user wishes to retrieve (recall) at a later time. The scenarios, of course, are not perfectly isomorphic. While in a cued-recall context, a subject is presented with the cue, and must retrieve from memory the associated item(s), in a tagging context the user may often do the opposite, recalling the cue, which triggers automatic retrieval (by the tagging system) of the associated items "for free" with no memory cost to the user. Furthermore, it is likely true in many cases that a user may not remember the specific items they have tagged with a given term at all. Instead, a tag might capture some relevant aspect of the item it is assigned to, such that it can serve to retrieve a set of items sharing that attribute (with no particular resource being sought). As an example, a user might tag upbeat, high-energy songs with the word "happy", and then later use that tag to listen to upbeat, happy songs. In such a case, the user may have no particular song in mind when using the tag for retrieval, as would be expected in a typical cued-recall scenario.

These observations reveal that, even when assuming tags serve a retrieval function, how exactly that function plays out in user behavior can take various forms. Nonetheless, we take the position that an effective tag – if and when that tag serves as retrieval cue – should share attributes of memory cues shown to be effective in the cued recall literature. In particular, we echo Earhard's (1967) claim that "the efficiency of a cue for retrieval is dependent upon the number of items for which it must act, and that an efficient strategy for remembering must be some compromise between the number of cues used and the number of items assigned to each cue" (p. 257). We base this on the assumption that tags, whether used for search, browsing, or any other retrieval-centric purpose, still serve as cue-resource associates in much the same way as in cued recall research; useful tags should connect a user with desired resources in way that is efficient and does not impose unreasonable cognitive load.

In cases of tagging for future retrieval, this should manifest as a balance between the number of unique tags (cues) a user employs, and the number of items which are labeled with each of those tags. Some classic research on cued recall would argue against such a balancing act, with various studies suggesting that recall performance reliably increases as a function of cue distinctiveness (Moscovitch & Craik, 1976). This phenomenon is sometimes explained by the cue-overload effect (Rutherford, 2004; Watkins & Watkins, 1975), under which increasing numbers of targets associated with a cue will "overload" the cue such that its effectiveness for recalling those items declines. In other words, the more distinctive a cue is (in terms of being associated with fewer items), the better. But when researchers have considered not only the number of items associated with a cue, but also the total number of cues a subject must remember, results have demonstrated that at both extremes – too many distinct cues or too many items per cue – recall performance suffers. Various studies support this perspective (e.g. Weist, 1970; Hunt & Seta, 1984), with two particularly notable cued recall studies being those by Earhard (1967), who found recall performance to be an *increasing* function of the number of items per cue, but a *decreasing* function of total number of cues, and Tulving & Pearlstone (1966), who found that subjects were able to remember a larger proportion of a set of cues, but fewer targets per cue, as the number of targets associated with each cue increased.

Two aspects of tagging for future retrieval that are not well-captured by existing work are (a) the fact that, in tagging, cues are self-generated and (b) differences in scale (the number of items to be remembered and tags used far exceed, in many cases by orders of magnitude, the number of cues and items utilized in cued recall studies). Tullis & Benjamin (2014) have recently begun to explore the question of self-generated cues in experiments where subjects are explicitly asked to generate cues for later recall of associated items, and their findings are generally consistent with the account of cued recall described here. Results suggest that people are sensitive to the set of items to be remembered in their choice of cues, and that their choices generally support the view that cue distinctiveness aids in recall. The issue of scale remains unaddressed, however.

In sum, the case of online tagging has important distinctions from the paradigms used in cued recall research, but we nonetheless find the cued recall framework to be a useful one for generating the specific hypotheses we explore below.

Problem Formalization and Approach

Stated formally, our overarching research question is this: By jointly examining when and how people tag resources, along with their patterns of interaction over time with those same resources, can we find quantitative evidence supporting or refuting the prevailing hypothesis that tags tend to serve as memory cues? In this section we address the challenges associated with answering this question, describe our dataset and how it provides an opportunity for insight into this topic, and outline specific hypotheses.

The Challenge

As discussed above, there is no shortage of ideas as to why people tag, but actually finding empirical evidence supporting the prevalent memory cue hypothesis – or any other possible tagging motivation, for that matter – is difficult. The simple fact of the matter is that there is plenty of data logging what, when, and with which terms people tag content in social tagging systems, but to our knowledge there are no publicly available datasets that reveal how those tags are subsequently used for item retrieval (or for any other reason). Of the various ways a user might interact with or be exposed to a tag after she has assigned it to an item (either by using it as a search term, clicking it in a list, simply seeing it onscreen, etc.), none are open to direct study. This is not impossible in principle, as a web service could log such information, but such data is not present in publicly available datasets or possible to scrape from any existing tagging systems.

Thus, we face the problem of making inferences about why a user tagged an item based only on the history of what, how, and when that user has tagged, without any ability to test if future use of the tag matches our inferences. It may seem, then, that survey approaches that directly ask users why they tag might necessarily be our best option, but we find this especially problematic. Not only are such self-reported motivations not wholly reliable, we are more interested in whether tags actually function as memory cues than whether users intend to use them as such. With all this in mind, we now turn to describing the dataset with which we are currently working, and why we believe it provides a partial resolution to these challenges.

Dataset

Our current work revolves around data crawled over the course of 2013 and 2014 from the social music website Last.fm. The core functionality of the site (a free service) is tracking listening habits in a process known as "scrobbling", wherein each timestamped, logged instance of listening to a song is a "scrobble". Listening data is used to generate music recommendations for users, as well as to connect them with other users with similar listening habits on the site's social network. Listening statistics are also summarized on a user's public profile page (showing the user's recently listened tracks, most listened artists, and so on). Although users can listen to music on the site itself using its radio feature, they can also track their listening in external media software and devices (e.g. iTunes, Windows Media Player, etc.), in which case listening is tracked with a software plugin, as well as on other online streaming sites (such as Spotify and Grooveshark). Because the site tracks listening across various sources, we can be confident that we have a representative – if not complete – record of users' listening habits.

