CONSUMPTION OF CONTENT ON THE WEB:
AN ECOLOGICALLY INSPIRED PERSPECTIVE

Jared Lorince

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements for the degree

Doctor of Philosophy

in the Department of Psychological and Brain Sciences,

and the Cognitive Science Program

Indiana University

September 2016

Accepted by the Graduate Faculty, Indiana University,

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Peter M. Todd, Ph.D.

Filippo Menczer, Ph.D.

Edward J. Castronova, Ph.D.

Robert Goldstone, Ph.D.

August 16, 2016

To my father, Jay Lorince.

I made it, Dad.

Acknowledgements

The six years it has taken me to complete my PhD have been eventful, to say the least. It seems I have compressed many of life's major events, both good and bad, into this span of time, and as such I find myself particularly grateful to the various people who have supported me through the process.

My various co-authors and collaborators must be mentioned, as they have directly contributed to elements of the work presented in this thesis. In particular, I thank Kenny Joseph, Sam Zorowitz, and Jaimie Murdock for their help on papers that now form part of this thesis, and Saurabh Malviya, who helped show me the ropes of database design and other technical issues when I was still a novice programmer.

Debora Donato is another co-author, but much more than that. As my supervisor at two internships (first at Yahoo! Labs, then at StumbleUpon) she has opened up opportunities to me for which I am deeply grateful, and has offered invaluable academic, intellectual, and professional guidance over the years.

Indiana University has provided a stimulating and welcoming environment for my research and academic training, and I can't begin to count the number of conversations I've had with faculty, graduate students, and post-docs – both within and outside of my home department – that have contributed to the work in this thesis and my general academic development. I count myself lucky to have studied at a university with such a strong culture of collaboration and openness. While I am surely unfairly forgetting some people, I extend my thanks (in no particular order), to Tim Rubin, Chris Harshaw, Simon DeDeo, Johan Bollen, David Crandall, Alessandro Flammini, Filippo Radicchi, Mike Jones, Josh de Leeuw, and the various members of the Concepts and Percepts Lab. My

committee members Ted Castronova, Fil Menczer, and Rob Goldstone have been mentors of particular value. Special thanks go to my colleagues in the Adaptive Behavior and Cognition Lab. To Sam Cohen and Sam Nordli in particular, I couldn't have asked for better labmates, or friends.

I of course should mention the funding sources that have permitted to me focus on my research over the years. Most recently I was supported by the "What drives human cognitive evolution" grant provided by the John Templeton Foundation, and for my first three years of graduate study I was funded as an NSF IGERT research fellow.

I thank my parents, Jan and Jay, without whose love and support I never would have made it this far. My father, who passed away during the course of my studies, isn't here to see me graduate, but I know how important it was to him and how proud he would be to read these words.

My wife, Elena, has been a pillar of love and support that I have leaned upon more times than I care to admit. She has been my friend, confidant, motivator, and nurse (though mountain bike accidents, nasal surgery, and more). She has given me the confidence to keep moving forward when I have doubted myself. She has been the embodiment of love, safety, and family for me. And together we brought our son Ivar into this world, one accomplishment I am more proud of than my PhD. I love them both more than words can say.

Finally, I must thank my advisor, Peter Todd. A PhD student's advisor of course always plays an important role, but Peter is special. He has been a role model to me in every sense of the word, and has guided me through the PhD process while simultaneously encouraging me to develop my own academic identity. He has been such

an integral and positive part of my graduate experience that I can honestly not imagine what grad school would have been without him. Peter, I am deeply and sincerely honored to call you my boss, my colleague, and my friend.

Jared Lorince

CONSUMPTION OF CONTENT ON THE WEB:

AN ECOLOGICALLY INSPIRED PERSPECTIVE

How do people search for and organize informational resources on the web? This dissertation explores this question in the context of music tagging and listening, incorporating web data mining and machine learning methods with theory from cognitive science. In particular, I draw inspiration from information foraging theory (and the earlier work on animal foraging on which it is based) and the ecological rationality research program.

The three principal research questions my thesis addresses are the following:

RQ1: Can we apply an ecological perspective to online music listening, and what novel insights does such a perspective offer? (Chapter 2)

RQ2: How and why do people tag content in social Web environments? Specifically, I explore motivational factors around the highly variable levels of contribution to collaborative tagging systems, and present models of tagging built upon simple imitation heuristics. (Chapter 3)

RQ3: How do patterns of content consumption and classification interact? As a case study of the interplay between tagging and consumption, I test whether tagging serves as an aid for future retrieval of content. (Chapter 4)

Taken together, these research questions represent an integrative view of how humans interact with digital resources on the web, specifically with respect to music. There already exists a substantial literature on the two types of data I focus on here

(music tagging and listening), but this dissertation makes a significant new contribution by framing both my research questions and analytic approaches in a psychologically and ecologically grounded manner. This link is most obvious in Chapter 2, where I offer a foraging-inspired perspective on music consumption, but is present throughout all my work, such as examining ecologically-based mechanisms of imitation in tagging behavior (Chapter 3) or exploring whether proposed motivational mechanisms behind tagging are consistent with observed interactions between tagging and listening (Chapter 4).

Peter M. Todd, Ph.D.

Filippo Menczer, Ph.D.

Edward J. Castronova, Ph.D.

Robert Goldstone, Ph.D.

Table of Contents

List of Appendices

List of Figures

List of Tables

Chapter 1: An Ecological Perspective on Information Search and Organization

How do people search for and organize informational resources on the web? This dissertation explores this question in the context of music tagging and listening, incorporating web data mining and machine learning methods with theory from cognitive science. In particular, I draw inspiration from information foraging theory (and the earlier work on animal foraging on which it is based), as well as the ecological rationality research program.

The three principal research questions my thesis addresses are the following:

RQ1: Can we apply an ecological perspective to online music listening, and what novel insights does such a perspective offer?

RQ2: How and why do people tag content in social Web environments?

RQ3: How do patterns of content consumption and classification interact?

The current chapter provides a review of information foraging theory, a well-established framework for linking spatial foraging processes to search in abstract informational spaces on the web. This establishes the foundation for thinking of information search in terms of foraging generally, and I expand on this work to motivate application of the perspective to music consumption in particular. I then move on to present empirical work on RQ1 in Chapter 2, first establishing a latent feature space for describing musical artists (a prerequisite for subsequent analyses), then testing the applicability of a variety of foraging-inspired analyses to music listening behavior.

Chapter 3 focuses on RQ2, presenting a review of existing work on behavior in collaborative tagging systems (especially with respect to tagging motivation), as well as a review of social imitation in web contexts. I then present two studies, the first of which

uses multi-agent modeling to test if simple imitation heuristics can explain high-level patterns of tagging activity in a collaborative tagging system. The second explores the phenomena of "supertaggers" (the minority of users who engage in tagging far more than other users), testing for difference in tagging patterns between supertaggers and other users, as well as motivational and expertise differences.

Finally, Chapter 4 addresses RQ3, presenting a case study of how content tagging and consumption interact. This focuses on tests of the common – but unvalidated – assumption that users tag content for the purposes of future retrieval. I present a variety of analytic approaches, and focus on a study utilizing regression methods to test whether tagging on Last.fm leads to increased rates of listening to the artists tagged.

Dataset

The majority of original work discussed in this thesis uses a dataset that I collected over the course of 2013 and 2014 from the social music service Last.fm (refer to the appendix for details of the crawling process). 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), 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 (Vander Wal, 2005, more details to follow), 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").

Though the site has undergone a major redesign since the data presented here was collected, and some features are now defunct, at the time of crawling it additionally included a variety of secondary features. This included a social network (users can establish bi-directional friendship relations, comment on each other's profile pages, etc.), user-created groups (similar to Facebook-style groups), and the ability to "love" or "ban" tracks (akin to thumbing up or thumbing down on Pandora).

Data Summary

Our final dataset consists of complete tagging and supplemental data from ~1.9 million users, and complete listening histories from a ~167,000-user subsample of those users, totaling over 4 billion individual scrobbles. All data is limited to the time period of July 2005 through December 2012. See Table 1 for a summary. Note that the scrobble API (at least when we collected data) included only artist and song information, not the album a

particular song appeared on. Thus the count of albums appearing below only includes albums that were explicitly tagged.

Table 1: Summary of complete Last.fm dataset. Item data includes all songs/artists encountered (whether tagged, scrobbled, or both), but only albums that were explicitly tagged.

Tagging data	
Total users	1,884,597
Total annotations	50,372,893
Total unique tags	1,034,684

Item data	
Total artists	7,333,724
Total songs	76,622,538
Total albums	415,096

Supplementary data	
Loved tracks	161,427,452
Banned tracks	22,102,864
Group memberships	5,458,935
Unique groups	117,663
Friendship relations	24,320,919

Listening data	
Total users	167,244
Total scrobbles	4,691,766,834

Because data collection was extended over a period of approximately two years, the various projects described below employ different intermediate versions of the data described here. I describe the specifics of these various incarnations of the data as appropriate.[1]

Information Foraging Theory

In a time in which navigating and searching in information environments is so important in our lives, a scientific understanding of how we search for information seems a valuable goal, but is a challenging one to achieve. The ways we seek information are constantly expanding and changing; it can be difficult for experienced Internet users to keep up with the evolving Web, let alone the scientists seeking to understand the mechanisms by which

[1] Note that data cleaning over the course of the two-year crawling process has led to some minor inconsistencies between the numbers reported here and in some of the studies below (e.g. some users who were included in data for certain studies, but not reflected in the final dataset). These differences are all minor, however, and Table 1 represents the final, fully-cleaned dataset.

we pursue our informational goals. Nonetheless, researchers in human-computer interaction, cognitive science, and computer science have made efforts to develop frameworks for studying and understanding human information-seeking behavior. The most well-known and elaborated of these frameworks is information foraging theory (hereafter IFT). Proposed and largely developed by Peter Pirolli, the theory borrows heavily from ecological models of optimal foraging behavior, which were developed to understand how animals seek out food in physical environments. It is an intuitively appealing theory, in which researchers apply ecological models of patch-based foraging and diet selection to an information environment, replacing food resources with documents or Web pages, and calories with informational content.

Core Assumptions and Theoretical Commitments of IFT:

The first paper on IFT (Pirolli & Card, 1995) defines it as the "activities associated with assessing, seeking, and handling information sources" (p. 51). The use of the term "foraging" functioned on two levels, serving both to "conjure up the metaphor of organisms browsing for sustenance" (p. 52) and to draw a connection to formal optimal foraging theories developed by ecologists. Stated simply, the goal was to apply formal optimal foraging models describing how organisms navigate space in search of food to an informational context, substituting the fundamental currency of energy (in the caloric sense) used in ecological models with so-called "information value".

Before describing the specifics of IFT, I orient the reader to the broad theoretical basis and assumptions of the theory. We can broadly separate the assumptions into two categories, those drawn from the optimal foraging literature, and those from the

psychological literature, specifically Anderson's notion of rational analysis. I will discuss each in turn.

Ecological assumptions – Optimal foraging theory: The ecological assumptions of IFT derive from optimal foraging theory. The theory, championed by David Stephens and John Krebs (Stephens & Krebs, 1987) describes optimized models of how animals make decisions in food foraging contexts. IFT draws in particular on the conventional patch model, which describes how foragers allocate their time while seeking resources in a patchy environment, and the diet model, which describes the optimal subset of resources a forager should consume when the available resources vary in terms of their caloric value and handling costs. These models are described in some detail below, but these brief descriptions are enough to lay out the core assumptions required to apply optimal foraging theory to the information search context.

First, to apply a patch model of foraging obviously requires that the resources being sought are in fact distributed in a patchy fashion. A simple example of a patchy resource distribution in the animal foraging context would be that of a bird that feeds on berries. Berries grow on bushes, which form discrete resource patches (in the sense that the boundaries of a patch are well-defined) that vary in quality (i.e. how many berries the bush contains). The task of the bird, then, is to determine, based on the quality of a patch and the time and energy required to seek out another patch, how long to remain at a given bush before moving on to another. It is a classic explore/exploit problem: Do I stay here and continue to reap the diminishing returns of this patch, or pay the travel and search costs of seeking out another patch?

To view information search in similar terms requires that information environments are homologous to the kinds of patchy natural environments described in optimal foraging theory. Pirolli is deliberately vague about the precise definition of "patch" in the information foraging sense, as it can take several forms depending on the type of information-seeking task at hand. Thus, depending on the task, a patch could be a list of search engine results, an entire Web site, grouped links on a Web portal, or even a set of topically similar linked Web pages that may span several sites. In the earliest work on IFT, the patchy structure of the Web was simply asserted: "A user's encounters with valuable or relevant information will typically have a clumpy structure over space and time", (Pirolli & Card, 1995, p. 53). In subsequent papers Pirolli argued for this structure, but provided no direct empirical support ("We assume a scenario in which a user forages for relevant, valuable information at some web locality, meaning some collection of related WWW pages. Perhaps the pages are related because they are at some particular physical site or WWW server, or perhaps related because they have been collected by a particular community or organization", Pirolli, Pitkow, & Rao, 1996, p. 118). In more recent work (Pirolli, 2005, 2007) Pirolli cites two studies that empirically support a patchy Web structure. Eiron & McCurley (2003) showed that approximately 75% of Web hyperlinks from a sample of approximately 500,000 crawled in 2002 were intra-website (as opposed to external links). These results are consistent with the common-sense observation that Web sites typically have a hierarchical structure, evident in their URLs.[2] Davison (2000) showed that pages near one another on the link graph have greater lexical similarity (that is, page similarity decreases as a function of link distance). This was

[2] An idealized example: "www.website.com/blogs/blogtitle/year/month/day/title.html".

provided as evidence that "the Web is organized into a hierarchy of patches and Web content is arranged into topically related patches" (Pirolli, 2005, p. 351). Also see Menczer (2002, 2004).

The second major theoretical assumption borrowed from optimal foraging theory is that of the validity of optimization models themselves. Falling under the adaptationist research program in ecology, optimization models are used to develop quantitative, testable hypotheses about animal behavior built around the assumption that such behaviors are adaptive responses to the pressures of natural selection. Pirolli's contention is that these same methods can be applied to human search behavior on the Web, such that "Human-information interaction systems will tend to maximize the value of external knowledge gained relative to the cost of interaction" (Pirolli, 2007, p. 14). Hand in hand with this claim is the hypothesis that human information-seeking behaviors are exaptations of evolved spatial foraging strategies; that is, that the ways we search for information online co-opt cognitive capacities that evolved for foraging. If this is in fact the case, we should expect the analytical tools applied to optimal foraging to be effective in describing information search strategies, much as they have been for describing human food foraging behavior (Smith & Winterhalder, 1992).

Psychological assumptions – rational analysis: While the predominant ecological influence on IFT comes from Stephens and Krebs, the psychological grandfather of the theory is John R. Anderson. His theory of rational analysis (Anderson, 1990) figures prominently in Pirolli's theoretical and methodological approach to studying information foraging, and his ACT-R cognitive model (Anderson et al., 2004) forms the basis of Pirolli's ACT-IF and SNIF-ACT models. Rational analysis is a variety of methodological

adaptationism, which in the psychological literature arose in response to ad hoc mechanistic models of cognition that were highly data-centric: Researchers would collect data from human participants, then fit a model to match that data. Methodological adaptationism, on the other hand, takes a task-centric approach in many ways similar to that used by optimal foraging theorists. That is, the goal is to first examine the environmental problem solved by some behavior, develop cognitive strategies that could solve that problem, and then compare those to human performance.

Anderson's rational analysis is a variety of methodological adaptationism, a formal approach to studying cognition that serves as a guiding framework for Pirolli's development of IFT. The approach can be summarized in six steps (Anderson, 1990): (1) Specify the goals of the agent or cognitive system; (2) develop a formal model of the environment to which the agent is adapted; (3) make minimal assumptions about the computational costs of the goal to be achieved; (4) determine the optimal behavior for the agent, given steps 1-3; (5) compare the model predictions to empirical data; and (6) iterate as appropriate. Given this framework, and the assumptions about the connection between spatial foraging and information search, the applicability of optimal foraging models as described above is clear.

(Information) foraging models

We can now move on to a more detailed description of the foraging models introduced above, and how they are applied to information-seeking within IFT. In this section I review the two major elements of foraging research that are central to information foraging theory, the conventional patch model and the diet model, as well as the major conceptual addition to foraging theory introduced by Pirolli and Card, information scent.

Alternative models of optimal patch use have been presented (e.g. McNamara, 1982), and Pirolli has described a variant of his information foraging model based upon McNamara's work.[3] The majority of work in IFT, however, is based on the conventional patch model (Stephens & Krebs, 1987), so that is what I describe here. The cognitive models used in IFT research substantially elaborate on the relatively simple models described here, but these capture the key intuitions upon which the more complex models are based. The patch and diet models are based upon the core principles of average-rate maximization (optimality is defined in terms of the maximum average rate of return over an extended period) and exclusivity of exploring and exploiting. Taken together, they frame foraging decisions in terms of the principle of lost opportunity, under which the decision to exploit a resource is balanced against the possible increase in return to be gained by searching for a better resource.

The conventional patch model: This model formalizes the optimal amount of time a forager should spend in a resource patch before moving on to a new one. This entails the assumption that a forager splits its time between two mutually exclusive activities: exploiting within a patch (e.g. searching for and eating berries on a particular bush) and searching between patches (looking for a new bush to exploit). The basis for this model (and the diet model) is Holling's disk equation[4] (Holling, 1959):

$$R = \frac{G}{T_B + T_W}$$

[3] McNamara's model is stochastic, and is thus formulated differently than the conventional patch model in mathematical terms. The key notion it captures – that a forager should remain in a patch so long as the expected within-patch returns exceed expected returns from switching patches – is, however, the same as in the conventional model.

[4] The use of the word "disk" has no mathematical relevance. The term originates from an experiment in which a blindfolded participant searched for cardboard disks arranged on the ground.

The equation describes the net rate of gain R while foraging as the quotient of the total gain of value (calories, information content, etc.), G, and the total time spent foraging, which can be broken down into the total time spent searching for patches (T_B) and the total time spent exploiting within patches (T_W). In an animal foraging context, this could be expressed in calories per hour, or some other equivalent manner. Similar to Pirolli's definition of a patch, the units of "value" here are not formally defined for the information seeking case, and are again context dependent. Generally speaking, however, the value of an information resource can be thought of as its relevance to an information-seeking task, or the amount of valuable information it contains.

In the form above, the disk equation only tells us the mean rate of return of a foraging bout, which is not particularly useful, so let us make several changes. First, we will assume that there is a set of unique patch types that the forager can identify, and that each type has its own gain function, describing the net gain from that patch as a function of the time spent exploiting within the patch. Second, we will assume the forager can identify these different patch types, and can have unique "strategies" for each (simply meaning the agent will spend differing lengths of time in patches of different quality). Third, we express the values from the original disk equation as averages. With this in mind, we can express the patch model using the following variant of Holling's equation. I exclude the details of the derivation for brevity, but see Pirolli (2007) and Stephens & Krebs (1987) for details.

$$R = \frac{\sum_{i=1}^{P} \lambda_i g_i(t_{w_i})}{1 + \sum_{i=1}^{P} \lambda_i t_{w_i}}$$

The equation now expresses a forager's net rate of gain while foraging among different patch types indexed from i=1 to i=p in terms of the encounter rate of a patch

type, λ_i (the number of encounters of that patch type per unit time), the patch residence time, t_{w_i} (how long the forager remains in patches of type *i*), and the gain function, $g_i(t_{w_i})$ (the expected net gain from a patch as a function of time spent foraging patches of type *i*).

For this equation to describe *optimal* foraging requires that we formalize the optimal values of t_{w_i}, how long the forager remains in patches of each type. This can be deduced most simply in the case of linear within-patch returns, where a forager encounters a constant rate of return within a patch until it is exhausted. This is presumably a rare occurrence in nature, where patches typically show diminishing returns (in the berry-seeking bird example above, the forager will encounter less berries per unit time as the bush is depleted). In information foraging a constant rate of return is more plausible, with examples being a randomly ordered list of documents (such that relevant documents are roughly uniformly distributed) or in a case where a patch is made up entirely of relevant content. In this case, the optimal behavior is to remain in the patch until it is depleted. Leaving before or after this point is suboptimal, as is illustrated in Figure 1 for the simplified case of a single patch.

But even in the information foraging context, the constant-gains scenario is implausible. When viewing a search engine results page (SERP hereafter), items are ordered such that the most relevant results are at the top of the list, and when browsing a web page, more relevant items are usually more salient. Thus we should expect diminishing returns – that is, a cumulative gain curve with decreasing slope – while information foraging, just as in patch-based food foraging.

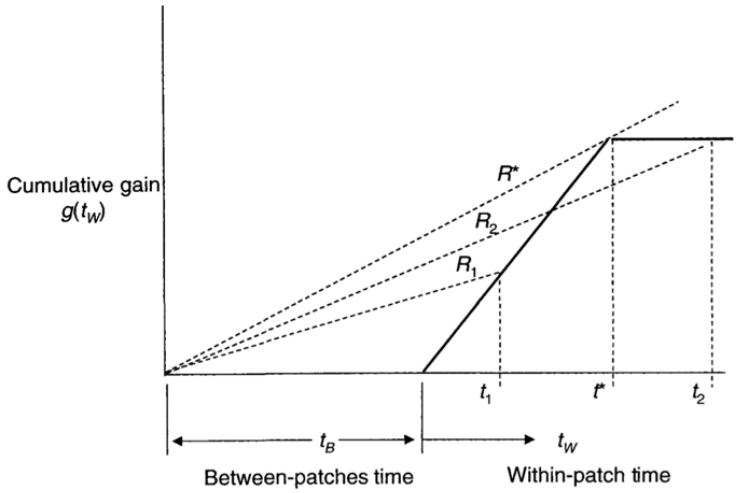


Figure 1: Illustration of optimal time allocation in a patch with constant returns. The bold line shows the cumulative gain as a function of time, which increases linearly until the patch is depleted, at which point the line flattens as no further gains are possible. The overall rate of gain (across time spent within and between patches) is indicated by the slope of the line linking the origin to the cumulative gain function at the time in which the forager leaves the patch. The optimal time allocation (leaving the patch as soon as it is depleted) yields a rate of R^*, while leaving the patch early (R_1) or late (R_2) yields a sub-optimal rate of return. Figure reproduced from Pirolli & Card (1999).

To determine the optimum policy in the case of diminishing returns we now turn to the marginal value theorem (Charnov, 1976). Charnov's theorem tells us that the optimal time to spend in a particular patch satisfies the equation $g'(t^*) = R(t^*)$; a forager should remain in a patch until the derivative of the gain curve equals the overall rate of gain across patches. In other words, given a mean rate of return across all patches as calculated above, the forager should remain in the patch until the instantaneous rate of return in that patch drops to the mean rate of return across all patches. This is best illustrated visually (see Figure 2a).

Figure 2 also shows the effects of within- and between-patch enrichment. Enrichment is a behavior typically not attributed to foragers in food foraging models, but which is afforded to information foragers under IFT. It comes in two forms, between-

patch (Figure 2b) and within-patch (Figure 2c). In the between-patch case, a forager can modify the environment so as to reduce the search times between patches. If seeking information in physical documents, this could be achieved by arranging stacks of documents so as to make switching between patches more efficient. A similar case in Web search might involve efficient organization of browser tabs or bookmarks, for instance. Within-patch enrichment is achieved by increasing the rate of gain within a particular patch. An information forager could accomplish this, for example, by issuing better queries to a search engine that result in a greater proportion of relevant resources in the search results.

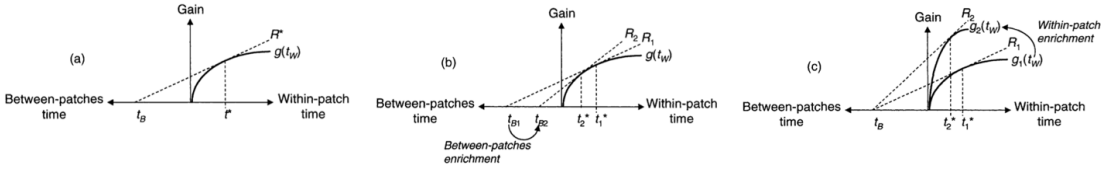


Figure 2: Illustration of marginal value theorem in the single-patch case. The mean rate of return across patches is indicate by the line marked R^*, and the optimal time to leave the patch occurs where R^* is tangential to the gain curve for the patch. (b) Between-patch enrichment reduces the optimal time in a patch while increasing the overall rate of gain by shortening the time spent searching for patches. (c) Within-patch enrichment has a similar effect by increasing the rate of gain within a patch. Figure reproduced from Pirolli & Card (1999)

The diet model: Rather than describing how long to remain in a particular resource patch, the diet model describes which resources an agent should incorporate into its diet. The classic example from the animal foraging literature is that of a predator faced with a variety of possible prey, varying in handling cost (e.g. how difficult a prey item is to catch) and value (e.g. caloric content). The diet model (also known as the prey model) describes which among the set of possible prey the predator should pursue upon encountering them.

When Pirolli and Card introduced the diet model, they deliberately left out a precise definition of what exactly makes up an information diet: "We are purposely ambiguous in our interchangeable use of item and patch. It may sometimes be more natural to think of things like documents as items and collections of documents as patches; however, one could conceivably develop diet models that treat collections as items or develop patch models that treat documents as patches of content that require time-allocation decisions" (Pirolli & Card, 1999, p. 654).

The formalization of the diet model is also based on Holling's disk equation. The model assumes a set of unique resource types such that each has a rate of encounter, λ_I, an average gain, *i*, and a handling time, t_{w_i}. Handling time in the food foraging case would reflect, for example, the effort required to capture a particular prey type, while in the information foraging context might reflect the effort required to extract valuable information from a resource. Given these definitions, the rate of gain from a diet consisting of items in the set *D* is:

$$R = \frac{\sum_{i \in D} \lambda_i g_i}{1 + \sum_{i \in D} \lambda_i t_{w_i}}$$

To determine the optimal set, *D*, of resources among those available to include in the diet (under the assumption that the time required to recognize the type of a resource is negligible), the following procedure is used. First, all possible resources are ranked in order of descending profitability, where π_i is the profitability of resource *i*, and profitability is defined as the ratio of the resource's value to its handling time ($\pi_i = \frac{g_i}{t_{w_i}}$). Starting with a diet consisting of only the most profitable resources, then considering a diet of the top two resources, and so on, the process iterates until $R(k) > \pi_{k+1}$. In

words, the algorithm terminates when the mean rate of return for a diet consisting of the top k items exceeds the profitability of the k+1st resource type. So, in the predator-prey example, if the inequality holds for k=3, the predator should always pursue the top three most profitable prey, and always ignore less profitable prey. The fact that the optimal diet is generated by including resources in an all-or-none manner is known as the "zero-one" rule.

The patch and diet models are highly idealized, and though they have proven useful in animal foraging research, they have several important deficiencies. Chief among these are that they are static models and that they assume complete information. With respect to the first point, they involve fixed behavioral strategies on the part of foragers (i.e. an unchanging repertoire of patch residence times depending on patch type, or a fixed optimal diet), and are not modulated by the forager's state or information gained while foraging. As for the second, the models assume that "the forager knows, or behaves as if it knows, the rules of the model" (Stephens & Krebs, 1987, p. 11). Though Stephens and Krebs do go on to describe both dynamic foraging models and models assuming *in*complete information in the later chapters of the book so much cited by Pirolli, IFT takes a unique approach to enriching the static, complete-information models of basic foraging theory: Information scent.

Information scent: The definition of information scent (Pirolli, 1997) varies slightly across the IFT literature, as it is again a highly task-dependent concept, but is most clearly stated as "the (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues, such as bibliographic citation, WWW links, or icons representing the sources" (Pirolli & Card, 1999, p. 646). The canonical

example is the information available describing a Web resource in a SERP, which serves as a proximal cue as to the value of the linked page itself. Though not a concept directly borrowed from food foraging research, the use of the term "scent" is a deliberate analogy to the kind of cue a food forager could use to locate or evaluate a resource.

The formalization of information scent depends on representing a proximal cue, *P* (e.g. the snippet of text describing a link in a SERP), and the informational (or distal) goal, *D*, as sets of features and applying a Bayesian analysis to predict the probability that a given proximal cue indicates the presence of content related to particular informational goal. Because information scent is not directly relevant to the work presented in this chapter, however, I do not discuss it further here.

Review of empirical work

The first paper on IFT (Pirolli & Card, 1995) was largely theoretical, outlining the patch and diet models detailed above, and justifying the use of optimal foraging models in the context of information search. The authors also present three example applications of the theory, all in the context of an experimental document search interface called "Scatter/Gather", which allowed users to select ("gather") clusters of documents that seemed relevant, and then re-cluster ("scatter") all the documents from the selected clusters into a new set clusters (Cutting, Karger, Pedersen, & Tukey, 1992).

The analog to patch-based foraging is clear here, as the document clusters can be viewed as resource patches. Pirolli & Card (1995) describe how a patch model, a diet model, and a dynamic foraging model could each be applied to the interface. They argue that the patch model could predict search times within different clusters (upon selecting to view the documents a cluster contains), while the diet model could predict choices of

which clusters a user would view. These were abstract proposals, however, used to illustrate the optimal foraging models and did not make any quantitative predictions. The dynamic foraging model, on the other hand, utilized quantitative estimates of the times involved in different Scatter/Gather subtasks to generate a state-space representation of the possible states of a forager using Scatter/Gather.

Subsequent work on IFT in the 1990s increased the focus on the environmental structure of the Web, introduced the notion of information scent, and saw the advent of ACT-IF, the first large-scale cognitive model to come out of the research program. Pirolli et al. (1996) for instance, focused on the problem of extracting usable navigation structures from the relatively unstructured WWW (from the perspective of a user seeking a resource). This involved a formalization of a web locality (a collection of related Web pages, roughly corresponding to a resource patch) and an early taxonomy of page types, including head pages (typically website home pages), reference pages, and content pages. The key insight from this paper, however, was that there are multiple ways to characterize the associations between sets of Web pages, each of which can be represented as a network structure: link topology, text similarity between pages, and usage paths (i.e. measurements of which links users actually follow). The paper presents some basic methods for applying spreading activation methods to these graphs as means of determining the most relevant pages. Such spreading activation models, applied differently, would be key elements of the ACT-Scent architecture.

The next major contribution to IFT was the introduction of information scent. Noting the need to describe a document or large set of documents in just a few words (as in Scatter/Gather or a SERP), Pirolli and colleagues describe, in the first use of the term,

such “terse representations of content as a kind of information scent whose trail leads to information of interest” (Pirolli, 1997, p. 3). This paper was also the first to introduce the ACT-R cognitive model (Anderson et al., 2004) as a framework for studying information foraging behavior, and Pirolli & Card (1999) would eventually develop a modification of ACT-R called ACT-IF (Adaptive Control of Thought in Information Foraging).

The next major development in IFT research was the introduction of SNIF-ACT (Scent-based Navigation and Information Foraging in the ACT architecture; Pirolli, Fu, Reeder, & Card, 2002). Unlike ACT-IF, which modeled behavior in the restricted Scatter/Gather environment, SNIF-ACT aimed to model more general information-seeking on websites. The primary goal in developing SNIF-ACT was to model decisions of (a) which links users would click on a Web page, given a particular information goal, and (b) when a user would abandon a page. The model utilizes a spreading activation mechanism, assigning informational values to links on a page by calculating information scent based on the words appearing in the link text and in the query. Decisions of which link to click were probabilistically related to information scent (i.e. higher-scent links were more likely to be clicked, but not in deterministic fashion). Page-leaving was modeled by the ability to backtrack (i.e. clicking the “back” button to return to the previous page) when the benefits of staying on the current page were outweighed by the benefit of exploring a new page. This corresponds to leaving a patch when within-patch gains drop below the overall rate of gain. When analyzing user data, site-leaving actions also included any time the user typed in a new URL or used a bookmark to go directly to another page (typically a search engine).

Looking forward: Pirolli's overview book on IFT (Pirolli, 2007), closes with a chapter on future directions for IFT research. He sketches four general directions for work: "upward", studies of social information foraging in collaborative tagging systems and other social contexts; "downward", development of finer-grained models of foraging behavior that include predictions of eye movements and other low-level information-seeking behaviors; "inward", general improvements to models like SNIF-ACT; and "outward", further work on understanding the task environments of information seeking, particularly in the context of sense-making.[5]

There has certainly been work furthering IFT research[6] along some of these lines, with examples such as a basic social information foraging model (Pirolli, 2009) and applications of IFT concepts to topics as diverse as goal attainment on long tail Web sites (McCart, Padmanabhan, & Berndt, 2013), image retrieval (Liu, Mulholland, Song, Uren, & Rüger, 2010), and "discovery browsing" (Goodwin, Cohen, & Rindflesch, 2012), to name but a few examples. However, recent work applying IFT to new problems has done little to empirically test[7] the validity of optimal foraging methods in information search. It seems information scent has been the most enduring concept of IFT, but (somewhat ironically) this stands as the major component of IFT not drawn from traditional optimal foraging theory. Of particular note is that the elaboration of SNIF-ACT that Pirolli calls for in his book does not seem to have occurred. Reviewing IFT (and SNIF-ACT in

---

[5] These naming of these four directions is an allusion to a review article on foraging theory entitled "Foraging Theory: Up, Down, and Sideways" (Stephens, 1990).
[6] A Google Scholar search for articles mentioning "Information Foraging Theory" since 2007 yields more than 600 results.
[7] The social information foraging paper does evaluate the applicability of models from social foraging theory to information search, but my focus here is on models of individual foraging.

particular) in a recent book chapter on search behavior (Fu, 2012) presents no new evidence supporting SNIF-ACT that was not already presented in Fu & Pirolli (2007). In short, IFT has definitely inspired research in recent years – much of it surely fruitful – but little of it bears on the central question of this paper.

Limitations of IFT

Theoretical issues: While connections between animal foraging and information search are intuitively appealing, we must cast a critical eye on the theoretical notions underlying IFT. The foraging metaphor is, again, quite compelling at first glance, and that arguably justifies testing its applicability to information search. But let us briefly consider some problems with how this connection is justified.

Recall that the key argument underlying the connection between animal food foraging and human information foraging is that, just as animals have evolved foraging strategies as adaptations to ancestral environments, humans demonstrate adaptive information seeking strategies on the Web. To quote one of many formulations of this idea: "similar to foraging behavior by animals, the decisions on where to search for information (food) and when to stop searching (patch-leaving policy) are assumed to be adapted to the statistical structures of the information environment. The detection and utilization of the structures are characterized as an adaptive response to the demands and constraints imposed by the information environments" (Fu, 2012, p. 287). This position is inextricably linked to the claim that human information seeking behaviors are exaptations of food foraging mechanisms: "Information Foraging Theory assumes that modern-day information foragers use perceptual and cognitive mechanisms that carry over from the evolution of food-foraging adaptations" (Pirolli, 2007, p. 14).

These appear to be plausible claims, but there is a distinct lack of clarity as to precisely what Pirolli means when he discusses information foraging in "adaptive" terms. Keeping in mind that adaptationist thinking is at times a problematic issue in biology in general (Gould & Lewontin, 1979; Mayr, 1983) and optimal foraging in particular[8] (Pierce & Ollason, 1987; Stearns & Schmid-Hempel, 1987; Stephens & Krebs, 1987) we are faced with two conflicting treatments of adaptationism in IFT. On the one hand, information foraging strategies are presented as exaptations, based on the logic that commonalities between the structure of information environments and food foraging environments (e.g. patchy structure) elicit in the former search strategies that evolved for the latter. This would seem to imply information search co-opts evolved, likely unconscious, strategies for resource-seeking. Under this view, search strategies would be *adaptive* in the sense that they are capacities that emerged over evolutionary time scales in ancestral environments, but turned out to be useful in Web information seeking.

On the other hand, Pirolli applies the notion of adaptation to his rational analysis of Web search itself, under the assumption that "the cognitive system optimizes the adaptation of the behavior of the organism" (Pirolli, 2007, p. 23). In other words, humans are "adapted" to information search such that "cognitive systems engaged in information foraging will exhibit…adaptive tendencies" (ibid, p. 14). This is problematic, as argued by proponents of the bounded rationality program, because rational analysis and the optimization under constraints methods it entails do "not directly address the question of what mental mechanisms could possibly yield behavior approaching the optimal norm"

---

[8] Even the torchbearer of optimal foraging theory admits that "Foraging theory is nothing if not controversial" (Stephens, 1990, p. 454).

(Gigerenzer & Todd, 2000, pp. 11–12). They instead provide a standard for evaluating proposed cognitive mechanisms. Pirolli does seem to recognize this, contending that use of rational analysis "does not mean that one assumes that the cognitive agent is performing the same calculations as the optimization models" (Pirolli, 2007, p. 24) and that simple heuristics are often good explanations of adaptive behavior. Yet precisely these kinds of optimization calculations are built into the cognitive models of IFT, and Pirolli fails to address this inconsistency.[9] In the context of animal foraging this would be less problematic, as the assumption there is that evolution has selected for mechanisms that approach optimality, and that these are typically enacted by simple proximate mechanisms. But in Web search this is of course not possible; it was a practically non-existent activity twenty years ago, and there cannot exist adaptations to it in the evolutionary sense. It may very well be the case that the strategies humans utilize in information-seeking are well-approximated by optimal foraging-inspired models, but if the exaptation argument is flawed, the mechanisms at play could be very different.

Further, IFT researchers present no experimental evidence to support, a priori, the claim that an evolutionary link exists between food foraging and information foraging: "We would like to think that the information foraging adaptions we observe are exaptations of the behavioral plasticity that humans evolved for food foraging, but it is unlikely that we will be able to obtain data relevant to tracing this evolution" (Pirolli & Card, 1999, p. 644). Interestingly, independent cognitive science research (Hills, 2006;

[9] It is unclear if Pirolli is one of the many researchers who erroneously equates bounded rationality and optimization under constraints ("a (mis)use we strongly reject", Gigerenzer & Todd, 2000, p. 12) or if he only fails to elaborate on what the actual cognitive mechanisms approximated by ACT-IF and SNIF-ACT might be.

Hills, Jones, & Todd, 2012; Hills, Todd, & Goldstone, 2008; Todd, Hills, & Robbins, 2012) does in fact support such a link. Combining behavioral and molecular evidence, this research program argues that spatial foraging capacities, in particular what is known as area-restricted search (Benhamou, 1992), are the evolutionary antecedents of information search strategies. Area restricted search is a more general foraging behavior, typically utilized when there do not exist discrete boundaries between patches (or by organisms who may not be adapted to detect such boundaries). In area-restricted search, a forager increases search effort upon encountering a resource, typically by searching more thoroughly nearby, and decreases effort when it goes a relatively long time without finding a resource, typically by moving in roughly linear fashion away from its current location. Whether this simple strategy could inform more effective models of information seeking, and how cognitive science research linking area-restricted foraging patterns to cognitive search might inform IFT more generally, are unexplored questions within the theory.

IFT and the modern Web: IFT, much like optimal foraging theory, stresses the importance of analyzing the environment in which Web search takes place for clues as to the strategies people use in that environment, and the constraints on those strategies. In so doing, a crucial motivation for applying optimal foraging analytical methods to Web search rests on structural commonalities between Web environments and food foraging environments, notably a patchy structure. While Pirolli and colleagues do provide evidence that the link structure of the web is patchy, more recent work casts some doubt on this view, as link structure does not necessarily constrain the way people navigate and search online as much we might think.

Given IFT's reliance on link structure as evidence of the Web's patchiness, which in turn underlies the applicability of optimal foraging models, this point is crucial to explore. There is of course the observation that foragers in information environments are not limited by link topology in the same way that physical foragers are limited by the topology of their environments; Physical topology (distances between patches, passability of terrain, and so on) fully constrains the movements of a food forager, but an information forager can "teleport", so to speak, moving directly to high-value resources (e.g. upon issuing an effective query on a search engine that returns the desired resource at the top of the results list). This distinction is fairly obvious, and IFT research does not ignore it, often equating "teleportation" to another page or search engine with the decision to abandon a patch and look elsewhere. But nonetheless, the validity of topology-based arguments for patchy Web structure rests on the assumption that such topology is the primary constraint on how information foragers navigate the Web. If this constraint does not hold, the appropriateness of patch-based foraging models comes into question.

While the explosive growth of the Web in recent years has arguably made large-scale studies of its link topology more difficult, the same period has seen technical capabilities such as toolbar-based activity tracking (Kumar & Tomkins, 2010) and HTTP packet recording (Meiss, Duncan, Gonçalves, Ramasco, & Menczer, 2009; Meiss, Menczer, Fortunato, Flammini, & Vespignani, 2008; Qiu, Liu, & Cho, 2005) enable large scale examination of human behavior online. Rather than focusing on the underlying link structure of the WWW, such work provides crucial insights into how people actually navigate online, with results that can be surprising.

Through several studies comparing the link topology of the web to actual navigation patterns of users, Meiss, et al. (2009) have found that "The link structure of the Web can differ greatly from the set of paths that are actually navigated, and it tells us little about the behavior of individual users." (p. 2). To be fair, Web behavior likely *was* much more constrained by link structure in the 1990s, when IFT was first developed, but changes in Web technologies and search patterns have clearly changed this.

In the early days of the Web, "surfing" – navigating from one page to another predominantly by clicking links – was the dominant metaphor for information seeking and consumption, but recent research indicates that this metaphor is much less appropriate today. This change can be attributed to a variety of factors, including improvements in search engines (modern search engines have greatly reduced the amount of link-following and manual search that Web users must perform), the rise of social media (permitting new avenues of information seeking, filtering, and discovery not well-captured by the IFT framework), and many others. But whatever the cause, Web navigation is much different today than from how it was when the framework for IFT was laid. IFT fundamentally relies on the supposed patchy structure of the Web, but the research described here suggests that such a patchy structure (if it in fact still exists) may not actually place strong constraints on how people navigate the Web. Despite IFT's undeniable contributions, the extent to which typical human searching and browsing patterns are consistent with patch-based foraging largely remains an unanswered research question.