Last.fm also incorporates tagging features, and users can tag any artist, album, or song with arbitrary strings. Being a broad folksonomy, multiple users can tag the same item (with as many distinct tags as they desire), and users can view the distribution of tags assigned to any given item. In addition to seeing all the tags that have been assigned to a given item, users are also able to search through their own tags (e.g. to see all the songs that one has tagged "favorites") or view the items tagged with a particular term by the community at large. From there, they can also listen to collections of music tagged with that term (e.g. on the page for the tag "progressive metal" there is a link to "play progressive metal tag").

The current version of our dataset consists of complete listening and tagging histories for over 90,000 Last.fm users for the time period of July 2005 through December 2012, amounting to over 1.6 billion individual scrobbles and nearly 27 million individual annotations (tuples representing a user's assignment of a particular tag to a particular item at a particular time). See Table 1 for a high level summary. All data was collected either via the Last.fm API or direct scraping of publicly-available user profile pages. We originally collected a larger sample of tagging data from users (approximately 1.9 million), and the data described here represents the subsample of those for which we have so far collected listening data. See our previous work using the larger tagging dataset (Lorince & Todd, 2013; Lorince et al., 2014; Lorince, Zorowitz, et al., 2015) for technical details of the crawling process.

The value of this data is that it provides not only a large sample of user tagging decisions, as in many other such datasets, but also patterns of interaction over time with the items users have tagged. Thus, for any given artist or song[2] a user has listened to, we can determine if the user tagged that same item and when, permitting a variety of analyses that explore the interplay between interaction with an object (in our case, by listening to it) and tagging it. This places us in a unique position to test if tagging a resource affects subsequent interaction with it in a way consistent with the memory cue hypothesis.

We of course face limitations. While these data present a new window on our questions of interest, they cannot establish a causal relationship between tagging and any future listening, and there may be peculiarities of music listening that limit the applicability of any findings to other tagging domains (e.g. web bookmarks, photos, etc.). Nonetheless, we find ourselves in a unique position to examine the complex interplay between music tagging and listening that can provide insight into whether or not people

[2] When crawling a user's listening history, we are able to determine song names and the associated artist names, but not the corresponding albums names.

tag for future retrieval, and tagging motivation more generally.

Hypotheses

As we clearly cannot measure motivation directly, we seek to establish a set of anticipated relationships between music tagging and listening that should hold if the memory cue hypothesis is correct, or at least in a subset of cases in which it applies. The overarching prediction of the memory cue hypothesis is that tags facilitate re-finding music in the future, which should manifest here as increased levels of listening to tagged music than we would find in the absence of tagging. Here we outline two concrete hypotheses:

Hypothesis 1. *If a user tags an item, this should increase the probability that a user listens to it in the future. Specifically, assignment of tags to a particular artist/song should correlate with greater rates of listening to that artist/song later.*

If tagging does serve as a retrieval aid, it should increase the chance that a user interacts with the tagged resource in the future. We would expect that increases in tagging an artist, on average, should correlate with and *precede* increased probability of listening to that artist. This would suggest that increased tagging is predictive of future listening, which is consistent with the application of tags facilitating later retrieval of a resource.

Hypothesis 2. *Those tags that are most associated with increased future listening (i.e. those that most likely function as memory cues) should occupy a "sweet spot" of specificity that makes them useful as retrieval aids.*

Even if the memory cue hypothesis holds, it is presumably the case that not all tags serve as memory cues. Those that do, as evidenced by a predictive relationship with future listening, should demonstrate moderate levels of information content (in the information theoretic sense, Shannon, 1948). A tag that is overly specific (for example, one that uniquely identifies a particular song) is likely of little use in most cases,[3] as the user may as

[3] This is not to say that such tags are *never* useful. We can imagine the generation of highly specific cues

well recall the item directly, while one that is overly broad (one that applies to many different items) is also of little value, for it picks out too broad a set of items to effectively aid retrieval. Thus we hypothesize that the specificity of tags (as measured by Shannon entropy) should be more likely on average to fall in a “sweet spot” between these extremes in those cases where tagging facilitates future listening.

Analytic Approaches

In this section we describe some of the analytic approaches we are employing to test the memory cue hypothesis, and a selection of early findings. We discuss, in turn, time series analysis methods including visualization and clustering, information theoretic analyses of tags, and other approaches to be explored in future work including modeling the causal influence (or lack thereof) of tagging on subsequent listening.

Central to the analyses presented below are user-artist *listening* time series and user-artist *tagging* time series. The former consist of the monthly scrobble frequencies for each user-artist pair in our data (i.e. for every user, there exists one time series of monthly playcounts for each unique artist she has listened to) in the July 2005 through December 2012 range. We similarly define tagging time series, which reflect the number of times a particular user tagged a particular artist each month. Although listening data is available at a higher time resolution than what we use for analysis, users’ historical tagging data is only available at monthly time resolution. Thus we down-sample all listening data to monthly playcounts to facilitate comparative analysis with tagging.

While it is possible in principle to define these time series at the level of particular songs as opposed to artists, the analysis we present here is limited to the artist level. For this first phase of research we have taken this approach because (a) the number of unique songs is much larger than the number of unique artists, greatly increasing the

(such as “favorite song of 1973”) that are associated with one or only a few targets, but are still useful for retrieval. As we will see below, however, such high specificity tags are not strongly associated with increased listening on average.

computational demands of analysis, and (b) the listening and tagging data (especially the latter) for any particular song in our dataset is typically very sparse. Thus, for the purposes of the work presented here, we associate with a given artist all annotations assigned directly to that artist, or to any of the artist's albums or songs.

Listening time series are normalized to account for variation in baseline levels of listening across users. We accomplish this by dividing a users's playcount for a given artist in a given month by that user's total playcount (across all artists) for that month. This effectively converts raw listening counts to the proportion of a user's listening in a given time period allocated to any given artist. After all pre-processing, our data consists of 78,271,211 untagged listening time series (i.e. user-artist pairings in which the user never tagged the corresponding artist), and 5,336,702 tagged time series (user-artist pairings in which the user tagged the artist at least once in the data collection period).