From IFT to Music Foraging

Is optimal foraging the right metaphor?

I have raised some substantial criticisms of Information Foraging Theory, noting imprecise applications of adaptationist thinking and a failure of the theory to evolve along with the changing Web. Some caveats with respect to these criticisms are in order, however.

First, Pirolli's hypothesis that the study of animal foraging could inform our understanding of how humans seek out information was in many ways ahead of its time. There have long been commonsense parallels between search in the spatial, informational, and cognitive domains. "Memory palace" mnemonics (in which one can remember a long sequence of items by mapping each to a physical location in an imagined "palace" or other structure), or the design of computer operating systems that organize documents into hierarchical "folders" are but two examples of how there exist symmetries between the organization of content in physical, digital, and cognitive spaces. But to directly connect analytic methods of proven use in the animal foraging literature to human information search was a novel and promising development, and one that was almost prescient in light of modern developments suggesting a deep evolutionary link between search in physical and informational space (Hills et al., 2008; Todd et al., 2012). There are of course still serious problems with this analogy as presented by Pirolli, chief among them the fact that a shared evolutionary antecedent between spatial and information foraging strategies does not necessarily mean that established models of spatial foraging will describe patterns of information foraging. Animal foraging models describe behavior that has emerged over evolutionary timescales, and even though

humans likely leverage evolved spatial search capacities in information search scenarios, the differences between informational and ancestral environments differ sufficiently that it is unclear how exactly these co-opted cognitive-behavioral capacities will manifest. IFT has not been able to conclusively demonstrate that they manifest in a way consistent with optimal foraging models, but the theory nonetheless deserves credit for hypothesizing and exploring the connection.

Second, Pirolli's attempt to take an ecological approach to information search, framing the information seeker as an organism adapted to its digital environment, deserves merit. It is an ambitious goal, to say the least. While an ecologist can study a species that has lived in roughly the same environment for thousands of years, the ecology of the Web does not abide by the same rules. Studying the human information forager is like studying an organism that can instantly teleport from one habitat to another and has a constantly changing and often ill-defined set of dietary needs, all the while moving between habitats that are not only radically different from one another, but undergo major internal changes on a regular basis.[10] Against this backdrop, developing any sort of general theory of information foraging, especially one of any lasting relevance, is a huge challenge. But even if particular models and predictions developed under the umbrella of IFT are of limited value (at least with respect to modern Web behavior), Pirolli's approach to studying human Web users in the same way as one might study an animal in its natural habitat is of immense value. We may not be as well adapted

[10] Any regular Facebook or Google user who has had to "adapt" to changes in these sites' interfaces and services over the years will be acutely aware of this fact.

to information environments as Pirolli contends, but studying information seeking behavior from an adaptive behavior perspective is certainly a worthwhile pursuit.

Third, we must recall that "traditional" Web search, in which a user predominantly engages in link-following, moving from one "patch" of resources to another, has likely not disappeared from the behavioral repertoire of modern Web users entirely. I have argued that the typical mode of modern Web browsing does not match this "traditional" picture, but the operative word here is "typical". Plenty of search tasks require extensive manual search and browsing, especially in the domain of content that is not well-indexed by search engines (e.g. proprietary databases). Even if IFT only applies to these atypical browsing scenarios, the theory still deserves credit for its contributions to understanding these kinds of complex, deliberative searches.

We must keep in mind a key point about optimal foraging itself as we evaluate how well it translates to the information search domain: As the old mantra goes, all models are wrong[11] (but some are useful), and this is particularly true of optimization models of behavior. Not only do they vastly simplify the behavior and decision-making of the organisms they are designed to study, but they also do so under the unrealistic assumptions of optimality. No ecologist would realistically expect to find true optimality in a real organism, but optimal foraging theory entails no such expectation. Optimal foraging theorists develop models which assume animal behaviors are "designed" for foraging in a way that at least approaches optimality, and then explore which constraints exist that prevent organisms from achieving such optimality: "Even if they serve no other purpose, well-formulated design models are needed to identify constraints: without a

[11] A quote attributed to British mathematician George Box.

design hypothesis there would be no basis for postulating any kind of constraint!" (Stephens & Krebs, 1987, p. 212).

In light of this view of optimality models, there may still be value in IFT's application of optimal foraging models to information foraging. Perhaps food foraging models do not lend themselves to precise quantitative predictions about human behavior in information environments in general, but they can guide theorizing about information seeking. It may also be the case that particular sub-domains of information search are more amenable to a foraging-inspired perspective than others.

Music listening as foraging

Above I raised concerns about the fact that modern Web structure and behavior do not align well with the IFT framework of patch-based foraging, but there is reason to be believe that certain information seeking tasks are more amenable to such analyses. In particular, web-based music listening represents an ideal candidate, for two principal reasons:

- It is inherently a consumptive process. Unlike many Web search tasks, which are self-contained and have well-defined and discrete endpoints, music listening (looking past the narrow scope of searching for a particular song/artist/album) is much more akin to food consumption. A listener is never "done", but rather engages in repeated listening sessions ("foraging bouts") over time; experiences "hunger" and satiation; has "diet" preference; and can assign values to different music resources/patches. Operationalizing these constructs is not always straightforward, but the analogy is quite clear.

- There is intuitive reason to believe the space of music is patchy or clumpy. Although of much higher-dimensionality than the 3D spaces through which organisms search in physical space, we can expect clusters of self-similar resources (songs or artists) to be distributed through the environment in a patchy manner. This will need to be tested quantitatively, but aligns with a classical genre-based conceptualization of music. While there are certainly genre-spanning artists, we can typically clump an artist along with similar artists in a genre that is comparatively dissimilar from others.

Determining how people move among the patches in music artist space is of both theoretical and applied interest. On the theoretical side, we aim to contribute to the literature linking spatial and information search strategies, and increase our understanding of the cognitive processes under- lying exploration in high-dimensional, virtual environments such as that of musical artists. On the applied side, a patch-based characterization of listening patterns can provide a new granularity at which to describe listening trends and the evolution of music listeners' preferences. Rather than the low-level (and challenging) problem of predicting or recommending music at the song-by-song level, or more generic recommendations of artists based on a user's aggregated listening history, patch-based analyses can identify the most behaviorally meaningful units of listening activity and provide novel avenues for prediction and understanding in the context of music consumption. A patch-based recommender could, for example, generate predictions about when a user is ready to switch from exploiting one genre of music to exploring a new one, enabling it to not only provide the right recommendation, but do so at the right time.

Chapter 2: Digital Resource Consumption as a Foraging Process

This chapter presents work on ecologically-inspired analyses of resource consumption, specifically within the domain of online music listening. I first describe the latent factor modeling process that allows us to quantify distances between artists in an underlying "music space", then present analyses demonstrating that this space is in fact patchy (a necessity for the patch-based foraging perspective we apply here). I then detail our approach to characterizing patches in this high-dimensional space, and finally present two empirical analyses of digital music consumption, providing evidence of area restricted search in users' music listening and examining patterns of exploration and exploitation. I close with a discussion of the next steps in this ongoing research program.

Latent factor modeling of musical artists

To determine whether people move from music patch to music patch, we first need to know what patches exist in music space. In this section, I present methods for building, evaluating, and visualizing a latent feature space that allows for identification of these patches.

Previous work on modeling musical artists

Various authors have explored the question of how to model musical artists and listening behavior. Most commonly this involves modeling and prediction of playlists or explicit sequences of songs listened for the purposes of recommendation (Dias & Fonseca, 2013; Liebman, Saar-Tsechansky, & Stone, 2015; McFee & Lanckriet, 2011). There also exist some large-scale general studies of music listening patterns (Carneiro, 2012; Lambiotte & Ausloos, 2005) but to our knowledge no authors have explored the applicability of patch- or foraging-inspired models of music listening. Previous approaches to computing

distances between artists have leveraged methods similar to what use here, such as extensions of LDA (Zheleva, Guiver, Mendes Rodrigues, & Milić-Frayling, 2010), or analysis of playlist streams (Ragno, Burges, & Herley, 2005), and there are also more general efforts to achieve automated genre classification of music (Pachet, Cazaly, & others, 2000; Panagakis, Benetos, & Kotropoulos, 2008), as well as multi-label tagging (Mandel et al., 2011).

Dataset and modeling approach

A variety of features can be used to define a music feature space, including acoustic features of songs and textual descriptions (e.g. tags). The former is the most intuitive – different songs or artists are presumably similar insofar as they sound similar – but technically challenging (Berenzweig, Logan, Ellis, & Whitman, 2004; McKay & Fujinaga, 2006; Sturm, 2014a, 2014b), and acquiring the audio files for such analysis at the scale explored here is problematic. The latter approach, using textual data from tags or elsewhere, may be promising, but suffers sparsity issues in that most artists end up sharing very few or no tags, making similarity computations difficult.

Here, we instead leverage co-listening data, applying a collaborative filtering approach typically used for recommendation purposes that allows for calculation of similarity between artists. Note that for all analyses described here, we consider only artists, not individual songs. That is, we draw no distinctions between different songs by the same artist. Thus our approach cannot capture variation within an artist's recordings, but makes analysis more tractable. We have also observed qualitatively in human evaluations of our model (described below) that people generally have a sense of "the

sound" of a band, and that they are able to evaluate the similarity of two artists in a way our model effectively captures.

For these analyses we use the final version of the dataset described in Chapter 1, which consists of complete listening histories from ~167,000 users (for the period of July 2005 - December 2012) The 4.7 billion total listens in the raw data are represented as user-artist-timestamp tuples. Frequency distributions of per-user and per-artist listening activity follow long-tailed distributions (see Figure 3), and as such we limit our analysis to the A artists with at least 1,000 total listens distributed over at least 100 different users (112,312), and then consider only those users U with at least 1,000 total listens across the limited artist set (145,148).[12] The final data can be represented as a $U \times A$ matrix in which each cell is the total number of times user u has listened to artist a.

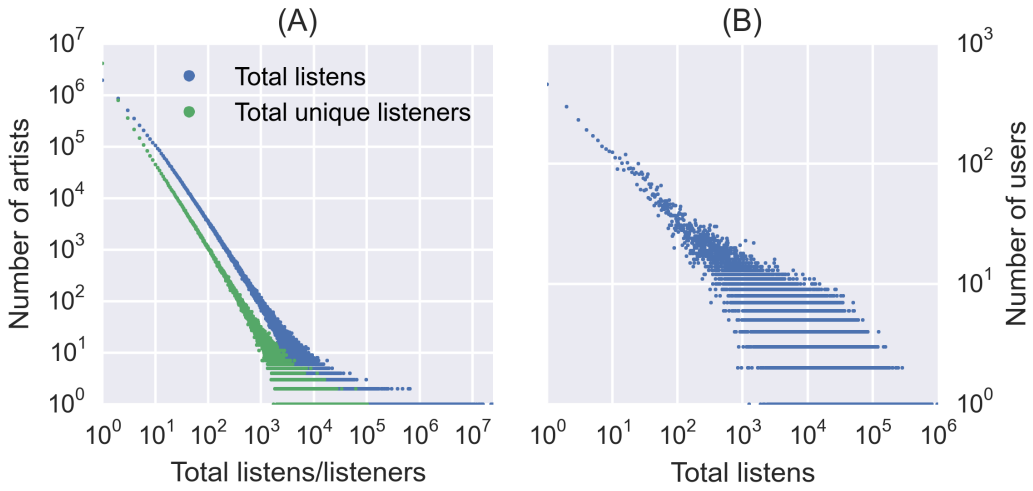


Figure 3: Frequency distributions of artist (A) and user (B) listening data, both on log-log scales. Y- axis shows the number of artists (A) and users (B) with the number of listen(er)s indicated by the corresponding point on the x-axis.

[12] Of course, many of these users will have at least some listens to artists not analyzed in our latent factor models, complicating analysis of their listening histories. To ameliorate this, all patch-based foraging analyses (described in subsequent sections) are limited to a further trimmed set of users (91,882) whose listening histories are at least 95% constituted by artists included in our latent factor models.

LDA topic modeling (Blei & Lafferty, 2009) is typically used to learn latent topics within a document corpus, where each topic is represented as a multinomial distribution over the unique words in the corpus, and each document as a multinomial distribution over *K* latent topics. LDA has been applied to a variety of contexts (including music, Zheleva et al., 2010), with various non-text-based entities taking the place of "words" and "documents". For our purposes, we ran two variants, the user basis model ("user-lda"), in which users were "documents" and artists were "words" (LDA is a bag-of-words model, under which word order does not matter, so documents are represented simply as rows from the $U \times A$ matrix), and the artist basis model ("artist-lda"), in which artists were "documents" and users were "words" (i.e. treating columns in *the* $U \times A$ matrix as documents. The first version was an attempt to derive a model in which latent features are human interpretable, each being represented as a distribution over artists, which could be interpreted as a type of "genre", but has the downside that it requires renormalization of the document-topic matrix. The second case does not generate directly meaningful latent topics (each topic is essentially a cluster of similar users), but does not require such manipulations. In both cases, we treat the multinomial distribution over topics for a given artist as a feature vector for all distance calculations.

The renormalization procedure for user-lda was as follows. Our modeling software generates two matrices, the user-topic matrix (in which each user is a multinomial distribution over topics) and the topic-artist matrix (in which each topic is a multinomial distribution over artists). We first multiply the user-topic matrix by the number of listens for each user, converting values to approximate occurrences (rather

than probabilities) of each topic for each user. The column-wise sum of this matrix then gives us the total number of occurrences of each topic over the full corpus. We can then multiply the topic-artist matrix by these counts to convert the matrix to approximate counts of the number of times each artist ("word") occurred under each topic. Finally, we can normalize this matrix artist-wise, rather than topic-wise, to achieve probability distributions over topics for each artist. This whole process amounts to converting the artist-topic matrix from vales of *P(artist|topic)* to values of *P(topic|artist)*.[13] In both cases, we treat the multinomial distribution over topics for a given artist as a feature vector for all distance calculations.

The second method we applied was a variant of matrix factorization (MF). MF begins with an input matrix in which rows represent users, columns represent items, and cells contain the ratings users have given to the corresponding items. As users typically have not rated all items (movies, products, songs, etc.), the goal is to generate a complete approximation of this sparse input matrix which can be used to predict ratings for items a user has not rated. This is accomplished by learning a $U \times K$ matrix and a $K \times I$ matrix (via one of several possible inference methods, see Koren, Bell, & Volinsky, 2009), in which U is the number of users, I is the number of items, and K is the number of latent factors (a model parameter), such that the product of the two matrices generates as close an approximation of the known values in the input matrix as possible. The columns of the $K \times I$ matrix can be used as low-dimensional feature vectors describing items (in our

[13] This method is of course approximate, and implementing our own Gibbs sampler for LDA (we employ Graphlab Create's LDA library) in which we save the final topic assignments for each token would allow us to calculate the relevant probabilities exactly. Our method, however, should not lead to any systematic errors.

case, artists) and permitting distance calculations. As commonly done, we constrain the learned latent factors to be greater than or equal to zero (non-negative MF, or NMF), which makes latent factor vectors more clearly interpretable and bounds the cosine similarity between any two vectors to the range [0 – 1].[14]

We employ a variant of NMF for implicit rating datasets, under which we binarize ratings (0 if the user has never listened to an artist, 1 otherwise) and treat the number of listens to an artist as a confidence value on that rating (Hu, Koren, & Volinsky, 2008). The method is well-established for similar contexts and greatly outperformed standard MF techniques in small-scale tests on our data. For all methods, we ran models for $K \in \{10, 20, \ldots, 200\}$, using standard parameter values.[15]

Feature space validation and selection

The challenging problem here is determining which of these candidate methods generates a "good" model of the latent space of musical artists, approximating ground truth of artist similarity. Internal measures of model quality (e.g. perplexity for LDA or RMSE of prediction for MF) are of little help here, as they measure a model's predictive capacity, not its approximation of a perceptual space. As we are interested in developing a space of musical artists that aligns with how humans perceive artist similarity, human judgments should be the basis of our evaluation approach. Various factors make this challenging (collecting a sufficient number of human judgments without massive resource

[14] For the analyses presented here, we experimented with several distance metrics, including cosine distance, Euclidean distance, and Jensen Shannon distance, and found that cosine performed best.
[15] LDA calculated using collapsed Gibbs sampling, with $\alpha = 50/K$ and $\beta = 0.1$. Implicit ratings MF calculated using alternating least squares with $\alpha = 0.01$ and regularization penalty $\lambda = 0.01$.

expenditure, variance in knowledge and opinion of artists, etc.), but we believe our novel method provides useful results.

Human raters (recruited informally from within our laboratory) performed a two-step computer-based evaluation process. They first completed a recognition task, indicating whether or not they were familiar with musical artists. Raters viewed artists in order of decreasing popularity (in terms of total listens within our dataset)[16] until they had built a set of approximately 200 recognized artists. Once this set of known artists was generated, we randomly generated triplets of artists from each rater's set and presented them in the form "Is *A* more like *B* or *C*?" (where *A*, *B*, and *C* are all artist names). Raters were instructed to "go with their gut", basing judgments on whatever knowledge was most salient for a given grouping of artists. Users were instructed to respond with *B* or *C*,[17] and could also skip questions that they felt they could not answer. We collected a total of 2,062 judgments across 11 independent raters.

A simple, rational model of the probability that a human rater will agree with a given latent space representation is given by Luce's choice axiom (Luce, 1959), under which the probability of selecting an item *i* from a set of *j* choices is

$$P(i) = \frac{w_i}{\sum_j w_j}$$

where *w* indicates the weight on or value of a particular item. We use the Softmax decision rule (Sutton & Barto, 1998), which adds a parameter γ that modulates the determinism of the choice rule (and must be estimated from the data): For sufficiently

[16] This obviously biases our method towards more popular artists, but is a necessary limitation, as most individuals have little familiarity with the long tail of less popular artists.
[17] In the first version of the task, users could say that the artists were equally similar, but we have excluded these responses in the analyses below and will use a forced choice in any future evaluation.

high γ it predicts deterministic selection of the higher weight choice, while very small γ values push the model towards random selection (regardless of the weights). Now the probability that a human rater will agree with the model is

$$P(agree) = min\left(\frac{\max\left(AB^{\gamma}, BC^{\gamma}\right)}{AB^{\gamma} + BC^{\gamma}}, 0.999\right)$$

where *AB* is the similarity between artist *A* and artist *B* and *AC* is the similarity between *A* and *C*. Because we use cosine distance, which is bounded [0-1], the similarity between two artists *A* and *B* is simply $1 - dist(A, B)$. Agreement with the model constitutes selecting the more similar option, hence using the max term in the numerator. We cap probabilities at 0.999 to allow for calculation of log-likelihoods (see below) and to avoid any complete determinism. This choice rule captures (a) the inherent stochasticity in human judgments, and (b) how this stochasticity should vary with the relative similarities between alternatives. When similarities are close to one another, humans should be near chance with respect to agreement with a feature space, and when one similarity is much greater, they should be near perfect agreement. The model also captures the fact that proportional differences in similarity are more important than absolute differences.[18]

Because such a framework allows us to calculate the probability of a response under a given a feature space (and value of γ), it affords a natural and intuitive evaluation process:

[18] Consider one case where *AB* = 0.9 and *AC* = 0.8, and a second where *AB* = 0.15 and *AC* = 0.05. In both the absolute difference in similarities is the same, but in the second the probability of agreeing with the model (selecting B as the more similar option) is higher than the first (75% vs. 53%, with γ = 1.0). This makes intuitive sense, as the proportional difference in similarities is small in the first case, but large in the second.

1. For each judgment $j \in J$ (the set of all judgments) calculate the *AB* and *AC* similarities and the probability of agreement, $P(agree|f,\gamma)$, according to feature space *f* and a decision rule with some γ.
2. Calculate the probability of each observed response under *f*, which is $P(j|f,\gamma)$ when a human agrees with *f*, and $1 - P(j|f,\gamma)$ otherwise. Agreement constitutes selecting the smaller of *AB* and *AC* (i.e. matching *f* in terms of whether *B* or *C* is more similar to *A*).
3. Determine the log-likelihood of *J*, the set of observed judgments, which is $\log P(J|f,\gamma) = \sum_{j=1}^{J} \log P(j|f,\gamma)$.

Under this framework, we simply select the feature space *f* and γ value that maximizes $\log P(J|f,\gamma)$.

For each feature space type (NMF, user-lda, and artist-lda) and each number of latent factors/topics ($K \in \{10, 20, \ldots, 200\}$), we optimized γ via maximum likelihood estimation, and then selected the best performing model of each type: NMF with 130 factors, γ = 1.05; user-LDA with 170 topics, γ = 0.46; and artist-LDA with 190 topics γ = 0.76. Figure 4 plots logistic regressions showing the relationship between predicted and observed human/model agreement, as well as mean observed agreement (with 95% CI) within each of five evenly-spaced bins of predicted agreement. The better a feature space, the closer it should approximate a perfect relationship between predicted and observed agreement (i.e. fitting the diagonal, indicated by the dashed line), and under visual inspection all three feature space types perform reasonably well. When examining the log-likelihood of each model, however, artist-lda (LL = -1200.50) is the clear winner over user-lda (LL = -1258.53) and NMF (LL = -1249.91). The proportion of judgments on

which humans and the model agree also follows this pattern: 69.0% for arist-lda, 66.2% for user-lda, and 66.5% for NMF.[19]

It is important to keep in mind when viewing Figure 4 that any given datapoint can appear in different locations for different feature spaces (i.e. both the predicted and observed agreement can differ from one space to the next). We include in the figure kernel density estimates of predicted agreement values for each model to illustrate this fact. This raises the question of whether we have based our evaluation on a "good" set of artist comparisons. Ideally we would like to collect responses across the spectrum of easy to difficult judgments (i.e. low to high predicted agreement), but of course we do not have this information ahead of time, and the only principled approach was to generate comparisons randomly as described above. We do not believe this issue can be avoided, but it is important to keep in mind that it may introduce some bias. The best solution is likely to simply collect more rating data, but the clear trends we observe suggest that more data is unlikely to meaningfully change our results.

A conceptually and methodologically simpler approach than what we have described here would be, for a pre-determined set of artist comparisons, have multiple human raters perform each judgment. This would allow us to calculate an empirical proportion of agreement for each judgment that could then be compared to a given feature space's predictions, thereby eliminating a major deficiency of our approach (because each rater's set of judgments is generated randomly, we cannot directly measure

[19] Although none of these values are particularly high, for many judgments we expect human raters to be near chance (generating comparisons randomly leads to more "hard" than "easy" judgments, see upper panel of Figure 4). Thus we do not expect that overall human-model agreement should be exceptionally strong, though this is a possible source of bias.

inter-rater variation for the same judgments). This, however, would require (a) more human raters, and (b) presenting each rater with the same set of judgments, with no guarantees that each rater would be familiar with all artists presented.

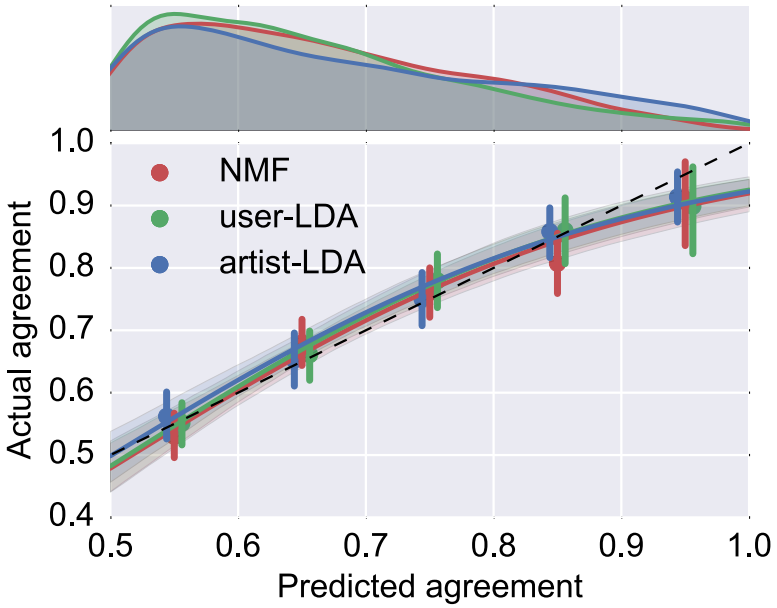


Figure 4: Logistic regressions of predicted versus observed agreement, with bootstrapped 95% CI, for the best parameterization of each model type. Also shown is mean observed agreement (binned, with 95% CI); x-axis locations indicate bin centers. Upper panel: Kernel density estimates of predicted agreement values under each model.

Is musical artist space patchy?

A preliminary question is whether the musical artist space is in fact distributed in patchy fashion, a necessary condition for the application of any patch-based analyses. There is a clear intuitive argument in favor of patchiness, namely that artists generally tend to fall into one of a limited set of well-defined genres. There certainly exist genre-spanning artists, but arguably this is the exception rather than the rule and artists are not uniformly distributed over the space of possible "sounds". Nonetheless, we here we examine the patchiness of the space quantitatively.

We begin by visualizing the space of musical artists (based on the 190-dimensional artist-lda model) in Figure 5 using t-Distributed Stochastic Neighbor Embedding (t-SNE, Van der Maaten & Hinton, 2008). This method achieves a 2D representation of data that maintains local distance relationships in many dimensions, although the absolute position of points is not meaningful. Colors indicate top-level genre classifications of each artist according to Gracenote,[20] but were applied post-hoc and did not affect model training or visualization in any way. Though necessarily an approximation, the visualization suggests a "clumpy" structure and that the patches visible in our artist space roughly align with established genres. Music of the same genre tends to cluster together, though given the fairly coarse classification we can visualize, it is not surprising that there are multiple separate clusters of the same genre. The fact that artist similarities are based on co-listening data does generate some peculiarities, however. For example, the elements of the small, highly heterogeneous clusters on the lower periphery of the figure tend to be united by sociological or geographic factors, rather than musical style.[21]

Examining the distribution of pairwise distances between artists (Figure 6) provides further evidence of a clustered structure. There is a preponderance of very large distances, rapid decay for smaller distances, and – critically – an increase for very small distances. This is consistent with a clustered feature space with high intra-cluster similarity but low inter-cluster similarity. For comparison, a uniform random feature

[20] http://www.gracenote.com/

[21] As concrete examples, the four small clusters starting at roughly "9 o'clock" in the image and moving counter-clockwise, respectively correspond to Christian, Scandinavian, Brazilian, and Eastern European artists.

space (i.e. with no clustering) of the same dimensionality shows a roughly Gaussian pairwise distance distribution (green dashed line in Figure 6). Such a uniform distribution is of course impossible under an LDA model[22], but is useful for visualizing how distances deviate from what we would expect under a uniformly distributed artist space (this is of interest here, because effective search strategies differ when resources are randomly versus non-randomly distributed). A more plausible null model is to randomly shuffle values of each latent dimension across artists, and then calculate the pairwise distances, the result of which is shown by the red dashed line in Figure 6. Under this null model, we still observe a preponderance of large distances and decay for smaller distances, but no uptick for very small distances (in fact, there is a large proportional drop for the smallest distances).

While the above analysis describes the "topology" of the latent feature space (i.e. the distribution of *possible* distances), a final question is whether our empirical data support that users actually navigate the space in a manner consistent with patchy distribution of artists. That is, we wish to test if the patchy structure of the environment actually constrains listeners' movement in that space in a meaningful way. To do so we plot in Figure 7 the probability distribution of "jump distances" across all users in our data. "Jumps" refer to anytime a user has switched from listening to one artist to another (this thus excludes consecutive listens to the same artist, or cases where the jump was to an artist not included in our latent factor model). Here we treat the empirical distribution of pairwise artist distances from Figure 6 as a null model, as random listening choices

[22] The Dirichlet prior used in LDA results in the topic distribution for a given item being concentrated on a relatively small number of topics (Blei, Ng, & Jordan, 2003).

should align with that distribution. In Figure 7A we see that user jumps follow the same qualitative form as the null distribution, with many large jumps (distances close to 1.0), fewer moderate jumps, and an uptick for very small distances, but the data is clearly much less skewed than the null model. Figure 7B offers a clearer picture of the data, plotting the probability of jump distances relative to the null model (i.e. empirical probability divided by probability for the null). This demonstrates that, although the qualitative distributions are similar, users are systematically more likely to make small jumps than larger jumps as compared to a null model.

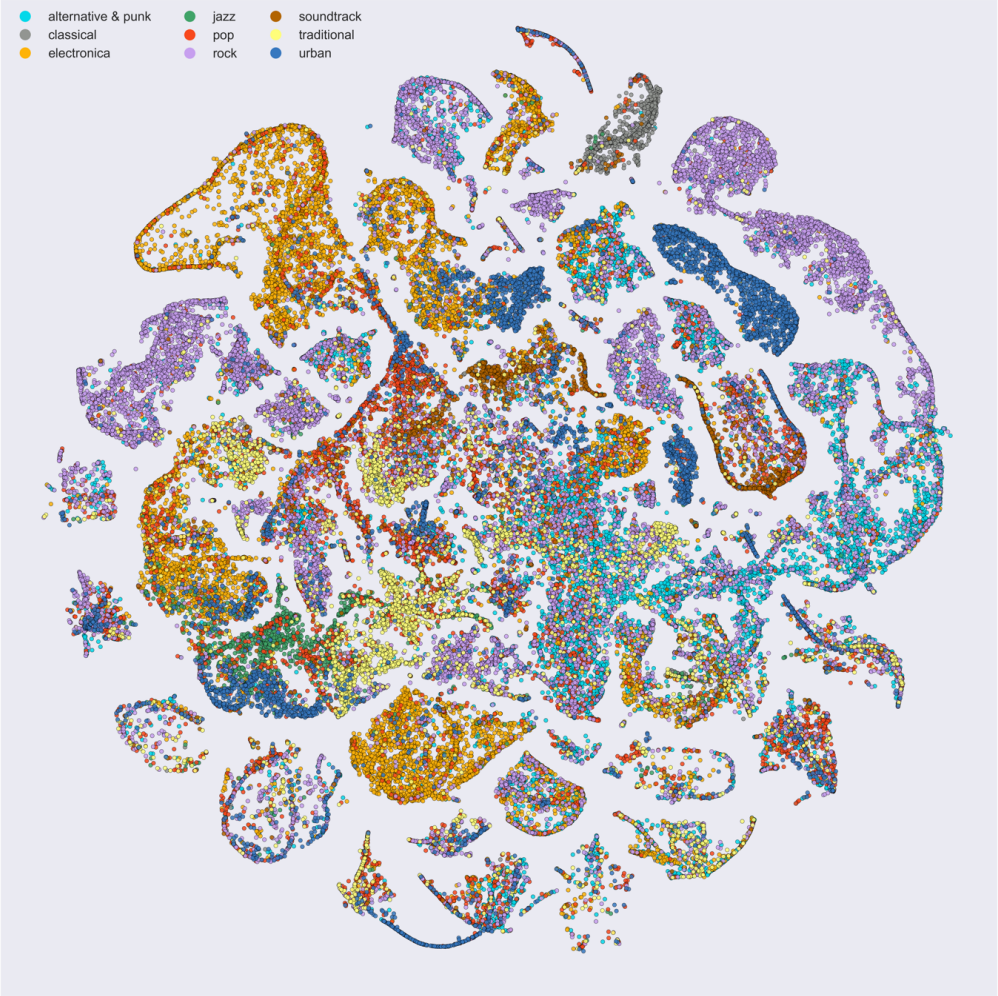


Figure 5: Projection of the 190-dimensional artist feature space to two dimensions using t-SNE. Colors indicate Gracenote genre classifications.

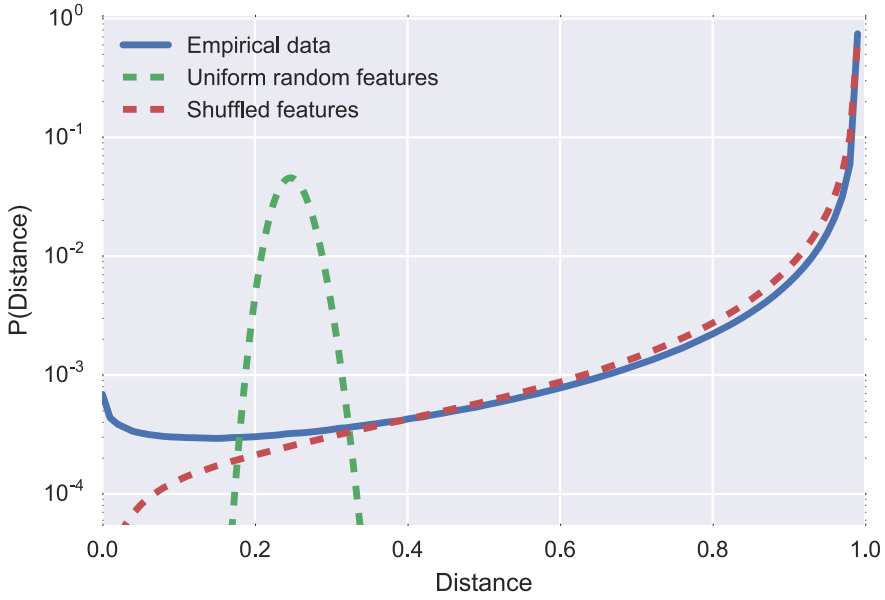


Figure 6: Probability distribution all possible pairwise artist distances under our feature space, as well as under two null models, log-linear scale.

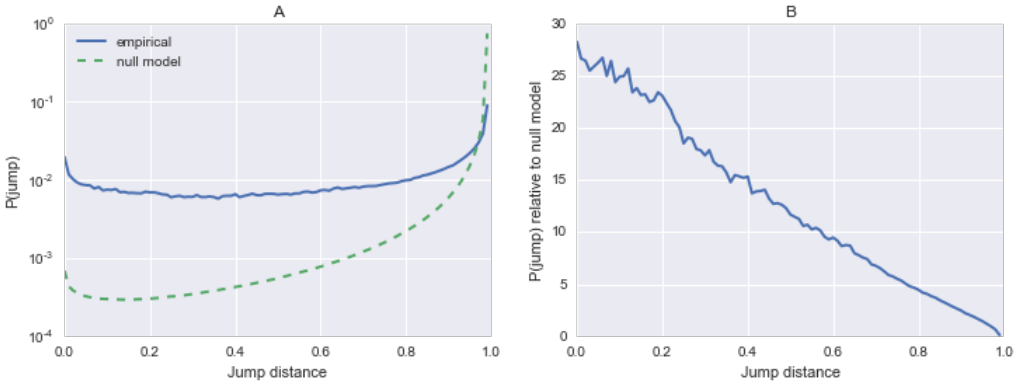


Figure 7: Comparison of empirical distribution of jump distances across all users and a null model, log-linear scale (A), and empirical jump distance distribution relative *to null model (B). (i.e., $P_{empirical}(jump) / P_{null}(jump)$).*

Taken together, this evidence suggests that the music artist space is indeed patchy, and that we have a reasonable basis for applying a patch-based foraging perspective to exploration in this space.

Characterizing patches

Having defined a feature space that permits distance calculations between artists and providing evidence of an underlying patchy structure, I now move to the problem of characterizing patches in this high-dimensional space. Four possible approaches to defining patches are the following:

- Patches as listening sessions: The conceptually simplest approach is to simply treat each discrete listening session as a patch. A session can itself be defined as a sequence of temporally contiguous listening, and a simple temporal threshold can be used to segment listening into sessions (i.e. if more than T minutes pass between two listens, we mark that the user has started a new session). Because it is very unlikely for any two sessions to be identical, this approach requires that we define session-level features (e.g. mean feature vector over listens) so as to cluster similar sessions if we wish to make claims about when users return to a given patch. The major disadvantage of this method is that is unable to capture any movement between patches that might occur within a single session.
- Patches as individual artists: Another simple method is to treat each artist as a discrete patch, such that every switch from one artist to another defines a new patch. This can be combined with the session-based model, such that a session break between two listens to the same artist constitutes a patch threshold (i.e. the user visited the same patch twice in a row). This method simplifies the patch return problem, as returning to the same patch is explicitly defined as returning to the same artist. The weakness in this case is that we cannot capture meaningful

"chunks" of listening that span different artists (e.g. listening to a playlist of thematically similar artists).

- Patches as localities in artist space: A more sophisticated and, we propose, more useful method that leverages our latent space representation is to define a patch as a locality in the artist feature space. That is, a cluster of similar artists in the feature space represents a patch. When segmenting a user's listening, this constitutes drawing a patch boundary every time a user moves between two artists whose distance is over some threshold (details below). We can roughly think of this threshold as a radius in high dimensional space around a given artist defining its neighborhood. Jumps between artists below this threshold constitute movement within a patch.[23]
- Patches as bouts of exploration or exploitation: While the previous method is the most directly analogous to the spatial notion of a resource patch, there are other possible *behavioral* patch definitions that it cannot capture. The most salient example is "shuffle" listening, where a user listens to a randomly ordered sequence of different artists within a constrained time window. Under the previous definition, such behavior would constitute multiple successive jumps between patches, but shuffle listening could arguably represent a qualitatively different kind of listening, namely *exploratory* behavior. Thus this fourth patch definition segments a user's listening into contiguous bouts of behaviorally consistent listening behavior, be it localized in the feature space (exploitation) or

[23] Note that this definition does not require that *all* listening within a patch fall within the neighborhood of a given artist. It is possible to observe a "chain" of artists, where each artist is similar to the previous one, but artists at the beginning and end of the chain are not necessarily similar.

shuffle-like behavior (exploration). This is the definition upon which we will focus.

Caveats

An obstacle to patch segmentation is the necessary incompleteness of our data. While we have extensive listening data for users, this likely does not capture *all* of their music consumption, and we are left with the problem of how to handle inter-session breaks. For instance, if a user is within a given patch, then stops listening for some substantial length of time (e.g. a day), then resumes listening within the same patch, do we consider that an instance of patch return or something else? Because we cannot know if or when users have explored other music (i.e. via sources not logged in their Last.fm profiles) during these breaks, we are left with inherent uncertainty about this question. Thus, rather than assign theoretical importance to breaks, we take the more parsimonious approach of ignoring them, basing patch segmentation only on the sequence of artists listened and disregarding the time between listens.[24]

Area-restricted search

A first question in testing the applicability of a foraging framework is whether we see evidence of area-restricted search. In real foraging environments, this typically manifests as an organism making small movements to stay in a restricted area upon discovering resources, and making longer-range movements to explore other areas when it has not recently encountered a resource (Bell, 1991). This simple strategy is effective in clumpy resource distributions, as the organism can use large jumps to explore the space, then use

[24] Thus we are not using the session-based patch definition in any capacity.

small jumps to exploit resources once it has encountered a patch (or more generally, an area with higher resource density).

If listeners regularly engage in ARS-like behavior, they should tend to remain within a confined locality of music space longer than expected under random music choice. Thus, one approach to testing for evidence of area restricted search is to check if listens that occur temporally close to another tend to be closer in the underlying feature space than listens that are farther apart. To answer this, we compute, for each listen L_i by user U, $cosine\_distance(L_i, L_{i+n})$ for $n \in \{1,2,3, \ldots 100\}$. Thus we know, for every song listened to by that user, the distance in music space between the i^{th} artist listened to and the artist listened to for each of the subsequent 100 listens. Averaging these vectors yields the mean artist distance between listens that are 1,2,3,…100 listens apart. Finally, averaging this result over all users, we have the result shown in Figure 8.

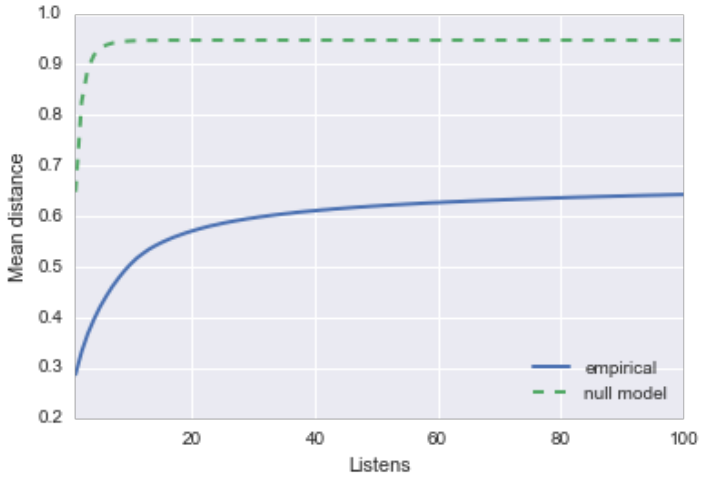


Figure 8: Mean distance between artists occurring X listens apart.

There is a clear effect of listens that occur closer together showing smaller distances, and for comparison we present a random choice null model. The null model is implemented using artificial random sequences of listening (see appendix for details), then performing

the same procedure as above. While there is a small effect of very nearby (roughly less than five listens apart) listens showing smaller distances, the null model does not show the gradual increase in distance in artist space evidenced in the empirical data, and we see that even artists 100 listens apart show systematically lower distances than the null model. One problem is that this result is surely biased by clustered listening to the same artist (e.g. listening to an album). Thus we repeat the analysis at the level of artist blocks (i.e. ignoring repeated listens to the same artist and only computing distance between subsequent artists, not listens). Though the difference is smaller, we observe a similar effect (Figure 9).[25] We utilize the same null model for comparison, computing mean distances between blocks in the random listening sequences.