Time Series Analysis

With our time series thus defined, a number of analyses become possible to address our first hypothesis defined above. In most cases, such time series analysis at the level of the individual is very difficult, as listening and tagging data (especially the latter) tend to be sparse for any single user. But by aggregating many time series together, we can determine if user behavior, on average, is consistent with our hypotheses. Tagging data is not sparse for all users, however, and some users are in fact prolific taggers with thousands of annotations. As is clear from Figure 1, tagging levels show a long tailed distribution in which most users tag very little, and a small number tag a great deal. Although we average across users for the analyses presented here, these discrepancies between typical taggers and "supertaggers" – the implications of which we directly examine in other work (Lorince et al., 2014; Lorince, Zorowitz, et al., 2015) – suggest that future work may benefit from analyzing different groups of taggers separately.

A first, high level perspective is to compare the overall average listening of tagged versus untagged listening time series (that is, comparing listening patterns on average for user-artist pairs in which the user has tagged that artist, and those in which she has not), to see if they match the intuitions set forth in H1. As is apparent in Figure 2, they do. Here, after temporally aligning all time series to the first month in which a user listened to a given artist, we plot the mean normalized playcount (i.e. proportion of a user's listening in a given month) among all untagged (solid line) and tagged (dashed line) time series. As predicted, tagging is correlated with increased listening to an artist after the tag is applied (and also within the month the tag is applied), as evidenced by the higher peak and slower decay of listening for tagged time series. Note that the tagged time series analyzed here are limited to those tagged in the first month a user listens to a given artist. We ignore cases where a user only tagged an artist in the preceding or subsequent months, as there is no principled way to align the tagged and untagged time series for comparison under these circumstances. However, tagging is by far most common in the first month a user listens to an artist (more than 52% of tagged time series have an annotation the month of the first listen), so this analysis still captures a majority of the data. While these are results are correlational (we cannot know if increased listening levels are *caused* by tagging, or if users are simply more likely to tag the artists they are more likely listen to), aggregate listening patterns are at least consistent with H1.

In concurrent work,[4] we are exploring canonical forms of music listening patterns by applying standard vector clustering methods from computer science to identify groups of similarly-shaped listening time series. The precise methodological details are not relevant here, but involve representing each time series as a simple numeric vector, and feeding many such time series into an algorithm (k-means) that arbitrarily defines k distinct cluster centroids. Vectors are each assigned to the cluster to whose centroid they are most similar (as measured by euclidean distance), and a new centroid is defined for each cluster

[4] This work is not yet published, but see the following URL for some methodological details: `https://dl.dropboxusercontent.com/u/625604/talks/Chasm2014.pdf`

as the mean of all its constituent vectors. This process repeats iteratively until the distribution of vectors over clusters stabilizes.[5] In Figure 3 we show results of one of various clustering analyses, showing cluster centroids and associated probability distributions of tagging for $k = 9$ clusters. Plotted are the mean probability distributions of listening in each cluster, as well as the temporally aligned probability distribution of tagging for all user-artist pairs in the cluster. Consideration of the clustered results is useful for two reasons. First, it demonstrates that tagging is, on average, most likely in the first month a user listens to an artist even when the user's listening peaks in a later month, which is impossible to see in Figure 2. Second, it provides further evidence that increases in tagging correlate with and precede increases in listening. This is demonstrated by the qualitatively similar shapes of of the tagging and listening distributions, but more importantly by the fact that the tagging distributions are shifted leftward (that is, earlier in time) compared to the listening distributions.

We have established that, on average, the relative behavior of listening and tagging time series are in line with our expectations, but an additional useful analysis is to explore if the probability of listening to an artist increases with the number of times that artist is tagged. Tagged time series should demonstrate more listening, as we have shown, but presumably the more times a user has tagged an artist, the more pronounced this effect should be. Figure 4 confirms the hypothesis, plotting the mean probability of listening to an artist as a function of the number of months since a user first listened to that artist, separated into the number of times the user has tagged the artist (or associated songs/albums). Formally, given that a user has listened to an artist for the first time at T_0, what is the probability that she listened to the artist one or more times in month $T_1, T_2, ..., T_n$. Tagged time series show greater listening as compared to untagged series,

[5] For these analyses, we also applied a gaussian smoothing kernel to all time series, and performed clustering on a random subset of 1 million time series, owing to computational constraints. Qualitative results hold over various random samples, however.

with listening probabilities increasing with the total number of times they are tagged.

Taken together, these preliminary comparisons of tagging and listening behavior demonstrate that tagging behavior is associated with increased probability of interaction with the tagged item, consistent with but not confirming H1.

In the next section we describe some of the information theoretic methods used to explore H2.

Information Theoretic Analyses

We have discussed the hypothesized importance of tag specificity in whether or not it serves as an effective retrieval aid, and here describe some analyses testing the hypothesis that the tags used as retrieval cues[6] should have moderate levels of specificity. A useful mathematical formalization of "specificity" for our purposes here is the information theoretic notion of entropy, as defined by Shannon (1948). Entropy (H) is effectively a measure of uncertainty in the possible outcomes of a random variable. It is defined as

$$H(X) = -\sum_{i} P(x_i) \log_b P(x_i) \tag{1}$$

where $P(x_i)$ is the probability of random variable X having outcome x_i, and b is the base of the logarithm. We follow the convention of using $b = 2$ such that values of H are measured in bits. The greater the value of H, the greater the uncertainty in the outcome of X. We can thus define the entropy of a tag by thinking of it as a random variable, whose possible "outcomes" are the different artists to which it is assigned. The more artists a tag is assigned to, and the more evenly it is distributed over those artists, the higher its entropy. H thus provides just the sort of specificity measure we need. High values of H correspond to *low* specificity, and low values of H indicate *high* specificity ($H = 0$ for a tag assigned to

[6] This is not to say that all tags *are* used as retrieval cues, only that that those are the ones that this hypothesis applies to. How to determine which tags are used as retrieval cues and which are not is a separate question we do not tackle here; for the purposes of these analyses we assume that such tags exist in sufficient numbers for us to see the proposed pattern in the data when considering all tags.

only one artist, as there is zero uncertainty as to which artist the tag is associated with).

We can define tag entropy at the level of an individual user's vocabulary, where H for a given tag is calculated over the artists to which that user has assigned it, and did so for each of every user's tags. We then binned all tags by their entropy (with a bin width of 0.5 bits), and for each bin retrieved all listening time series associated with tags in that bin. We then determined the mean probability of listening to those artists each month relative to the month when the tag was applied.

The results appears in Figure 5. Each line shows the average probability of a user listening to an artist at a time X months before or after tagging it, given that the user annotated that artist with a tag in a given entropy range. Entropies are binned in 0.5 bit increments, and entropy values are indicated by the color of each line. Two obvious large-scale trends should be noted. First, consistent with the earlier finding that tagging overwhelmingly occurs in the first month a user listens to an artist, the probability of listening to an artist peaks in the month it is tagged, and is greater in the months following the annotation than preceding it. Second, there is a general trend of overall lower listening probabilities with higher entropy, consistent with findings suggesting that greater tag specificity ought to facilitate retrieval. But, in support of our "sweet spot" hypothesis, this trend is not wholly monotonic. Tags with the lowest entropy (between 0.0 and 0.5 bits, dashed bold line) are *not* associated with the highest listening probabilities; tags with low, but not *too* low, entropy (between 0.5 and 1.0 bits, solid bold line) have the highest rates of listening.

The left hand inset plot is the probability distribution of total listening by binned entropy (i.e. the mean sum total of normalized listening within each bin). This is, effectively, a measure of the total amount of listening, on average, associated with artists labeled with a tag in a given entropy bin, and makes clear the peak for tags in the 0.5 to 1.0 bit range. Also of note is the relative stability of total listening (excepting the aforementioned peak) up to around 7 bits of entropy, after which total listening drops off rapidly. The right hand inset plot is the probability distribution of listening time series across tag entropy bins – or in other words, the distribution of rates of tag use versus tag entropy. Very low entropy tags (0 to 0.5 bits) are clearly the most common, indicating the existence of many "singleton" and low-use tags – that is, tags a user applies to only one, or very few, unique artists. Ignoring these tags, however, we observe a unimodal, relatively symmetric distribution peaked on the 5.0–5.5 bit entropy bin (marked with a vertical dashed line) that corresponds more or less directly to the stable region of total listening in the left hand inset plot. Precisely what drives the preponderance of "singleton" tags is not totally clear, but excluding them, these data do suggest that users demonstrate a preference for moderate-entropy tags associated with relatively high listening probabilities.

These results do not strongly suggest the existence of a single "sweet spot" in entropy (the peak in the 0.5–1.0 bit bin may be partly due to noise, given the relatively low frequency of tags in that entropy range), but do demonstrate that there is *not* a simple, monotonic relationship between increased listening and lower entropy values. Instead, we observed a range of entropy values (from 0.0 to approximately 7.0 bits) that are associated with higher listening rates. We must be cautious in drawing strong conclusions from these results, however. Because we are collapsing tagging and listening activity by artist, we cannot know the number of particular songs a user might retrieve with a given tag. Thus there may exist dependencies between tag entropy and the number of associated songs that drive mean listening rates higher or lower in a misleading manner. For example, a tag that tends to only be associated with a small number of songs may show low mean listening rates not because it is an ineffective retrieval cue, but because a small set of songs may generate low listening rates compared with a larger set.

This is just one of various difficulties in interpreting large scale data such as these. When considering the average behavior of many, heterogenous users, normalization and other transformations (such as our normalization of playcounts to account for variation in users' overall listening levels) are necessary, but can interact with derived measures (such as our entropy calculations) in complex, sometimes unexpected ways. As we continue this research program, we will need to further evaluate and refine the normalization methods we employ. Nonetheless, these early results are suggestive of systematic, meaningful relationships between listening habits and tag specificity.

Next steps: Causal analyses

The major shortcoming of the results we have presented thus far is that they cannot provide a *causal* argument in support of the memory cue hypothesis. Tagging is certainly correlated with listening, and early results suggest that observed tagging/listening relationships are, on average, in line with our hypotheses, but this is insufficient to make a strong causal argument. There is no simple method to address the critical question here: Does tagging an artist result in a user's listening to that artist being measurably different *than it would have been had the user not tagged the artist*?

Without addressing the philosophical problems surrounding claims about "true" causality, we are still tasked with testing if any sort of predictive causality exists between tagging and subsequent changes in listening behavior. Several relevant statistical methods exist, such as Granger causality (Granger, 1969), which tests for a causal relationship between two time series, as well as new methods like Bayesian structural time-series models (Brodersen et al., 2014), which estimate the causal impact of an intervention on time series data as compared to control data without an intervention. Though these and related methods are powerful, their applicability to our case appears limited for two reasons: First, tagging data is very sparse for any particular user-artist pair (typically consisting of one, or only a few, annotations), making methods that measure the impact of one time series on another, like Granger causality, untenable. Second, and more importantly, it is currently difficult to determine – even if tagging shows a predictive relationship with future listening – whether tagging actually facilitates retrieval of resources, thereby increasing listening, or if it is simply the case that users are more likely to tag those artists which they are independently more likely to listen to. Methods like Granger causality are useful when only two variables are involved, but cannot eliminate the possibility of a third variable driving both processes (in our case, intrinsic interest in an artist on the part of a user might increase both listening and the probability of tagging that artist).

We are currently exploring methods to sidestep this problem, but it is without doubt a challenging one. One possible approach may employ clustering methods similar to those described above to identify similar *partial* listening time series. If there exists a sufficient number of time series that show similar forms during the first N months a user listens to an artist, and if enough of those time series are tagged in month N, we can compare if and how tagged time series tend to diverge from untagged time series once a tag is applied. This poses some computational hurdles, and it is unclear if the sparsity of tagging data will permit such an analysis, but we hope the approach will prove fruitful. We also aim to expand our analysis to employ standard machine learning algorithms (such as support vector machines and logistic regression models) to develop a classifier for categorizing tagged and untagged time series. If a high-performing classifier based on listening behavior can be developed, it would indicate that there are systematic differences in listening behavior for tagged time series. This would suggest that tagging is not simply more likely for those artists a user is likely to listen to anyway, but instead associated with distinctive patterns of listening.