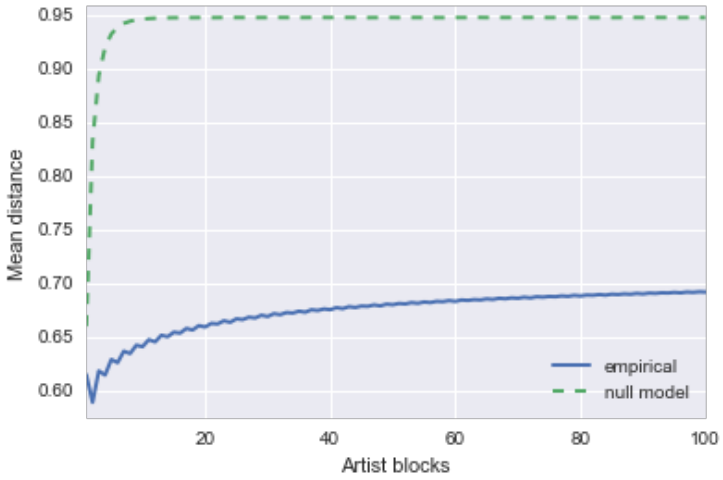


Figure 9: Mean distance between artists occurring X blocks (rather than listens) apart.

Finally, we can repeat the analysis at various temporal resolutions. To do so, we replicate the analysis, but utilize average feature vectors over different time resolutions (days,

[25] The scalloping observed in Figure 9 is an artifact of the fact that the two artists in adjacent blocks must be different, and thus cannot have a distance of zero, while blocks that are two apart can be from the same artist (e.g. listening to artist A, then artist B, then artist A again). The effect does not affect interpretation of the general trend, however.

weeks, and months). For example, to calculate the distance between month *M* and month *M+1*, we compute the mean feature vector for all listens in each month, then compute the cosine distance between these two feature vectors.[26] Figure 10 shows the result. Again, we compare results to the same analyses performed on our random choice null models. As before, we find similar monotonically increasing distance as a function of time, demonstrating that the observed behavior arises at multiple timescales.

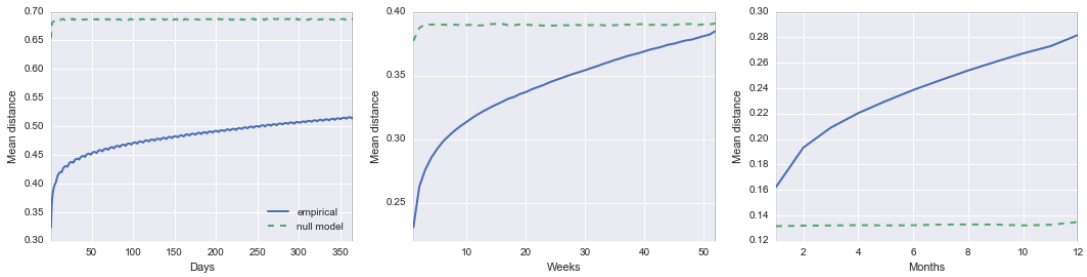


Figure 10: Mean distance between listens occurring X days, weeks, or months apart. Distances calculated between mean artist feature vectors within each time period. (Note that y-axis ranges differ across plots.)

Note that the distance at which we observe an asymptote in mean distance drops as we consider coarser time resolutions. This effect is likely due to the fact that, at coarser time resolutions, we are averaging feature vectors over more listens, and thus observing a kind of regression toward the mean of all possible feature vectors. Therefore, the distance between two mean feature vectors is necessarily reduced (on average). This is particularly evident when considering the characteristic distance under the null model; even choosing listens at random, distances between periods systematically drop as the length of these period increases (at a monthly resolution, distances between months are in fact *lower* under the null model than for the empirical data). Also of note is the scalloping

[26] This process is repeated for each user, and the plotted results show the average across users.

in the day resolution plot, which suggests the presence of week-level cyclicality in typical listening behavior (the observed valleys are exactly seven days apart). While seasonal patterns are not a focus of the current analysis, we ran a second analysis to confirm that this pattern was not artefactual, checking if users are more likely to return to a given artist at weekly intervals than at other times. For each artist listened by each user, we computed the probability distribution of inter-block temporal gaps, where a gap is defined as the time difference between the first listen of a listening block B_i and the first listen of block B_{i+1}. Figure 11 shows the mean result across all users, and while the general trend is of lower probabilities for longer gaps, those gaps corresponding to multiples of seven (indicated by vertical dashed lines) are disproportionately likely. Thus, users tend to exhibit week-level cyclicality in listening at the artist level, and this is likely driving the effect seen in Figure 10.[27]

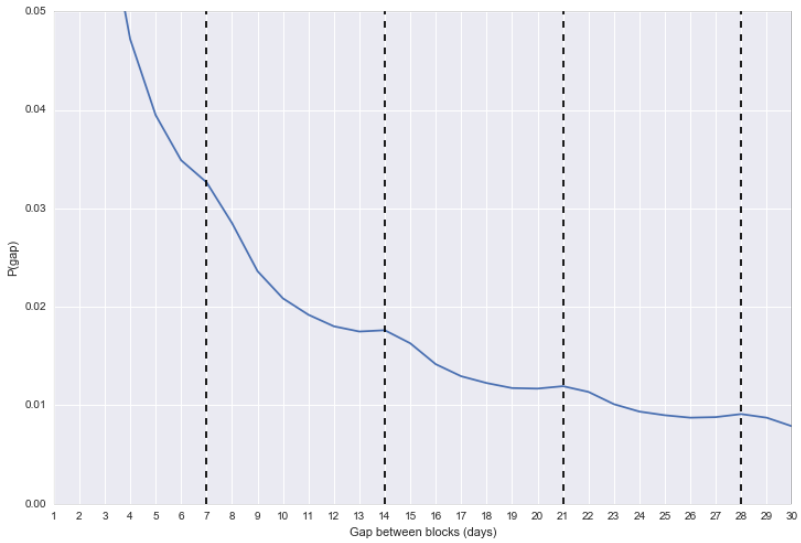


Figure 11: Probability distribution of temporal gaps (in days) between blocks of listening to the same artist. Vertical dashed lined indicate week-long periods (7, 14, and 21 days).

[27] Note this effect is not simply a "weekend effect" (or other effect driven by a particular day of the week, as our analysis considers gaps of a certain length, irrespective of the particular day of week.

Finally, we must question whether the observed trend at the coarsest time resolutions in fact suggests ARS, or a separate (but analogous) phenomenon. The observed effect at the individual listen, block, and day level arguably are consistent with ARS, but observing the same phenomena when comparing typical listening over weeks or months may indicate an effect of longer term exploration on top of short-term ARS. This could, alternatively, be considered a form of long-term ARS, and warrants more examination in future work. The next section examines patterns of exploration and exploitation in more detail.

Exploration and exploitation

Modulating the tradeoff between *exploiting* known resources and *exploring* in search of different or more valuable resources is a key behavioral challenge for foraging organisms (in particular when resources are distributed in a clumpy fashion). More generally, it is a framework for thinking about how humans and other organisms balance "the need to obtain new knowledge and the need to use that knowledge to improve performance" (Berger-Tal, Nathan, Meron, & Saltz, 2014, p. e95693; see also Christian & Griffiths, 2016; Mehlhorn et al., 2015).

Music consumption, too, can be viewed through the lens of exploration and exploitation: Is a listener remaining in a confined area of artist space for an extended period (exploiting) or trying out new or more diverse kinds of music (exploring)? By identifying periods of self-similar listening (i.e. listening within a locality of artist space) and the periods of shuffle-like listening behavior between them, we can characterize when users are engaging in either exploration or exploitation. We do so by implementing the last of the patch segmentation algorithms described above, first segmenting users'

activity into periods of exploitative versus exploratory listening, then clustering the exploitative segments such that we can identify when a user has returned to the "same" patch (see the appendix for technical details of this process).

Once the segmentation and clustering process is complete, each user's listening is represented as a sequence of listening bouts that are either exploratory/shuffle-like in nature, or are visits to one of a constrained set of patches (these patches are idiosyncratic to each user and cannot be directly compared from one user to another; see appendix).

Having formalized definitions of exploration and exploitation, how can they usefully inform our understanding of music consumption patterns? One central question is what governs switching between the two behaviors: Are there systematic patterns of when a listener is more or less likely to switch from exploration to exploitation or vice versa?

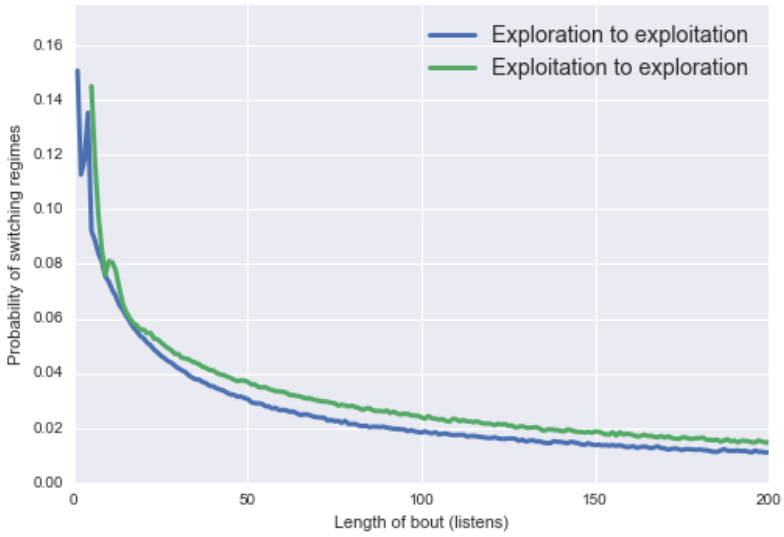


Figure 12: Probabilities of regime switching. Blue line shows the probability of switching from exploration to exploitation as function of how long a listener has been exploring.

Green line shows the probability of switching from exploitation to exploration as a function of how long a listener has been exploiting.[28]

To answer this, we calculate for each user the probability of switching from exploration to exploitation as a function of how long (in listens) the user has been exploring. We also calculate the corresponding probabilities for switching from exploitation to exploration. Figure 12 shows the mean results across users for switching between exploration and exploitation regimes.

Regime switches in both directions follow a qualitatively similar pattern of monotonically decreasing switch probability as bout length increases.[29] In other words, the longer a listener remains in one regime, the more likely she is to remain in that regime. This result emerges naturally from the observed distribution of segment lengths across all users, a long-tailed distribution with many short listening bouts and few longer bouts, for both regimes (see Figure 13). The most notable aspect of these results is the strong qualitative similarity between exploratory and exploitative listening behavior, a point discussed more below.

We also perform a conceptually simpler analysis at the artist block level. Here we characterize exploitation simply as the decision to continue listening to the same artist more than once, and an exploratory period as any made up of consecutive single listens to

---

[28] For the purposes of this analysis, the length of an exploit bout is the total amount of exploitation (even across multiple patches) occurring between two periods of exploration. An alternative method would be to consider only the exploitative listens within the single patch immediately preceding a regime switch. That analysis shows qualitatively similar results, with increased switch probabilities for lower exploration segment lengths.

[29] Note that when switching from exploration to exploitation, values are undefined for bouts shorter than five listens (our definition of patches enforces this, and also leads to the artefactual spike at x=4 listens in the exploration-to-exploitation plot). We also find a "bump" for bouts of around 10 listens. We hypothesize that this corresponds to the typical length of an album.

different artists. This serves as a sanity check on the previous result; if the effect is genuine (as opposed to an artifact of exploration/exploitation definitions) we should expect qualitatively similar effects at the artist level. As is visible in Figure 14, this is what we find. There is clearly a bias towards shorter bouts (for both regimes, but the difference is more pronounced for exploitation-to-exploration switches), which is a natural consequence of the stricter definitions of exploration and exploitation used here.

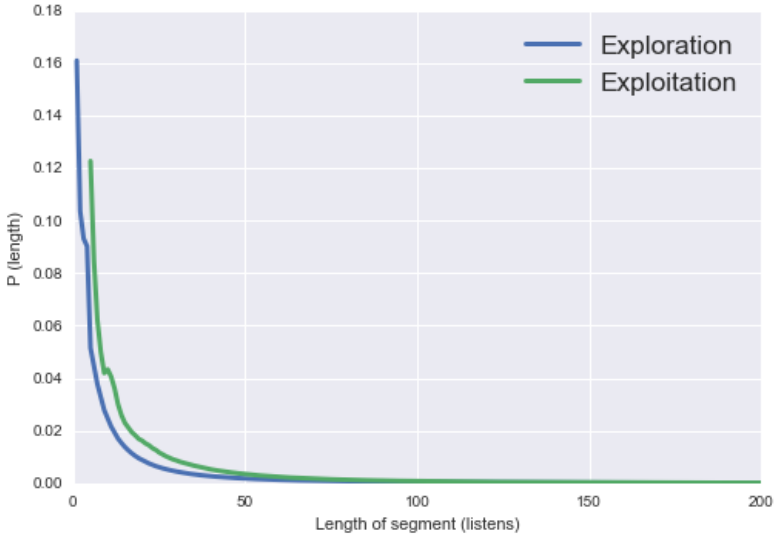


Figure 13: Probability distributions of listening bout lengths. Blue=exploration, green=exploitation.

The fact that we observe decreasing switch probabilities as a function of how long a listener has been in a given regime suggests a kind of momentum in listening, wherein the longer a user is in one regime, the more likely she is to remain in that regime. This is consistent – at least for exploitation – with a coherency maximizing interpretation of listening behavior (Riefer & Love, 2015), under which extended exploitation can lead to an increase in preference for the exploited resource. More work is required, however, to fully understand the mechanism at play here (and we must also consider the role of platform algorithms – i.e. recommendation algorithms – which presumably drive at least

some at least some of the listening behavior in our data). But what is particularly surprising about our results is that the observed trend holds for both exploration and exploitation, suggesting that there are some crucial differences between music foraging and animal foraging in physical environments.

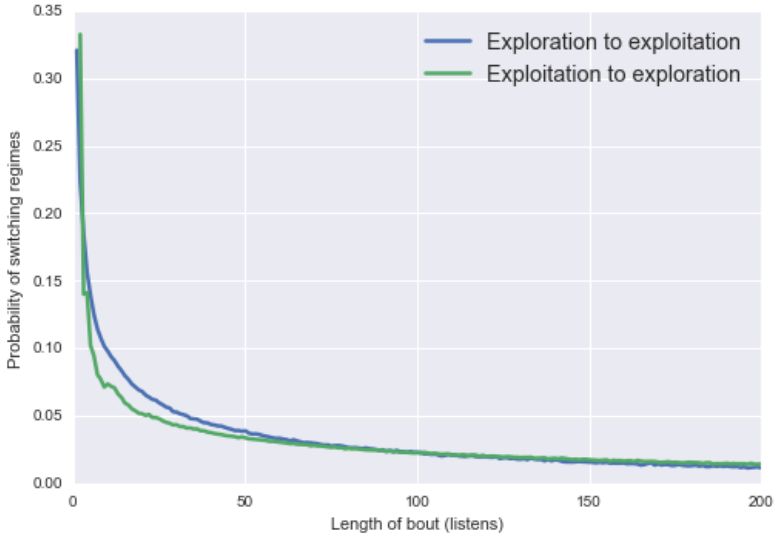


Figure 14: Probabilities of regime switching (artist block level). Blue line: probability of switching from exploration (1 or more consecutive single listens to different artists) to exploitation (listening to the same artist 2 or more times). Green line: probability of switching from exploitation to exploration.

When an organism is foraging for food, a mate, or any other resource, exploration is inherently costly. This is not to say that it should be avoided – on the contrary, exploration provides critical information about the environment that may enable future exploitation of valuable resources – but the exploration itself does not provide any direct benefit. Thus, while our exploitation results are unsurprising, we would expect diminishing returns on exploration behavior that would greatly reduce the number of long exploration bouts we observe. Instead, however, we see a surprising near-symmetry in exploratory and exploitative behavior. In other words, the fact that switching probability

drops with increasing exploitation bout lengths is not surprising, but the equivalent trend for exploration *is*.

The exploration-exploitation symmetry is further supported by simply examining how much time, overall, users spend in each of the two regimes. In Figure 15, we plot the probability distribution of users spending differing lengths of time exploring (measured as a proportion of overall listens). Under the patch-based definition of exploration and exploitation, the modal value is for users to allocate ~25% of their listening to exploration, but it is clear that there is substantial individual variation in how listeners allocate time across regimes, with many users (41%) spending over half their listening time exploring. Considering the stricter definition of exploration for artist blocks (even two repeated listens to the same artists constitutes exploitation), we unsurprisingly see behavior more strongly biased towards less exploration time, excepting a major spike for users with almost pure exploratory behavior (i.e. very rarely listening to the same artist two or more times in a row). The fact that we do not see this at the patch level highlights the importance of examining listening behavior at the patch level, as there clearly exist bouts of listening to similar music that cannot be identified by only considering the sequence of artists listened.

Three possible explanations for these high rates of exploration present themselves. The first is related to the fundamental differences between the space of musical artists and a physical environment. Web-based music consumption offers up an effectively limitless set of possible resources to consume, with little to no travel cost involved when moving from one resource to another (e.g. typing a search query, or simply waiting for the next song to be played by a music recommender). A music listener

can always "teleport" instantly to a known resource patch of high value, such that costs of exploration, insofar as they exist, are easily avoidable. In physical space, on the other hand, the costs of traveling in search of a new resource/patch are very real (in terms of physical energy and forgoing whatever resources are available in known patches). Further, because the space of musical resources is far larger than any one listener could hope to explore (and is constantly growing as new music is produced), she does not necessarily encounter the diminishing value of exploration (or, at least, encounters it at a much lower rate) that would be expected in a physical environment. There is always new music to discover.

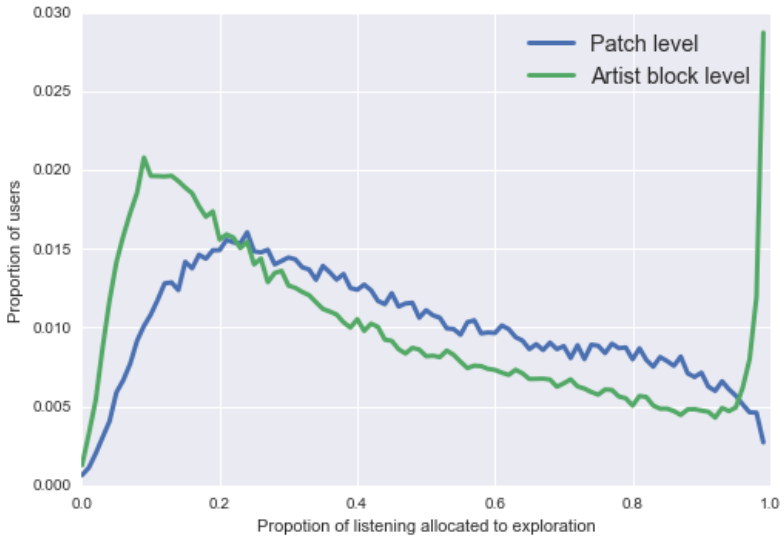


Figure 15: Distributions of time allocation (in listens) to exploration and exploitation. X-axis is the proportion of listening a listener allocates to exploring, and Y-axis indicates the proportion of users with that time allocation. Results shown for the patch-based (blue) and artist block-based (green) definitions of exploration and exploitation. Note that ~41% of users spend 50% or more of their time exploring (at the patch level).

A second possible reason for high exploration rates is that music exploration is a process akin to exploration in a multi-armed bandit problem (Mahajan & Teneketzis, 2008). In these problems, an individual chooses from a set of alternatives of initially

unknown value. Exploitation consists of repeatedly choosing an option of known value, whereas exploration is the choice to try an untested option. The important connection here is that untested alternatives still provide some value, which may be greater or less than the known alternatives (contrast this with an organism exploring a physical environment for new resources; the time spent exploring may not yield any reward, and certainly does not *until* the organism has found something). Similarly, a music listener is presumably receiving some hedonic value from music, even when in an exploratory bout.

Finally, a more speculative possibility is that music exploration may in and of itself hold value for listeners. Does the decision to put one's music library on shuffle, for instance, imply a goal of finding (or rediscovering) an artist/patch to exploit, or might the shuffle listening itself be of value? This is difficult to test empirically, but may partly explain the surprising amount of exploratory behavior observed in our data. [30]

Moving forward

This chapter has laid the groundwork for studying music consumption as a foraging process, developing a latent feature space permitting calculation of distances between artists, describing methods for evaluation of that feature space, and presenting some early results. In particular, we find evidence of area-restricted search and an interesting symmetry in patterns of exploration and exploitation among Last.fm listeners. The work presented, however, represents only a first foray into a foraging-inspired approach to studying music consumption, and I conclude the chapter with a discussion of some of the possible avenues for future work in this area.

---

[30] Anecdotal evidence suggests that people at least sometimes do this without the goal of (re-)finding a resource to exploit, but rather for the sake of shuffle listening itself.

Modeling approaches

While the foraging-inspired analyses presented here are of theoretical and practical interest, further work is necessary to establish whether or not listeners are in fact engaging in foraging-like behavior, a prerequisite to drawing any strong conclusions from the specific results regarding, e.g., ARS or exploration/exploitation tradeoffs. This can be best achieved via modeling of listening behavior under several explanatory accounts, including but not limited to a patch-based exploration model. We can then simulate listening behavior under these different models, and see which best align with the behavior in the empirical data. In this section I briefly describe and comment on several such candidate models.

Random walk models: A major assumption of the empirical results presented in this chapter is that users are actually "moving" through the musical artist space in a meaningful way (as opposed to say, simply "teleporting" to whatever music they wish to listen to independent of the underlying space of musical artists). While the analysis of the listening jump distribution across users shown in Figure 7 suggests that this is the case, a more robust approach is to actually develop random walk-inspired models of listening and compare artificial listening sequences generated under these models to the empirical data. This will allow us to establish whether or not simple random walks on the space are sufficient to explain listening choices. If so, the more involved questions around foraging strategies may not apply.

A random walk model in high dimensional space is somewhat more involved than on, e.g., a grid or 2D plane. The general approach would be the following:

1. Determine a start artist, A_0, for the random walk (this could be selected completely randomly, probabilistically using the artist popularity distribution, or using observed first-listened artists for users).
2. Determine a jump distance, D, by drawing from an appropriate distribution (which can take several forms, see below), or using a fixed jump distance.
3. Because a listener must jump to a discrete artist, determine the set of artists that are approximately distance D from A_0 (i.e. all artists for which $(D - \varepsilon) < Dist(A_i, A_0) < (D + \varepsilon)$, where ε is a sensitivity parameter), then randomly select A_1 from that set of artists (alternatively, simply select the artist for which $Dist(A_i, A_0)$ is closest to the sampled distance D).
4. Repeat steps 2-3 until a listening sequence of the desired length has been generated.

The interesting element of this model to modulate will be the distribution from which jump distances are drawn (or the particular jump distance used if using a fixed value). Reasonable candidate distributions are Lévy flight distributions (Baronchelli & Radicchi, 2013), power law distributions of jump distances observed in some real-world foraging contexts; the empirical distribution of jump distances observed in our own data; and a uniform distribution. The second option is equivalent to the method used to generate a null model for our ARS analyses (described in the appendix), and given that it generated very different ARS results than the empirical listening data, it is reasonable to assume that no simple random walk method as described here will be sufficient to explain listening behavior. Nonetheless, these represent an important class of models to examine and compare to other approaches.

Explore/exploit regime switching: The analyses presented on exploration/exploitation tradeoffs above are built on a second major assumption of this chapter: that listeners are in fact alternating between these two behavioral regimes. But the fact that listening can be *described* using the language of exploration and exploitation does not guarantee that these behaviors are in fact generating the sequences of listening we observed. Thus it is important to implement models explicitly including exploration/exploitation regime switching to compare to our observed results.

The primary parameters in this form of model will determine (a) when a (simulated) listener switches between the two behavioral regimes, and (b) what constitutes an exploratory versus exploitative behavior. With respect to (a), two approaches are of interest. In the first, regime switching is determined by a single parameter that can take three forms:

- A fixed probability of exploring, *P*, at each time step (such that the probability of exploiting is 1-*P*).
- Fixed probabilities of switching regimes at each time step (either a single switch probability, or separate probabilities for exploration-to-exploitation and exploitation-to-exploration).
- Probabilities of switching that are dependent on how long a listener has been in a regime. This is in line with the analyses on explore/exploit tradeoffs presented above, and requires that we define a mathematical relationship between length of time in a regime and the probability of leaving that regime. This could be drawn from our empirical analyses, though other methods are possible.

The second aspect of the model to formalize is how we define exploration and exploitation. The analyses presented above utilize an ad-hoc definition based on a cosine distance threshold of 0.2, taking steps between artists smaller than this threshold for exploitation and larger than this threshold for exploration, but further tests could be performed (see the future directions section below).

Patch-leaving models: Any patch-based account of foraging, in its simplest form, boils down to an organism deciding how long to spend in a patch before leaving in search of a new patch. As with the explore/exploit analyses, it is relatively straightforward to describe listening behavior under this paradigm (describing patches as localities in the musical artist space), but we have not established that this is actually a reasonable account of people's listening decisions.

The obvious choice here would be to build a model inspired by the conventional patch model described in Chapter 1, but the challenge with this and similar approaches is that they depend on being able to measure resource value, something to which we have no immediate access in the music listening case (see next section).[31] As a first approximation, however, we could implement a simplified patch model under which we assert that patches of different types have particular values (plausible estimates of these values could be derived by examining users' allocation of time across different musical genres/artists in the empirical data). While patch depletion obviously does not occur in the same manner here as in real-world foraging contexts (unlike a food resource, music can be "consumed" as many times as a listener wishes), a reasonable analogue here is

[31] This is true in the case of models inspired by optimal foraging theory, at least. Heuristic patch-leaving rules (giving-up time, item count, etc.) could also form the basis of models in this category, and are a possible avenue for future work.

satiation; Rather than instantaneous patch value dropping as the patch is depleted, value should drop as the listener becomes satiated, with the speed of satiation being a function of the overall value of the patch. This of course raises the question of how patch replenishment should be handled, and it will be necessary to develop plausible mathematical formalizations of both patch depletion and replenishment.

There are further details that demand further attention in such a model, including listeners' knowledge of patch values at any given time (a variety of the complete information problem from traditional foraging theory, see Stephens & Krebs, 1987) . If we assume music listeners are rate maximizing foragers, their predicted behavior will strongly depend on the accuracy and completeness of their knowledge both of the range of possible patch values and the instantaneous values of these patches (i.e. levels of depletion) at any given time. This and further model elements will require further development in future work.

Multi-armed bandit models: Another possibility mentioned in this chapter is the multi-armed bandit model, which can be thought of as a much simplified version of the problem faced by a foraging animal. As discussed briefly above, a multi-armed bandit problem is an extension of the one-armed bandit (i.e. a slot machine) where there exist multiple options (arms), each providing a different, typically stochastic, reward. A "player" of a multi-armed bandit problem thus must choose between exploiting by pulling an arm whose value (or distribution of values) is known or exploring by testing other arms.

The analogy to patch-based foraging should be fairly clear here, with the simplification that travel costs and other issues related to the topology of the underlying

decision space need not be introduced. Considering the discussion above regarding the low cost of "travel" in music space, it is reasonable to test models in which music listening is posed as a multi-, or possibly infinite-armed bandit problem. Further work is of course necessary to operationalize the value of the various arms, but this approach represents a useful avenue for modeling.

Diet models: Chapter 1 introduced the diet model in a review of information foraging theory, but we have yet to directly consider its applicability to music listening. Might it be the case that music listening can be explained under such a framework?

The simple diet model described in Chapter 1 deals with three main variables: resource value, handling cost, and encounter rate. Because music listeners can presumably listen to any music type whenever they wish (i.e. they can control the encounter rate of resources in a way foraging animals cannot), a very basic diet model for music listening is similar to a multi-armed bandit model (though only after the values of the different arms are known).

Diet models may be of greater interest and value if we consider more sophisticated versions in which the forager/listener's goal is not simply to maximize overall resource value, but instead to achieve a balanced diet with an appropriate ratio of different "nutrients" (see Stephens & Krebs, 1987, p. 116). How best to implement such a model is an open research question, but several aspects will be important to specify.

First, how are dietary preference / nutrition requirements defined? The simplest approach here may be to think of dietary needs as preferences. In other words, a listener has a certain set of preferences, codified as relative values of different artists, genres, or some combination of the two. Intuitively, we would expect these dietary needs to be

idiosyncratic to each user, but another possibility is that there exist either a constrained set of canonical interest profiles or perhaps even a globally defined "core diet" that can be realized in multiple ways.

Making sense of this "core diet" idea would require defining dietary needs differently than just preferences for particular artists or genres, as there is incredible heterogeneity in users' listening behavior. Thus it raises the possibility that "nutrients" in this type of model may be, for example, lower level musical attributes. For instance, listeners might seek a certain balance of high energy versus calmer music. Much work remains to be done to characterize such attributes and how they might fit into a diet model, but even descriptive work in this domain should be of practical and theoretical interest.

A further question is how to account for exploratory and exploitative behavior in a diet model. Do we assume listeners have explicit knowledge of their dietary needs, or that exploration is a necessary activity for a listener to establish her listening diet needs? Or should novelty and familiarity themselves be considered nutrients in a diet model? These and related questions will be necessary to answer if a diet model is to explain when and how listeners sample new musical resources.

Coherency maximization: Finally, coherency maximizing (Riefer & Love, 2015) presents a possible account of listening behavior under which listeners' preferences come to align with their choices over time. In other words, listening choices are not dictated by a pre-existing set of preferences, but those preferences instead emerge over time in response to earlier listening choices. Some existing work on music choice (Salganik, Dodds, & Watts, 2006) is consistent with such an interpretation. To implement such a

model, we would need to formalize both (a) what governs decisions early in a listener's history (i.e. before preferences have been well-established), and (b) the precise mechanism by which preferences emerge from listening choices.

A very simple first approximation would be to model music listening as a Pólya urn process (Johnson & Kotz, 1977). In such models, we imagine an urn containing differing numbers of balls of different colors (for the music listening case, each color would correspond to a given artist or patch/genre), and the decision process for each listening event consists of drawing a ball from the urn, then replacing it and adding an additional N (a model parameter) balls of the drawn color to the urn.

Such models are often used to model "rich get richer" processes, and we can imagine many modifications/variants of the simpler version described here (e.g. Do preferences eventually become fixed? Is each listen modeled as a draw from the urn, or instead does a draw determine the artist for each listening *block*, with the length of the block determined in some other manner? What initial ratio of balls exists in the urn?). This approach represents a simple process under which we can model listening choices to compare to more complicated foraging-inspired models.

Hybrid models and other possibilities: It is of course feasible to imagine that music listening is best explained by a hybrid model incorporating some subset of the mechanisms proposed above. For instance, animals foraging in real-world environments in many cases may have to simultaneously solve the problems of patch-based foraging and diet selection, and we can certainly posit a model incorporating both these elements. Likewise, coherency maximization might coexist with some of the other mechanisms

discussed, and various other combinations of the proposals may be feasible to implement and test.

There are also other modeling approaches that we will not discuss in detail, including cognitive-inspired models (e.g. models including a spreading activation component, where listening to artists in one region of music space may increase the likelihood of listening to nearby artists) and models that more explicitly incorporate temporal dynamics (wherein, for example, the likelihood of exploring versus exploiting is related to the time a listener plans to allocate to listening). Further work on individual differences in listening behavior will also be an important area for work.

I have been deliberately vague in describing what constitutes a resource in the models described here, as it is conceivable to develop versions at different levels, where resources are music types (i.e. genres or patches), artists, or even individual songs. We can even imagine multi-level models, wherein low granularity decisions regarding the style of music to be listened to are mediated by different processes than high granularity decisions like the particular artist or song listened.

Future directions

Perhaps the most critical question for moving forward is how to quantify resource value in the space of musical artists. Many ecological foraging models depend on knowledge of the value of a given resource to an organism (e.g. calories, preference level, etc.), but measuring value in this context is challenging. One possibility is to treat the total proportion of a listener's listening to an artist, either cumulatively up to some point in time T or across her complete listening history, as a proxy for the value of that artist to the listener. The latter equates to the assumption that a user's distribution of listening will

eventually stabilize to reflect her underlying preferences for different artists. There is a strong element of circularity to both of these definitions, however, and alternate measures of value should be explored.

A second avenue for further work is to refine our methods for identifying exploratory versus exploitative behavior. In particular, it is reasonable to assume that users at times consume collections of diverse music tied together by some higher level relations (e.g., a workout playlist, a movie soundtrack). It remains to be seen how common such behavior is, but our current approach would almost definitely classify this is as exploratory behavior, when it arguably should be thought of as exploitation.

Third, and most ambitious, is to apply our foraging-inspired perspective to prediction and recommendation. There is much work to be done before this becomes practical, but it presents many exciting opportunities. Already, our findings on area-restricted search could help a music recommender incorporate the typical time a user prefers similar content before being ready for something new, and an understanding of a listener's typical patterns of exploration and exploitation is of course valuable information when generating recommendations. Eventually, this work could lead to concrete predictions of, for instance, the most likely length of time after listening to a given artist or genre of music before a listener will want to return to that artist/genre.

I have mentioned only a few directions here, and work in areas as diverse as seasonality patterns, satiation dynamics, and user- or context-specific measures of artist similarity are all viable possibilities under this framework. In closing, however, the analyses presented here support the applicability of a foraging-inspired framework (while

also highlighting some important differences between animal foraging and music consumption) and open a broad variety of future research directions.

Chapter 3: Motivation in collaborative tagging

I now turn from content consumption to content organization on the Web. There are of course various manners of organizing digital content, but here I explore how it manifests in collaborative tagging systems. I focus specifically on the question of motivation, presenting two studies. The first considers the role of social imitation heuristics in tagging, while the second explores the case of "supertaggers", the minority of users who engage in tagging to a much greater extent than other users. But to lay a theoretical framework, I first review the literatures on tagging motivation in general, and, maintaining the ecological focus of this thesis, discuss evolved imitation strategies that may play in a role in tagging systems.

Why Do People Tag?

"Everything is miscellaneous" is the title and mantra of David Weinberger's (2008) ode to the "new digital disorder", in which he reviews the power of digital technology to upend our traditional notions of how information has to be organized. The power of this "disorder" lies in our capacity to organize digital content free from the restrictions of physical objects. Chief among these restrictions is that any given physical object can only be in one place at one, a painfully obvious but crucial observation. A collection of books can be organized by author *or* by subject *or* by binding color *or* whatever scheme the shelver desires, but of course the key word here is "or". We cannot multiply categorize the same physical object in space, and though libraries helped to circumvent this issue by keeping separate catalogs along different dimensions (title, author, subject, etc.), our ability to find a given book was always constrained by whatever curated infrastructure

was developed to organize the library.[32] Similar points can be made about grocery stores, music catalogs, or any other collection of physical artifacts.

In digital media, however, things are radically different. Content can be searched, sorted, annotated and otherwise organized along any dimension that can be digitally encoded, allowing for a multiplicity of organization schemes to exist simultaneously for the same content.[33] When we log on to Amazon.com to find a book, for example, we can search for or browse products using any of the many attributes Amazon records at practically instantaneous speeds. All the while, our searches can be enriched by second-order information from other users, such as when we receive purchase suggestions ("customers who bought the items in your cart also bought…") and other recommendations. Perhaps most importantly, none of this organization is static. Users can apply their own labels and organization schemes, and share those with others, constantly creating new ways to find and organize digital objects.

Algorithmic advances have been a driving force in this change, with full-text indexing by search engines, for instance, able to identify the most important terms in a document and allow us to find relevant items without knowing anything about the classical metadata, like author, date, or title, assigned to it. Many other technological

---

[32] Such curated systems are rife with the biases of their designers. Though rarely malicious, organizational schemes reflect the social, cultural, and technological milieu in which they are developed. Consider the Dewey Decimal System. Upon its introduction in 1876, eight of the nine top-level categories under its religion classification related to various aspects of Christianity, while all "other religions" were relegated to single section, and "Paranormal phenomena" warranted their own top-level classification within the "Philosophy and Psychology" classification. Many similar oddities exist in Dewey's system (see Weinberger, 2008).

[33] It would be disingenuous to claim that this transformation completely eliminates the types of biases we see in curated systems, however. The methodology used to organize content, even when it is based on as much data as possible and not explicitly curated, can still introduce bias. The debate over if and how Google's ranking algorithm unfairly points traffic to already popular sites Menczer, Fortunato, Flammini, & Vespignani (2006) is an excellent example from the digital age.

changes could be cited here, but another driving force has been largely human-powered: Collaborative tagging. In collaborative tagging systems, which have been implemented for content as diverse as music, academic papers, photos, and Web bookmarks, users make independent decisions about textual labels to assign to content. These individual tagging decisions are then aggregated, resulting in an emergent, bottom-up classification scheme commonly referred to as a folksonomy. These systems are my focus here.

It is easy – and unfortunately common – to think of these tagging systems in "black box" terms: The users of a system make their tagging decisions, and with enough users, the "wisdom of the crowd" kicks in, and you are left with a functional folksonomy. In many cases, both when designing and studying tagging systems, little thought is paid to users' motivations for tagging, or else the reason for tagging is simply assumed (tagging an item for future retrieval being the most common assumed motivation). But the question of why users tag is a crucial one, with implications for the emergent organizational schemes that system designers aim to generate and researchers to study.

Many possible motivations exist for taggers, from organizing one's music collection to labeling Web bookmarks to be found again later, to name but a few. But of course these are *collaborative* systems, thereby introducing an array of possible social motivators, as well. Thus there exists no short, single answer to why people tag, as the research points to it being a behavior heavily influenced by a variety of social and environmental factors. In what follows, however, I attempt to synthesize the various strands of research addressing this question so as to provide as clear an answer to this question as possible.

This section will begin with an overview of collaborative tagging systems, and from there move on to a review of the literature on tagging motivation, discussing taxonomies of tagging motivations and the approaches used to elucidate and study these motivations. Finally, I review some of the under-explored approaches to studying tagging motivation.

What is Tagging?

Before examining the question of why people tag, I begin with a general overview of social tagging, in which I define the key concepts surrounding tagging systems and behavior, and provide a variety of examples of different collaborative tagging systems. In so doing, I will highlight both the key attributes that social tagging systems share, and the dimensions along which they most commonly vary.

I will not be focused on motivations for tagging in non-social environments. This would include tagging systems that built solely as personal information management tools, such as tagging files on one's local computer operating system or other software that allows for content tagging (as in, for instance, tags assigned to documents in the Evernote note-taking software, or ID3 tags assigned to songs in a personal digital music library). To be clear, such individual-centric, personal information management strategies do appear in many social tagging environments, and as such will be discussed to some extent.

Definitions: Collaborative tagging is, under the simplest possible definition, the labeling of objects in a shared information space. The details differ from system to system, but all social tagging systems enable users to assign tags to resources that are publicly visible to other users. Most commonly, distributions of tags are visualized using what is referred to

as a "tag cloud", which shows the relative frequencies of the tags users have assigned to some particular resource or within some particular timeframe (Figure 16). "Resource" is a deliberately vague term, meant to capture any of the many types of digital objects tagging systems have been designed to organize. As I will discuss below, these include items as diverse as images, songs, books, academic papers, and Web bookmarks

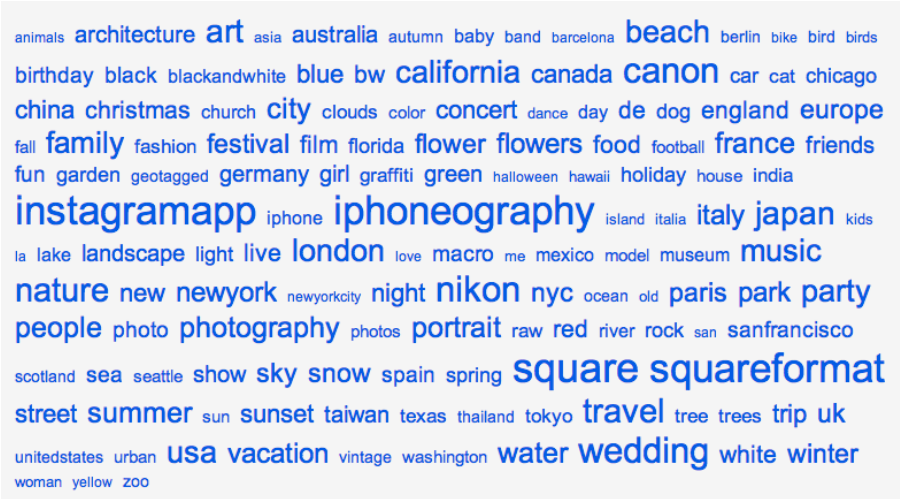


Figure 16: Example tag cloud showing the most popular tags across all time and images on the photo-sharing site Flickr.com. The font size of each tag indicates its relative popularity. Image from http://www.flickr.com/photos/tags/.

The emergent structure that develops as more and more users tag content is commonly known as a *folksonomy*. The term, typically attributed to Thomas Vander Wal (2007), is a portmanteau[34] of "folk" and "taxonomy", capturing the fact that is a taxonomy-like structure collectively generated by a group of people. There are some critical differences between a folksonomy and a traditional taxonomy, however, which I will mention below.

[34] The terms combines the concepts "folk" ("people"), "taxis" ("classification"), and "nomos" ("management") (Gupta, Li, Yin, & Han, 2010)

The formalization of a folksonomy most common in the literature was introduced by Hotho, Jäschke, Schmitz, & Stumme (2006b). They define a folksonomy, *F*, as a quadruple in the form $F := (U, T, R, Y)$. *U*, *T*, and *R* are finite sets defining, respectively, the users, unique tags, and resources in the folksonomy. *Y* is a ternary relation between the other three elements, $Y \subseteq U \times T \times R$, representing the tag assignments in the system (alternatively known as "annotations"). Stated simply, this formal definition captures the fact that there exist triples describing the assignment of a particular tag to a particular item by a particular user. They also define the term "personomy", which is a structure in essentially the same form, but only describing the tag assignments of a particular user, and the associated tags and resources.

One element that is absent from this definition is time; that is, *when* tag assignments are made. Though surprising at first glance, excluding time data for annotations from this formal definition is consistent with the fact that such temporal data is not typically made available to users of tagging systems. That is, users can view data on tag distributions for particular items, but little or nothing about how that distribution has changed over time. By extension, it is difficult or impossible in most contexts for researchers to incorporate temporal information into analyses of tagging systems.[35]

Folksonomy Versus Traditional Classification: Folksonomies differ in fundamental ways from traditional taxonomies and related classification schemes, as one might find in

[35] This is not always the case however. On Last.fm, for example, although no aggregate data on tagging patterns over time is made available to users, individual users' tagging histories are recorded in their profiles. By crawling and aggregating such data, temporal aspects of tagging can be revealed (Lorince & Todd, 2013). In other cases, researchers have developed and built their own tagging systems (Hotho, Jäschke, Schmitz, & Stumme, 2006a) or worked in conjunction with system designers to obtain temporal tagging data (Floeck, Putzke, Steinfels, Fischbach, & Schoder, 2011).

library or record store. Though they can serve a similar functional role (providing a structure for browsing, searching, and otherwise organizing some collection of resources), they achieve this end in very different ways. The distinction, stated as simply as possible, is that a taxonomy is a "top-down" structure, while a folksonomy is "bottom-up". Traditional taxonomies are made up of a discrete set of pre-existing, often expert-generated, categories to which resources are assigned (examples include the Dewey Decimal System, the All Music Guide's genre classifications[36], and Amazon's product category structure[37]). A folksonomy, on the other hand, is the aggregation of many individual tagging decisions without any (or at least limited) imposed structure. Availability of other users' tagging decision can of course generate feedback effects, and characteristics of tagging system design (e.g. tag recommendations, number of socially generated tags displayed, etc.) certainly influence users' tagging decisions. Thus tagging decisions are not made independently in the strict sense of the term, but the contrast with a formal categorization system, where a small number of experts (relative to the size of the user base) take sole responsibility for content classification, is clear.

Social tagging gained popularity in the mid-2000s, earning praise in the media and popular science literature (Shirky, 2005; Sterling, 2005; Weinberger, 2008). This came alongside an explosion in academic research on tagging systems, spearheaded by Golder and Huberman's (2006) classic study of tagging patterns on the social bookmarking platform Delicious. Proponents of social tagging cited benefits both at the system level and for individual usability. The arguments at the systems level were based

---

[36] http://www.allmusic.com/genres
[37] http://www.amazon.com/gp/site-directory/ref=sa_menu_top_fullstore

principally on the notion that the ability to multiply classify resources could allow for easier retrieval, providing multiple points of access to a resource. For example, different user communities might use varying terms to describe the same type of music, but if that music is classified in a fixed category within a formal taxonomy, the ability of users who describe it differently is inhibited. By allowing users to freely tag resources, the classification scheme becomes flexible and effectively allows resources to be classified in as many ways as there are users. Taxonomies often implement multiple classifications[38], such that an item appears in multiple categories, but this still has to be implemented in a top-down manner by a system administrator. A folksonomy, on the other hand, is free to mutate in response to the tagging decisions of its users, without any oversight. Also, proponents have argued, social tagging is relatively low cost. By "crowdsourcing" classification, the labor-intensive process it distributed among many unpaid users. Examples of user-centric benefits cited in the literature include the alleviation of difficult categorization decision (users can apply multiple tags to a resource, rather than assigning it to a single category) and the social sharing features built into most tagging systems (Heckner, Heilemann, & Wolff, 2009; Sinha, 2005).

The enthusiasm for social tagging was not universal, however, with critics in both the academic and non-academic domain (Davis, 2005; Merholtz, 2005) pointing to deficiencies in social tagging systems. Macgregor & McCulloch (2006) in particular

---

[38] The Open Directory Project (http://www.dmoz.org/) provides an example. The site maintains a taxonomy of Web pages, organized in a hierarchical structure, but allows for link between leaves of the hierarchy. For example, the category "computer science" is categorized under the top-level category "computer", but is also appears (as a link) under the top-level category "science". The Library of Congress, assigns items a single classification (similar to the Dewey Decimal System), but supplements this with subject headings, additional metadata not directly tied to the hierarchical taxonomy.

address strengths of formal taxonomies that social tagging systems cannot emulate. Formal taxonomies and other controlled vocabularies can, for example, control synonym use by ensuring there is only one term for any given concept, discriminate between homonyms (Java the country versus Java the programming language versus java the bean), and relate similar terms (typically via hierarchal structures). These are all difficult tasks for folksonomic systems to accomplish, but even the staunchest supporters of controlled vocabularies recognize that such systems cannot keep up with the exploding quantities of digital information available today: "the proliferation of digital libraries and the Web precedes the ability of any one authority to use traditional methods of metadata creation and indexing" (Macgregor & McCulloch, 2006, p. 3).

What becomes clear in reviewing this literature is that both proponents and critics of social tagging tend to envision a hybridized future, "a middle ground, somewhere between the pure democracy of bottom-up tagging and the empirical determinism of top-down controlled vocabularies" (Wright, 2004). What exact form this middle ground should take is debated, but plausible is one "where end-users could freely create, adopt or reject terms stored in a distributed repository that gets administered by a representative authority that 'owns' the vocabulary: normalizing terms, identifying and mapping semantic relationships (determining synonyms, related terms, parent-child relationships and such)" (Ibid.).

Types and examples of tagging systems: Collaborative tagging functionality can, in principle, be applied to any collection of digital content. As such, examples of tagging systems abound, for content as diverse as Web bookmarks (Delicious.com, StumbleUpon.com), music (Last.fm), academic papers (CiteUlike.org, Bibsonomy.org,

Mendeley.com, Connotea.org), programming questions (StackOverflow.com), images (Flickr.com, 500px.com), video (YouTube.com), consumer products (Amazon.com), books (librarything.com), and more. While they all fit the broad definition presented above, there is much variation in exactly how these and other systems implement collaborative tagging. Based on an analysis presented in one of the first through reviews of the social tagging phenomenon (Marlow, Naaman, Boyd, & Davis, 2006), I review the principal dimensions along which social tagging systems differ.

Marlow, et al. describe two organizational taxonomies, one for system design and attributes, and the other for user incentives. I will return to the user incentives they describe below, but here focus on the system attribute taxonomy. The authors' taxonomy is made up of seven components: Tagging rights, tagging support, aggregation, object type, material source, resource connectivity, and social connectivity.

- *Tagging rights* deal with which users can tag particular content. In some systems, the norm is for users to only tag their own content, while other systems allow many users to tag the same resource. On Flickr, for instance, the default settings for photos only allows for the user who uploaded an image to tag it, while Last.fm allows the same artist, album, or song to be tagged by any number of users. This distinction maps onto Vander Wals's (2005) distinction between "broad" (allowing multiple users to tag the same content) and "narrow" (in which the tags on a given item come only from the item's creator/uploader) folksonomies. Note that the distinction here is not dichotomous; It it possible to implement permissions systems such that the content creator can decide who can tag his or

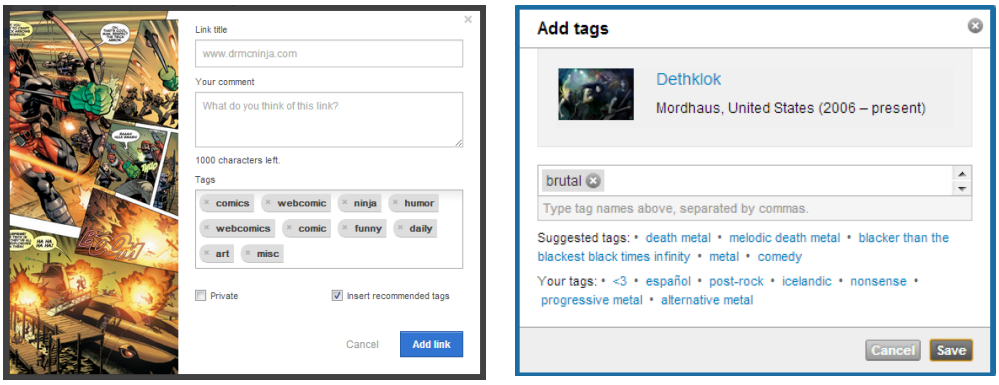
her resources. For example, a Flickr user can set the privacy settings for an image such that members on her friends list can tag the photos she uploads.

Tagging support *corresponds to the specifics of how the user tagging interface is implemented, and Marlow et al. describe three general classes (which are not mutually mutually exclusive): Blind tagging, viewable tagging, and suggestive tagging. In a blind blind tagging system, the user is unable to see how others have tagged a particular item item when she elects to tag it (e.g. Delicious). Viewable tagging, such as on Last.fm or Last.fm or Bibsonomy, permits the user to see the existing tags on an item when adding adding her own tags. Finally, suggestive tagging recommends tags to a user through the through the tagging interface. These suggestions can be generated from how others have others have tagged the resource, from a user's own tagging history, or from automatic automatic analysis of the content tagged it itself.*

- Figure 17 shows two example tagging interfaces
- *Aggregation*: There are two principal methods of aggregating the tag decisions of users (when multiple users are allowed to tag the same content), which Marlow et al. (2006) refer to as the "bag-model" and the "set-model". The bag model allows for multiple users to assign the same tag, which enables visualizations like that in Figure 16, where the most popular tags for an item are easily identifiable. The set model, on the other hand, only allows a given tag to be applied to an item one time, as is the case with Flickr.

- *Type of object* simply reflects what type of resource users tag on the system, and includes any of the examples cited above (music, bookmarks, and so on).
- *Source of material:* Whatever kind of content it may be, the resources tagged by users of a tagging system typically come from one of three sources. The content is either supplied by users (i.e. the content is user-generated, as on Flickr or YouTube), open to any Web based material (as on social bookmarking systems like Delicious), or is supplied by the system. This last category can refer to systems that index non-user-generated content, but allow tagging (like Last.fm), or systems that explicitly present content to users for the purposes of tagging. An example of the latter is the ESP game (Von Ahn & Dabbish, 2004), which leveraged players' tagging decisions to determine appropriate labels for images.
- *Resource connectivity:* Various types of intra-item connections can exist between the resources in a tagging system, independent of how they are tagged. These connections typically take the form of links (e.g. between Web pages) or groupings (e.g. sets of Flickr photos). Some tagging system have no support for explicit links between items.
- *Social connectivity*: The different classes of possible social links (for those systems that have an explicit social network component) roughly follow that of other, non-tagging social systems. Thus social connectivity can come in the form of links and groups. Links can be symmetric (like Last.fm or Facebook friendships) or asymmetric (like Twitter's follower model). These links can also be typed or untyped. Flickr provides a good example of typed social links, where a user can distinguish between – and assign differing permissions for viewing,

tagging, and commenting on photos to – contacts, friends and family, and the Flickr community at large. Untyped social links correspond to standard friendship or contact relationships on social networks. On many systems, users can also join groups built around a shared topic or interest.



Flickr, Last.fm, and other systems have this functionality.

Figure 17: Two example tagging interfaces. On Delicious (left) recommendations are drawn from other users' tagging decisions. Delicious would still fall under the "blind tagging" category, however, as the user cannot view the existing distribution of tags for a particular item. On Last.fm (right) recommendations are drawn from other users (the most popular tags for the item among all users) and from the user's own tagging history (the most commonly used tags by that user). Last.fm falls under the "viewable tagging" category, as users can view the tag clouds for any item they wish to tag.

Having overviewed the key attributes and types of social tagging systems, I can now move on to address the core question of what motivates users to tag in these kinds of systems. As I proceed, many, though not all, of these system attributes will prove to be

important factors in understanding what motivates people's behavior in collaborative tagging environments.

Basic taxonomies

If anything should be clear from the previous section, it is that tagging is not a simple process than manifests uniformly across different systems. Thus we cannot simply ask why people choose to tag in the abstract (though that is surely part of the process), but must also consider how motivations differ and are mediated by the kinds of system attributes discussed above. Do Flickr users, tagging their own photos, do so for different reasons than Last.fm users who join the large crowd of existing taggers when they assign a label to a musical artist? How and why do people tag differently when they have more or fewer friends within the tagging system's social network? These are the types of specific questions we must keep in mind as we move forward, all the while attempting to synthesize their answers into some sort of general framework. To do so, I begin by discussing the basic taxonomies of tagging motivations that exist in the literature, review the methods for arriving at and studying such classifications of motivation, and end with a discussion of some of the problems with existing approaches to studying tagging motivation.

A survey of social tagging techniques and systems (Gupta, Li, Yin, & Han, 2010) presents ten proposed user tagging motivations: Future retrieval of content, contribution and sharing, attracting attention to one's own resources, play and competition (as in the ESP game), self-presentation, opinion expression, task organization, social signaling, earning money (e.g. when a user tags resources for pay on a service like Mechanical Turk), and technological ease (i.e. when software greatly reduces the effort required to

tag content). Their work does little to connect or contextualize this long list of motivations, but I present it to point out that there are many possible reasons for which a user might tag a resource. Furthermore, it is reasonable to expect that many of these motivations could exist simultaneously in a user, or that a user might have differing motivations in different contexts. Let us dig deeper, then, and explore some of the well-developed attempts to tease apart different reasons people tag, and how they relate to one another.

Self versus others: A fundamental aspect of determining why a user has assigned a particular tag is the tag's intended audience, with the fundamental distinction being whether the tag is the for the benefit of the user who assigned it or other members of the tagging community. That is to say, does the tag play some functional role for the user herself, or does it have a socially directed purpose? This distinction maps onto what Heckner et al. (2009) refer to as personal information management (PIM) and resource sharing. PIM (Jones, 2007; Teevan, Jones, & Bederson, 2006) is a major topic of research in its own right, particularly within the library and information science community, and broadly covers "the activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use, and distribute the information needed to complete tasks…and fulfill various roles and responsibilities" (Jones, 2007, p. 453). For the moment, though, it suffices to say that self-directed tagging is generally construed as being for some sort of PIM goal, with future retrieval (i.e. tagging a resource find it again later) being the most commonly cited motivation. We can point to many examples of tagging scenarios that fall under the umbrella of PIM (varying in how aligned they are with the "future retrieval" motivation), such as when a Last.fm user tags music with the

word “workout” to have exercise-appropriate music at his fingertips, when a Delicious user maintains a “to-read” tag to mark Web articles she wishes to read at a later time, or when a Flickr user tags photos with the names of family members so as to easily find all the pictures of a particular person.

Contrast this with resource sharing, in which a tag is applied not to organize or otherwise manage one’s own resources, but instead as a social signal to other users. Under Heckner and colleagues’ definition, resource sharing generally takes the form of a user applying tags to facilitate other people’s finding of his or her content. Canonical examples of this are when Flickr users tag images with their own username as a means of self-promotion, or when YouTube users assign multiple, synonymous tags to uploaded videos to increase the chances that they show up in other users’ search results. The authors refer to this second class of social tagging as “overtagging”, a phenomena introduced in earlier work (Heckner, Mühlbacher, & Wolff, 2008). In the 2009 study, the authors investigated their proposed categories through a survey performed via the Mechanical Turk platform. Qualitative coding of participant descriptions of recent tagging episodes were consistent with PIM/sharing dichotomy, and across the four tagging platforms, seemed to align with the broad/narrow folskonomy distinction proposed by (Vander Wal, 2005).Tagging on Flickr and YouTube, clear narrow folksonomies (where users predominantly tag their own content) tended to be motivated by resource sharing, while broad folksonmies (Delicious and Connotea) were used predominantly for PIM.

Though Hecker and colleagues lump all social tagging behavior under the umbrella of resource sharing, there are other plausible, social motivations for tagging.

Zollers (2007), for instance, describes three social motivations for tagging: opinion expression, performance, and activism. Opinion expression could equivalently be described as the application of evaluative tags, and include such common terms as "great", "awesome", "garbage", and so on. Performance tagging (in the sense of performing for the benefit of others) is often similar to opinion expression. It is distinct, however, "in that it is contextually dependent and interpreted, with the goal of accomplishing an informational exchange, speech act, or interpersonal bonding" (Zollers, 2007, p. 4). Zollers presents examples of performance tags such as "Maybe that is why i sometimes still dont feel like a grown woman-music" and "makes me wanna smash the radio". Though these tags do arguably reflect opinions, they also suggest a secondary motivation to leave one's mark on the item tagged. Finally, activism tagging, reflects purposive tagging of content in support of a socially defined goal. Zollers' example of this is a campaign spearheaded by the anti-DRM advocacy group Defective by Design, in which the group successfully encouraged Amazon users to tag products incorporating DRM with the word "defectivebydesign". A similar breed of socially-motivated tagging is apparent when user groups apply distinctive tags to content a means of expressing a relationship between the group and that particular content. For instance, the members of the Last.fm group "Altar of the metal gods" have applied tags like "altar of the metal gods death metal" or "altar of the metal gods black metal" to songs or artists as a means of giving the resource a stamp of approval, so to speak.

Another social motivation for tagging not covered by these taxonomies is play and competition (Marlow et al., 2006). In particular they cite "gamified" environments like the ESP game (Von Ahn & Dabbish, 2004). In the ESP game, two simultaneous

players assign textual labels to the same image, and neither can see the labels assigned by the other. When both players assign the same label, however, both receive points and move on to another image. Doing so, players attempt to maximize their score within a set time limit. Though entertaining for players, the underlying goal of the system was to determine appropriate labels for images by determining which terms arose from spontaneous player consensus.

The ESP game only loosely fits the standard definition of collaborative tagging, but is a good example of a game-like environment that proved motivating to taggers. Work on gamified approaches to social tagging has remained relatively stagnant in the years since the ESP game was introduced, with the notable exception of recent work by Weng & Menczer (2010). Built as a supplement to the GiveALink social tagging system (Stoilova, Holloway, Markines, Maguitman, & Menczer, 2005), the authors describe a gamified system for motivating users to tag Web pages called GiveALink Slider. In the game, users are presented with a start page and a target page, and must build a chain between the two. Players assign tags that are relevant to the current page and can then select from a sample of other pages to which that same tag is relevant. They continue iteratively until able to tag the current page with a label relevant to the target page, all the while earning points for the number, trustworthiness (i.e. have other users tagged the item similarly?), and novelty of tags they assign.

Finally, there is the possibility of social imitation in tagging. It is debatable whether social imitation should be considered a motivation in the strict sense of the term, because a user may be influenced by other users' tagging decisions without any explicit motivation to copy those decisions. It is plausible, however, that a user could be partly

motivated by imitation, so as to match an agreed-upon community vocabulary, for example. The question of imitation is covered in more detail in the next section of this thesis.

Sociality and function: While Heckner et al. (2009) see the classification of tagging motivation as fundamentally one-dimensional, Ames & Naaman (2007) offer a more nuanced, two-dimensional perspective. Based on a qualitative analysis of a survey of Flickr users, they argue that people's reasons for tagging vary in terms of *sociality* and *function*. Sociality distinguishes between tags that are self- or socially-directed (much like Heckner and colleagues' distinction), while the *function* of tagging can be organizational or communicative. This schema allows a user's tagging motivation to be classified into one of the four quadrants in Figure 18. To contextualize the motivations discussed in the previous section, traditional PIM motivation would fall in the upper-left quadrant (self organization) resource sharing would fall in the bottom-left quadrant (social organization), and the motivations described by Zollers would likely fall in the lower-right (social communication).

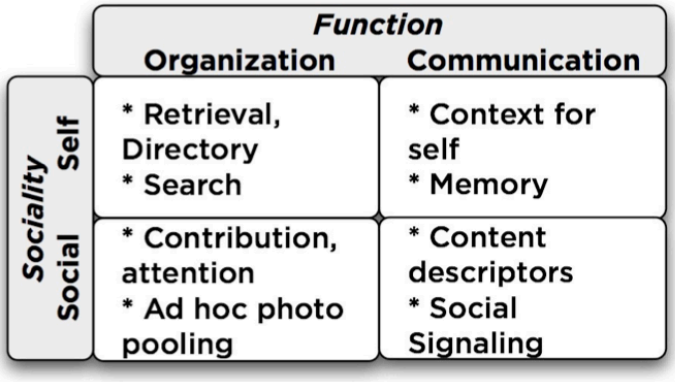


Figure 18: Ames and Naaman's taxonomy of tagging motivations (Ames & Naaman, 2007).

In the upper half of the diagram, self communication could likely be classified under the umbrella of PIM, but Ames and Naaman distinguish this category from organizational PIM activities (in the upper-left quadrant), which are focused on future retrieval. Tags used for self-communication are intended as contextual reminders about attributes of a tagged resource but are unlikely to be used to find the picture. For instance, a user can tag a photo with specific details of the event it captures, even though such context-specific tags are unlikely to be used for future retrieval. The lower half of the diagram helps differentiate the social motivations introduced by Heckner and Zollers by distinguishing organizational social tags (which help other users' find the tagger's photos) and communicative social tags (which convey messages about the content of the resource or about the tagger herself to other users).

Categorizers versus describers: The two means of classifying tagging motivations just discussed incorporated sociality as a central component, but some approaches eschew this dimension entirely to focus only on the functional differences in how people tag. Perhaps the most developed research program along these lines is that developed by Körner and colleagues. These authors argue that the fundamental distinction to be made with respect to how people tag is that between *categorizers* and *describers*. Under this framework, categorizers "tag because they want to construct and maintain a navigational aid to the resources for later browsing", while describers "tag because they want to produce annotations that are useful for later searching" (Strohmaier, Körner, & Kern, 2010).

This difference in motivation manifests in several measurable ways, all of which revolve around categorizers tending to stick to relatively small, structured vocabularies, and describers utilizing broad vocabularies of descriptive terms: "Categorizers typically

use a well-defined set of tags as a replacement for hierarchical classification schemes, while describers are annotating resources with a wealth of freely associated, descriptive keywords" (Körner, Benz, Hotho, Strohmaier, & Stumme, 2010).

In Strohmaier et al. (2010) and Körner, Kern, Grahsl, & Strohmaier (2010), the authors present a variety of specific user-centric metrics for quantifying differences between categorizers and describers, including tag/resource ratio (the number of unique tags used by a user divided by the total number of resources tagged; a value that is small for categorizers and large for describers), orphaned tag ratio (the proportion of tags used on only a few resources; large for describers), conditional tag entropy (relative use of different tags; large for categorizers), and overlap factor (how much overlap of tags there is across resources; low for categorizers, as their tags are more discriminative). Körner and colleagues perform a variety of analyses of large-scale tagging data to classify users as categorizers or describers, aggregating the measures above to give individuals a single categorizer/describer score. Thus the distinction is not binary; users fall somewhere on a continuum between pure categorizers and pure describers.

Approaches to Studying Tagging Motivations

I have presented some of the prominent taxonomies of tagging motivation in the experimental literature, and now turn to the question of how such taxonomies can be developed. Identifying motivations, it turns out, is a tricky business. It is comparatively simple to *observe* how taggers annotate resources, or *ask* them why they tag in a particular manner, but both approaches can be problematic. Observation can ultimately only answer with certainty the *how* question, not the *why* question, while self-report (i.e. as in surveys) is only as reliable as the respondents' ability to understand and effectively

communicate their underlying motivations. Approaches to studying motivation broadly fall in one of the four following categories: logical system analysis, tag analysis, usage pattern analysis, and qualitative approaches.

Logical system analysis: With the term "logical system analysis" I refer to approaches that explore the features, constraints, and other attributes of a tagging system to develop hypotheses about how and why users of such a system tag. The approach, though not described as such, bears high-level similarity to methodological adaptationism (Anderson, 1990), as it analyzes the task environment as a first step in understanding how people behave in that environment.

In the context of tagging, this specifically involves examining system properties like those outlined by Marlow et al. (2006). These authors are, in fact, acutely aware of the importance of studying system design to gain insights into tagging motivation: "A large part of the motivations and influences of tagging system users is determined by the system design and the method by which they are exposed to inherent tagging practices" (Marlow et al., 2006, p. 5). As a minimal example, one can include or exclude social motivations for tagging based on whether or not a system employs blind tagging, as we can safely assume users are not tagging for the benefit of others when each user's tags are visible only to herself.

A more developed example can be found in Floeck, Putzke, Steinfels, Fischbach, & Schoder (2011). Analyzing the social bookmarking site BobrDobr.ru[39], the authors refer to a phenomenon first observed by Golder & Huberman (2006) in which the relative distribution of tags assigned to a given item on Delicious tends to stabilize over time.

[39] Roughly speaking, this is Russia's take on Delicious.

Golder and Huberman proposed shared, pre-existing knowledge as a possible account for this, but Floeck and colleagues suggest that social imitation could also be a cause, and set out to test that hypothesis empirically. The relevance of this study to the present discussion is that the authors demonstrated a sensitivity to the design specifics of BobrRobr.ru in designing their study, and leveraged that information in their experimental design. To examine the role of social imitation, they noted that the site enabled two methods of saving bookmarks, one "internal" (adding a bookmark already saved by another user to one's library) and the other "external" (using a toolbar button to add the page currently being viewed to one's bookmarks). In both cases, the user could add tags to the bookmark, but in the case of internal tagging the system displays the top-five tags assigned to the item by other users, while when tagging externally the user has no social data on which to base tagging decisions (even if other users have previously bookmarked the page). The authors furthermore had access to usage data from BobrDobr.ru, such that they could compare tagging behavior across users for bookmarks saved internally (those that allowed for social imitation) versus those saved externally (such that social imitation was impossible). Their analyses suggest strong effects of social imitation in social tagging.

As noted above, work on imitation is far from the best example of tagging motivation research. Unfortunately, though, few studies directly focused on the question motivation perform detailed analyses of tagging interfaces and system attributes in developing their hypotheses. This study nonetheless demonstrates an example of how such information can be leveraged in studies of tagging behavior.

Tag analysis: Another method that can help elucidate tagging motivations is to analyze the tags themselves. In other words, researchers will analyze semantic or lexical attributes of tags to make inferences about why users may have applied those tags. Zollers' (2007) study is exemplary of this kind of work. A variety of other studies (Al-Khalifa & Davis, 2007; Heckner et al., 2008; Sen et al., 2006) have utilized tag analysis and categorization, though not always with an explicit focus on questions of motivation.

The most typical methodology is to start with a pre-established taxonomy of tag types, and then manually categorize a sample of tags following that scheme. Sen et al. (2006), for example, classify tags as "factual" (identifying attributes of the item tagged), "subjective" (expressing a user's opinion), or "personal" (self-directed tags, typically used for content organization). Heckner et al. (2008) utilized a more elaborate two-level system. Tags were first classified as either subject-related or personal. Subject-related tags were then divided into resource-related (usually corresponding to metadata about the item tagged, such as author or date) and content-related (descriptions of the content of a resource or other categorization attempts such as labeling an item with a genre or topic). Personal tags, on the other hand, were classified as either affective or time and task related.

The above classification schemes can all be considered *functional* in that they ostensibly address the function of the tag (e.g. to describe item attributes, express an opinion, etc.). Some researchers, however, have also studied attributes of tags independently of their function. The categorization scheme of Heckner et al. (2008) described above was actually just one of three modules in their classification system. They also developed a linguistic category model and "tag to text" category model, the

former classifying tags according to linguistic features (multi- versus single-word, part of speech, etc.) and the latter indicating whether the tag came from the text of the resource being tagged (their study analyzed tagging of academic papers). Veres (2006) performed a similar analysis of tags used on Delicious.

Studies of this type, though they can useful for studying tagging motivation, typically do not frame their analyses in terms of motivation per se. Thus certain distinctions (e.g. subjective versus personal, as in Sen et al., 2006) correspond to motivations, while others (e.g. functional versus origin collocations, as in Veres, 2006) do not. Furthermore, such studies can beg the question with respect to tagging motivation, as the authors will often develop an a priori taxonomy of tag types that presupposes certain classes of motivation before beginning their analysis. Nonetheless, analysis of tag types remains a useful means of classifying users' tagging activity into interpretable motivational.

Analysis of tagging patterns: Analysis of tagging patterns is likely the most common type of social tagging research in general, if not with respect to motivation in particular. I consider any study that engages in large-scale, quantitative analysis of user tagging behavior in an attempt to pick out differing motivations under this umbrella.

The various studies Körner and colleagues, as well as the work of Floeck et al. (2011) are both examples of tagging pattern analysis, but there are otherwise few studies directly leveraging large scale analytic methods to explore the question of why users tag. The challenge is that this requires identifying which quantitative signals correspond to different tagging motivations, and no quantitative analysis can reveal this directly.

Körner's work, for instance, rests on an initial qualitative analysis suggesting the importance of the categorizer/describer distinction.

This is not to reduce the discussion to a philosophical critique of the ability of quantitative research to illuminate questions around motivation. I do, however, stress the importance of informing quantitative analysis methods with qualitative and theoretical work There are few examples of such an integrative approach in the literature, though one we can point out is the study by Nov et al. (2008) mentioned above. Their work combined a survey with a quantitative measure of tagging level (in this case, number of unique tags). The survey measured the extent to which Flickr users were motivated to tag for the benefit themselves, their friends and family, and the general Flickr community. Combining this with quantitative measures of social presence (number of contacts and number of groups), the authors found that stated motivations of tagging for one's self and for the public were significant predictors of tagging level (though not for friends and family), as were the social presence indicators. Social presence was shown to have a stronger effect however, suggesting that social connectivity may be a major, and possibly implicit, contributor to tagging motivation. This is consistent with other work pointing to the importance of fostering strong social ties as means of increasing user participation in virtual communities

There is work on tagging pattern analysis that does not explicitly address tagging motivation, but is still useful in furthering our understanding of it. Schifanella et al. (2010), for example, present several findings that inform the present discussion. In an analysis of two social tagging platforms (Flickr and Last.fm), the authors found strong correlations across a variety of user metrics: number of friends/contacts, total number of

annotations, number of unique tags utilized, and the number of social groups of which users are members. They furthermore found that the closer the social distance (i.e. degrees of separation) between two users on either site, the more similar their tagging behaviors are.

This work suggests that tagging should not be considered independently of other kinds of activity on social tagging platforms, and that it is deeply linked to social connectivity. There is not conclusive evidence of a causal link, but it is certainly the case that users who are more socially involved on these platforms (in terms of the number of friends they have and the number of groups they join) are more active taggers, using more unique tags and generating more total annotations. Furthermore, increasing similarity in tagging vocabulary as a function of social distance suggests that users' social contacts influence tagging decisions (though this cannot be clearly distinguished from a homophily effect). Although the authors did not address tagging motivation per se, their work further highlights the importance of considering social influence in any theory of why people tag.

Another example of relevant research is a study by Meo, Ferrara, Abel, Aroyo, & Houben (2013). The authors collected a dataset of users with accounts on multiple social sharing services (Flickr, Delicious, and StumbleUpon), and were thus able to explore how tagging habits differed across different systems within the same set of users. They compared metrics such as number of friends, number of tag assignments, entropy of users' tag vocabularies, and others, finding major differences in usage patterns across the different tagging systems. Their results do not speak directly to the issue of motivation, but provide quantitative evidence of people's tagging habits being influenced by tagging

system design. Coupled with logical analyses of system design, such a comparative approach is likely to be a useful tool in understanding user motivation across different social tagging systems.

Qualitative studies: Every method described thus far can really only make inferences about tagging motivation; *why* a user engages in some behavior is rarely explicit in a record of that behavior. What instead can be a most direct way to learn of a user's motivation for tagging is to simply ask him or her.

The studies by Ames & Naaman (2007) and Nov et al. (2008)[40] discussed above are examples of the qualitative approach, with the former utilizing interviews and the latter a survey. The benefit of this approach is obvious: The researcher can directly ask the question "why do you tag?". But these approaches can also be problematic, not only because they still often rely on a pre-determined taxonomy of tagging motivation (a more serious issue in surveys than in unstructured interviews), but also because they depend on users providing accurate stated accounts of why they tag. The fact that Ames & Naaman found that social presence indicators were better indicators of tagging level than stated motivations suggests that users may not always respond to surveys in a way that accurately reflects their motivations. This could also point, however, to researchers having developed questionnaires (or coded free responses) in a way ill-suited for accurately capturing motivational factors.

Qualitative work remains crucial in studying tagging motivation, however, not only for developing intuitions about possible tagging motivations, but also for informing quantitative work that looks for measurable signals of different motivations. I reiterate

---

[40] These results are further discussed and summarized in (Nov & Ye, 2010).

the importance of an integrated approach if we are to fully understand what drives tagging behavior.

Problems with existing approaches

A common thread among the taxonomic distinctions outlined above is that they tend to present themselves as capturing *the* fundamental categories of tagging motivation. Every study of course has its caveats, typically being focused on one particular tagging system with unique design features that may be reflected in users' motivations. Nonetheless, overzealousness in "binning" users into the taxonomy proposed by a particular research program is the norm. Consider the work by Körner and colleagues, for instance. Their first paper on the categorizer/describer distinction states that "*at least* two vastly different types of tagging behavior can be found in tagging systems" (Körner, 2009, emphasis added) but in subsequent work the authors work to classify entire datasets of tagging system users into this dichotomy. Is this justified? Does it even make sense to classify every user along the categorizer-describer continuum, or might some taggers break from this entirely? The complete exclusion of social factors involved in tagging motivation, if nothing else, is suspect here.

This is not to pick out Körner's research specifically. Most of the work I have discussed has similar deficiencies, from the lumping of all social motivations for tagging under "resource sharing" (Heckner et al., 2009) to a complete focus on tag audience (Ames & Naaman, 2007). It is of course important to focus research on specific phenomena – no single study can be a treatise on all the possible reasons users tag – but it is also important to contextualize a proposed motivational taxonomy, integrating it with others that classify users along different dimensions. For example, how does a user

Körner would classify as a describer tag differently along the PIM and resource sharing dimension proposed by Heckner et al. (2009)? No such integrative work yet exists.

A second issue is that almost all work on tagging motivation is post-hoc, taking observed tagging behavior and attempting to classify as being "for" one reason or another. But few researchers have started from the other direction, asking foundational questions about why users choose to participate in tagging systems in the first place, or why tagging can be motivating beyond the functional roles it can play. These types of questions can do much to inform how we classify users' reasons for tagging and the analytic approaches we bring to tagging datasets, but have been largely unexplored. In the following section, I will address a number of such under-developed perspectives.

Underexplored perspectives

How are existing tags used? There is clearly much research on how people tag – that is, how they assign tags to content – but a dearth of research on how those tags are actually utilized by users once they exist. How and when do users view tag clouds on tagging systems? How and when do users use their own tags to retrieve resources? How are different types of tags evaluated differently by different users?

I could continue the list of questions, but the point is that we know very little about how the tags assigned to resources on social tagging systems are actually viewed and utilized. Existing qualitative work can provide some insights, but it has focused largely on the process of tagging itself, not how existing tags fit into everyday usage patterns.

The lack of research in this regard can, in large part, be attributed to availability of data. A snapshot of the folksonomy on a tagging system is a comparatively simple

structure to collect and analyze, but other usage patterns are more complex to work with. Web services in principle can collect data on page view times, click rates, search queries, and so on, but this data is rarely made available to researchers. It is possible to implement small in-lab studies examining how human subjects interact with a tagging system (using mouse tracking software, keyloggers, and so on), but large-scale analysis of tag usage patterns remains a challenge. One notable exception is Bibsonomy.org, a social tagging system for bookmarks and academic papers expressly developed for research purposes. Though at present only folksonomy data appears to be publicly available[41], Mitzlaff, Benz, Stumme, & Hotho (2010) utilized Bibsonomy server and click log data in a social network analysis of Bibsonomy users. Unfortunately, the data has yet to be applied to questions around tagging motivation.

If future researchers can acquire and leverage this kind of data, it will allow them to paint a more complete picture of what drives tagging behavior. Not only will we have further quantitative signals corresponding to different tagging motivations, but such data could help provide further insight into the reliability of people's stated reasons for tagging.

Why tag to begin in the first place? Existing tagging studies that deals with the question "why do people tag?" more accurately address the question "why do *taggers* tag?" This may appear to be a distinction without a difference, but the issue I raise is that the literature principally addresses the question of what motivations for tagging exist among those people who choose to tag in the first place. Comparatively little attention has been paid to determining what motivates the transition from non-tagger to tagger.

---

[41] See http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

Social tagging, like many Web-based collaborative activities, faces a poverty of participation. Well captured, in spirit if not numerical exactitude, by the 80/20 rule (80% of activity comes from 20% of users, also known as the Pareto principle), the simple fact is that folksonomies tend to be built by the activity of a minority of a system's users. In one sample of almost two million users of Last.fm, less than 30% tagged even a single resource, and among those who did, 46% applied a tag on five or fewer occasions (Lorince & Todd, 2013).[42]

Similar patterns are visible in other tagging systems, and the implications are substantial. The emergent folksonomies in such systems are generated by a relatively small subset of their users, and thus it is unclear how representative the emergent classification is of a system's overall user base. It is surprising, then, that relatively little work has explored what distinguishes those users who choose to tag from those who do not.

As mentioned above, we tend to have little data on how existing tags are used, so it remains unclear if the majority of users who tag very little, if at all, actually utilize the tags applied by the minority of users doing most of the work for search or any other reason. In other words, we do not even know the full scope of the question; are a minority of users the only ones who annotate content with tags *and* the only ones to utilize those tags at a later time, or does that minority define a folksonomy utilized by a majority of uses?

---

[42] To precisely frame this in terms of the 80/20 rule, we can attribute 80% of all annotations (instances of a particular user assigning a particular tag to a particular item) in our dataset to the 35,703 most prolific taggers in our sample, a number making up only 2% of the users in the sample, and 7% of those who tagged at least once.

In either case, the question of what motivates users to participate in the tagging process to begin with is an important one for tagging researchers and system designers. Though little has been done to address it directly, research pointing to the importance of social connectedness is relevant here. Strong correlations between measures of social involvement and tagging activity are suggestive of the importance of social pressure, but it remains unknown how and when this encourages participation in tagging systems. One fruitful direction may be research comparing usage patterns of taggers versus non-taggers within the same system. Most social tagging systems can still be useful to "lurkers", users who utilize certain aspects of the system but do not engage in tagging. Determining what measurable signals, if any, differentiate taggers from users who never choose to tag may provide clues as to what encourage users to become involved in tagging at the outset.

Variation within individuals: Marlow et al. (2006) present a list of possible tagging motivations with the caveat that "they are not intended to be mutually exclusive; instead we expect that most users are motivated by a number of them simultaneously" (ibid., p.5). Though many researchers seem aware that users are likely to be simultaneously motivated by various factors, or that the motivation for applying a tag may differ for a user from one annotation to the next, there do not exist formal analyses that take this into account. Large scale analyses instead tend to generate some kind of characterization of a user's tagging behavior across all their annotations (such as Körner's categorizer-describer score) without thought towards how those annotations might be differentially motivated.

Identifying multiple concurrent motivations or classifying individual annotations both present methodological challenges, but still warrant attention. It may simply not

make sense, at least in some cases, to determine that a user is of one "type" or another. There may not be a single, characteristic motivation for many individuals' tagging behaviors, and developing analytic methods to tease these apart is an important future direction. Pursuing work on tag usage patterns may prove to be useful in this regard, enriching the data characterizing any given instance of tagging.

Psychological and cognitive perspectives: Trivial though it may be to state, tagging is a human behavior. What is not trivial is that work in the cognitive and psychological sciences that could do much to inform our understanding of tagging behavior has been largely absent from research on the topic, which mostly has been pursued by the computer science and library and information science communities.

Though topic is generally speaking unexplored, a small number of researchers have taken note of the mutual informativeness of work in collaborative tagging and cognitive science.[43] Robert Glushko in particular is a proponent of this perspective, arguing for treating social tagging systems as instances of "categorization in the wild" (Glushko, Maglio, Matlock, & Barsalou, 2008). By the same token, the vast psychological literature on categorization likely has applications to understanding tagging behavior. Glushko is also a proponent of an interdisciplinary approach to studying how people organize information (2013), combining methodologies from the cognitive, computer, and library and information sciences. Though social tagging is only one topic

[43] Note that I exclude from this discussion work on psychological factors that may implicitly impact how and why people tag. Though work in psychology on, for example, implicit association might be useful in understanding tagging decisions at a granular level, they likely say little about the high-level factors motivation tagging behavior.

to fall under the umbrella of "organizing", this bodes well for future interdisciplinary attempts to understand tagging behavior.

A second sense in which psychological methods may come to bear on the study of tagging motivation (and the most speculative of the perspectives I present here) is in understanding its possible hedonic value. No existing work on why people tag has addressed the question of if and why tagging might be intrinsically motivating, independent (at least partly so) of its social, PIM, or other functions described in this paper. In other words, the question I propose is the following: Does the act of tagging content, of assigning labels to resources for organizational or other purposes, satisfy some intrinsic desire to categorize, "put things in their place," or otherwise describe content? With no existing research in existence, it is difficult to say a priori whether this question can be of substantive value, but an understanding of if, and if so when, the act of tagging itself is motivating would certainly paint a clearer picture of why users tag. Might there exist users who simply derive pleasure from the process of reviewing and tagging resources? If so, what attributes differentiate those users from non-taggers, and what environmental or other factors modulate the extent to which they enjoy the process? Methods from the positive psychology literature, or on motivation in play environments would be effective starting points for answering such questions.

Mechanisms of social influence: It should be clear from the discussion above that social factors play heavily into people's tagging behaviors. This encompasses both situations in which users expressly tag in a communicative manner, and the findings that social connectivity and involvement is a strong predictor of people's level of tagging activity. Though the research points unequivocally to the strong social component of tagging,

work remains to be done on exactly *how* social factors come to influence tagging behavior.

We know, for instance, that users sometimes apply tags with the apparent goal of helping others find or learn more about content, or to express their opinions to others, but what factors encourage this behavior as opposed to more self-focused tagging? Is such activity largely a byproduct of personality differences among users, or do certain system designs foster greater desire for socially-motivated tagging than others?