One approach that has born fruit in ongoing work, building upon the time series analysis methods described above, is the use of a regression model that predicts future listening rates as a function of past listening rates and whether or not a user-artist listening time series has been tagged (Lorince, Joseph, & Todd, 2015). Using a Generalized Additive Model (GAM, Hastie & Tibshirani 1990), our dependent variable in the regression is the logarithm of the sum of all listens in the six months after a tag has been applied, to capture the possible effect of tagging over a wide temporal window (the results are qualitatively the same when testing listening for each individual month, however), while our independent variables are a binary indicator of whether or not the time-series has been tagged, as well seven continuous-valued predictors, one each for the logarithm of total listens in the month of peak listening[7] in the time series and in each of the six previous months. The regression equation is as follows, where m corresponds to the month of peak listening, L is the number of listens in any given month, T is the binary tagged/untagged indicator, and f represents the exponential-family functions calculated in the GAM (there is a unique function f for each pre-peak month, see Hastie & Tibshirani 1990 for details):

$$\log \sum_{i=1}^{6} L_{m+i} = b_0 + b_1 T + \sum_{i=0}^{6} f(\log L_{m-i}) \tag{2}$$

We refer the reader to the full paper for further details, but the model (after imposing various constraints that permit temporal alignment of tagged and untagged time series data), allows us to measure the effect of tagging an artist on future listening, while controlling for users' past listening rates. Our early results suggest that tagging has a measurable, but quite small, effect on future listening. As we cannot visualize the regression results for all model variables at once, Figure 6 instead displays the predicted difference in listening corresponding to tagging as a function of the number of peak listens, calculated with a similar model which considers only the effect of listening in the peak month on post-peak listening. This plot suggests and the full model confirms that, controlling for all previous listening behavior, a tag increases the logarithm of post-peak listens by .147 (95% CI = [.144,.150]). In other words, the effect of a tag is associated with around 1.15 more listens over six months, on average, than if it were not to have been applied. These results thus suggest that tagging does lead to increases in listening, but only very small ones. Further analysis comparing the predictiveness of different tags for future listening (again, see the full paper for details) furthermore indicates that only a small subset of tags analyzed have any significant effect on future listening. Taken together, these tentative results provide evidence that tags certainly do not always function as memory

[7] Our methods align all time series to month of peak listening, and consider only tagged time series where the tag was applied in that peak month.

cues, and that facilitating later retrieval may actually be an uncommon tagging motivation.

Summary and Conclusions

In this chapter, we have made the following concrete contributions:

- A description of collaborative tagging systems, and how they offer valuable data on people's use of external memory cues in their day-to-day lives;
- a description of the "memory cue hypothesis", and the value of empirically testing it both for researchers specifically interested in tagging systems and cognitive scientists interested in human memory cue use;
- a review of the challenges associated with testing the "memory cue hypothesis" and a description of a new dataset that can help address them;
- two concrete hypotheses with respect to tagging and listening behavior that should hold if tags do in fact serve as memory cues; and
- a set of analytic methods for exploring those hypotheses.

Studying human cognition "in the wild" simultaneously presents great promise and difficult challenges. Big data like that described here permit correlational analysis on a large scale, with often compelling results, but can leave causal relationships difficult to discern. The time series and information theoretic analysis methods we have introduced do provide evidence that, on average, music tagging and listening behavior interact in a way consistent with the memory cue hypothesis insofar as tagging is associated with greater levels of listening and that moderate entropy tags are most strongly correlated with high listening probabilities. But as we have discussed, much work remains to be done to determine whether a compelling causal case can be made: Does tagging actually *cause* increases in listening that would not have occurred otherwise, specifically by facilitating retrieval? Our early results using a regression model suggest otherwise.

A second issue, particularly relevant to our data, but problematic in any study of "choice" in web environments, is the pervasiveness of recommendation systems. In comparing listening and tagging patterns, we have made the tacit assumption that users are making (more or less) intentional decisions about their music listening. In reality, however, an unknown proportion of users' listening is driven not by the active choice to listen to a particular artist (whether or not it is mediated by usage of a tag), but instead by the algorithms of a recommendation engine.[8]

These are challenges faced in any "big data" scenario, but a secondary issue is particularly relevant for psychologists and other researchers interested in making claims about individual cognitive processes. By analyzing and averaging data from many thousands of users, we are essentially describing the activity of an "average user", but must be hesitant to claim that any *particular* user behaves in the manner our results suggest. Even if aggregate data suggest that tags do (or do not) function as memory cues, we must remain sensitive to the limits on the conclusions we can draw from such findings. Large scale data analysis is a valuable tool for psychological researchers, but must be interpreted with care. This is particularly important given the non-normal distribution of tagging behavior observed in our data.

Though our results are tentative, we have presented an informative case study of human memory cue use in a real-world environment (digital though it may be), and a suite of tools for analyzing it. Our hope is that this work has provided evidence of the usefulness of collaborative tagging data for studying human memory and categorization, an introduction to some of the methods we can employ for research in this domain, and more generally an example of the power of big data as a resource for cognitive scientists.

[8] Because the Last.fm software can track listening from various sources, a given scrobble can represent a direct choice to listen to a particular song/artist, a recommendation generated by Last.fm, or a recommendation from another source, such as Pandora or Grooveshark.

References

Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 971–980). ACM.

Block, L. G., & Morwitz, V. G. (1999). Shopping lists as an external memory aid for grocery shopping: Influences on list writing and list fulfillment. *Journal of Consumer Psychology*, *8*(4), 343–375.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2014). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*.

Cattuto, C., Baldassarri, A., Servedio, V. D., & Loreto, V. (2007). Vocabulary growth in collaborative tagging systems. *arXiv Preprint*.

Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, *104*(5), 1461–1464.

Earhard, M. (1967). Cued recall and free recall as a function of the number of items per cue. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 257–263.

Floeck, F., Putzke, J., Steinfels, S., Fischbach, K., & Schoder, D. (2011). Imitation and quality of tags in social bookmarking systems–collective intelligence leading to folksonomies. In *On collective intelligence* (pp. 75–91). Springer International Publishing.

Glushko, R. J., Maglio, P. P., Matlock, T., & Barsalou, L. W. (2008). Categorization in the wild. *Trends in Cognitive Sciences*, *12*(4), 129–135.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of information science*, *32*(2), 198–208.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, *12*(1), 58–72.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web* (pp. 211–220). ACM.

Harris, J. E. (1980). Memory aids people use: Two interview studies. *Memory & Cognition*, *8*(1), 31–38.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (Vol. 43). CRC Press.

Heckner, M., Mühlbacher, S., & Wolff, C. (2008). Tagging tagging: analysing user keywords in scientific bibliography management systems. *Journal of digital information (JODI)*, *9*(2).

Higbee, K. L. (1979). Recent research on visual mnemonics: Historical roots and educational fruits. *Review of Educational Research*, *49*(4), 611–629.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of 3rd European Semantic Web Confernece (ESWC)* (pp. 411–426). Springer International Publishing.

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 454.

Hunter, I. M. L. (1979). Memory in Everyday Life. In M. M. Gruneberg & P. E. Morris (Eds.), *Applied Problems in Memory.* Academic Press.

Intons-Peterson, M. J., & Fournier, J. (1986). External and internal memory aids: When and how often do we use them? *Journal of Experimental Psychology: General*, *115*(3), 267.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2007). Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007* (pp. 506–514). Springer International Publishing.

Kausler, D. H., & Kausler, D. H. (1974). *Psychology of verbal learning and memory.* Academic Press New York.

Körner, C., Benz, D., Hotho, A., Strohmaier, M., & Stumme, G. (2010). Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web* (pp. 521–530). ACM.

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

Lorince, J., Joseph, K., & Todd, P. M. (2015). Analysis of music tagging and listening patterns: Do tags really function as retrieval aids? In *Proceedings of the 8th Annual Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP 2015).* Washington, D.C.: Springer International Publishing International Publishing.

Lorince, J., & Todd, P. M. (2013). Can simple social copying heuristics explain tag popularity in a collaborative tagging system? In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 215–224). ACM.

Lorince, J., Zorowitz, S., Murdock, J., & Todd, P. M. (2014). "Supertagger" behavior in building folksonomies. In *Proceedings of the 6th Annual ACM Web Science Conference* (pp. 129–138). ACM.

Lorince, J., Zorowitz, S., Murdock, J., & Todd, P. M. (2015). The Wisdom of the Few? "Supertaggers" in Collaborative Tagging Systems. *arXiv preprint*.

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library review*, *55*(5), 291–300.

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

Moscovitch, M., & Craik, F. I. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior*, *15*(4), 447–458.

Noll, M. G., Au Yeung, C.-m., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). Telling experts from spammers: expertise ranking in folksonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 612–619). ACM.

Nov, O., Naaman, M., & Ye, C. (2008). What drives content tagging: the case of photos on Flickr. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1097–1100). ACM.

Robu, V., Halpin, H., & Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web (TWEB)*, *3*(4), 1–34.

Rutherford, A. (2004). Environmental context-dependent recognition memory effects: An examination of ICE model and cue-overload hypotheses. *The Quarterly Journal of Experimental Psychology Section A*, *57*(1), 107–127.

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

Seitlinger, P., Ley, T., & Albert, D. (2013). An Implicit-Semantic Tag Recommendation Mechanism for Socio-Semantic Learning Systems. In *Open and Social Technologies for Networked Learning* (pp. 41–46). Springer International Publishing.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., … Riedl, J. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 181–190). ACM.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags.* Retrieved from `http://www.shirky.com/writings/ontology_overrated.html`

Sterling, B. (2005). Order Out of Chaos. *Wired Magazine*, *13*(4).

Tullis, J. G., & Benjamin, A. S. (2014). Cueing others' memories. *Memory & cognition*, 1–13.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381–391.

Vander Wal, T. (2007). *Folksonomy Coinage and Definition.* Retrieved 2014-07-29, from `www.vanderwal.net/folksonomy.html`

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 319–326). ACM.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 442.

Weinberger, D. (2008). *Everything Is Miscellaneous: The Power of the New Digital Disorder* (First Edition ed.). Holt Paperbacks.

Weist, R. M. (1970). Optimal versus nonoptimal conditions for retrieval. *Journal of Verbal Learning and Verbal Behavior*, *9*(3), 311–316.

Weng, L., & Menczer, F. (2010). GiveALink tagging game: an incentive for social annotation. In *Proceedings of the acm sigkdd workshop on human computation* (pp. 26–29). ACM.

Weng, L., Schifanella, R., & Menczer, F. (2011). The chain model for social tagging game design. In *Proceedings of the 6th International Conference on Foundations of Digital Games* (pp. 295–297). ACM.

Yeung, C.-m. A., Noll, M. G., Gibbins, N., Meinel, C., & Shadbolt, N. (2011). SPEAR: Spamming-Resistant Expertise Analysis and Ranking in Collaborative Tagging Systems. *Computational Intelligence*, *27*(3), 458–488.

Zollers, A. (2007). Emerging motivations for tagging: Expression, performance, and activism. In *Workshop on Tagging and Metadata for Social Information Organization, held at the 16th International World Wide Web Conference.*

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

Table 1

Dataset summary. Per-user medians in parentheses.

Measure	Count (per-user median)
Total users	90,603
Total scrobbles	1,666,954,788 (7,163)
Unique artists scrobbled	3,922,349 (486)
Total annotations	26,937,952 (37)
Total unique tags	551,898 (16)
Unique artists tagged	620,534 (16)