We also know that users with more social connections tag more often and with more unique tags, and that they furthermore tend to use tagging vocabularies similar to those of their friends. With respect to the former, are tagging activity and social connectivity simply two correlated facets of a higher level variable (i.e. investment in the tagging system), or does having more friends somehow encourage more tagging in another way? As for the latter, do such patterns of vocabulary only reflect homophily, or are there implicit mechanisms whereby users are influenced by the tagging habits of their friends?

In short, social influence on tagging is undeniable, but we have little functional understanding of the mechanisms that underlie it. And this is without even introducing other possible social motivators beyond those discussed above. Consider the possible role of prestige in tagging systems, for example. We can hypothesize that users might be partly motivated by a desire to be particularly prolific in their tagging, for example. This, however, like the questions mentioned above, remains unaddressed in the literature.

Conclusions

I began this section with what appeared a simple question: Why do people tag? The further we explored that question, however, the more complex it turned out to be. First overviewing the attributes of tagging systems, and then covering both the prominent taxonomies of tagging motivations in the literature and methods for developing them, it became clear that the landscape of tagging systems and users' motives for using them are far from homogenous. I furthermore suggested a variety of approaches to studying tagging motivation that remain to be exploited, demonstrating that our understanding of the topic is far from complete.

It is thus not possible to provide any succinct answer to the question I have posed. People tag to organize content for browsing, to search for and find resources at a future time, to send social signals to other users, and for other reasons. They are likely motivated by multiple factors simultaneously, in ways that may change from one annotation to another, and we have yet to develop effective methods for teasing apart these complex motivations. Other factors beyond those examined in the existing literature may motivate users to tag, and we are faced with technical obstacles to fully understanding if hypothesized motivations actually play out in how people utilize existing tags. In short, people's reasons for tagging are complex and modulated by a variety of individual and system-level factors.

The remainder of this chapter, after a brief overview of social imitation and its manifestation in collaborative tagging, presents two studies directly contribution to further understanding these motivational questions.

Social imitation (on the Web)

Social creatures that we are, humans are capable of great feats of collaborative action, of learning from others to pursue our own interests and those of the group, and otherwise achieving goals that no individual could alone. Social behavior can of course take many forms, but what may be its fundamental manifestation – simultaneously the simplest and most critical of social behaviors – is imitation. Our ability to observe and copy the actions of our conspecifics is the building block for higher forms of complex social interaction, but is a phenomenon that has been the topic of much research itself.

Examples of imitation in the natural world abound. Birds and fish are often the first example to come to mind, as the intricate movements of flocks and schools depends on the ability of individuals to follow the movements of their neighbors. But many other examples of imitation behavior across a variety of species exist, in contexts as diverse as foraging decisions, mate selection, habitat assessment, and more.

Humans, however, are the truly great imitators. Even as infants we mimic the facial expressions of our parents, and as adults we engage in far more complex forms of imitation, from coordinated movement, to the mimicry of style and preferences that leads to cultural "fads", to copying of mate choice preferences. While imitation behavior can of course have its downsides (the propagation of gossip and popular delusions being obvious examples), it can best be understood as an adaptive solution to a complex world where other individuals have different – and often better – information than ourselves.

The question I raise here is this: How do these evolved, adaptive tendencies for imitation manifest in the virtual environments of the World Wide Web? Modern, Web-based technologies not only create a vast new diversity of opportunities for social

imitation, but leverage the complexities of this and other forms of social interaction in the creation of so-called "collective intelligence." We look to analysis of trends on Twitter for an understanding of the topics people care about, utilize collaborative tagging systems to organize large corpora of content, and leverage purchasing decisions and user reviews to generate recommendations on e-commerce systems, to name only a few examples. Entire industries rest on the power of online collective intelligence, but in interpreting behavior in such systems, relatively little attention has been paid to how such behavior can be reconciled with our evolutionary understanding of imitation and social learning. The Web is radically different than the ancestral environments out of which our social capacities emerged; what are the implications of these differences?

Imitation: A brief overview

To begin I review the ecological phenomenon of social imitation, focusing first on the benefits of imitation and its evolution as an adaptive behavior, and then considering its possible downsides. My treatments of these topics is necessarily brief, focusing on developing a general framework for understanding imitation behavior to which I will relate to social tagging. I will, for instance, not discuss the low-level neurological mechanisms of imitation (e.g. the mirror neuron system, as proposed by Rizzolatti & Craighero, 2004). For a thorough examination of imitation behavior, I point the reader to Hurley and Chater's volumes on the topic, which exhaustively cover the mechanisms, evolution, and implications of imitation in both humans and animals (Hurley & Chater, 2005a, 2005b).

The Good – Understanding Imitation and its Benefits: When an organism engages in social imitation, it copies a decision or action of another individual or aggregation of

individuals, leveraging social information to guide its own behavior. In most, but not all cases, we can roughly equate social imitation with social learning (when an individual learns which behaviors to avoid by observing others, for instance, would presumably be considered social learning, but not social imitation). Note that in adopting this stance, I use a relatively inclusive definition of imitation that may not be considered "true imitation" by some authors. The most restrictive sense of the term "requires that a novel action be learned by observing another performing it, and in addition to novelty, requires a means/ends structure. You copy the other's means of achieving her goal not just her goal or just her movements" (Hurley & Chater, 2005a, p. 14).[44]

Stated simply, the advantage of social imitation or learning is that it allows an individual to reap the benefits of others' effort – be they resources discovered, skills acquired, information learned, or otherwise – without having to put in the "heavy lifting" itself. When a fish follows the movements of others nearby it in a school, it can successfully avoid a predator without needing to have detected it personally. When a person reads movie reviews or checks the box office listings to see what is popular, she can decide which movie to see without having to evaluate all the options herself. Disparate though these examples may seem, they are conceptually quite similar.

It is trivial to concoct examples wherein imitation is beneficial and individuals "benefit by copying because by doing so they take a shortcut to acquiring adaptive

---

[44] These authors point out three specific behaviors that are not "true imitation", but would fall under the looser definition I employ: *stimulus enhancement*, whereby one individual's actions draw another individual's attention to a stimulus and triggers a previously learned response; *emulation*, in which one individual observes another achieving some goal, and attempts to achieve that same goal (but not necessarily by copying the observed individual's strategy); and *response priming*, where bodily movements are copied but not as learned means of achieving a goal (e.g. when birds copy movements of neighbors in the flock).

information, saving themselves the costs of asocial learning" (Laland, 2004, p. 4). Individuals can follow others to find resources, determine good items to eat based on others' choices, avoid dangers without needing to stay ever-vigilant themselves, and more. But let us formalize what we mean by imitation. When and how do individuals engage in such behavior? It cannot be the case that imitation is always superior to non-social strategies, or else all individuals would adopt the behavior (i.e. would be information *scroungers*, as opposed to information *producers*) and there would never emerge novel behavior to copy. Laland eloquently elaborates this point:

> Either some individuals in the population must be consistent information producers and rely exclusively on asocial learning or, more realistically, individuals must use social learning selectively and directly sample the environment through their own asocial learning some of the time. It is precisely because individuals do not use social learning indiscriminately and engage in asocial sampling of environments that social learning is typically adaptive (ibid, p. 4).

Laland presents a variety of social learning strategies, which he broadly categorizes into "when" strategies and "who" strategies. The first class of strategies determine *when* an individual should engage in imitation, as opposed to acquiring information on its own, and the second dictate *whose* behavior should be copied. There is surely much variation across species and circumstances as to when organisms engage in imitation, but Laland's framework is a useful one. Under it, there are three principal reasons an individual would adopt an imitation strategy.

First, and simplest, an individual might copy others' behavior if its established or unlearned behavior is unproductive. This is best illustrated in the context of producer-scrounger games (Barnard & Sibly, 1981), in which foraging animals can either produce (i.e. discover resources) or scrounge (exploit the resources discovered by others).

Scrounging is an unlearned behavior to which organisms will often default, but empirical evidence (Giraldeau & Beauchamp, 1999) suggests that when scrounging is unproductive (e.g. when there are too many scroungers in a group, such that relatively little food is being discovered), scroungers will imitate novel food-finding behavior of producers. A second "when" strategy is to imitate when asocial learning is costly. Here I point out a uniquely human example: picking a restaurant at which to eat in a new town. Here the costs of asocial learning include reading reviews and doing other research, or alternatively trying each of the restaurants to see which is best. Imitation, on the other hand, is a much "cheaper" strategy. One can simply walk downtown and observe which restaurant is the busiest, inferring that it is of the highest quality because so many other people have chosen to eat there. The third example, "copy when uncertain", is a plausible response when an individual does not have a clear means of determining the proper behavior in an environment. Rats, for example, have been shown to imitate conspecifics' choice when simultaneously presented with two novel food items. A human might perform analogously when exposed to a novel cuisine at a restaurant.

Assuming an individual does engage in social imitation, how should she go about it? Laland describes a variety of possible strategies, the applicability of which varies across social groups and circumstances, for determining who an individual should imitate. We can summarize by segregating the "who" strategies into three classes. The first two ("copy-the-majority" and "copy-if rare") are aggregative strategies, in which an individual considers the group as a whole and imitates behavior that is either particularly widespread or particularly rare. The former is adaptive in the more common case where beneficial behaviors are more likely to persist and be common in the population, while

the latter would arise in cases where novelty is particularly adaptive. The second class are what I refer to as evaluative strategies, in which an individual evaluates the success of others in deciding who to copy. Thus an individual might copy those who perform particularly well at a given task ("copy-if-successful"), anyone who simply performs better than the individual ("copy-if-better"), or those who have demonstrated previous success in social learning ("copy-good-social learners"). The third class of "who" strategies I refer to as reputation strategies. I do not employ the term "reputation" in the formal sense, but as a means of distinguishing strategies where the individual actually evaluates the behavior of others from those in which it bases its decision to imitate or not based on cues not directly reflecting performance. Examples include cases where individuals imitate kin, "friends" (referring to individuals with whom one has exchanged reciprocal altruistic acts), or older conspecifics. In all these cases, the individual does not evaluate how well another performs with respect to a particular behavior, but instead bases the imitation decision on a cue that can indirectly indicate value (e.g. older individuals have more accumulated knowledge, and therefore may be more likely to have adopted effective behaviors).

Garcia-Retamero & Dhami (2009) argue that, in humans, such social copying mechanisms can be understood within the framework of fast and frugal decision-making heuristics. The fast and frugal heuristics research program (Gigerenzer & Todd, 2000) explores how simple decision strategies, or "rules of thumb", can function both more quickly and frugally (in terms of the cognitive resources they require) than more complex decision strategies, and as such are adaptive solutions to decision-making under the constraints of time, information, and cognitive processing capacity. Imitation strategies

like those described by Laland are examples of such heuristics, and Garcia-Retamero & Dhami (2009) in particular focus on what they call "imitate-the-best" and "imitate-the-majority". They present several example scenarios in which the use of simple imitation strategies like these can be used to solve social inference problems, including the selection of a mate.

Many of the evolutionarily significant problems faced by organisms can be described in terms of explore-exploit tradeoffs. Foraging animals, for instance, must choose between exploiting the current resource patch, or exploring to find a new one. When searching for a mate, humans must decide whether to pursue an available potential mate, or to keep looking (likely sacrificing the opportunity presented by the current prospect). Speaking generally, organisms face a constant tradeoff between continuing to approach a problem with the current strategy (exploiting) and seeking out a novel solution to the problem (exploring).

For social organisms, imitation can serve to modulate such an explore-exploit tradeoff at the group level. In social foragers, for example, imitation behavior (i.e. scrounging) is effectively a manifestation of exploitation, while producers who seek out new resources embody the exploration side of the tradeoff. But these forces are not necessarily at odds, as sometimes suggested by producer-scrounger models (i.e. the implication that scroungers are "freeloaders" that do not contribute to the group). In many contexts, a balance of imitative and non-imitative behavior can increase the overall fitness of a group. In the context of social foraging, for instance, there can be advantages to group foraging, in which some individuals act as producers and others as scroungers (Clark & Mangel, 1986). Benefits include increased vigilance against predators,

improved information use (i.e. if resources are hard to find, having multiple searchers increase the mean return per group member), increased ability to capture prey (i.e. pack-hunting), and more.

The power of social imitation for the group is that it can allow advantageous behaviors to spread through the group, and for members to capitalize on information they need not accumulate on their own. Consider the power of flocking fish to avoid predators. Not only does the large number of fish lead to a confusion effect on the part of the predator (J. Krause, Ruxton, & Krause, 2010), evasive movements in response to predator detection by a minority of individuals propagate through the group via social imitation (an information cascade of this sort can be disadvantageous, however, as I will discuss below).

The ability of group members to engage in social imitation and otherwise leverage social information allows for what is often referred to as swarm or collective intelligence (Katsikopoulos & King, 2010; J. Krause et al., 2010; S. Krause, James, Faria, Ruxton, & Krause, 2011). Group-level intelligence can emerge in different ways, but typically arises from the fact that there exists a diversity of information sources within a group, as individuals vary in the the information locally available to them, the quality of their strategies for the task at hand, and so on. Group members' ability to copy the strategies of individuals, or to acquire individual or socially aggregated information from others, enables the group to perform feats far outweighing the capacities of any individual. Striking examples of swarm intelligence exist in both the animal kingdom (e.g. the complexity of ant or termite colonies) and in humans (e.g. the classic finding that the average of many individual guesses of the weight of an ox tends to be remarkably close

to the true value). A large body of research has explored this topic from various perspectives, of which I offer few examples from the human domain.

Katsikopoulos & King (2010) for example, examined when groups perform better by copying specific "leaders" or "experts" within the groups (similar to the "copy-successful-individuals" strategy from the previous section) or by aggregating the decisions of the group (similar to the "copy-the-majority" strategy). Using a set of simple computational models, they found that in one-shot scenarios, expert imitation led to greater performance on average, while across repeated decisions the collective strategy was superior.

S. Krause et al. (2011) explored some of the factors that determine when collective decision-making strategies do and do not perform better than individuals. They compared human performance on two swarm intelligence tasks: Estimating the number of marbles in a jar, and estimating how many times one would have to flip a coin such that the probability of all flips being heads was equal to that of winning the German lottery.[45] The authors found that on the first task, the average of many individuals' guesses was quite close to the true number of marbles in the jar (within 1.5%), but on the second task the collective did not come close to arriving at the correct value. The authors explain this by noting that, on the marble task, there is a variance in people's responses, but no systematic bias. On the coin-flipping problem, however, a comparatively high level of expert knowledge (i.e. of probability theory) was required to provide a reasonable estimate, and people tended to vastly overestimate the correct value. This was

[45] The correct answer was about 24, as the probability of flipping 24 heads in a row is approximately 1 in 17 million, which roughly lined up with the chances of winning the lottery at the time.

borne out by analyses showing that collective accuracy improved with larger group sizes for the first task, but not the second. Clearly, some tasks simply are not amenable to collective intelligence (though, presumably, if people could accurately identify experts in such a scenario, an imitation strategy would be useful).

Goldstone and colleagues have performed a variety of studies on collective behavior. One example is the analysis of group path formation (Goldstone, Jones, & Roberts, 2006), exemplified by the spontaneously arising paths cutting through grass on a college campus. Human behavior in these environments, they found, is well-modeled by the Active Walker Model (Helbing, Schweitzer, Keltsch, & Molnár, 1997), which incorporates a component whereby individuals are attracted to the paths left by others in the environment (consistent with stigmergic path formation via pheromones, as in ants, and facilitation of travel through an environment, as when human paths through the snow facilitate future travelers' movement).

The bad – Dangers of imitation: We often do well to copy the behaviors or decisions of others, but conventional wisdom tells us that this is certainly not the best course of action under all circumstances. Many of us, after attempting to assuage our parents' concerns about some less than prudent course of action by claiming that "everyone's doing it", were met with the response: "And if everyone jumped off a bridge, would you do that too?". Though perhaps trite[46], this bit of parental wisdom is relevant to the present discussion. In effect, the lesson parents attempt to impart to their children with this comment is that socially acquired information is fallible. The wisdom of the crowd (or of

[46] Not to mention that the point of the argument is easily deflated. If all one's friends were in fact jumping off a bridge, there would be strong reason to suspect that some catastrophic event prompted their behavior. See http://xkcd.com/1170/.

friends, or experts, or anyone else) is not by definition rational to follow. There are cases where one's privately acquired information is superior to that copied from the group, and social imitation can in fact have major negative consequences for the imitator or the group at large.

The previous section introduced the possibility that overuse of social information in a group can represent a sub-optimal explore-exploit tradeoff at the group level. Giraldeau, Valone, & Templeton (2002), however, point out two additional possible disadvantages of social imitation: incompatibility of socially and individually acquired information, and the emergence of information cascades. Here I discuss only the latter, as it is more prevalent in human social contexts.

Information cascades occur "when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behavior of the preceding individual without regard to his own information." (Bikhchandani, Hirshleifer, & Welch, 1998, p. 994). To understand how this can occur, consider this example (Giraldeau et al., 2002).

Imagine a group of social foragers that encounters resources that can be either nutritious or harmful, and that the resource value can only be determined by tasting the resource. Further assume that there is a 0.5 base probability of it being either nutritious or harmful. Finally, assume that the signal obtained after tasting the resource is informative, but non-deterministic. That is, if the resource tastes good, there is a probability $P>0.5$ that the resource is nutritious and a probability $1- P$ that it is actually harmful. Conversely, if it tastes bad, it really is harmful with probability P and in fact nutritious with probability $1 - P$. Given this information, imagine that the group of foragers encounters such a resource for the first time. The first individual to samples the resource (individual A)

tastes it, finds that it is good-tasting, and thus decides to eat it (because it is more likely nutritious than not). Now consider the second individual to try the resource, B. B tastes the resource, also finds that it tastes good, and decides to eat it, as well. B is more certain of the quality of the resource, however, as he has two pieces of evidence supporting its quality: his own tasting it, and the observation that A decided to eat it (assume for simplicity that the foragers weigh their personal experience of resource taste and observations of others' decisions to eat it or not equally). Now a third forager, C, tastes the resource, but she finds that it tastes bad. How is C to respond? Keeping in mind that these foragers weigh social and private information equally, she will disregard her private information and eat the resource, as she has seen two other individuals consume it. At this point the information cascade has begun, as all subsequent foragers will decide to consume the resource regardless of how the resource tastes, because the aggregated social information indicating that it is nutritious will outweigh their private information. Individuals have come to uniformly ignore their private information and follow the crowd.

The key element of this simplified scenario, and of any information cascade, is that social information is not in the form of relevant cues (in this case, the value of a resource), but instead takes the form of behavioral responses *to* relevant cues. It is this that permits such cascades of conformity to occur: Instead of simple aggregation of data, social information gains a self-reinforcing sequential dependence. The danger of such cascades is that, under the right circumstances, they can easily lead to conformist adoption of a deleterious behavior. A mathematical analysis of the simplified scenario described above (Bikhchandani, Hirshleifer, & Welch, 1992) showed that the chance of a

cascade occurring in favor of the incorrect behavior (i.e. eating the resource when it is harmful, or vice versa) decreases as the discrimination value, *P*, increases, but remains surprisingly high. At *P*=0.9, the chance of an erroneous cascade is greater than 5%, and with *P*=0.75 it is greater than 15%. They further demonstrate that cascades tend to start sooner when *P* is greater, and rapidly increase in probability with the number of individuals. Even with an uninformative *P* of 0.5, there is a less than 0.1% chance of a cascade *not* occurring after ten individuals.

I have a presented a heavily simplified, idealized account of information cascades, and in realistic situations – in humans or otherwise – things are considerably more complicated. For one, decisions that may lead to cascades can be more complex. Models are typically presented as simple adoption or rejection scenarios, but there are of course situations where individuals decide from a much larger set of alternatives. Though cascades can still occur under such circumstances, they tend to be delayed. Individual variability is also a crucial component to understanding information cascades. Individuals can vary in their relative weighting of private versus social information, and certain individuals can be more influential than others (a celebrity adopting a new style will likely elicit more imitation than would the average person, for instance). Such variation, coupled with their idiosyncratic development (recall that the emergence of a cascade depends not only on aggregated social information, but the possibly random order of that information) lends a component of fragility to information cascades. Though they can occur with relative ease, they can also quickly be "broken" by the introduction of more informed individuals, releases of public information, and so on.

There are a multitude of examples of information cascades in humans and animals. Howell (1979), for instance, describes how a flock of nectar-feeding bats will all "follow the leader" as soon as one bat elects to stop feeding at a particular plant, without any attempt to obtain private information about it. Stamps (1988) demonstrated that territorial species disadvantageously cluster their territories in a manner suggestive of an imitation-based cascade. Humans demonstrate cascading behavior in a variety of domains (Bikhchandani et al., 1992). These include politics (where polls can influence people's voting decisions, thus generating a cascade), finance (such as when the actions of a few early investors in a company's IPO create a cascade of investment), and the emergence of cultural "fads". The propagation of gossip, urban legends, and the like can also be understood in terms of information cascades (where coming to hold a particular belief constitutes copying a behavior).

Information cascades are not, by definition, dangerous. We should keep in mind, for instance, that cascades generally do favor the "right" choice, and we can propose many scenarios where an information cascade is inarguably beneficial. The ease with which a cascade might develop when individuals flee a burning building (even if many of those fleeing have not actually encountered the fire directly), is one such example. In short, information cascades are a manifestation of social imitation that surely can be dangerous, but more often is not.

Imitation on the Web

I have presented a broad overview of social imitation as an adaptive human behavior, but the relevant question for this thesis is if and how this framework can improve our understanding of behavior on the Web. Adoption of the WWW has been rapid and

widespread over the past two decades, and it is now a deeply-integrated part of many people's lives. Though some common Web activities are simply more efficient proxies for offline activities (online banking, for instance) much of our interaction with the Web is inherently social. This includes the obvious cases of social networks like Twitter or Facebook, but also the constantly deepening integration of social features into other activities, like e-commerce, search, and news consumption.

We leverage the collective intelligence of humans operating in these social Web environments – the wisdom of the crowds – in various ways, from the use of collaborative filtering to recommend movies on Netflix or products on Amazon, to the analysis of how "memes" spread on Twitter and elsewhere. These are no doubt powerful sources of information, but relatively little thought has been paid to how (or if) our deeply imitative nature should affect the interpretation of such Web behaviors. Social copying strategies were adaptations to ecological problems posed to our ancestors in social environments that shared little resemblance to those we access through our Web browsers or mobile apps; We have, in many ways, the same brains that evolved to solve problems that are thousands of years old. It is also the case, however, that those same capacities for imitation have enabled cultural learning that can operate at speeds far outpacing those of biological evolution. Even assuming that cultural evolution has "kept up", so to speak, with the explosion of the Web, it remains unclear how our interpretations of collective Web behavior accurately account for what we know about human tendencies and capacities for imitation.

In light of this, it is worthwhile to question how our evolved capacities for social imitation might manifest in Web environments, and how this perspective might lead to

differing interpretation of collective behavior online. This is a speculative question to which there are no concrete answers, but one that can inform our understanding of web behavior generally and the dynamics of collaborative tagging in particular.

The power of social tagging systems, it is generally assumed, lies in their ability to aggregate the many individual decisions people make when assigning labels to content. Tagging systems, however, commonly attempt to facilitate the tagging process by adding tag recommendations drawn from other users' tagging choices (typically the most common tags assigned to the item), thereby enabling social imitation when a user chooses to act on a recommendation and engage in copying behavior. Even when no explicit recommendations are made, if a user can view the top tags or tag cloud for a given resource, she can easily engage in imitative behavior. Whether enabled by recommendations or simply viewing the top tags for an item, social imitation in these contexts appears most likely to employ something like a "copy-the-majority" strategy.

Whether imitation is beneficial or detrimental to these systems can be difficult to determine. On the one hand, enabling users to see others' decisions (and therefore imitate them) can allow for consensus effects that may be informative and presumably makes the tagging process easier for users. On the other, tag imitation can also result in information cascades in which the first few taggers have a disproportionate impact on the final tag distribution and any final consensus is in fact spurious. In the end, the desirable level of imitation in a tagging system is dependent on the design features and goals of a particular system.

Imitation is typically assumed to occur in tagging systems in at least some capacity, and has been shown to effectively model the dynamics of tagging systems.

Cattuto, Loreto, & Pietronero (2007) for example, developed a mathematical model of tag co-occurrence and evolution where new tags were added to a resource with probability proportional to their frequency[47], consistent with imitation behavior on the part of users, and my own work (described in detail below) involves a multi-agent model in which agents preferentially copy popular tags. Both approaches have been able to replicate statistical regularities in tag distributions without incorporating semantics, similarity relations between tags or resources, or other features. Though imitation is clearly important in the dynamics of tagging systems, there is relatively little work on how behaviorally plausible imitation strategies play into tagging behavior. Beyond my own work, in which I suggest the importance of social copying heuristics in tagging behavior, the only study to explicitly study tag imitation is one by Floeck et al. (2011). These authors compared how users tag when imitation was either possible (i.e. with top tags visible) or impossible (no existing tags visible). They found interesting differences in the tagging dynamics between the two groups, but did not relate them to behavioral imitation strategies.

An understanding of how behavioral imitation manifests in the tagging context is, however, an important aspect of interpreting behavior in this domain. We can begin by positing that the adaptive benefit of imitation is that it saves the user from needing to select a tag from the effectively infinite space of possibilities (assuming the system allows free-form tagging). But users clearly do not *always* engage in imitation, so what governs the trade-off? In ecological contexts, there is some benefit to imitation that can

[47] They technically employed a Yule-Simon model, modified with a memory kernel such that more recently applied tags were more likely to be added at any given point.

be saturated, so to speak, if too many members of a group are using the imitation strategy. It is this that necessitates selective imitation strategies. But in the context of tagging it is less clear what that benefit is. There is a clear analog in that the benefit of the folksonomy in terms of search and information organization would be lost if every user only copied others' tagging decisions, but what is unclear is the mechanism enabling this. In an ecological context – foraging, for instance – there is a measurable difference in the benefit reaped from a social versus individual strategy at any moment in time (e.g. calories of food acquired), but for a tagger there are no apparent differential benefits with respect to a particular tagging decision.

Three general possibilities present themselves here. First is that sufficient variety in tagging is introduced into the system by users who simply have no interest in imitation, and independently maintain their own idiosyncratic tagging vocabularies. In other words, to use Laland's terminology, there may exist individuals who are "consistent information producers" in the tagging context. Second, there may exist sufficient intrinsic tag-resource relationships that are external to the dynamics of the tagging system so as to maintain tag variability. In other words, users might already generally agree on how items ought to be tagged, based on outside knowledge and independent of social influence. On a music tagging platform like Last.fm, for example, there may be agreement about the genres of different artists built on external, shared knowledge that overwhelm tendencies for tag copying. The third possibility is that users *are* in fact sensitive to the distributions of tags assigned to different resources, such that they are more likely to explore the tag space (i.e. innovate by applying a new tag) when many users are exploiting (i.e. copying existing tags). Evidence for this could come in the form

of an inverse relationship between the probability of imitating an existing tag for a given resource and the the diversity (i.e. entropy) of the existing tag distribution for that resource. Also possible (as in the case of Twitter above) would be the the application of producer-scrounger models to innovate versus imitate decisions in tagging.

It is unlikely that any one of these factors is driving tagging behavior on its own, and future research will have to explore how and to what extent these different processes interact in collaborative tagging systems. Whatever the case, the social imitation framework discussed in this section serves as a useful theoretical guide for interpreting human behavior in this domain.

Study 1: Simple imitation heuristics in a collaborative tagging system[48]

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

[48] Adapted from Lorince & Todd (2013).

models, and compare their results to the extracted parameters from the tagging dataset. Results indicate that simple social copying mechanisms can generate surprisingly good fits to the empirical data, with implications for the design and study of tagging systems.

Introduction

Collaborative tagging systems have been the topic of serious research within the computer science community since the mid-2000s, gaining particular ground when Thomas Vander Wal (2007) coined the term "folksonomy" to refer to the bottom-up, user-generated organizational structure tagging systems enable. From early descriptive work (Golder & Huberman, 2006) to computational models (Cattuto, Loreto, et al., 2007) and prediction of social links from tagging behavior (Schifanella et al., 2010), our understanding of the dynamics of tagging systems has increased greatly. Despite these advances, there has been little work exploring the decision-making processes of users in tagging environments. The high-level properties of collaborative tagging systems (e.g. their emergent vocabularies, distributions of tag frequencies, and so on) are the result of human behaviors in a dynamic environment where individuals' tagging decisions structure the information environment in which subsequent decisions are made. As such, the cognitive sciences can do much to inform our understanding of how and why people tag. This work is an early step in that direction, exploring the extent to which a simple multi- agent model of heuristic decision-making can account for the aggregate distribution of tag use in a collaborative tagging system. Though tagging is clearly a complex process, with both social influence (i.e. copying of other tags) and evaluation of the tagged content playing a role, the present study explores the strength of social copying in producing global patterns of tag use of the kind we see in Last.fm.

How do People Tag?: Golder and Huberman (2006) performed one of the first in-depth studies on the dynamics of a tagging system, exploring global patterns of tag use, changes in individuals' tagging vocabularies, and classification of different types of tags on the social bookmarking website Delicious. Subsequent work has elaborated all of these research threads. Al-Khalifa and Davis and Marlow, et al. (2007), for example, have developed taxonomies of tag types and motivations for tagging (e.g. tagging for future retrieval, for contribution and sharing, or for opinion expression), and studied their implications for the study and design of tagging systems. Farooq, et al. (2007) reviews a number of other metrics for evaluating tagging systems, including tag discrimination and tag reuse, using them to develop a set of design heuristics for collaborative tagging systems. More recently, Schifanella, et al. (2010) studied the role of social relationships in tagging systems, discovering that social activity (i.e. friendships and group memberships) correlates with tagging activity, and furthermore that tagging habits were predictive of the existence of social ties between users.

This is only a sampling of the substantial body of work on tagging, but illustrates the expanding understanding of tagging systems. It remains a challenge, however, to connect high-level analyses of tagging patterns to a lower-level understanding of human decision processes in these environments. Making strong claims about cognitive processes based on Web data, whether the data is large- or small-scale, is a difficult problem (Kraut et al., 2004), but one that has attracted increasing attention among cognitive scientists in recent years (Glushko et al., 2008). Though datasets like our own do provide ecologically valid data — that is, signals of behavior carried out in people's day-to-day lives without the artificiality of a lab context — they also suffer from a lack of

experimental control. We can only make indirect inferences as to exactly what information was available to a user at the time of a tagging decision, and how this information did (or did not) factor into that decision. An understanding of why people tag to begin with is not even fully developed; theories and taxonomies of tagging motivations, as mentioned above and developed from survey research (Nov & Ye, 2010) are helping, but it is still often unclear what motivated the decisions to generate the annotations we crawl from the Web, or how differences in motivations might imply different generating processes for the choice of which tag to apply. This of course does not make the development of models of tagging behavior a hopeless pursuit, and several notable efforts have been made. For example, Cattuto et al. (2007) developed a modified Yule-Simon "rich-get-richer" model of tagging (Simon, 1955; Yule, 1925), in which new tags are added to the folksonomy with a probability P, and with probability $1 - P$ a tag is copied from the existing distribution proportional to its frequency. This was then modified to include a memory kernel that resulted in more recently used tags being more likely to be copied. Other models have focused more directly on cognitive plausibility, such as Fu, et al.'s (2009) semantic imitation model. In their model, users infer from the existing tags for a document the topics it contains, and assign semantically related tags based on that inference.

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

Ecological Rationality: Research on ecological rationality, spearheaded by the Adaptive Behavior and Cognition Group (Gigerenzer & Todd, 2000; Todd & Gigerenzer, 2012), approaches decision making from a different perspective than traditional judgment and decision-making theory. Rather than developing models that start with the assumption that people are rational, optimal decision makers, and then modifying them to account for humans' observed deviation from rational principles (a strategy especially common in behavioral economics, Berg & Gigerenzer, 2010), the ecological rationality approach takes a decidedly different view of what it means for an agent to be rational. To be ecologically rational is not to achieve optimality, or even optimality under constraints, but rather to utilize simple decision-making heuristics — rules of thumb — derived from evolutionary adaptations to problems humans faced in ancestral environments. These adaptive strategies allow humans to make quick decisions under the simultaneous constraints of time, available information, and cognitive processing capacity. For decades, this kind of decision making was lumped under the category "heuristics and biases" (Tversky & Kahneman, 1974), with the negative connotations of "bias" overshadowing the fact that heuristic strategies are often effective under the right circumstances: "Heuristics are efficient cognitive processes that ignore information. In contrast to the widely held view that less processing reduces accuracy, the study of heuristics shows that less information, computation, and time can in fact improve accuracy" (Gigerenzer & Brighton, 2009, p. 107).

There exists a large variety of cognitive heuristics, collectively making up the mind's so-called "adaptive toolbox", that vary in their appropriate domain of application. A thorough review is beyond the scope of this paper (but see Gigerenzer & Gaissmaier,

2011; Gigerenzer & Todd, 2000; Todd & Gigerenzer, 2012), and we specifically wish to focus on social heuristics. These strategies, such as imitate-the-successful or imitate-the-majority (Boyd & Richerson, 2005), leverage available social information to make quick decisions. Such heuristics of course can be problematic when misapplied (e.g. by allowing the spread of misinformation within a population), but have immense adaptive value as tools for making decisions in uncertain environments. When first-hand information is scarce or costly to acquire, emulating the behavior of successful individuals or of large groups can be an efficacious strategy.

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

tagging patterns can be explained by the use of copying heuristics, we also help to identify the extent to which other factors, including semantics, also play a role.

Last.fm: Our analysis is of an earlier version of Last.fm data described in Chapter 1. The choice of Last.fm as our object of study was, to a certain degree, arbitrary, as our questions of interest could be asked of any broad folksonomy. In a narrow folksonomy, like the photo-sharing website Flickr, users predominantly only tag content they have uploaded themselves, making questions around social copying minimally applicable. However, even among broad folksnomies, the music tagging domain is especially appropriate for our investigation. Classification of musical content is notoriously challenging, so much so that work has been published in the music informatics community explicitly asking if musical genre classification is even a problem worth pursuing (McKay & Fujinaga, 2006). We thus propose that in such an uncertain domain, social copying may play a particularly important role.

Terminological Notes: Throughout this section, an "annotation" refers to a given instance of a user assigning a tag to an item at a particular time, and can be thought of as a unique four-element tuple in the form user-item-tag-time. An "item" is a generic term refer ring to an atomic target of tagging activity on Last.fm, and can be an artist, album, or song. Last.fm maintains distinct tag distributions for each unique item (even if one is a sub-category of another, e.g. an album by a particular artist).

Dataset

Table 2 presents a detailed summary of the data used in this study[49]. Consistent with other large-scale tagging datasets, we find long-tailed distributions[50] for several key summary metrics of the dataset, visualized in Figure 19.

Table 2: Summary of annotation data. Active taggers are users with at least one annotation.

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

Simple Tagging Heuristics

Exploring the interface: In considering heuristic decision-making strategies people may use in a tagging environment, we seek possibilities that are both simple (from both the cognitive processing and computational modeling perspectives) and psychologically plausible. In the tagging domain this can be challenging, because the environment is quite complex, and the available data tell us little about precisely how users were interacting with the system when they made their tagging decisions. There are various ways for users to arrive at the tagging interface, including via the radio feature (Figure 20A), an item information page (Figure 20B), or the tag cloud for an item (Figure 20C). Furthermore, the site's API supports methods for assigning tags via external applications and services,

[49] Recall that this represents only the subset of the complete dataset describe in Chapter 1 that had been collected at the time that this project was completed.

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

and we can only speculate as to how and to what extent users tag content with such services.

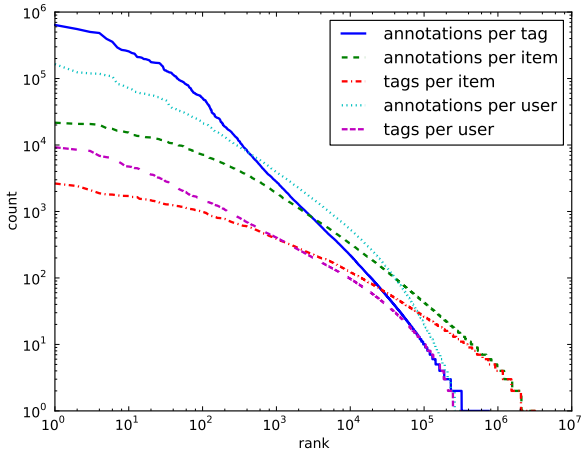
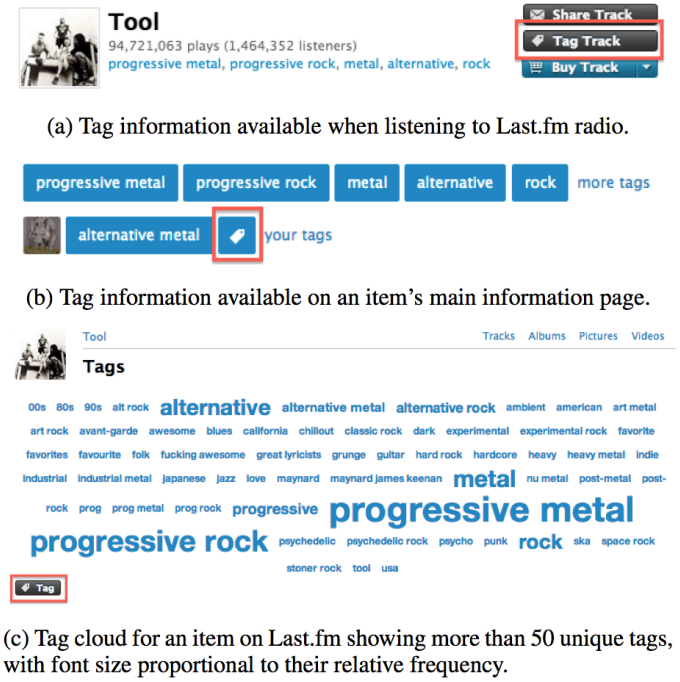


Figure 19: Frequency-rank plots of key metrics for the dataset, on a log-log scale. Displayed are the total number of times each tag was used across the dataset ("annotations per tag"), the total number of times each item was tagged ("annotations per item"), the total number of times each item was tagged ("annotations per item"), the total number of unique tags assigned to each item ("tags per item"), the total number annotations made by each user ("annotations per user"), and the total number of unique tags used by each user ("tags per user"). In all cases, values are ranked from highest to lowest ("rank" on x-axis), while the count for each metric is shown on the y-axis, and one point is plotted per user/item.

Nonetheless, we can still develop plausible hypotheses as to people's tagging strategies based on knowledge of the Last.fm interface. Though there is variance in the presentation of tag information on Last.fm, the web-based interface via which users can actually assign tags to items has a standard format (Figure 21) that prominently displays the top five most popular tags for the item being tagged, ordered by frequency, and describes them as "suggested tags". Thus when a user decides to tag an item, (s)he is presented with aggregated social data on other individuals' tagging decisions in a way that facilitates copying those decisions, explicitly presenting them as suggestions, and making their assignment to an item a simple matter of clicking the name of the tag, rather

than typing it in. Re-use of one's own commonly-applied tags ("your tags" in Figure 21) is facilitated in a similar manner, with a user's most commonly used tags listed in overall frequency order (i.e. "español" is the most commonly-used tag across all items for the user in Figure 21).



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

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

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

Figure 20: Different presentations of tagging information on Last.fm. In each image, the button triggering the display of the tagging interface is highlighted in red.

Figure 21: Web interface for tagging an item on Last.fm.

Description of possible heuristics: If people do in fact utilize simple social copying heuristics in deciding how to tag, what form might these heuristics take? Our goal here is not to develop a complex cognitive model, but rather a simple strategy that, if embodied by agents in a tagging system, would approximate at least some of the behavior we observe in our dataset. Let us formalize the decision strategy as a stochastic process in which a user, upon deciding to tag an item (we will leave aside, for the present model, how a user decides *which* items to tag) engages in copying behavior with probability P, and novel tagging behavior (that is, generating a new tag rather than copying) with probability $1 - P$. A user copies at time T by re-using a tag from the cumulative tag distribution for an item existing at time $T - 1$. We begin our discussion with three simple strategies:

1. Uniform heuristic: The simplest possible copying strategy, from an algorithmic perspective, would be to randomly pick a tag from the existing tag distribution at time $T - 1$ for the item being tagged. Though simple, this is not likely behaviorally, as there is no way for a user to review the complete tag distribution for a given item.
2. Normalized heuristic: Instead of sampling randomly, we could assume that the probability of copying any particular tag is proportional to its relative frequency for that item at time $T - 1$. Though unrealistic for the same reason as the uniform heuristic, this would at least capture the idea that more frequent tags are more likely to be copied (as they are displayed more prominently in an item's tag cloud). This model is akin to a preferential attachment on the level of items.

3. Top-5 heuristic: Rather than make the unrealistic assumption that users have access to the entire distribution of tags for a particular item (or that they would be able to effectively use that information, were it available), the top-5 heuristic assumes that, when copying, a user selects randomly among the top 5 most popular tags (the "suggested tags" in Figure 21) for an item. This remains agnostic as to how precisely the user selects which of the five top tags to copy, but provides a plausible tagging heuristic that is based on social information that is (a) directly available and (b) of manageable size. We thus hypothesized that this heuristic would be the top performer when modeled.

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

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

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

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

We calculated the copy index *P* for all possible items and months. It was of course impossible to calculate values for (a) the first month in which we had recorded annotation data for a particular item, and (b) for those items for which we only had annotation data from a single month. We were thus limited to considering 1,119,345 of the items in our dataset. There was enough variance across items and months that simple summary statistics were uninformative, so we instead considered how the three copying metrics varied as a function both of time and increases in annotations. Figure 22 shows the three copy indexes as a function of total annotations, with each point in the scatterplot indicating the copy index (y-axis) for all items with the corresponding total number of

annotations shown on the x-axis, averaged across all months for which the index was calculated. Thus each point is the average value of the copy index across all time for all items with a particular number of annotations. The lines show moving averages of the points in the scatter plot, and clearly indicate a convergence of copy index as items acquire progressively more annotations. The data is noisy for items with few annotations, but the convergence of copy index values as annotations increase allows for an estimate of the underlying value of the copy index.

The data as presented in Figure 22 may overrepresent those items with the most annotations, so we also calculated the copy index metrics as a function of time. Figure 23 shows the same three copy indexes over items' lifetimes. We calculated, for all items, the average copy index each month after the first month in which we had tagging data for that item. This was then averaged across items for each month index. If an item had no annotations in a given month, it was excluded from the calculation for that month. For example, the copy index where item age equals 30 is the average copy index of all all items in the 30^{th} month after their first annotation. Thus we have a picture of the average progression of the copy indexes over the life of items. Copy indexes for the largest values of item age show greater error because relatively few items have existed across the full 89 months for which we have data.

Making a precise estimate of the underlying value of *P* from these data is a challenge, as it shows a clear temporal dependence; all three metrics increase as items accumulate more annotations over time. This analysis does, however, provide us with a range of plausible values to focus on in the development of our multi-agent models (in the range of approximately 0.6 to 0.9 for the uniform heuristic, 0.3 to 0.5 for the

normalized heuristic, and 0.4 to 0.6 for the top-5 heuristic). Even with these estimates, it remains to be seen if models using the described heuristics as generating mechanisms can reproduce the patterns of tag popularity observed in our dataset. In the following section, we address precisely that.

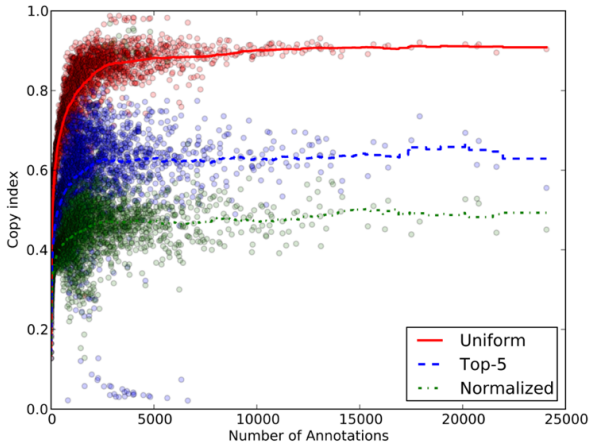


Figure 22: Copy index as a function of total annotations. Each point indicates the mean copy index across all months for items with the corresponding number of annotations. Solid lines are moving averages of increasing window size equal to $a^{0.9}$, where a is the total number of annotations shown on the x-axis.

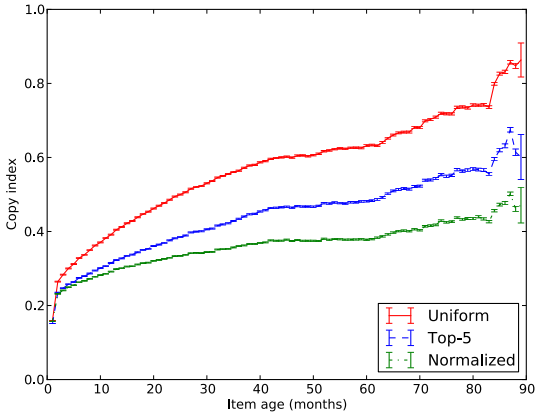


Figure 23: Copy index as a function of time. The x-axis shows the "age" of an item (months since first annotation), while the y-axis shows the mean copy index for all items x months after their first annotation. The average for a particular month does not incorporate those items with 0 annotations in that month. Error bars show +/- one standard error.

Modeling

Core model description: With a set of candidate heuristic models in hand, we developed a basic multi-agent simulation framework in which to test them. The core model required a single parameter *P* and the metric by which we evaluated the models' performance was the overall distribution of tag use it generated (c.f. the solid blue line in Figure 19). In a complex environment consisting of one million or more agents, difficult-to-predict, emergent effects are to be expected, so we constrained the model to emulate the environmental structure of the empirical dataset as much as possible, but not so much so that we trivially ensured that it would mirror the true distribution of tags. We thus ran all simulations with the same number of users, items, and possible unique tags as in our crawled data (see Table 2). Furthermore, we constrained it such that the distributions of total annotations per item and total annotations per user were the same as in the real data. We accomplished this by generating random user-item-timestep triplets that mirrored the distributions just mentioned. This amounts to the assumption that taggers' activity levels (i.e. how many items they tag), item popularity (i.e. how many times a particular item is tagged), and taggers' decisions of which items to tag are all processes independent of the heuristic process we modeled of deciding how to tag an item. This is of course a gross simplification, but permits a focused analysis of the effects of the simple heuristics described above.

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

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

The generation of a novel tag, in this simple model, was simply the assignment of a random tag from the set of possible tags (i.e. the number of unique tags in our crawled dataset).[51] The program maintained a data structure for each item containing the frequency of each item assigned to it, and tagging decisions at any given time step were based upon the distribution of tags existing for the item at that time. Tags lacked any semantic content or relationship to one another, being represented simply as integers. The same held for items and users.

Our model, especially when using the normalized heuristic, is akin to a multi-agent version of the Yule-Simon model upon which Cattuto, Loreto, et al. (2007) based their model. The key difference is that our model, in an effort to represent the information available to real users when tagging, operates on the level of item tag distributions (complete distributions for the normalized heuristic, or only the most popular tags in those distributions for the top-5 heuristic). A true Yule-Simon model, on the other hand, operates on the level of the global tag distribution, but we cannot assume that users have access to that overall distribution.

Results of basic models: We ran the first version of the model a total of 27 times, 9 times across a sampling of *P* values (from 0.1 to 0.9, in increments of 0.1) for each of the

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

three basic tagging heuristics. For each run we calculated and plotted the distribution of total tag use as in Figure 19. Figure 24 shows the results. No quantitative analysis is required to see that none of the three models – with any possible parameter value – generate reasonable fits to the data. With all three heuristic strategies, and across the range of possible *P* values, we see an underrepresentation of the most popular tags and overrepresentation of rare tags, as compared to the empirical distribution. However, it is clear, as we expected, that the top-5 heuristic comes closest to matching the empirical data among these three basic models. Although the normalized heuristic, for high values of *P*, shows a similar pattern as the top-5 heuristic, the latter model is more robust over a wider variety of parameter values, including those previously estimated. The other two models, for the range of parameter values estimated, show much poorer fits.

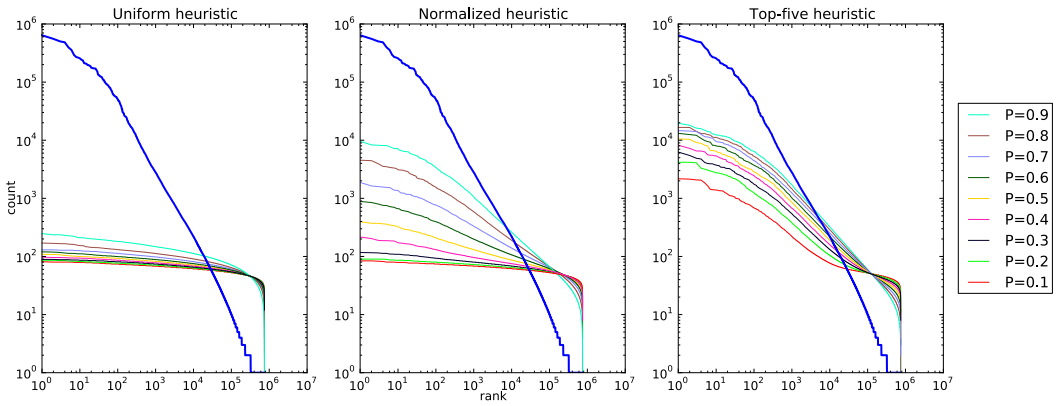


Figure 24: Frequency-rank plots of overall tag use for each of the three basic models, across a sampling of P values. The bold blue line shows the empirical tag distribution.

Model refinement: Leaving aside the somewhat unrealistic uniform and normalized heuristic models, we explored a variety of extensions and modifications of the top-5 model. With such a simple model, there are countless possible changes that could have led to better fits to our empirical data, but we only explored those options that retained

our focus on simple copying heuristics (without making assumptions about, e.g., individuals assigning semantically appropriate tags to content) and did not require that users have any knowledge about the global distribution of tags. We experimented with several model modifications, such as:

- Varying the number of suggested tags displayed to users (effectively making the "top-5" heuristic, for example, a "top-10" or "top-15" heuristic). Though not in line with our interface-based predictions, if users often explore the tag clouds (Figure 20C) of items, it is possible that the effective set of suggested tags is in fact larger than the five we hypothesized. This modification resulted in minimal change to the overall tag popularity distribution.
- implementing a fixed set of common tags that were more likely to be used than a random tag when users did not copy. This amounted to there existing a set of popular or well-known tags that all users have in mind, and are more likely to use. This modification was not successful either, as it resulted in those top tags dominating the distribution while being used with roughly equal frequency to one another.

After extensive experimentation, we found that implementing a secondary form of copying within the tagging decision process led to substantially better fits to our data. Leaving the principal copying mechanic of the top-5 heuristic as it was, we added the assumption that users keep track of the tags they have been exposed to every time they tag any item (i.e. the five top tags suggested for any item seen). This information was used to modify the novel tag generation process such that, rather than simply assigning a random tag, users would engage in one of two behaviors. With probability Q, they would

re-use (i.e., copy from a different item) a tag they had previously seen with probability proportional to the number of times they had encountered it, and with probability $1 - Q$ they would assign a random tag. This new, two-parameter version of the model embodies the simple assumption that users, when selecting a new tag to assign to an item, are more likely to use tags they are more familiar with. In pseudocode:

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

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

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

The tag rank-frequency plot generated by this distribution is shown in Figure 25. We also show, for comparison, the best fit of the standard top-5 model within the estimated range ($P = 0.6$), the empirical distribution, and a null model distribution. The null model, which is equivalent to users tagging randomly across all 33 million simulated annotations, has an equal frequency across all possible tags (approximately 44 annotations per unique tag).

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

Though the relative fits of the of the basic top-5 model versus the modified version can be directly observed in Figure 25, we can quantify them by calculating the

root mean squared error (RMSE) of each as compared to the null model. The basic model has an RMSE of 2224.72 as compared to the empirical data, versus the null model's RMSE of 2296.46 as compared to the empirical data. This represent an improvement of only about 3%, but if we bear in mind the scale of the empirical data (the most popular tag is used close to one million times), a relatively small percent improvement in RMSE corresponds to the substantial improvement in model fit that is evident in the plot. The modified model performs much better, with an RMSE of 1898.88, a 17% improvement over the null model. The final model is still far from closely matching the empirical data, in particular underestimating the prevalence of the most popular tags. The fit across the middle portion and tail of the distribution is quite good, however. Given that we had no expectations that our simple model would fully explain the data, the modest fit we find is in fact meaningful, showing what copying alone can achieve, and what is left to be explained by other mechanisms. As we have argued from the beginning, copying is of course not the only process driving tagging decisions, and the deviance between the model and empirical data at the head of the distribution is likely due to their being content-based non-copying mechanisms that lead to consensus as to how the most popular tags are applied.

Conclusions

This study has presented the following:

- A novel methodology for crawling Last.fm that allows us to explore the temporal evolution of tags used within our dataset;
- A set of possible social copying heuristic models in a tagging environment;

- Methods for estimating from empirical data plausible parameter values for our heuristic models; and,
- A set of multi-agent models employing social copying heuristics that demonstrate the extent to which the patterns of tag popularity we see in our crawled data can be explained through copying behavior.

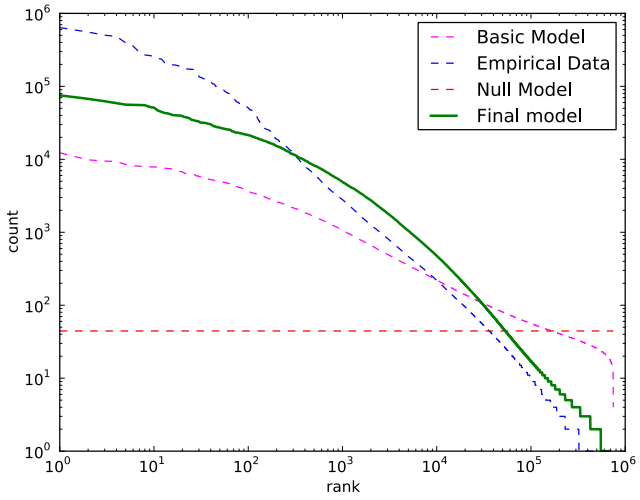


Figure 25: Frequency-rank plots of overall tag use for the fitted version of the final, two-parameter model (P = 0.6, Q = 0.9, solid line), the original top-5 model (P = 0.6, lower curved line), the null model (flat distribution), and the empirical data (topmost line).

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

extremely simple, psychologically plausible mechanisms are capable of generating data that account for a surprising amount of the empirical data as compared to a null model.

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

This work has several limitations, a number of which we have already discussed in the course of developing the model. Further unrealistic assumptions of the model could of course be mentioned (e.g. assuming homogeneity of P and Q across agents, lack of variance in agents' vocabulary sizes), but these kinds of assumptions were necessary for

the development of the simple model that interested us. One relatively small change that we would like to implement in a subsequent version of the model is a memory mechanism similar to that used in Cattuto, Loreto, et al. (2007). Having users "forget" tags that they have not seen recently may very well result in reinforcement of the popular tags that our model currently underestimates. Another possibility will be to incorporate preferential re-use of previously used tags on a user-by-user basis (assuming that the suggestion of "your tags", as seen in Figure 21, elicits a form of self- copying).

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

of tagging dynamics, and of the realism of the types of copying mechanisms we have proposed.

Study 2: Supertagger behavior in social tagging systems[52]

A folksonomy is ostensibly an information structure built up by the "wisdom of the crowd", but is the "crowd" really doing the work? Tagging is in fact a sharply skewed process in which a small minority of "supertagger" users generate an overwhelming majority of the annotations. Using data from three large-scale social tagging platforms, we explore (a) how to best quantify the imbalance in tagging behavior and formally define a supertagger, (b) how supertaggers differ from other users in their tagging patterns, and (c) if effects of motivation and expertise inform our understanding of what makes a supertagger. Our results indicate that such prolific users not only tag more than their counterparts, but in quantifiably different ways. Specifically, we find that supertaggers are more likely to label content in the long tail of less popular items, that they show differences in patterns of content tagged and terms utilized, and are measurably different with respect to tagging expertise and motivation. These findings suggest we should question the extent to which folksonomies achieve crowdsourced classification via the "wisdom of the crowd", especially for broad folksonomies like Last.fm as opposed to narrow folksonomies like Flickr.

Introduction

In social tagging systems, users annotate content with freeform textual tags that can facilitate organization, sharing, and discovery of resources. Each instance of tagging is

[52] Adapted from Lorince, Zorowitz, Murdock, & Todd (2015) and Lorince, Zorowitz, Murdock, & Todd (2014).

referred to as an annotation, and can be formally represented as a four element tuple (user-item-tag-time) indicating which user tagged which resource, the tag used, and the time of the annotation. Participation rates in these systems vary widely, from users who never tag to "supertaggers" who tag thousands of resources. This imbalance in contribution rates has important implications for how we interpret social tagging data, especially as most users are precisely that: users. They may use tags to search for or gain information about resources, but only some users actively contribute to the knowledge-generation process through tagging.

How effective or useful folksonomies are in general is not a topic we address here. Instead, our research questions the assumption that the "crowd" is at play in any meaningful way in collaborative tagging. Our results demonstrate that an overwhelming proportion of tagging is carried out by a minority of users, suggesting that the folksonomy does not necessarily represent the aggregated knowledge of its users, but is instead dominated by contributions from the few "supertaggers" among them.

Underlying this discrepancy is the fundamental issue of motivation — why do users contribute to social tagging systems in the first place? A substantial literature has explored this topic in terms of why users tag in one manner rather than another (Ames & Naaman, 2007; Nov & Ye, 2010; Strohmaier et al., 2010), but there is little work addressing what motivates some users to tag so much more than others. The differences we find between supertaggers and other users can be used to explore the motivational factors that may distinguish these two groups.

The relative importance that users place on tagging content versus other available activities is certainly a key factor here, and may explain variation in the relative

contributions of supertaggers from one tagging system to another. On the social bookmarking site Delicious, for instance, tagging and organizing bookmarks is the principle use case for the service. However, on other services tagging is a secondary feature. Systems like Last.fm and Flickr incorporate tagging features, but their principal use cases (learning about and listening to music on Last.fm, photo sharing and discovery on Flickr) do not involve tagging. Many active users never make any substantive contribution to these systems' folksonomies. Such cases, in which tagging is a deliberate choice with costs of time and effort outside the primary use of a service, partly account for the lack of tagging participation we observe.

In summary, the high-level question that interests us is this: *How does the disproportionate contribution to the folksonomy by a small number of users change the interpretation of the presumed crowdsourced nature of tagging?* In other words, does the folksonomy truly represent the collective knowledge of its users? We approach this by exploring three research questions:

- RQ1: How do we most usefully quantify the imbalanced tagging contributions observed in collaborative tagging systems, and how do we formally define the term "supertagger"?
- RQ2: How do observed patterns of tagging differ between supertaggers and other users? Do supertaggers simply tag more or do they tag differently?
- RQ3: What might be driving these differences? Can motivational or expertise effects, for instance, distinguish supertaggers from their counterparts, and how do such differences inform our interpretation of folksonomic data?

Here we address these questions across two additional large-scale tagging datasets from Delicious and Flickr. After presenting related work and an overview of the datasets, we formalize our definition of supertaggers and illustrate their disproportionate tagging contribution (RQ1). We then present our analyses of supertaggers' behaviors as compared to other users (RQ2). Next, we explore RQ3 by examining if supertaggers differ from other users in terms of motivation and expertise. We conclude by synthesizing our results and discussing future avenues for work.

Overall our findings demonstrate that a small proportion of users, the supertaggers, generate a disproportionate share of the tagging activity. This in and of itself may not be surprising, as long-tailed distributions in user activity (including, but not limited to, tagging) on the web are well-established. We also show, however, that their tagging patterns are quantifiably different than those of other users. This holds with respect to both the content they tag, most notably that supertaggers are more likely to label content in the long tail of less popular items, and the terms they use to tag it. Using established measures, as well as two novel methods, we also find that supertaggers show greater expertise and differing tagging motivations than other users. Precisely why some users tag so much more than others may be partly accounted for by describer-like, as opposed to categorizer-like, tagging motivations, but remains a direction for future work discussed further in the conclusion.

Related work

Measuring motivation in tagging: One factor that may differentiate supertaggers from other users, and may modulate levels of tagging in general, is tagging motivation: Different tagging goals may lead users to tag more or less. Though motivation in tagging

behaviors has been operationalized in numerous ways, one prominent approach (Körner, Benz, et al., 2010; Körner, Kern, et al., 2010) characterizes users as either categorizers or describers. When tagging, categorizers use a limited vocabulary to construct a personal taxonomy conducive to later browsing of tagged content. In contrast, describers do not constrain their vocabulary; instead, they freely use a variety of informative keywords to describe items, facilitating later keyword-based search. Körner and colleagues present several metrics with which to classify users according to this dichotomy, discussed more below. Other researchers have developed taxonomies of tagging motivation that can be broadly mapped onto dimensions of sociality (are tags self- or socially-directed?) and function (are tags used for organization or communication? (Ames & Naaman, 2007; Heckner et al., 2009). Methods for identifying these motivations programmatically in large-scale datasets have yet to be developed, however.

Measuring expertise in tagging: Another important consideration for studying user contributions to a folksonomy is expertise. Inevitably, some annotations will provide more useful information about an item than others. Expert users presumably generate higher quality annotations on average.

Though expertise has no single agreed-upon definition with respect to tagging, one noteworthy approach to expert detection is Spamming-Resistant Expertise Analysis and Ranking (SPEAR, Yeung et al., 2009; Yeung, Noll, Gibbins, Meinel, & Shadbolt, 2011). SPEAR as- signs an expertise score to users for each unique tag they use based on two principles. First, under a mutual reinforcement model, user expertise in a topic (as defined by a specific tag) is determined by the quality of items the user tags with that term, and item quality is in turn determined by the expertise of users who have tagged it.

Second, users who tend to tag items earlier receive higher expertise scores, as they identify new, high quality resources sooner than others. In this way, SPEAR is adept at weeding out spammers, who tend to indiscriminately annotate items with tags. The use of a spam-robust expertise measure is important, as Wetzker, Zimmermann, & Bauckhage (2008) found an overwhelming majority of the most prolific taggers in a large folksonomy were spammers.

We also consider a new expertise measure to supplement SPEAR. Because it evaluates users with respect to a given tag, SPEAR provides a useful measure of domain expertise, or knowledge of a particular topic. Our measure, on the other hand, is designed to evaluate general user expertise (i.e. across all annotations). In contrast to SPEAR, our approach evaluates users on an item-by-item basis, and assigns higher scores to users annotating an item in alignment with the consensus of annotations for that item. The details of this measure are discussed further below, but it essentially asks if supertaggers are more likely to assign "better" tags to an item, where the quality of a tag is defined in terms of how much agreement there is across multiple users that it should be assigned to an item.

A third approach to expertise is inspired by classic research on the structure of mental categories (Rogers & Patterson, 2007; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), which suggests that linguistic consensus emerges around labels/words indexing categories of an intermediate level of abstraction. For example, people prefer basic-level terms (e.g. "dog") over super- and sub-ordinate terms (e.g., "mammal" and "terrier", respectively) to refer to an object. In contrast to the consensus, experts in a given domain tend to deviate from this verbal behavior reliably (Tanaka & Taylor, 1991)

by applying more specific (subordinate) labels. Kubek, Nützel, & Zimmermann (2010) present a method based on conditional probabilities to automatically extract semantic taxonomies from folksonomic data, which allows us to calculate a depth score for each tag in the resulting taxonomy. We can then determine if supertaggers tend to use more subordinate terms, thereby showing evidence of greater expertise.

A similar approach was taken by Fu & Dong (2010), who applied a Latent Dirichlet Allocation (LDA) model to a subset of items from Bibsonomy and found that resources tagged by experts, as determined by SPEAR, contained tags more predictive of topics as compared to those by non-experts. Such an approach is not applicable here, however, due to the fact that items in our datasets are either non-linguistic in nature (photos on Flickr, music on Last.fm) or not directly available (we have only arbitrary IDs for the URLs tagged in the Delicious dataset).

Datasets

We performed our analyses on datasets from three different collaborative tagging systems, the social music site Last.fm, the photo-sharing site Flickr, and the social bookmarking tool Delicious. The targets of tagging are, respectively, music (users can tag artists, albums, or songs), photos (users tag the images they upload to the site), and web bookmarks (users save and tag links to webpages).

The Flickr and Delicious datasets were collected by Görlitz, Sizov, & Staab (2008) and are publicly available.[53] These datasets consist exclusively of annotation data

[53] http://www.uni- koblenz- landau.de/de/koblenz/fb4/ AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_ html

(i.e. tuples in the form user-resource-tag-date). The Last.fm dataset, on the other hand, is a subset of that described in Chapter 1.

Crawling methodology: Chapter 1 covers the collection of the Last.fm data, so here I describe only the collection of the Delicious and Flickr datasets.

The Delicious dataset consists of an effectively random sample of users for whom complete tag histories were collected, much like our Last.fm data. For this data, however, annotations are recorded at an increased temporal resolution (seconds). The Flickr dataset was crawled at the tag level (i.e. the complete sets of annotations associated with an effectively random set of tags were collected). While this means that we cannot guarantee that any particular user's tag history is complete, we can assume that the number of tags "missed" is approximately equal across all users, such that the relative annotation counts over users accurately represent the true distribution. We base this assumption on (a) the sheer size of the sample (113 million annotations across 1.6 million tags), and (b) the fact that the distribution of annotations per user is generally consistent with the other two datasets. Further details of how the Flickr and Delicious datasets were collected can be found in Görlitz et al. (2008).

Overall the distributions of tagging activity are consistent both across datasets (as is particularly evident in Figure 26) and with previous work examining similar datasets (e.g. Figueiredo et al., 2013).

Data Summary: Table 3 gives an overview of the tagging data from all three tagging systems. Note that the "taggers" column reflects the total number of users with ≥1 annotation. While the crawling methods used in Görlitz et al. (2008) are such that only users who have tagged are included, our Last.fm data also includes users who have never

tagged. Unless otherwise noted, however, all analyses of Last.fm presented here are limited to the subset of users who have tagged at least once. Across all systems, an "annotation" refers to a given instance of a user assigning a particular tag to a particular item at a particular time.

Table 3: Global tagging data summary.

Dataset	*Taggers*	*Tags*	*Resources*	*Annotations*
Last.fm	521,780	1,029,091	4,477,593	50,372,895
Flickr	319,686	1,607,879	28,153,045	112,900,000
Delicious	532,924	2,481,108	17,262,475	140,126,555

Table 4: Median number of annotations per user (A_u), tag (A_t), and resource (A_r) across datasets. Interquartile range (25th percentile – 75th percentile) in parentheses

Dataset	A_u	A_t	A_r
Last.fm	7 (2-29)	1 (1-4)	2 (1-6)
Flickr	41 (12-175)	2 (1-10)	3 (2-5)
Delicious	41 (9-188)	1 (1-4)	2 (1-5)

Even at this high level of description, substantial differences between these systems are apparent. Users clearly tag more overall on Flickr and Delicious (both with medians of 41 annotations per user) than on Last.fm (median of 7 annotations per user). Though the median numbers of annotations per item are similar across datasets, Last.fm has the greatest ratio of annotations to items tagged (11, versus 4 and 8 for Flickr and Delicious), suggesting a stronger trend towards popular, heavily tagged items. See Table 4 for a summary of per-user, per-item, and per-tag median numbers of annotations (given the long-tailed distributions of these measures, the mean does not accurately capture the central tendency of the data).

These observations are consistent with the design of these systems. Following Vander Wal's (2005) terminology, Last.fm is a broad folksonomy in which many users tag the same, publicly available resources (i.e. multiple individuals tagging the same artists, albums, and songs), while Flickr is a narrow folksonomy, in which users predominantly tag their own photos.[54] Delicious exists somewhere between these two extremes: On the one hand, users use the service to manage their own resources (in this case, Web bookmarks), much like Flickr. But on the other, multiple users can save and tag the same URL (either independently, or by exploring the bookmarks saved by other users on the site). Note that on Delicious and Last.fm, users receive the top five most popular tags for an item as recommendations when tagging it. On Last.fm (but not Delicious) users can browse the full tag distribution for an item, as well. On Flickr, where users tag only their own photos, such popularity-based recommendations are of course not possible.

The difference in sheer volume of tagging between the systems is also of note, with Last.fm having well less than half the total number of annotations of the other systems (despite having a comparable number of users). This is again consistent with how the systems are used. Tagging is a more central activity on both Flickr and Delicious, as users actively contribute and organize resources. Last.fm, on the other hand, is primarily used for music consumption, and tagging is generally speaking a non-primary activity for users.

---

[54] In the vast majority of instances, photos can only be tagged by the users who upload them, and in our data no single photo has been tagged by multiple users.

Supplemental Last.fm data: For Last.fm, we crawled a total of nearly 1.9 million users, of whom about 28% had tagged at least once. In addition to the tagging data, we recorded friendship relations, group memberships, loved/banned songs,[55] and self-reported demographic data. For a subset of our users, we also have collected full song listening (scrobble[56]) histories. Table 5 summarizes the supplemental data collected.

Table 5: Supplemental data summary for Last.fm

Total users	1,884,597
Friendship relations	24,320,919
Total loved tracks	162,788,213
Total banned tracks	23,321,347
Unique groups	117,663
Users with scrobbles recorded	73,251
Total scrobbles	1,181,674,857
Unique items scrobbled	32,864,795

Identifying "Supertaggers" and Measuring their Influence

Figure 26 presents the distribution of per-user annotation counts for our three datasets in a traditional manner. For a given number of annotations on the x-axis, the corresponding y-axis value indicates how many users have generated that many total annotations. Plotted on a log-log scale, the distributions take roughly linear forms consistent with long-tailed, power-law-like distributions.[57] Though Flickr is more variable for lower annotation counts than Last.fm and Delicious, the distributions generally decrease

[55] "Loving" a track is roughly equivalent to favoriting a tweet, or other similarly-defined activities, while "banning" allows a user to indicate disliked items and exclude them from any recommendations by Last.fm.
[56] Last.fm tracks users' listening for music recommendation purposes, and "scrobble" is the term for an instance of a user listening to a particular song at a particular time.
[57] We do not examine the precise mathematical form of the distributions, as it is not relevant to our analyses.

monotonically with increasing annotation counts, indicating that users with relatively small numbers of annotations are much more common in all three services.

These long-tailed distributions make it clear that there exist a relatively small number of prolific users generating many annotations and a large number of users generating only a few annotations. But to show exactly how pronounced this pattern is, we plot in Figure 27 the proportion of total annotations generated by the most prolific taggers against the proportion of top taggers considered (i.e. the proportion Y of annotations generated by the proportion X of top taggers, ranked by total number of annotations).

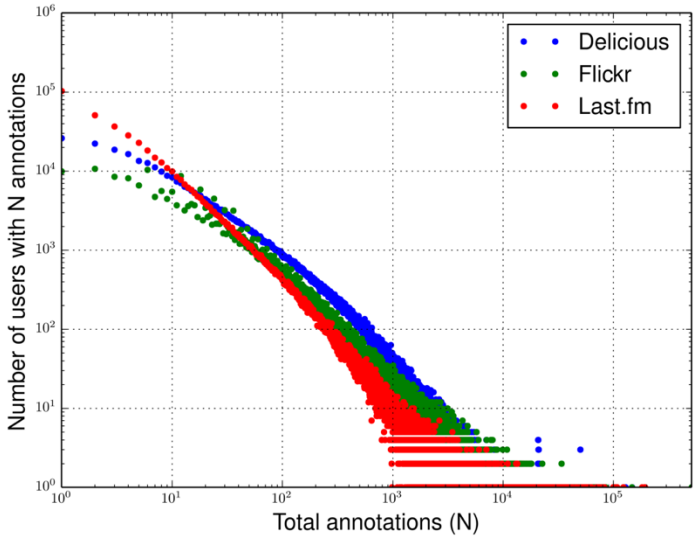


Figure 26: Frequency distributions of per-user annotation counts.

This representation highlights just how skewed these distributions are, even more so than predicted by the Pareto Principle (Newman, 2005), or 80-20 rule, under which we would expect 80% of annotations to come from the top 20% of users. Last.fm is the most extreme case, with 80% of all annotations generated by less than 7% of users, but both

Flickr and Delicious show similarly skewed patterns, with approximately 12% and 16% of users, respectively, responsible for 80% of all annotations. These findings are corroborated by calculating the Gini coefficient, a measure of income inequality:

$$G = \frac{2\sum_{i=1}^{n} i y_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

where values of y are individuals' "wealth" (here, their total numbers of annotations), indexed by i in non-decreasing order ($y_i \leq y_{i+1}$), and n is the total number of individuals. Values range from 0, indicating total equality (all individuals have equal wealth) to 1, total inequality (one individual has all the wealth). For all three datasets, we find high values of the Gini coefficient (0.806 for Delicious, 0.847 for Flickr, and 0.898 for Last.fm), indicating a small number of users performing most of the tagging. As expected, the effect is most pronounced on Last.fm.

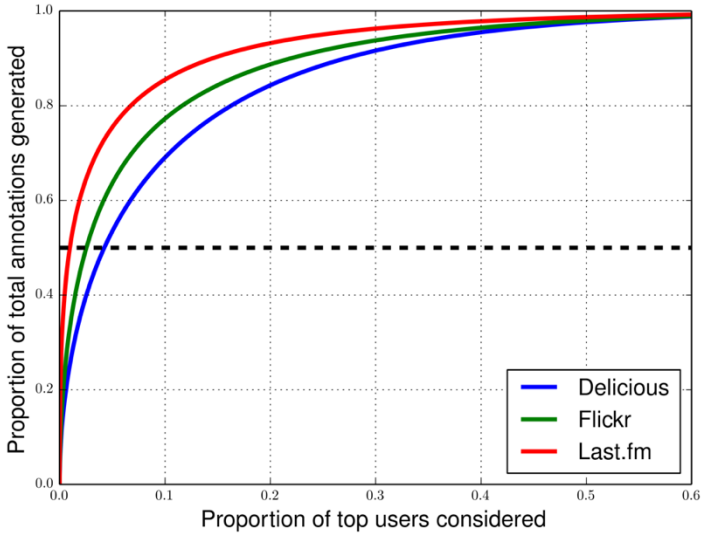


Figure 27: Proportions of total annotations generated by the most prolific taggers as a function of the proportion of top users considered. The dashed line shows the threshold used to identify supertaggers.

Our analyses do not reveal the existence of a clear split that naturally divides users into supertaggers and non-supertaggers, so an a priori definition of supertaggers based on annotation counts is necessarily arbitrary. One option would be to echo the Pareto principle, considering supertaggers to be the top 20% of users, but under this definition over 90% of tagging activity would be from supertaggers, making it difficult to compare the aggregated activity of supertaggers versus non-supertaggers. Instead, we elected to split the data in half with respect to annotations, allowing us to compare equally-sized sub-folksonomies (in terms of total annotations) from supertaggers and non-supertaggers (although it does mean the number of users we classify as supertaggers is much smaller). We hope these analyses will show the value of considering more prolific taggers separately, and can lead to more precise methods for identifying supertaggers in future work.

Thus, *we formally define "supertaggers" as the topmost prolific taggers accounting for half of all annotations in each dataset*. This split is marked by the horizontal dashed line in Figure 27. Under this definition, 5,086 users (0.97%) from Last.fm, 8,142 users (2.55%) from Flickr, and 22,630 users (4.25%) from Delicious are classified as supertaggers. These correspond to annotation thresholds (i.e. the number of annotations required to be a supertagger) of 1,457, 2,701, and 1,285, respectively.

We reiterate that the particular threshold used here is arbitrary, in that there is no special behavioral shift that occurs at this point. In fact, various behavioral measures show relatively smooth changes as we consider users with progressively more annotations, and as such, we present measures as a function of users' total annotation counts (rather than simply comparing averages for supertaggers and non-supertaggers)

wherever possible. Nonetheless, in those analyses that directly compare the annotations of the two groups we have defined, the 50 percent split is both clearly interpretable (in that it compares "normal" users to the most prolific ones), and analytically convenient, as it normalizes all analyses of the sub-folksonomies such that the total number of annotations in each is constant.

Differences in tagging patterns

This section presents analyses comparing the tagging patterns of supertaggers to those of their non-supertagger counterparts. Except where otherwise noted, analyses for each dataset were performed on two "sub-folksonomies", one containing all annotations by supertaggers, designated S*, and the other containing all annotations for non-designated* ¬S*. Summary measures for these groups across all datasets appear in*

Table 6, and
Table 7 shows the relevant per-user medians of these values. We analyze differences in tagging behavior from three perspectives:

- How similar is the vocabulary of supertaggers and non- supertaggers?
- How much does the content tagged by supertaggers and non-supertaggers overlap?
- How similarly do supertaggers and non-supertaggers tag particular content?

Variation in tagging vocabulary: *We first ask how similar the tag vocabularies are between supertaggers and non-supertaggers, independent of the content they are annotating. While the set of non-supertaggers clearly employ a larger aggregate vocabulary (see*

Table 6), the median number of unique tags per user is much greater for the supertaggers (see
Table 7). However, both groups' vocabularies are largely shared, with most annotations coming from the set of shared tags occurring at least once in both *S* and ¬*S* in each

dataset (95%, 88%, and 93%, respectively, for Delicious, Flickr, and Last.fm). This suggests the existence of many "singletons" – tags used only once – and other tags used only a small number of times. This is verified in Figure 28, which compares the distributions tagging activity over tag popularity for S and $\neg S$. The plot shows the proportion of total annotations within each sub-folksonomy allocated to tags with a given total annotation count. Singletons and other low-frequency tags are clearly very common, and the u-shaped distributions show that, across both S and $\neg S$, and also across datasets, tagging is concentrated on a few popular tags and the many singleton tags (with proportionally little use of moderate-popularity tags). Note that on Delicious very popular and very rare tags show similar overall proportions of use indicating a more varied vocabulary, while on Flickr and Last.fm popular tags are proportionally more common.

To directly measure the similarity between the vocabularies of S and $\neg S$, we use two simple summary measures: the rank correlation, Spearman's ρ, of tags for each folksonomy (measuring how similar the rank order popularity is between the two vocabularies) and the cosine similarity between the two global tag vocabularies (i.e. calculated across vectors of the frequency of each tag in each of the two folksonomies). Considering all tags, we find low rank correlations of $\rho = -0.402$ for Delicious, $\rho = -0.256$ for Flickr, and $\rho = -0.219$ for Last.fm. In contrast, cosine similarity between S and $\neg S$ is high, with values of 0.979 for Delicious, 0.963 for Flickr, and 0.872 for Last.fm. These give rather opposing impressions of the distribution similarities, so it is informative to consider these measures for smaller subsets of the data.

Table 6: Summary statistics. "Total Tags" represent all distinct tags used by each group, while "Unique Tags" are those tags appearing in only the supertagger folksonomy (S) or

the non-supertagger folksonomy (¬S). "Shared Tags" are those tags used at least once in both S and ¬S. Corresponding counts are shown for items.

	Delicious		*Flickr*		*Last.fm*	
	S	*¬S*	*S*	*¬S*	*S*	*¬S*
Users	22,630	510,294	8,142	311,544	5,086	516,694
Annotations	70,062,323	70,064,232	56,449,589	56,450,411	25,185,082	25,187,811
Total tags	1,210,748	1,698,863	803,722	1,094,358	399,552	797,784
Unique tags	782,245	1,270,360	513,521	804,107	231,307	629,539
Shared tags	428,503		290,251		168,245	
Total items	8,039,337	11,516,472	10,339,003	17,814,042	2,992,045	2,515,069
Unique items	5,746,003	9,223,138	10,339,003	17,814,042	1,962,522	1,485,546
Shared items	2,293,334		0		1,029,523	

Table 7: Per-user summary statistics. Shown are the median number of annotations, unique tags, and items tagged per user in S and ¬S, across datasets. Interquartile range (25th percentile - 75th percentile) in parentheses.

	Delicious		*Flickr*		*Last.fm*	
	S	*¬S*	*S*	*¬S*	*S*	*¬S*
Annotations	2,115 (1,597-3,227)	36 (8-153)	4,615 (3,417-7,315)	38 (11-153)	2,586 (1,860-4,457)	6 (2-27)
Tags	427 (265-644)	19 (6-60)	326 (129-670)	8 (3-24)	209 (97-405)	4 (2-12)
Items	641 (428-981)	14 (3-62)	952 (629-1,499)	16 (6-56)	882 (483-1,652)	3 (1-15)

We calculated both measures for the top N tags in both sub-folksonomies as a function of increasing N. For example, if $N = 100$, we consider tag-frequency vectors for the top 100 most frequent tags in each sub-folksonomy (considered independently) and then calculate the rank correlation and cosine similarity of these vectors between S and $\neg S$. Tags that appear in S but not $\neg S$ (and vice versa) are assumed to have rank $N + 1$ for the purposes of calculating the rank correlation, and frequency of zero for the cosine similarity calculation. This was repeated for N from 1 to 100,000.

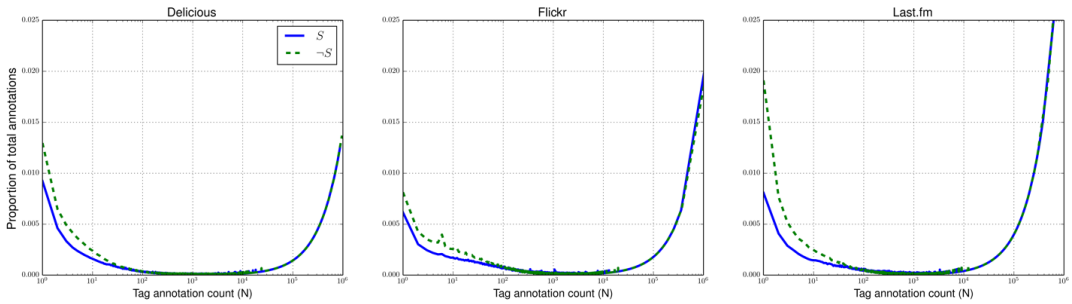


Figure 28: Distributions of tag usage for S (blue) and ¬S (green) for all datasets. Each point indicates the proportion of total annotations within a sub-folksonomy allocated to tags that have been used N total times.

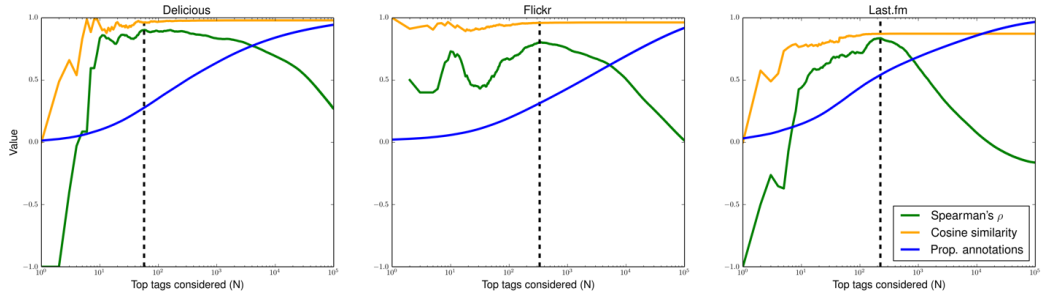


Figure 29: Spearman's ρ and cosine similarity between S and ¬S as a function of N, considering only the top N most popular tags overall from each sub-folksonomy. Also plotted is the proportion of all annotations across the full folksonomy using the combination of the top N tags from each group. The vertical dashed lines indicate the maximum values of ρ and occur at N = 57 (Delicious), N = 333 (Flickr), and N = 224 (Last.fm).