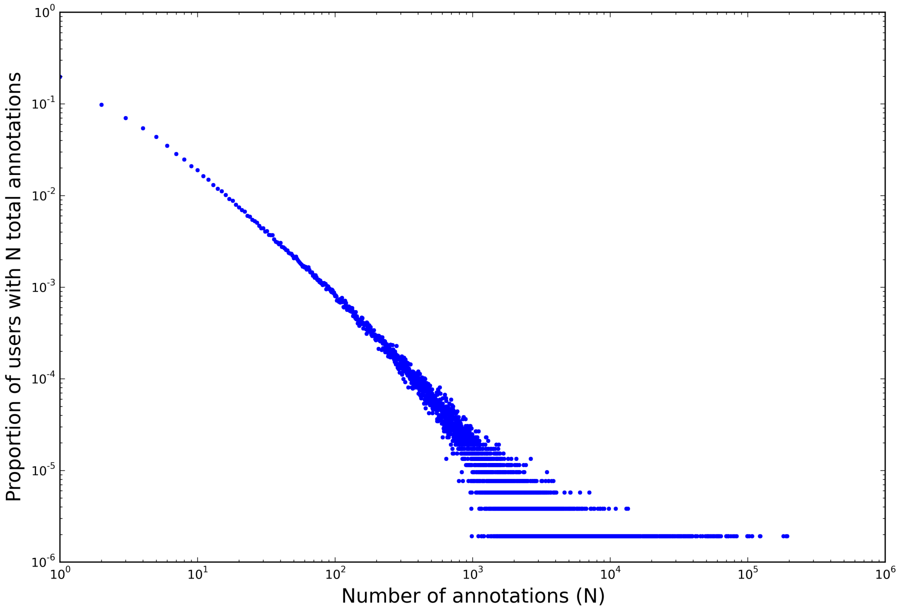

Figure 1. For a given total annotation count N, the proportion of users in our tagging dataset with a total of N annotations, on a log-log scale.

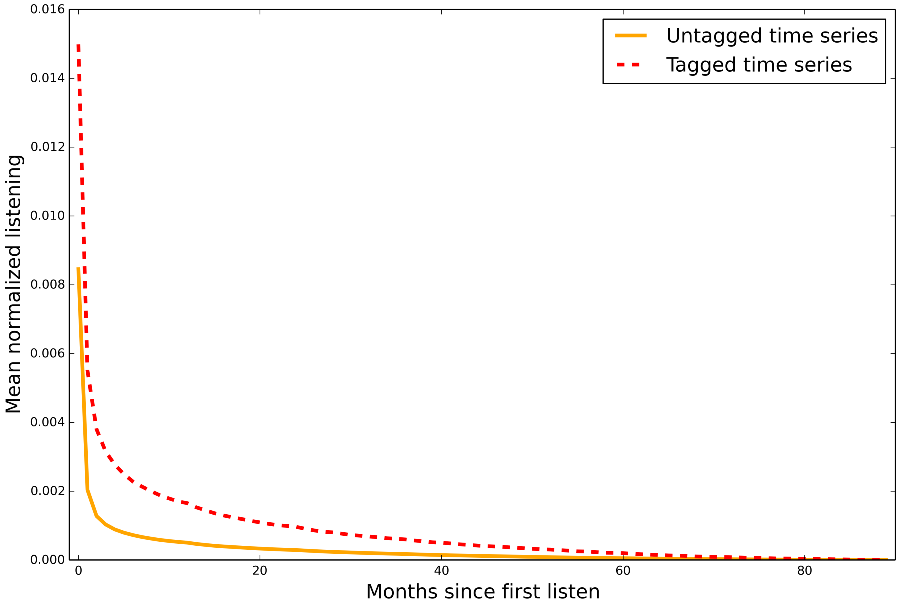

Figure 2. Mean normalized playcount each month (aligned to the month of first listen) for all listening time series in which the user never tagged the associated artist (solid line) and listening time series in which the user tagged the artist in the first month she listened to the artist (dashed line).

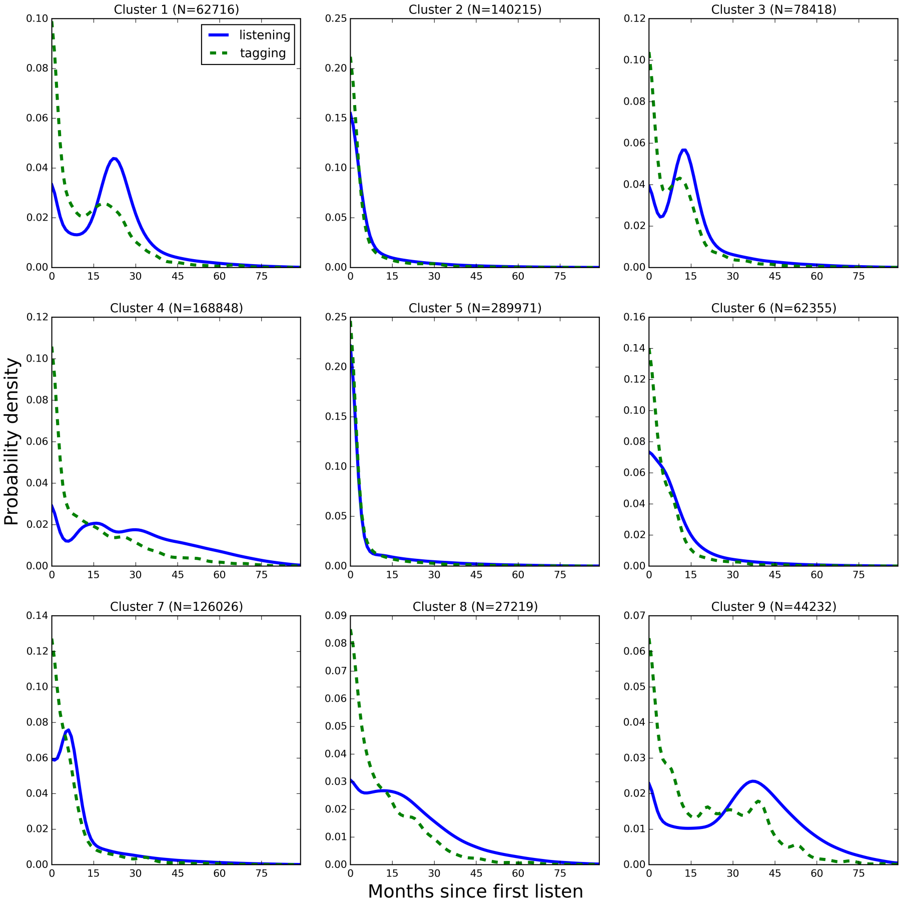

Figure 3. Clustering results for $k = 9$. Shown are mean normalized playcount (solid line) and mean number of annotations (dashed line), averaged over all the time series within each cluster. Time series are converted to probability densities, and aligned to the first month in which a user listened to a given artist. Clusters are labeled with the number of listening time series (out of 1 million) assigned to each cluster. Cluster numbering is arbitrary.