Figure 29 shows the results, additionally plotting the proportion of total annotations (i.e. across the full folksonomy) generated by the combination of the top N tags from S and $\neg S$ (e.g. for $N = 100$, what is the sum proportion of total annotations from the top 100 tags from S and the top 100 from ¬S?).[58] Unsurprisingly, the rank correlation is noisy for small N, but across all datasets has a distinct peak (0.902 for Delicious, 0.804

[58] Clearly, the top 100,000 tags make up an overwhelming majority of total annotations. But note that the combination of the top N tags from S and $\neg S$ contains more than N unique tags overall. The proportion shown for $N = 100{,}000$ tags, for example, corresponds to a total of 144,622, 155,702, and 160,470 unique tags, respectively, from the global folksonomies of Delicious, Flickr, and Last.fm.

for Flickr, and 0.836 for Last.fm) after which it decreases monotonically. Cosine similarity is also noisy for small N, but clearly stabilizes near the peak in ρ (at approximately 0.96 for Delicious, 0.96 for Flickr, and 0.87 for Last.fm).

These observations suggest a "core" vocabulary of common tags that S and $\neg S$ more or less agree on, the size of which is estimated by the index of the maximum value of ρ. The fact that cosine similarity stabilizes at approximately the same N for which ρ begins to decrease suggests that the cosine similarity between S and $\neg S$ is driven by the most popular tags (i.e. the top N occurring before the peak in ρ). Thus we find core vocabulary sizes of the datasets are 57, 333, and 224, respectively for Delicious, Flickr, and Last.fm (indicated by the dashed vertical lines in Figure 29). These core vocabularies in turn account for 28%, 31%, and 54% of all annotations (i.e. across S and $\neg S$) from the datasets. A possible explanation of the higher percentage observed for Last.fm is the existence of a relatively well-defined, constrained set of canonical music genres ("rock", "jazz", "classical", and so on) that are common in music tagging. There is not an obvious analog to these popular categories with respect to photos (Flickr) or web bookmarks (Delicious). Running contrary to this reasoning, is the fact that on on Flickr there is strong agreement as to the most popular tags (as evidenced by relatively high cosine similarity and ρ for low N). Despite this agreement, a relatively lower proportion of annotations come from these agreed-upon popular tags, suggesting that the "vocabulary" on Flickr is well-defined, but not broadly used. Thus, a second and possibly more important factor in the higher proportion of annotations from the core vocabulary on Last.fm may be that Last.fm facilitates observation of other user's tagging habits (through publicly viewable tag distributions on resources that are not available on Delicious or

Flickr). This may allow for the emergence of a large-scale, socially shared vocabulary responsible for most annotations.

The Delicious data is unique with respect to the slow decay of ρ, suggesting that the core vocabulary may be considerably larger. This is consistent with the distribution observed in Figure 28. Also note the second, earlier peak in ρ for Flickr. This suggests the existence of a smaller subset of popular tags (the top 10) within the larger core vocabulary. In fact, eight of the top 10 tags ("2004", "2005", "family", "friends", "japan", "party", "travel", and "wedding") are the same for S and $\neg S$.

Taken together, these results suggest that supertaggers share a core, popular vocabulary with other users, but deviate with respect to the many idiosyncratic and "singleton" tags.

Differences in tagged content: *Having explored aggregate differences in vocabulary, we now ask if the resources tagged by supertaggers differ from those tagged by other users. Supertaggers clearly tend to tag many more items than other users (by at least an order magnitude, see*

Table 6 and
Table 7). Last.fm is particularly notable here, as the total number of items tagged by supertaggers actually exceeds that tagged by other users. Overlap is substantial, however, with 66% of annotations for Delicious and 78% for Last.fm occurring for items tagged by both groups. Note that users in Flickr exclusively tag their own photos, so the sets of Flickr items tagged by S and $\neg S$ are totally disjoint.

Similar to the analyses of tags performed in the previous section, we compare how users' annotations are distributed over item popularity (as measured by total

annotation count). How- ever, Figure 30 shows the cumulative proportion of tagging in S and ¬S allocated to items tagged at least a given number of times.

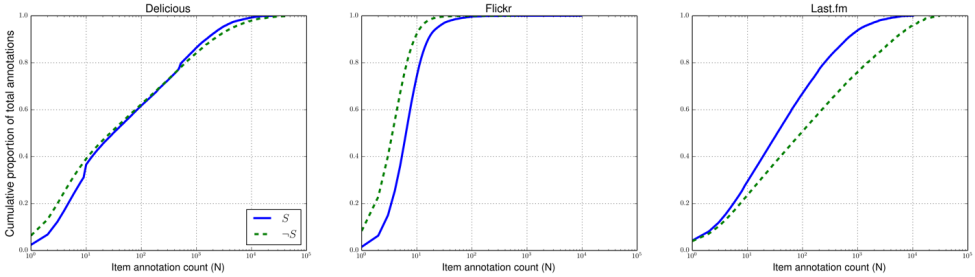


Figure 30: Distributions of item tagging for S (blue) and ¬S (green) for all datasets. Each point indicates the cumulative *proportion of total annotations within a sub-folksonomy allocated to items that have been tagged at least N total times.*

In contrast to tag usage, we find much more pronounced differences here, both between *S* and ¬*S*, and between datasets. On Delicious and Flickr, supertaggers assign proportionally fewer annotations to infrequently-tagged, less popular items (i.e., those with < 10 total annotations). As both are narrow folksonomies, this is consistent with users generally tagging their own, idiosyncratic content. In contrast, Last.fm supertaggers are proportionally *more* likely to tag less popular content. This suggests that most (non-supertagger) users are more likely to tag shared, popular content on a broad folksonomy like Last.fm.

Note that Flickr shows much greater dominance (across *S* and ¬*S*) of singleton and near-singleton tagging. This is confirmed by calculating the Gini coefficient over item annotation counts, which is quite low for Flickr (0.376 for both *S* and ¬*S*), suggesting high "equality" over items (i.e. most items are tagged a similar number of times). The corresponding Gini coefficient for Delicious is much higher (0.703 for *S*, 0.713 for ¬*S*). This finding is not surprising, though, given that on Flickr users are only tagging their own items. Thus the distribution of tagging over item popularity for Flickr

is, in effect, showing the number of times users in S and ¬S tend to tag each of their uploaded photos.

We also replicate the correlation and similarity analysis from the previous section, but this time comparing the distributions of tagging over items, as opposed to over tags (Figure 31) for the top 1,2,…,*N* items, following the procedure illustrated in Figure 29. From these results, we can conclude the following: First, the "core" set of tagged items, as operationalized by the peak in Spearman's ρ as in the previous analyses, is much larger for items than it is for tags (over 11,000 items for Delicious, and close to 1,000 for Last.fm). Second, although the cosine similarity of item vectors stabilizes at relatively high values (0.89 for Delicious, 0.76 for Last.fm), these similarities, as well as the corresponding peaks in rank correlation (0.681 versus for Delicious, and 0.540 for Last.fm), are substantially lower than the corresponding values for the tag distributions. Finally, the proportion of total tagging activity from the core set of items is lower than that from the core tags (21% for Delicious and 12% for Last.fm).

Taken together, these results indicate that there is quantifiably less similarity between *S* and ¬*S* with respect to what is tagged than which tags are used (for Delicious and Last.fm; again, there is no overlap for Flickr). Therefore, the "core" of heavily tagged items is much less clearly defined than the "core" of common tags.

One analysis permitted by our supplemental Last.fm data is to compare the popularity of the items tagged by *S* and ¬*S* using an exogenous (i.e. independent of tagging) measure of popularity, namely the global number of scrobbles (listens). In Figure 32A we plot the mean number of annotations in S and ¬S for items with a

particular global (i.e. across all users) number of scrobbles.[59] Though the overall shapes of the distributions are similar, there is a small but reliable effect of supertaggers being more likely to tag items with lower scrobble counts than other users. This is clarified Figure 32B, which shows the difference in mean number of annotations of items as a function of global scrobble count for *S* and ¬*S*. Thus, according to an exogenous popularity measure, we find that supertaggers are disproportionately likely to tag less popular content, while non-supertaggers are more likely to tag popular content. This is consistent with the findings with respect to non-exogenous popularity in Figure 30.

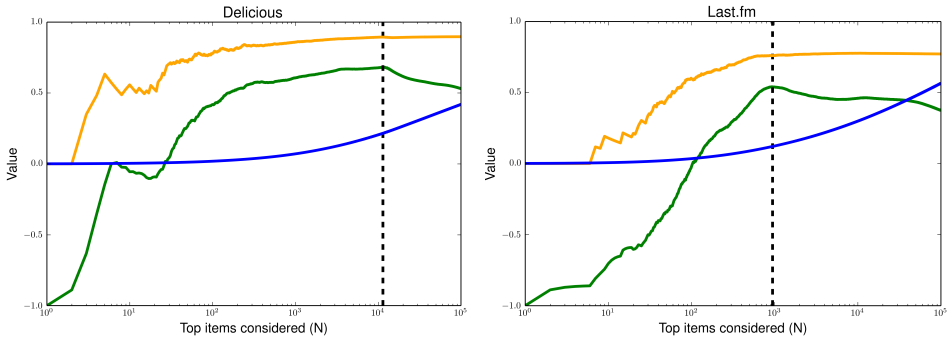


Figure 31: Spearman's ρ and cosine similarity between S and ¬S as a function of N, considering only the top N most popular items overall from each sub-folksonomy. Also plotted is the proportion of all annotations across the full folksonomy assigned to the top N items from both groups. The vertical dashed lines show the maximum values of ρ and occur at N = 11,434 (Delicious) and N = 943 (Last.fm).

Consensus effects: Having explored aggregate-level differences in items tagged and vocabulary used, it is reasonable to ask – for those items tagged in both *S* and ¬*S* – whether or not supertaggers agree with other users as to how particular items ought to be tagged. Various existing work has established that tagged resources tend to show consensus effects as they accumulate annotations, as measured by a stabilization of the

[59] This measure is limited to tagged songs, not albums or artists.

relative proportions of different tags assigned (Golder & Huberman, 2006; Robu, Halpin, & Shepherd, 2009). Our approach is different however, as we are curious if there is consensus between the two different groups of taggers we have defined for those items tagged by both groups.

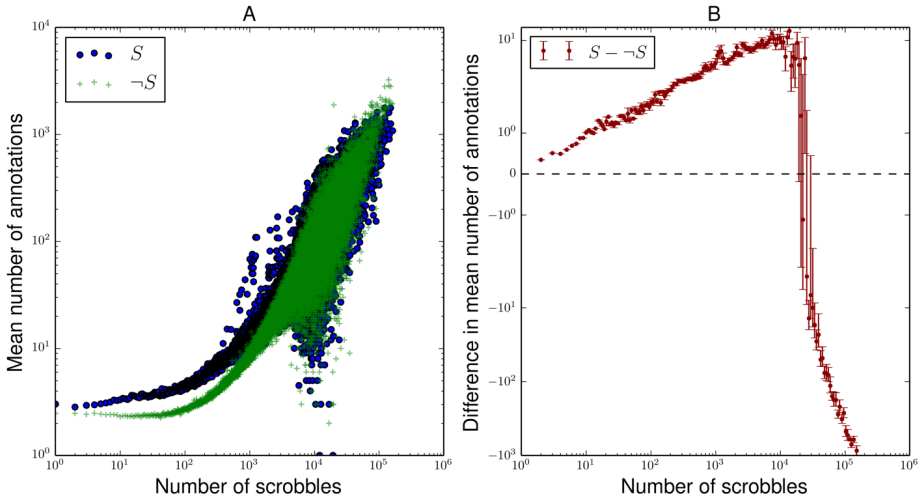


Figure 32: Mean number of annotations by S and ¬S on Last.fm for items with a given global *scrobble count (A), and difference in mean number of annotations between S and ¬S (B). Differences in B are plotted as function of logarithmically-binned scrobble count, and error bars show ±1 standard error.*

We present two formulations of consensus here. The first, and simplest, measures whether or not the most popular tag for a given item in *S* is the same as in ¬*S*. The second mea- sure is the cosine similarity between the distribution of tags assigned to an item in *S* and in ¬*S*. These allow us to measure consensus at two levels of granularity, with the first addressing the question of whether users in *S* and ¬*S* agree as to the single "best" tag for an item, and the second measuring the overall level of agreement between the two groups. Because we know resources' overall tag distributions tend to stabilize as they accumulate more annotations, we calculate these measures for all items, averaging over

items with similar numbers of annotations.[60] The results are shown in Figure 33 for Last.fm and Delicious (the analysis is impossible on the disjoint Flickr distributions).

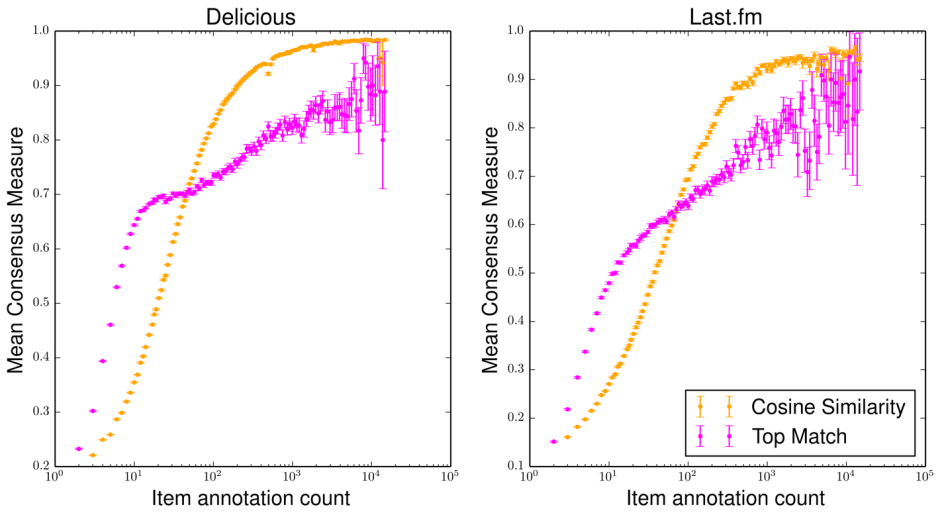


Figure 33: Mean proportion of items on which both groups agree as to the most popular tag ("Top Match", magenta) and cosine similarity (orange) of tag distributions, as a function of logarithmically- binned annotation count. Error bars show ±1 standard error.

The results are consistent with existing work, as items that have been tagged more show, on average, both greater similarity in the distributions of tags assigned to them and in the probability that S and $\neg S$ will agree as to the top tag. Thus it appears that, as items accumulate more total annotations, S and $\neg S$ tend to converge as to how those items ought to be tagged. It is notable, however, that on Last.fm the consensus values are uniformly noisier and lower on average than on Delicious. This is true despite the fact that the Last.fm data has a greater percentage of total items tagged by both S and $\neg S$

[60] Because data is sparse for high annotation counts, simply averaging over items with the same annotation count results in a noisy plot that does not make clear the general trends in the data. Thus we bin the data logarithmically (such that the bins for larger annotation counts are wider), and then average over the values within each bin. For this particular plot, the bins are cut at the values $2^i, i \in 0.0,0.1,0.2,...14.0$, and other logarithmically binned plots use a similar procedure (only varying the maximum value of i so as to accurately capture the range of the data).

(78% versus 66% for Delicious). Thus, despite being a broad folksonomy with more users socially tagging shared content, Last.fm demonstrates less pronounced consensus effects.

What makes a supertagger?

We have presented the differences in tagging habits between supertaggers and other users, but what might be driving them? Two reasonable questions to ask are (a) do supertaggers' motivations for tagging differ from those of other users, and (b) Are supertaggers "better", or more expert, taggers? Here we briefly address these questions quantitatively.

Motivational effects: Our characterization of user motivations follows that of Körner and colleagues (Körner, Benz, et al., 2010; Körner, Kern, et al., 2010), who locate users along the categorizer-describer spectrum. Categorizers are users who constrain their tagging vocabularies to construct personal taxonomies for later browsing; in contrast, describers annotate content freely with a wide assortment of tags to facilitate later keyword-based search. We quantified user motivation along this spectrum using three metrics developed by Körner and colleagues: tags per post (TPP), tag/resource ratio (TRR), and the orphan ratio (OR). TPP measures the number of distinct tags a user annotates an item with on average. Based on Körner's results, we expect describers to annotate items with more tags on average, and thus score higher on this measure. TRR is the ratio of the vocabulary size of a user to the total number of items tagged by that user. We expect categorizers to maintain their limited, personal taxonomies in tagging, and thus use fewer unique tags overall, thereby scoring lower on this measure. OR relates the vocabulary size of a user to the number of seldom-used tags for this user (i.e. what

proportion of a user's tags are "orphans"?). We expect describers to be less motivated to reuse tags, and thus score higher on this measure. Though there exist other measures of motivation, we limit our analyses to these three in light of previous research reporting high correlations between TPP, TRR, OR, and other measures (Zubiaga, Körner, & Strohmaier, 2011).

Figure 34 presents the TPP, TRR, and OR scores as a function of users' total annotation counts. Across all datasets, although the data is unsurprisingly noisy for high annotation counts, TPP scores tend to increase as total annotations increase. This suggests that users in *S* are not simply annotating more items; rather, they are, on average, annotating any given item with more tags than those in ¬*S*. We find a similar trend for OR scores: the number of orphaned tags in the vocabulary of a user increases as a function of that user's total annotation count. These two results suggest that supertaggers are more like describers than are non-supertaggers.

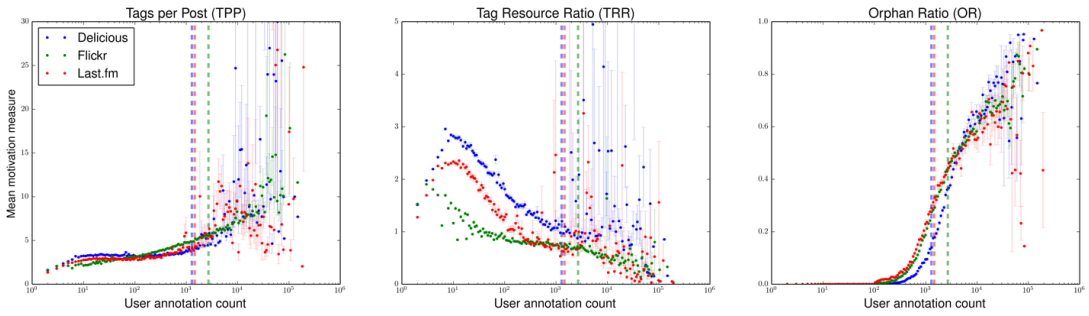


Figure 34: Mean categorizer/describer measures from Körner, Kern, et al. (2010) as a function of logarithmically binned annotation count. Shown are Tags Per Post (TPP), Tag-Resource Ratio (TRR), and Orphan Ratio (OR). Error bars show ±1 standard error. Vertical dashed lines show the supertagger/non-supertagger thresholds.

The trend of decreasing TRR scores as a function of total annotations across all datasets presents as a challenge to this interpretation. As explained above, greater TRR scores are characteristic of describers. We believe the discrepancy can be explained,

however, by the growth rate of user vocabularies. Cattuto, Loreto, et al. (2007) report sub-linear growth of user tag vocabularies as compared to the total number of annotated items, perhaps reflecting a saturation point in the number of unique tags a given user will employ.

In sum, then, our results suggest that supertaggers are better characterized as describers, who are more likely to tag the same items multiple times on average, while drawing from a more diverse set of tags. Non-supertaggers, on the other hand, are better characterized as categorizers, maintaining smaller, more structured tagsets. This suggests that supertaggers and non-supertaggers may differ with respect to motivations in tagging. This may be reflective of differences in approaches to item retrieval, with supertaggers annotating with more tags for easier lookup. There are of course other approaches to studying tagging motivation, and further work is needed to establish what factors might drive certain users to tag far more than others.

Are supertaggers expert taggers? A second feature by which to characterize the users of a folksonomy is their expertise, or quality of information provided in their tagging contributions. There is not, however, a single agreed-upon definition of expertise with respect to tagging. As such, we quantify expertise using three approaches. The first, SPEAR, is an established measure that determines expertise through two principles, mutual reinforcement and time of tagging (i.e. expert taggers tend to tag things earlier). Second, we introduce a new measure that, in contrast to SPEAR's tag-level approach, evaluates users on item-by-item basis, assigning higher scores to users annotating an item in agreement with the consensus of annotations for that item. A third measure, inspired

by longstanding psychological research, defines expertise as the increased usage of subordinate terms in semantic taxonomies. These approaches are detailed below.

Spamming-resistant Expertise Analysis and Ranking (SPEAR) has two core mechanisms. First, it is a mutual reinforcement model based on the HITS ranking algorithm (Kleinberg, 1999), in which user expertise in a topic (as defined by a particular tag) is based on the quality of the items tagged, and an item's quality is in turn based on the expertise of the users tagging it. Second, it incorporates a discoverer/follower mechanic by assuming that the first users to annotate an item with a particular tag are better at identifying high quality items and are more likely to be experts than those annotating after them. Thus SPEAR is able to rank users in terms of their expertise (or authority) in a topic (tag), favoring those users who are among the first to "discover" an item by tagging it. In other words, expertise in a tag is quantified such that users who are among the first to annotate high quality items with that tag are assigned the highest scores. For full details of the algorithm, see Yeung et al. (2009) and Yeung et al. (2011). We used the default parameters of the algorithm in the results presented here.

It is important to emphasize that SPEAR is by design a measure of domain expertise in that it provides user expertise scores for a particular tag. That is, each user receives one score per tag that is independent of her scores for all other tags. SPEAR's mutual reinforcement model does not result in the score for each of a user's tags being on the same scale, making the computation of an overall expertise score something of a challenge. As an attempt to address this, we standardized all scores corresponding to a given tag to mean = 0 and standard deviation = 1 (*z*-score), relative to the distributions of

scores for that tag across all individuals using it. Though this allows for a mean score per user, we reemphasize that this is not part of the original intentions for SPEAR.

For our SPEAR analyses, we used a subset of each dataset corresponding to the top 10,000 most popular tags overall that have at least 10 unique users. This decision was made in part due to computational limitations on calculating scores for all tags,[61] and also because SPEAR generates unreliable scores for tags that are very rarely used.[62] Despite including only a small proportion of total unique tags, this trimming still gives us a reasonable coverage of the annotation data; the top 10,000 tags account for approximately 82.6% of annotations from Last.fm and 83.6% from Delicious. Note that the Flickr dataset was excluded from this analysis as its design feature of exclusive self-tagging prohibits the mutual reinforcement necessary for SPEAR. Further, these calculations are performed over the full folksonomy for each dataset (i.e. not considering S and $\neg S$ independently). Figure 35 presents average user SPEAR expertise scores as a function of user annotation count. We find an overall positive relationship in both Last.fm and Delicious such that average user expertise increases with number of annotations, although the data is noisy for high annotation counts. This suggests that supertaggers are not only more prolific, but also more expert, in their tagging behavior, at least as defined by SPEAR.

Despite similar overall trends, there are noteworthy discrepancies in the shapes of the SPEAR score distributions between Delicious and Last.fm. Last.fm exhibits an earlier

---

[61] Exploratory analyses with various thresholds, however, yielded qualitatively similar results.
[62] SPEAR strongly rewards users for being among the first to use a given tag, so for tags used by only one or a small number of users, the algorithm necessarily leads to radically inflated, uninformative expertise scores.

peak in the growth of user expertise scores as total annotations increase, and scores remain consistently higher afterwards. Interpreting this finding, however, is complicated by the fact that SPEAR was not intended to be used as a measure of overall user expertise in the way we have presented. Though we took additional steps to make the user expertise scores across tags comparable (via the score standardization), we have still used SPEAR in a non-traditional manner and hesitate to make strong claims as to what may be driving the observed differences between datasets. A further complication arises as a result of the low temporal resolution of our Last.fm data, which makes it difficult to determine the true sequence in which tags were assigned to an item (because we only know the month in which an annotation was made, all annotations in the same month are necessarily treated as having been generated simultaneously). This issue is not relevant to the Delicious dataset.

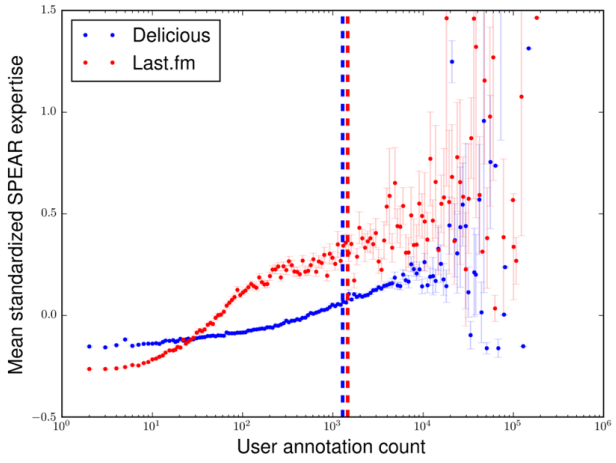


Figure 35: Users' mean standardized SPEAR expertise scores as a function of logarithmically binned annotation count. Vertical dashed lines show the supertagger thresholds for each dataset. Error bars show ±1 standard error.

Given these issues, we developed two novel measures of expertise targeted specifically at providing general expertise scores for each user. Our measures capture two

contrasting, but reasonable characterizations of expertise. The first, described in this section, assumes that an expert user will, on average, be in agreement with the consensus on how an item should be tagged. In other words, an expert should be in alignment with the "wisdom of the crowd", being more likely to assign the "correct" tag to a given item. Though SPEAR indirectly captures consensus through mutual reinforcement, our new measure explicitly does so by computing the popularity of a user's tag choice in annotating an item relative to the most popular tag for that item over the entire lifetime of the item, for each annotation the user generates. Thus, whereas SPEAR calculates expertise over tags, our measure generates a raw score for each of a user's annotations relative to the item tagged, without reference to how that tag is used globally. To calculate a user's overall average expertise score, the score for each of a user's annotations is weighted by the logarithm of the total annotation count for the item tagged. In this way, tagging items with very low annotation counts contributes little to a user's overall score, while annotations of heavily tagged items contribute more. This captures the intuition that the more total times an item has been tagged, the better defined the tagging consensus for that item is. Note that our novel measure also does not take time into consideration. For our measure, it matters less when a user annotates an item than it does that her tagging choice coincides with the eventual consensus of other users for that item.

The measure is formally defined as follows. For each annotation (user-item-tag triple, ignoring time) we calculate a raw expertise score, $E_{u,i,t}$:

$$E_{u,i,t} = \frac{F(t,i) - 1}{\max\left(F(x,i), x \in T(i)\right.}$$

where the expertise score assigned to a user u annotating item *i* with tag *t* is the frequency of that tag, *F(t, i)*, in the overall distribution for item *i*, divided by the frequency of the most popular tag for that item (max $(F(x,i), x \in T(i))$, where *T(i)* is the set of all tags assigned to item *i*). As this is a consensus-based metric, we ignore a user's own contributions to that tag distribution by subtracting 1 from the numerator.[63] To determine a user's average expertise score over all his or her annotations, $\bar{E}_u$, we calculate the following weighted mean:

$$\bar{E}_u = \frac{\sum_{i,t} E_{u,i,t} W_{u,i,t}}{\sum_{i,t} W_{u,i,t}}, W_{u,i,t} = \log\left[\left(\sum_{t=1}^{n} F(t,i)\right) - F(u,i)\right]$$

That is, we calculate a weighted mean of the expertise score for each of a user's annotations, where the weight, *W*, is equal to the the logarithm (base 10) of the total number of annotations across all users for the item tagged ($\sum_{t=1}^{n} F(t,i)$), minus the current user's contribution to that distribution (i.e. the total number of times that user tagged the item, *F(u, i)*). In cases where a user has assigned multiple tags to the same item, we only include the single highest expertise score for that item in the user's mean score. In this way, we capture whether or not the user knows the "best" tag for an item, without penalizing her if she additionally assigns other tags to it. This, combined with the low weighting of items tagged only a few times, sidesteps the issue that supertaggers tend to use many idiosyncratic tags (which necessarily are not "expert" tags under any consensus-based measure). In effect, this analysis allows us to determine whether supertaggers show expertise with relatively popular tags, while not considering their

[63] Though it is not shown here, in the event that the user has tagged an item with the most popular tag, we assign a consensus score of 1.

usage of idiosyncratic tags.[64] For the same reason as above, Flickr was excluded from this analysis.

Figure 36 presents mean consensus-based user expertise as a function of user annotation count. In contrast to SPEAR, and somewhat surprisingly, the results show an inverse-u shape. Expertise scores increase monotonically as a function of annotation count for $\neg S$ (left of the dashed lines in Figure 36); the growth of expertise scores, however, tapers off for S before de- creasing substantially for the most prolific taggers. Expertise, then, as defined by agreement with the consensus, increases with a user's number of annotations for all but the most prolific of taggers.

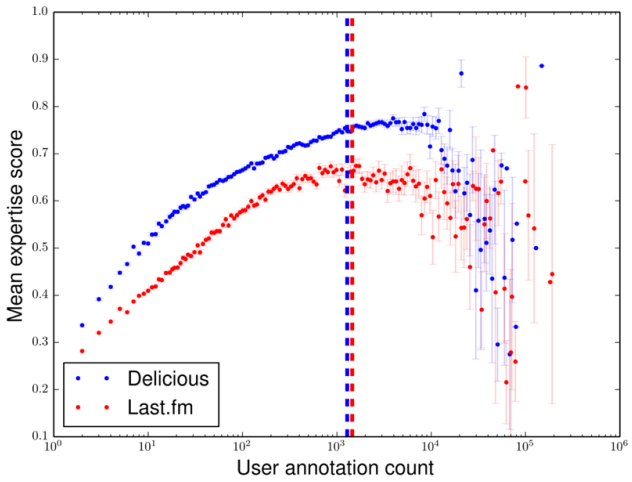


Figure 36: Users' mean consensus-based expertise scores as a function of logarithmically binned annotation count. The vertical dashed lines show the supertagger thresholds for each dataset. Error bars show ±1 standard error.

Of note, however, is the high variability in the consensus-based expertise scores of these most prolific taggers. Though this is in part due to a rapidly diminishing sample

[64] Additionally, we ran exploratory analyses in which we only considered tags with at least 10 unique users, thereby calculating expertise scores for only non-idiosyncratic tag use. The results were qualitatively similar.

size of users as total annotation count increases, it may also be indicative of divergent tagging away from the consensus in a subset of users in *S*. Whether this is a legitimate difference in tagging behavior, an artifact of spammers, or simply noise is beyond the scope of this paper, but again emphasizes the importance of investigating separately the behaviors of supertaggers and non-supertaggers.

There is also a clear difference between datasets for users with similar numbers of annotations in Figure 36. Users in Delicious show reliably higher expertise scores as compared to similarly prolific users on Last.fm. This result is not entirely surprising, however, given the results discussed above. We found that supertaggers and non-supertaggers showed greater agreement as to how items ought to be tagged on Delicious than Last.fm, so it is little surprise that, on average, consensus based expertise scores are higher for Delicious. Exactly what is driving this is less clear, however. Both Delicious and Last.fm provide frequency-based tag recommendations (i.e. suggesting the top five most popular tags for an item in the tagging interface) that presumably would encourage consensus effects, but Last.fm is unique in that users can choose to explore the full tag distributions for an item and thereby be exposed to more tags. Other factors beyond the scope of this paper are certainly at play, both social (e.g. do the publicly shared tag distributions for items somehow encourage greater tagging diversity?) and content-based (is music inherently more difficult to reliably classify than are webpages?), and deserve attention in future work.

An alternative, more psychologically grounded approach to measuring expertise is one based on classic research (Rogers & Patterson, 2007; Rosch et al., 1976) showing that people tend to prefer basic level categories to describe objects, whereas domain

experts are more likely to use subordinate labels to describe objects within in their area of expertise. For example, non-experts may refer simply to a "tree" (a basic level category), while a botanist is likely to identify that same tree at a lower level (e.g. "spruce", a sub-ordinate category). Neither group is likely to refer to it simply as a "plant" (a super-ordinate category). Applying this to our analyses, if supertaggers demonstrate more expertise than other users, we should expect from them to use more sub-ordinate terms on average than other users.

To test for this, we employ an method developed by Kubek et al. (2010) to create a tagging taxonomy for each folksonomy. Their algorithm measures the conditional probabilities of tag co-occurrence over items to infer when one tag is a sub-class another. For example, the term "classic rock" is more likely to co-occur with the term "rock" than vice-versa, and is therefore likely to be a sub-class of "rock". Full details of the algorithm appear in the original study, but it involves first calculating all pairwise conditional probabilities between tags (i.e. both $P(A|B)$ and $P(B|A)$), then, following thresholds defined in Kubek et al. (2010), defining the sets of sub- and super-classes for each tag. From these values, a taxonomy defining the hierarchal relationships between all tags can be extracted. Once this taxonomy is defined, we can use a given tag's depth in the tree as a proxy for how sub- or super-ordinate of a term it is (i.e. tags at root nodes are presumably super-ordinate terms, while tags further down the tree are sub-ordinate terms).

Because the method involves calculating all pairwise similarities between tags and is thus computationally intensive, and also because rarely used tags are unlikely to co-occur across enough items to effectively determine their associated conditional

probabilities, we limit the analysis to the subset of tags used in the above SPEAR analyses (the top 10,000 tags with at least 10 unique users across the full folksonomy). For this measure, however, Flickr data can be analyzed. As mentioned before, however, this gives us good coverage of annotations across datasets (82.6% for Last.fm, 83.6% for Delicious, and 67.6% for Flickr). The analysis results in a forest of taxonomies per folksonomy, as not all tags fall under a single root node. Each tag was part of one of these taxonomies or else a disconnected node. We filtered our results to exclude disconnected nodes, as these are tags without discernible relationships to other tags, and thus without well-defined taxonomy depth. This reduced our set of considered tags from 10,000 to 4,724 for Last.fm (covering 68.1% of all annotations), 2,224 for Flickr (29.3%), and 5,148 for Delicious (44.1%). Each tag is assigned a simple depth score based on the taxonomy (i.e. a root node has a score of zero, its children have a score of one, its children's children have a score of two, and so on). Furthermore, we normalize these raw scores, which ranged from zero to a maximum depth of five, by dividing them by the maximum depth of the branch the node was contained in, thus normalizing to a 0-1 scale. This normalized score is our measure of term-depth expertise, with the assumption that leaf nodes are more specific than root nodes.

After mapping each tag to its depth score, a user's average depth-score expertise is simply the average over annotations (i.e. over each instance of using a tag) of tag depth scores. In contrast to our other expertise measures, we do not observe systematic increases in expertise as annotation counts increase, and generally speaking there is no substantial difference in expertise between supertaggers and non-supertaggers (hence we have not included these results visually). We repeated the analysis, however, at the

vocabulary level, which means we averaged each user's tag depth scores over the unique tags she employs, regardless of how many times each tag was used. These results, presented in Figure 37, show a small but clear effect of increasing depth scores for the most prolific users.

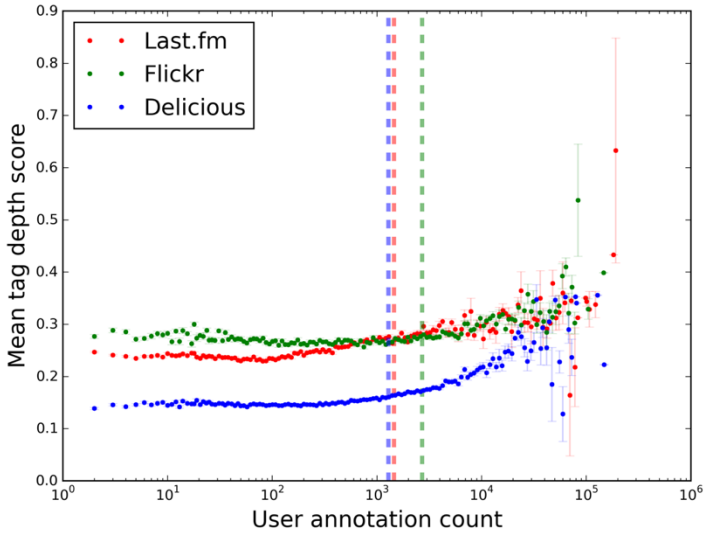


Figure 37: Users' vocabulary-level *mean term-depth expertise scores as a function of logarithmically binned annotation count. The vertical dashed lines show the supertagger thresholds for each dataset. Error bars show ±1 standard error.*

Thus we can conclude that, at least for the tags we were able to examine with this method, the most prolific taggers tend to have vocabularies consisting of more subordinate terms as compared to other users. This trend does not extend to their aggregated tagging activity – that is, when averaging depth scores across all of a users' annotations, rather than across only unique tags – as we observed with the previous two expertise measures. The method is also limited in that we can only make claims about a proportion of unique tags in the dataset. Future work may be able to use an information theoretic approach to compare the specificity of tags for the purposes of expertise analysis, and thereby improve on the method used here. Nonetheless, these findings are

suggestive that supertaggers introduce more terms associated with greater expertise into the folksonomy than other users, even if they do not make up a large proportion of their tagging activity.

Discussion and conclusions

The principal contributions of this work are the following:

- A formalization of the disproportionate contribution by "supertaggers" to a folksonomy;
- an analysis of the differences between these taggers and their non-prolific counterparts, at the levels of the users themselves and the folksonomic structures they generate; and
- an analysis of the role of expertise and tagging motivation in these differences, including novel metrics for tagging expertise.

Our results demonstrate that the most prolific taggers are not simply generating a greater volume of annotations in a manner consistent with "the crowd". Instead, their tagging patterns are quantifiably different from those of other users across datasets. With respect to tag vocabulary, we find that both groups use many of the same most popular tags, but disagree on the long tail of less common tags. With respect to items tagged, results differed across the datasets. On Last.fm, supertaggers allocate proportionally more annotations to less popular items than do other users, while on Flickr and Delicious the opposite trend held. On Last.fm, this suggests that the tagging of users in *S* is more exploratory, disproportionately tagging content in the long tail of obscure, unpopular items. This was confirmed by an exogenous measure of item popularity, as well (see Figure 32).

We hypothesize that these differences are driven by the different interaction paradigms of broad and narrow folksonomies. In narrow folksonomies, such as Delicious and Flickr, users are predominantly uploading and tagging their own content (photos and bookmarks). Distributions of tagging over item popularity are thus driven at least partly by the number of times users tend to tag any given item. In broad folksonomies, such as Last.fm, shared items are publicly tagged by multiple users, precisely the process that ostensibly allows for the "wisdom of the crowds" to emerge in collective classification. However, we found that users demonstrate less consensus about what is tagged and what tags are used than on Delicious or Flickr. On Delicious, multiple users may save and tag the same bookmarks, but presumably tag privately for their own organization of bookmarks. Thus one could reasonably expect lower consensus across users as to how those items should be tagged in contrast to Last.fm, where users knowingly tag publicly shared content, but we found exactly the opposite. This surprising result demands further work exploring precisely what factors make consensus more or less likely in collaborative tagging.

Supertaggers also differed in their scores on previously established measures of user motivation. Supertaggers are not only annotating more often but also using broader tag vocabularies and assigning more unique tags to each item. These results suggest not only that supertaggers are better characterized as describers than categorizers, but also that they may differ in the underlying reasons for their patterns in tagging, perhaps reflective of different information retrieval strategies.

We also found that average user expertise increases as a function of number of annotations made by the user, at least for the majority of users. This result was

consistently found across two measures of expertise, the SPEAR algorithm and our novel measure, both of which define expertise in part based on consensus in tagging with the majority of users. A third measure of expertise, based on the specificity of terms used (i.e. sub-ordinate versus super-ordinate), did not show a similarly clear relationship between amount of tagging and expertise, but does suggest supertaggers are more likely to have more expert tags in their vocabularies than other users. Across all three measures, we found expertise scores vary considerably for the most active of users. This was especially true for our consensus-based expertise measure, where expertise scores exhibited an overall decline for the most prolific taggers. Due to small numbers of the most active users it is difficult to discern if this finding is reflective of a discordant subset of supertaggers or simply a result of noise. Thus, refinements of expertise measurement in tagging remain a useful avenue future work.

Of course, expertise can be defined in many ways. Two of the measures used here assume a connection of expertise to consensus. Implicit in this, however, is a further assumption that the crowd will invariably arrive at an accurate description of an item. This may not be so, especially in the case of broad folksonomies where social interaction may have an important effect on the formation of tag distributions across items. For example, the first annotations of an item may be the most important to the resulting consensus of tags if social imitation is at work. If supertaggers are more often than not among the first users to annotate an item, as our SPEAR results suggest,[65] then it is

[65] This is confirmed directly in the case of Delicious by supplemental analyses (not presented here) examining only the average point at which supertaggers annotate items relative to other users. As expected, more prolific taggers do tend to tag items earlier than others. In the case of Last.fm, however, the low temporal resolution of our data makes the results of such an analysis unreliable and difficult to interpret.

possible that their early contributions may shape the resulting distributions in ways favorable to higher expertise scores as defined by our measures. Especially given that our non-consensus-based expertise measure did not show the same pattern as the other measures, our results suggest the need for exogenous measures of expertise. This can be explored in future work by defining a user's expertise in a musical genre based on her listening habits, and then exploring how this interacts with her patterns of tagging.

While tagging has been hailed as an example of the "wisdom of the crowd", we have shown that the majority of tagging is not done by "the crowd" at all. These results call for questioning just how much "collective classification" is actually happening in social tagging systems. While the most popular items are tagged by many users, the long tail of less popular items are being tagged mostly by supertaggers, especially in broad folksonomies like Last.fm. This is not necessarily an argument against folksonomies, especially if supertaggers are shown to be experts. We only claim that when designing or studying social tagging systems we need to be sensitive to not only variation in how much people tag, but the varying manners in which they do so. Determining whether the "division of labor" we see among taggers serves to generate a more (or less) usable semantic structure than would be created by users with more homogenized tagging strategies is a promising direction for future research.