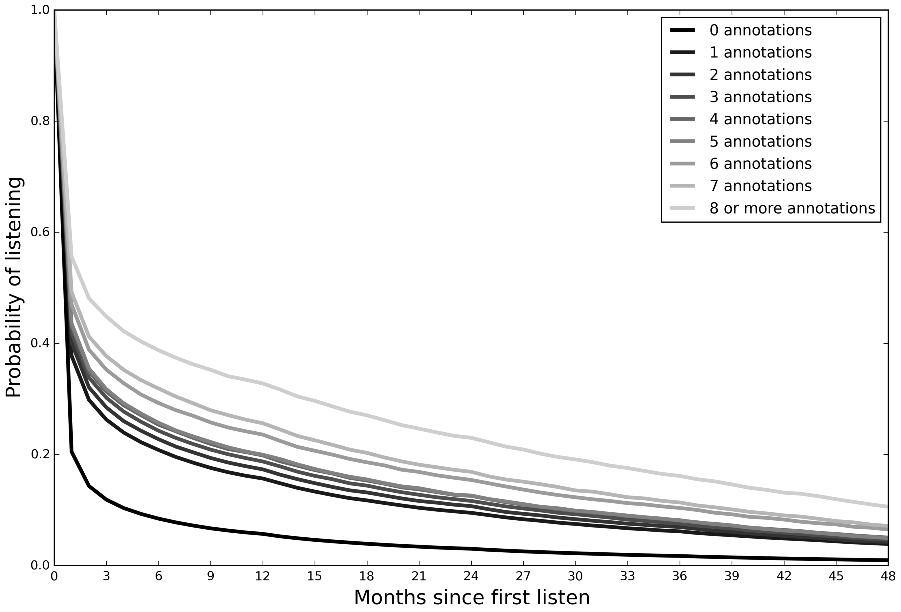

Figure 4. Mean normalized playcount for user-artist listening time series tagged a given number of times.

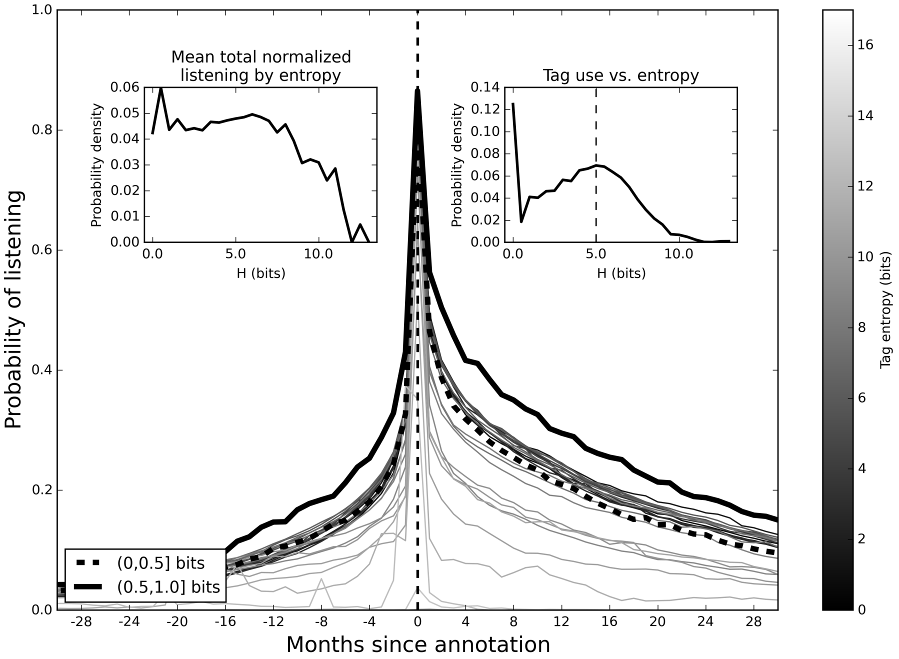

Figure 5. Mean probability of listening each month (relative to the month in which a tag is applied) for user-artist time series associated with tags of a given binned entropy (bin width of 0.5 bits). Each line represents the mean listening for a particular entropy bin, with line color indicating the entropy range for the bin (darker shades show lower entropy). Highlighted are the listening probabilities associated with 0.0-0.5 bit entropy tags (bold dashed line) and 0.5 to 1.0 bit entropy tags (bold solid line). The inset plots show the total mean listening (i.e. sum over all values in each line from the main plot) for each entropy bin (left), and the probability distribution of tags by entropy (right).

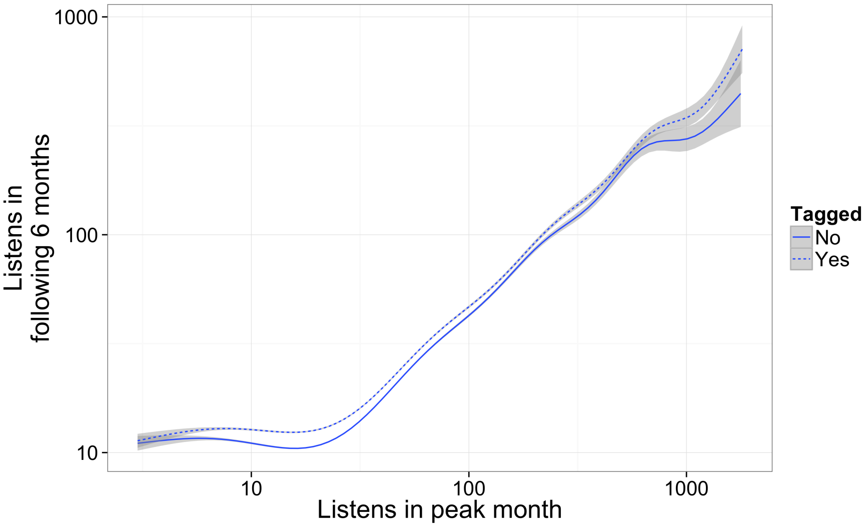

Figure 6. Regression model results, showing predicted sum total of listening in the 6 months after a tag is assigned as a function of the number of listens in the month of peak listening in a time series. Results shown on a log-log scale, and shaded regions indicated a bootstrapped 95% confidence interval. Figure replicated from Lorince, Joseph, & Todd (2015).