There are of course limitations to the methods we have used here. Most notable is our arbitrary partitioning of users into supertaggers and non-supertaggers (by splitting the data in half on total annotations), especially given that none of our results provide evidence of a strong qualitative division of users into these two groups. To be clear, we do not make the claim that there is anything inherently meaningful about the particular

threshold we used to define supertaggers. Rather, we used this division of users to demonstrate (a) the extreme skew in tagging contributions towards prolific taggers, and (b) that users at one end of the annotation spectrum tag quantifiably differently from users at the other end. But our division of users means that it could be the case, for instance, that the measured differences in motivation of supertaggers are partly a function of their tagging more obscure items. This might occur if more obscure items do not fit canonical musical categories and demand multiple classifications such that users tagging them appear more like describers than categorizers, even when this does not reflect a fundamental motivational difference. Relatedly, the motivations of supertaggers may not reflect internal, stable user traits but may instead result from interacting with the folksonomy over time. By virtue of discovering more obscure items through increasing use, users' motivations may

transition from resembling categorizer to describer behavior for the reasons described above. In fact, we do not yet understand whether any of the observed differences between supertaggers and non-supertaggers reflect anything inherently different about users, or might predictably emerge as users tag more. Much future work thus remains to be done to understand how tagging patterns evolve over time. There also is the question of the extent to which our results were affected by spam tagging, which we did not directly address here. Effective identification and elimination of prolific spam taggers might shift the dominance in annotation counts away from the most prolific taggers.

Finally, our analyses do not account for within-user variation, which may turn out to be crucial to this kind of work. For example, we generated a single, overall expertise score for every user, even though that may not be appropriate. Users presumably show

varying levels of expertise in different domains, such that the average across those domains may not be meaningful. A similar case might be made for the simple classification of users by annotation count. A user who tags many items within a single topical domain is behaving differently from one who tags the same number across a broad variety of domains.

A major question left unanswered in this work is why the differences we have observed exist. Why do some users become supertaggers, while others tag very little at all? We can only speculate at this point, but our results are suggestive of some possible explanations. First, differences in tagging motivations, as measured by Körner and colleagues' methods, suggest that supertaggers behave more like describers than categorizers, so it is possible that describer-like tendencies encourage users to tag more (though, of course, the reason for these tendencies remain unknown). Second, differences in expertise are suggestive that more expert users tend to tag more, but it remains unknown which way the causality runs: Is greater expertise a result of more tagging, or increased tagging rates a result of greater expertise? Studying trends in expertise over time within users could help shed light on this question.

Social tagging systems represent intriguing environments in which a subset of users are highly active in the absence of clear incentives for doing so (there are no explicit rewards for being a supertagger, social or otherwise). In some cases, spamming may be at play, but it is doubtful that this accounts for all or even most cases of supertagging observed in our data. Though we are unable to answer the question of why such pronounced differences in tagging activity exist, we believe our analysis of how

prolific and non-prolific taggers differ represents a substantial contribution to understanding such systems.

Despite the need for further investigation, our work nevertheless presents compelling evidence that the bulk of tagging activity comes from a minority of users whose tagging behavior is quantifiably distinct from that of other users. Thus, it is important for both researchers and designers of collaborative tagging systems to identify and differentially interpret the metadata generated by these supertaggers in order to understand and promote the use of these systems by all.

Chapter 4: Linking consumption and classification of content[66]

I have so far presented work on content consumption (music listening, Chapter 2) and content organization (tagging, Chapter 3), but this chapter will explore how these two processes interact. There are, of course, various possible relationships between consumption and classification, but the analyses here are aimed at answering a very specific question: Do tags function as retrieval aids? This work constitutes both an integration of my two main research areas, and specifically an example of exploring human memory cue use in an ecologically valid context.

Chapter 3 presented a detailed summary of the work on tagging motivation, but stepping back from the motivation literature specifically, there is a substantial literature on the dynamics of tagging behavior and related concepts. Research has covered topics as diverse as the relationship between social ties and tagging habits (Aiello et al., 2012; Schifanella et al., 2010), vocabulary evolution (Cattuto, Baldassarri, Servedio, & Loreto, 2007). mathematical and multi-agent modeling of tagging behaviors (Cattuto, Loreto, et al., 2007; Lorince & Todd, 2013), identification of expert taggers (Noll, Au Yeung, Gibbins, Meinel, & Shadbolt, 2009; Yeung et al., 2011), emergence of consensus among taggers (Halpin, Robu, & Shepherd, 2007; Robu et al., 2009), and tag recommendation (Jäschke, Marinho, Hotho, Schmidt-Thieme, & Stumme, 2007; Seitlinger, Ley, & Albert, 2013), among others. In this broader literature, there is little nuance with respect to tagging motivation, and 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 empirical evidence demonstrating that this is in fact the case. Here we present

---

[66] Adapted from Lorince & Todd (2016) and Lorince, Joseph, & Todd (2015).

several analytic methods we are using to explore how patterns of content tagging and interaction support or refute the hypothesis that tags function as retrieval cues. 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.

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 a 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.

We begin this discussion with a brief summary of 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. Next, I present an overview of three possible analytic approaches to addressing these questions. Finally, I present in detail a study employing the most promising of these methods.

Background

Chapter 3 covered collaborative tagging systems and the research on user motivation in such systems, so 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. Hunt & Seta, 1984; Weist, 1970) 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

The exploratory analyses presented in this chapter, employ a subset of the data described in chapter 1, summarized in

Table 8. The formal study presented below uses a slightly expanded version of that data, described therein.

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 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.

Table 8: 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)

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 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,[67] as the user may as 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

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

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

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 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 user'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. Tagging levels show a long tailed distribution in which most users tag very little, and a small number tag a great deal (qualitatively these distributions are similar to those shown in Chapter 3). Although we average across users for the analyses presented here, these discrepancies between typical taggers and "supertaggers" – the implications of which were discussed in Chapter 3 – 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 38, 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.

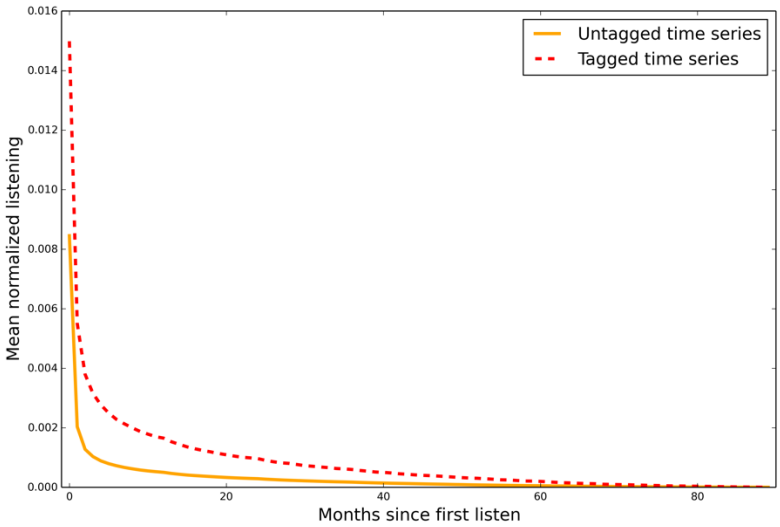


Figure 38: Comparison of tagged and untagged listening patterns. Shown is mean normalized playcount each month (aligned to the month of first listen) for all listening time series in which the user never tagged the associated artist (solid line) and listening

time series in which the user tagged the artist in the first month she listened to the artist (dashed line).

We have also explored 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, applying Gaussian kernel smoothing, and then feeding many such time series into a clustering 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 as the mean of all its constituent vectors. This process repeats iteratively until the distribution of vectors over clusters stabilizes. In Figure 39 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 38. 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.

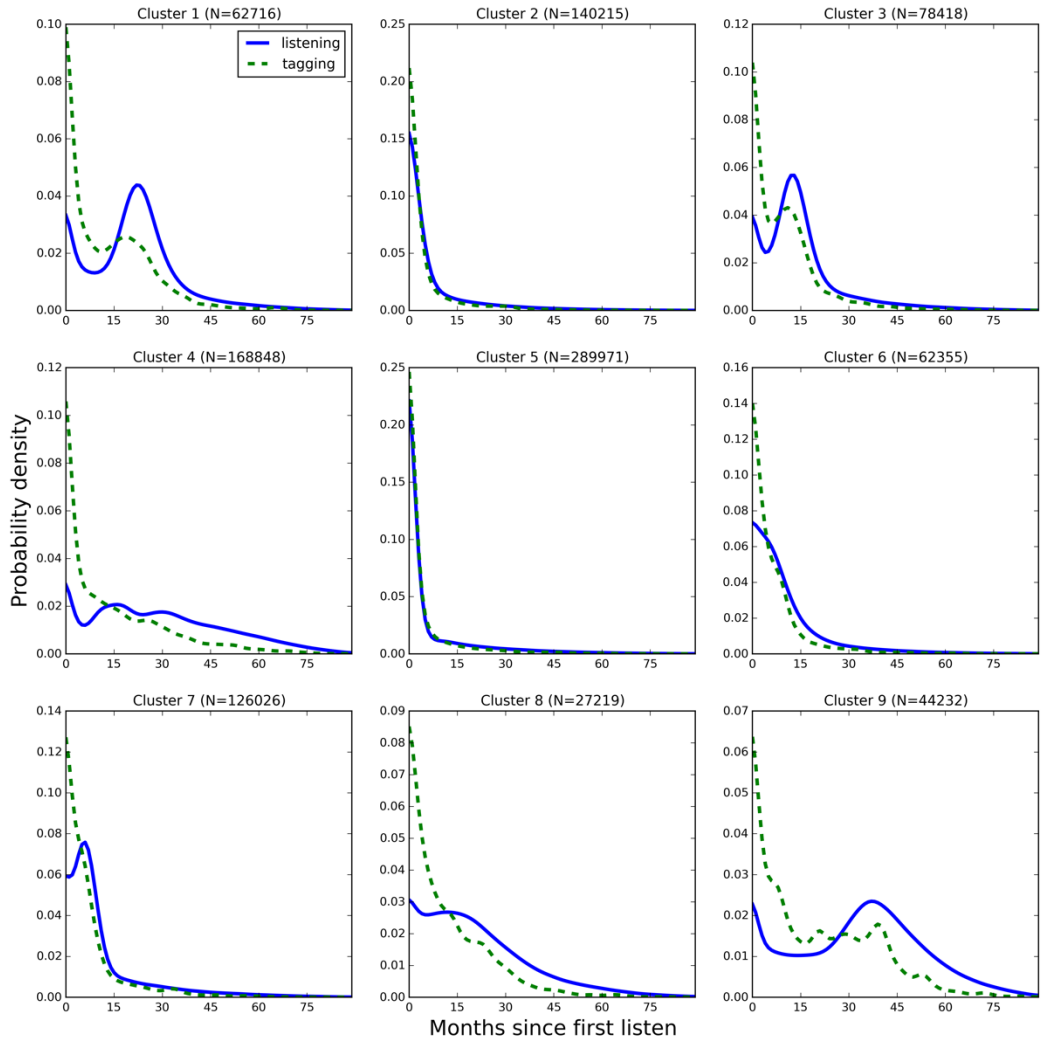


Figure 39: Clustering results for k = 9. Shown are mean normalized playcount (solid line) and mean number of annotations (dashed line), averaged over all the time series within each cluster. Time series are converted to probability densities, and aligned to the first month in which a user listened to 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.

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

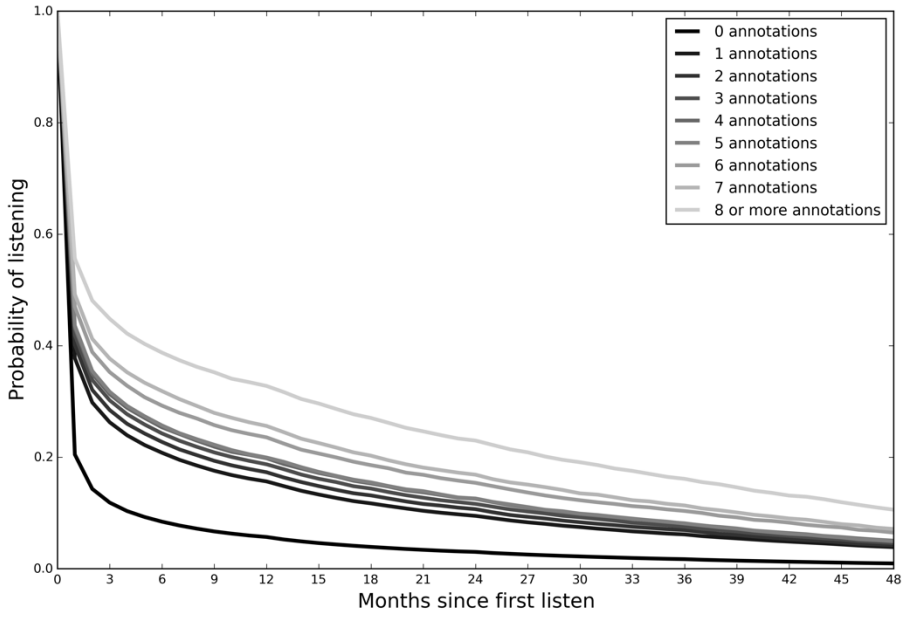


Figure 40: Mean normalized playcount for user-artist listening time series tagged a given number of times.

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[68] 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

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

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 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

[68] 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.

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 appear in Figure 41. 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.

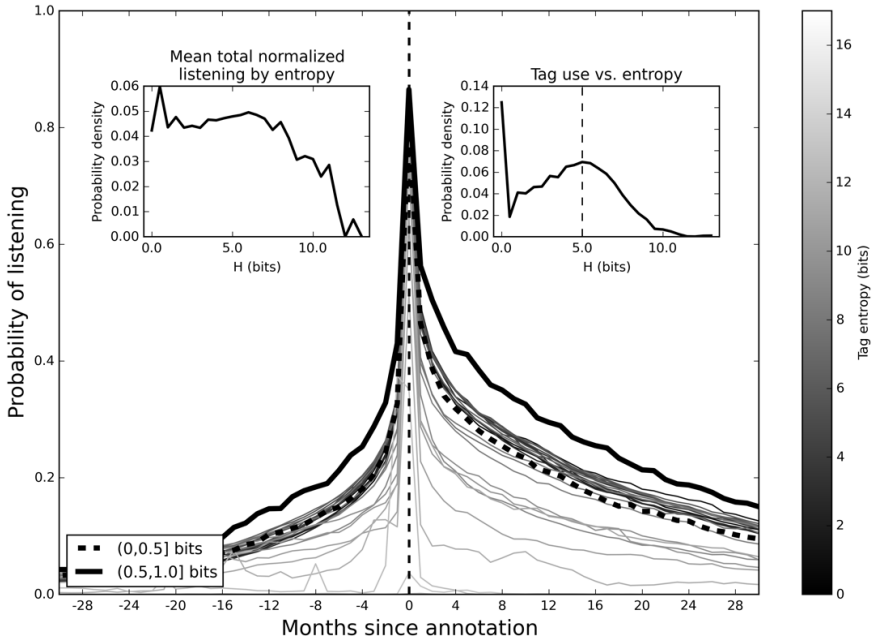


Figure 41: Mean probability of listening each month (relative to the month in which a tag is applied) for user-artist time series associated with tags of a given binned entropy (bin width of 0.5 bits). Each line represents the mean listening for a particular entropy bin, with line color indicating the 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).

This is just one of various difficulties in interpreting large scale data such as these. When considering the average behavior of many, heterogeneous 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.

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 may also expand our analysis to employ standard machine learning algorithms 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.

Study: Estimating the causal impact of tagging on future contact interaction

Having discussed several possible approaches to testing the memory cue hypothesis and some tentative findings, I now turn to a more developed study that attempts a causal analysis of the type described above.

While detailed information on how existing tags are utilized remains beyond our reach, an alternative approach is to examine how patterns of user interaction with tagged versus untagged content vary. If tags do serve as retrieval aids, we should expect users to be more likely to interact with a resource (e.g. visit bookmarked pages, listen to songs, view photos, etc.) once they have assigned a tag to it.

Here we test this hypothesis using a large-scale dataset consisting of complete listening and tagging histories from more than 100,000 users from the social music website Last.fm. From this dataset, we extract user-artist listening time-series, each of which represents the frequency of listening over 90 months to a particular artist by a particular user, and compare time-series in which the user has tagged the artist to those that are untagged. Specifically, we address the following two sub-questions:

- RQ1: Does tagging an artist lead to increased listening to that artist in the future, as shown by comparison of tagged versus untagged time-series?
- RQ2: Are certain tags particularly associated with increases in future listening, and if so, can we identify attributes of such "retrieval-targeted" tags as opposed to others?

As discussed in Chapter 3, data on how users actually use existing tags is simply not available to researchers through any tagging system APIs (or through other methods) that we are aware of. Thus the existing work on tagging motivation is limited to inferring why

people tag from how they tag, rather than from how they use their tags. In presenting our novel method, we are aware that it still represents an inferential approach. Our approach is distinct from those previously described, however, in that we test a concrete hypothesis about how tagging should affect a behavior on which we do have data: interaction with tagged content, in our case music listening.

Dataset

As in earlier presented studies, we utilize an earlier version of the data described in Chapter 1, and for our current purposes consider only those users for whom we have both tagging and listening histories. For each user, we extract one time-series for each unique artist listened to by that user, as described above.

The individual scrobbles in our dataset (a slightly larger one than that described earlier in this chapter) contain a total of approximately 95 million user-artist listening time-series. In about 6 million of these cases, the user has assigned at least one tag to the artist (or to a song or album by that artist) within the collection period (we refer to these as tagged time-series), while in the remaining cases (~89 million) the user has never tagged the artist. We summarize these high-level dataset statistics in Table 9. Comparison of these tagged and untagged listening time-series is the heart of the analyses presented in the next section.

Table 9: Dataset summary for causal tagging analysis study

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)

Analyses and results

RQ1 (comparison of tagged and untagged time series): Our principal research question is whether listening patterns for tagged content are consistent with the expectation that tags serve as memory cues. If so, we would expect to see an increase in a user's listening rate to musical artists after the user has tagged them, under the assumption that a tag facilitates retrieval and increases the chances of a user listening to a tagged artist.

Unfortunately, several factors combine to make such an analysis difficult. First and foremost, the desired counterfactual of the untagged "version" of a particular tagged series, which would allow a direct testing of how tagging changes listening behavior, does not, of course, exist. We thus must utilize untagged time-series in a way that allows them to approximate what a true counterfactual might look like. In searching for such samples, a second difficulty that arises is that listening rates for tagged time-series are much greater than for untagged time-series (the average number of total listens across time-series is 16.9 when untagged and 98.9 when tagged). While suggestive of the importance of tagging, this imbalance also suggests that controls must be incorporated in both sample selection and statistical analysis to account for previous listening behavior prior to tagging. Finally, the actual point in time at which tags are expected to increase listening behavior for any given user is unknown. Thus, we must formulate our analysis to account for this uncertainty.

To alleviate issues with the non-existence of a true counterfactual, we subselect from both the tagged and untagged series using the following formal procedure. We first select only those tagged time-series that have:

- more than 25 total listens;
- peak in listening at least 6 months from the edges of our data collection period (ensuring that the period from 6 months before to 6 months after the peak does not extend beyond the limits of our data range); and
- at least one listen in the 6 months prior to and after the peak (e.g. if the peak occurs in July, there should be at least one listen between January and June, and one between August and the following January).

We then select only those tagged time series where the tag was applied in the month of peak listening, and align all of those series at that peak point. Constraining our time-series in this manner, we are left with a total of 206,140 tagged time-series. Next we randomly select a same-sized sample of the 4.1M untagged time-series meeting the same criteria, and also align them at the peak of listening. Where the peak was reached in multiple months in any series, we chose one of these peaks at random to align on. All results below have been verified with multiple random samplings of the untagged data.[69]

After temporally aligning the tagged and untagged samples, we limited our analysis to a 13-month period extending from 6 months prior to the peak month to 6 months after the peak. This allows us to consider a manageable variety of ways in which listening prior to the tag may affect future behavior.

[69] Our method thus compares tagged and untagged data aggregated over many users. While it would be preferable to perform a within-subjects analysis (i.e. comparing tagged versus untagged data for each individual user), thereby accounting for much of the variability in listening across different users, the data for any particular user tends to be too sparse, as most individuals have tagged only a few times (if ever).

In Figure 42 we plot mean listens, with 95% normal confidence intervals, for each month across all tagged and untagged time-series in the subsampled data. All values are normalized by the peak number of listens for each series, and thus values at the peak month for both the tagged and untagged lines are unity. Comparing the line heights before and after the peak, Figure 42 shows that the mean normalized listening rate increases in the months after the peak for both tagged and untagged time-series. But there is also a small but reliable difference between the tagged and untagged series: The tagged time-series show proportionally higher mean normalized listening rates after the peak month (in which the tag was applied) as compared to untagged time-series. This is suggestive of an increase in listening as a result of tagging.[70]

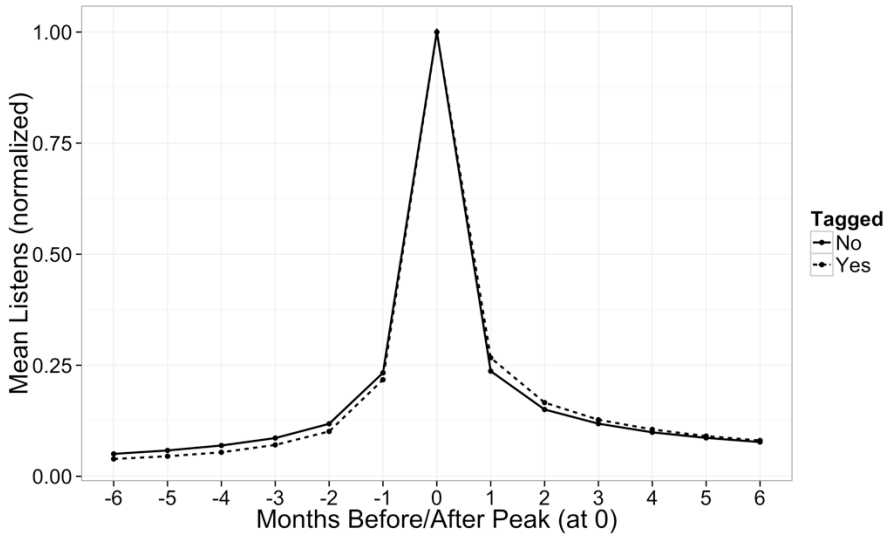


Figure 42: Comparison of tagged and untagged listening time-series. Mean normalized listens by month 95% CI included, but too small to be visible.

[70] There is also, however, a small but reliable lower rate of listening to tagged artists versus untagged artists prior to peak listening. This may indicate that songs that "catch on" for a user more quickly (rise faster in listening from before the peak to the peak) are more likely to be tagged, a possibility to be explored in future work.

While Figure 42 thus supports our hypothesis, there are two important caveats. First, as the distribution of the number of listens in any given month across all time series is heavily skewed, the mean is not fully representative of the data. Our further statistical analysis uses a log-transformed version of the listening counts to account for this. Second, the initial analysis does not control for the presumably important effect of pre-peak listening behavior on post-peak listening.

To test our hypothesis more robustly, we therefore use a regression model that incorporates previous listening behavior to predict post-peak behavior. Due to a lack of knowledge about the relationship between these variables and the volume of data we have, it was both unreasonable and unnecessary to assume a linear relationship between the dependent and independent variables. Because of this, we opted for a Generalized Additive Model (GAM; Hastie & Tibshirani, 1990) using the R package mgcv (Wood, 2001) applied to all of the tagged and untagged series in the selected sets described above. 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.) 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 peak month and 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):

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

The regression model, which explained approximately 30% of the variance in the data (adjusted R^2), indicated that the tagged/untagged indicator and the listening rate parameters (smoothed using thin-plate regression splines) for all seven previous months had a significant effect on post-peak listening behavior ($P \ll 0.0001$). As we cannot show the form of this effect for all model variables at once, Figure 43 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. The large confidence interval on the right hand side of Figure 43 reflects the small number of users who have extremely high listening rates for particular artists.

RQ2 –Tag analysis: To examine if and how different tags are associated with increased future listening, we ran a regression analysis similar to that described above, but with two important changes. First, instead of a single tagged/untagged indicator, we included binary (present/not present) regressors for the 2,290 unique tags that had at least five occurrences in our subsample.[71] Second, due to the data-hungry nature of the GAM and

[71] We chose a threshold of five to ensure that data was not too sparse for the regression model but was still inclusive of infrequently occurring tags. Again, qualitative results hold when using both more and less restrictive thresholds.

the large number of additional variables introduced by utilizing all tags as unique predictors, we chose to only control for listening in the peak month, and not the six prior months. This decision limited the computation associated with estimating a model of this size and did not appear to affect model fit substantially according to tests we ran on subsamples of the data. The same data were used as in the previous analysis (untagged time series, of course, had values of zero for all possible tags). Formally, the regression model can be represented as follows, where again m is peak month, L is the number of listens in a given month, T_i is the binary indicator for a given tag, and f is the exponential-family function calculated by the GAM:

$$\log \sum_{i=1}^{6} L_{m+i} = b_0 + f(\log L_m) + \sum_{i=1}^{2290} b_i T_i$$

After running the model, which explains approximately 28.5% of the variance in the data (adjusted R^2), 161 unique tags were statistically significant predictors at $\alpha = .001$, a threshold selected in order to account for the large number of comparisons against the null hypothesis being made in the regression model. We proceeded to examine which of these tags were relatively strong predictors in the model.

Unsurprisingly, most of the 161 tags tend to have a positive (albeit small) impact on future listening, as evidenced by positive regression coefficients and consistent with the small positive effect of tagging overall as found in the previous analysis. The most telling observation is that commonly-used genre tags (e.g. “pop”, “jazz”, and “hip-hop”) tend to be weak positive predictors of future listening. In contrast, relatively strong

predictors (both positive and negative) appear to be comparatively obscure, possibly idiosyncratic tags (e.g. “cd collection”, “mymusic”, “purchased 09”).[72]

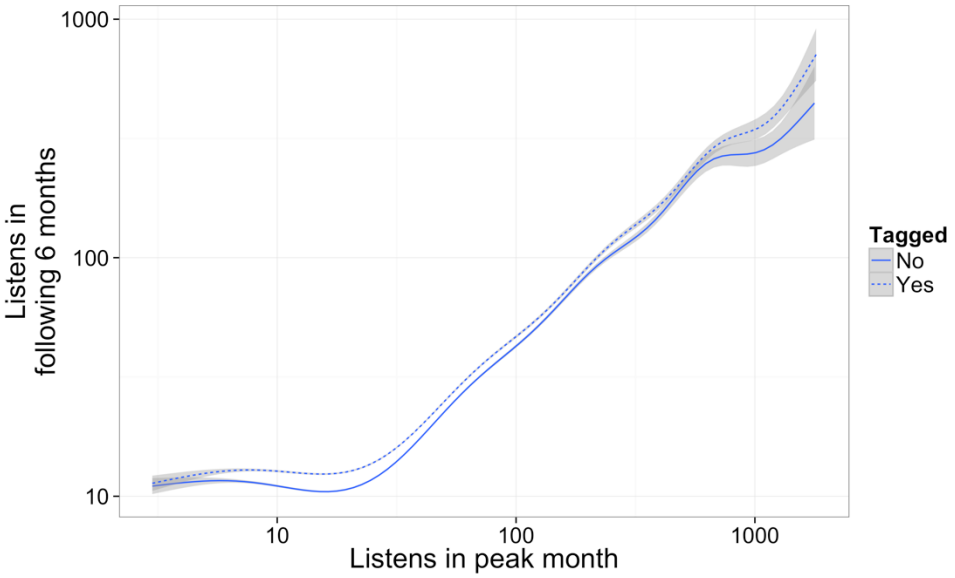


Figure 43: Regression model results, showing predicted sum total of listening in the 6 months after a tag is assigned as a function of the number of listens in the month of peak listening in a time series. Results shown on a log-log scale, and shaded regions indicated a bootstrapped 95% confidence interval.

To examine this trend quantitatively, we plot in Figure 44 the global tag usage (i.e. the total number of uses of a tag in our full dataset of approximately 50 million annotations) as a function of the tag’s impact on listening indicated by its coefficient in the regression model. Similarly, we plot in Figure 45 the unique number of users utilizing the tag, again as a function of its regression coefficient. The value e^c, where c is a tag’s regression coefficient, represents the number of listens we expect a user’s post-listening behavior to increase or decrease by if she were to apply a (thus the strongest predictors

[72] For a full listing of the regression coefficients across all tags in the model, see https://dl.dropboxusercontent.com/ u/625604/papers/lorince.joseph.todd.2015.sbp.supplemental/regression coefficients.txt

lead to an increase of fewer than 7 listens on average). Finally, in both plots, the horizontal red bands mark the upper and lower limits of a bootstrapped 95% confidence interval on the popularity of the 2,129 remaining tags that were not significant in the regression model.

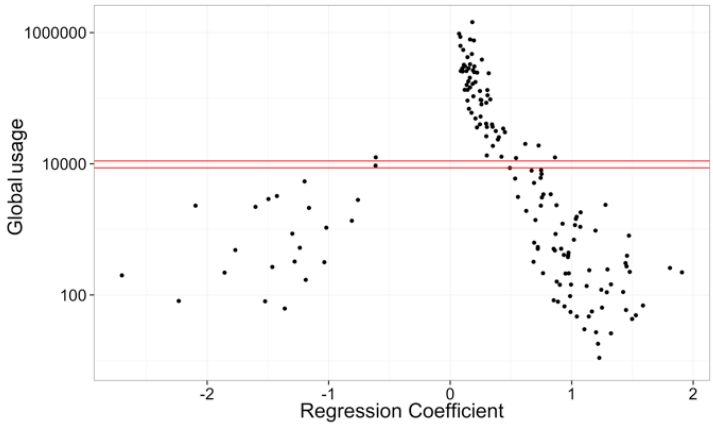


Figure 44: Each tag's global usage as a function of its regression coefficient.

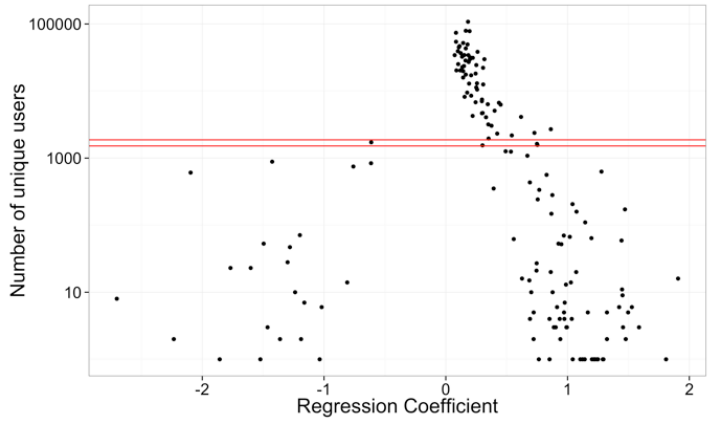


Figure 45: Number of unique users of each tag as a function of regression coefficient

The data suggest that the most popular tags on both metrics (i.e., those above the red lines) are significant, weakly positive predictors of future listening, while relatively unpopular tags (i.e., below the red lines) tend to have relatively strong positive (and, in some cases, strong negative) impacts on listening. Tags which were not significant in the

model (i.e., those that would fall between the red lines in each plot) had moderate popularity levels with respect to both metrics. The high statistical reliability but small regression coefficients for the most popular tags may be somewhat artefactual, primarily reflecting high variability in how predictive these tags are of listening across individuals. We believe this finding is still informative, however, as it indicates that popular tags are not consistently associated with future listening.

Conclusions

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

- 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;
- a set of candidate analytic methods for exploring those hypotheses; and
- a study applying one such method to test for a causal impact of tagging on future music listening

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 exploratory 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. However, a more detailed analysis paints a less compelling picture. Results suggest that tagging an artist does lead to an increase in listening, but that this increase is, on average, quite small (amounting to only 1 or 2 additional listens over a six month period). Given the various possible motivations for tagging, we expected only some tags to serve as retrieval cues, and thus tested the relative predictiveness of future listening for different tags. This analysis revealed systematic differences in how predictive the presence or absence of different tags was for future listening as a function of tag popularity. The data suggest that, at least for the small number of most highly significant tags that we consider, those that are globally popular have relatively little effect on future listening, and are generally associated with very small increases in post-tagging listening rates. The tags that seem to "matter" (i.e. those that are relatively strong predictors of whether or not a user will listen to an artist after tagging it) are generally much less popular. Even these stronger predictors, however, lead to relatively slight increases in listening. The strongest predictors are associated with a change of only about 7 listens over a six month period, on average.

Because we only analyzed a small sample of statistically influential tags, we are at this point tentative to make strong claims about which specific factors contribute to tags being better or worse predictors of increased listening, or even of decreased listening. The evidence nevertheless suggests that relatively uncommon (and in many cases idiosyncratic) tags are most predictive of future listening behavior. The intriguing flipside is that the descriptive, popular tags that are arguably most useful to the community at

large (i.e. genre labels and related tags) are not particularly associated with increases in listening for those who applied the tags, and thus are likely not functioning as memory cues.

Overall it appears that, while on average tagging an artist has a small positive effect on one's own future listening, the most common tagging activities are not strong predictors of future retrieval. We cannot be sure of the extent to which the many other possible tagging motivations are at play here, nor can we tell at this point if and when a tag is applied with the intention of being used for retrieval, while ultimately not being used for this purpose. That said, our results may indicate that the primary motivation for tagging on Last.fm is not for personal information management (tagging a resource for one's own retrieval), but rather may be socially or otherwise oriented, which may in turn result in tags that are useful for the community at large. This leads to the interesting possibility that a folksonomy can generate the useful, crowdsourced classification of content that proponents of collaborative tagging extol, even if this process is not strongly driven by the self-directed, retrieval-oriented tagging that is typically assumed in such systems.

While our results provide clues as to whether tags really function as retrieval aids, this remains early work addressing a hitherto unstudied research question. There is certainly room to refine and build upon the methods we present here for testing if and when tagging increases listening rates. In particular, our analysis at the level of artists (rather than the individual resources tagged) may be problematic, and we hope to develop models that operate directly at the level of the content tagged. There are also many factors we have not controlled for here that could be incorporated into future models,

such as exogenous influences on listening (e.g. when an artist goes on tour or releases a new album), and we should explore alternative methods for normalizing and controlling for user listening habits beyond our approach here of simply considering raw monthly listens.

It will be critical to expand on methods for understanding which tags serve as memory cues and under what circumstances. It is clearly the case that not all tags function as memory cues, so more robustly identifying which tags do serve as retrieval aids is a fruitful direction for future work. Incorporating research on human memory from the cognitive sciences can also further inform hypotheses and analytic approaches to these questions, something we are actively pursuing in ongoing research. A final limitation is that we are exploring tagging in a particular collaborative tagging system, which operates in the possibly idiosyncratic domain of music. Tagging habits may vary systematically in different content domains, but until usable data becomes available, we can only speculate as to exactly how.

One 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.[73]

---

[73] 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.

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.

In closing, to directly address the question of whether or not tags function as retrieval aids, the best answer for Last.fm at least would appear to be "sometimes, but usually not". While there is much work to be done on when and why particular tags serve this function and others do not, it is clear that the overarching retrieval assumption is far from universally valid: Tags certainly do not always function as memory cues, and our results suggest that facilitating later retrieval may actually be an uncommon tagging motivation.

Chapter 5: Conclusions

Summary

This thesis has presented an ecologically inspired perspective on two inexorably linked behaviors in digital environments, content consumption (music listening in particular) and content organization (specifically tagging).

Chapter 1 provided relevant background and motivation for applying an ecological perspective to content consumption on the web, and introduced our dataset of tagging and listening activity from Last.fm.

Chapter 2 focused on music consumption, first developing a framework for quantifying the distances between different artists in a latent feature space and methods for evaluating it. We then examined that space to ensure it was in fact patchy (a prerequisite for patch-based foraging perspective), and developed methods for characterizing patches. Finally, we presented a set of quantitative analyses, finding evidence of area-restricted search at multiple time scales in listeners' music choices, as well interesting commonalities between exploration and exploitation in music listening that set this domain apart from typical foraging in physical environments.

Chapter 3 focused on tagging behavior, opening with a review of research on tagging (with a focus on motivational factors) and social imitation behavior in web contexts. We then presented two research projects, on using multi-agent modeling to explore the role of simple imitation heuristics in collaborative tagging systems. Results demonstrated that large-scale properties of these systems may be driven by simple, ecologically plausible heuristic processes. The second project examined "supertaggers", the minority of users in tagging systems that generate an outsize share of the tagging

activity. After a set of descriptive analyses comparing tagging patterns of supertaggers to other users, we provided evidence that such users may differ from other users in terms of expertise and motivational factors.

Finally, Chapter 4 presented a case study of how organization and consumption can interact, concretely by testing the hypothesis that users predominantly tag content for the purposes of future retrieval (i.e. re-finding a tagged item at a later time). By examining how tagging patterns and listening patterns interact, we were able to empirically test this hypothesis. Results suggest that future retrieval may in fact be a rare motivation for tagging.

Limitations of the dataset

Before concluding, some general comments on our dataset are in order. While the data is at a massive scale (it is, to our knowledge, the largest such dataset available for academic research), it is not without limitations.

First, while we hope the data is a representative sample of last.fm users it is difficult to confirm that our traversal of the social network effectively sampled across the spectrum of Last.fm users. But perhaps more important is the fact that Last.fm likely represents a biased selection of music listeners in general. Last.fm users are typically young, American or British, tech-savvy web users that may or may not be representative of music listeners at large. Qualitatively, there are some notable differences between the most popular artists in our sample and artists typically found in, say, the Billboard rankings. Furthermore, some genres (e.g. country music) are underrepresented in our data (and on Last.fm in general). These issues do not invalidate our results by any means, but

it is important use caution when extending conclusions about music organization or listening to listeners at large.

A second issue is that some further data cleaning, ideally, should be applied to the data. The long tail of rarely occurring artists includes many misspelled, mislabeled, or otherwise noisy data, and while we have largely dealt with this by limiting the set of artists considered (as described in Chapter 1), some improved handling of such cases is desirable.

Third, the choice to perform all analyses at the level of artists, as opposed to individual songs, has clear implications. Though it made many analyses tractable that would otherwise be challenging (because of the sheer number of unique songs, as well as the sparsity of data for many of them), it also involves some strong assumptions about the homogeneity of an artist's work. Plenty of artists change stylistically over their career, and a listener's preference for an artist might be limited to a subset of that artist's songs or albums.

Finally, we must question the extent to which our sample of data for any given user accurately captures her patterns of music consumption. It is impossible to know how much music a user consumes via outlets not tracked in her Last.fm profile, and for some users it is entirely possible that the data we have is in fact a biased sample of their overall listening behavior. As with the previously mentioned limitations, this is not a critical flaw with our analyses (and is, after all, not resolvable), but is important to keep in mind when considering the generalizability of our results.

Closing remarks

This thesis represents an attempt to apply an ecological lens to the study of web users' consumption and organization of content. We applied a foraging perspective to music consumption, examined the role of ecologically and psychologically grounded mechanism to tagging decisions and motivation, studied the interplay between tagging and music listening.

The individual chapters highlight specific future directions for the different research projects described, but I wish to close by considering more generally the generalizability of this work. While the scale of our data is quite large, the scope is of course limited to one domain (music) and further to one web-based music service (Last.fm). How might our findings extend to other contexts?

With respect to tagging, we were at least able to consider several other tagging systems for a subset of our supertagger analyses, but there is reason to believe that tagging dynamics are strongly influenced by the particular system in which users are tagging. The analyses in Chapter 4, on the other hand, were necessarily limited to the case of Last.fm, and it remains an open question whether we would find stronger evidence of tagging serving the purposes of future retrieval on other systems or in other content domains. Speaking generally, it seems that unified accounts of tagging behavior will be challenging to come by in many cases.

As for content consumption, an intriguing question is how well the foraging-inspired perspective introduced here fits domains other than music consumption. Information search tasks such as those presented in the core work in information foraging theory are fundamentally different than music consumption. The former tend to be

discrete, isolated informational goals, whereas music consumption is an inherently hedonic, repeated search task without any clear parameters for "completion'. Thus, a reasonable hypothesis for future work is that the findings presented here will be applicable to other similar domains, such television or movie watching or particular subsets of web browsing behavior that lack a concrete search goal (e.g. browsing newsfeeds or similar tasks). All that said, music consumption represents an important activity for many individuals, one on which they are willing to spend time, energy, and money. As such, even if these results are unique to the music domain, they are of practical value.

In closing, this thesis has presented a large-scale dataset of music listening and tagging behavior (one that we plan to make publically available in the near future), along with in depth examination of how users interact with content in this domain. Aside from the specific contributions to our understanding of music consumption and tagging, my hope is that the lasting contribution of this work is to serve as a case study in how behavioral patterns with respect to these processes can be studied with large-scale analytic tools while maintaining a grounding in psychologically and ecologically grounded mechanisms.

References

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, *6*(2).

Al-Khalifa, H. S., & Davis, H. C. (2007). Towards better understanding of folksonomic patterns. In *Proceedings of the eighteenth conference on Hypertext and hypermedia* (pp. 163–166). ACM.

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.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, N.J: Psychology Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060.

Barnard, C. J., & Sibly, R. M. (1981). Producers and scroungers: a general model and its application to captive flocks of house sparrows. *Animal Behaviour*, *29*(2), 543–550.

Baronchelli, A., & Radicchi, F. (2013). Lévy flights in human behavior and cognition. *Chaos, Solitons & Fractals*, *56*, 101–105. http://doi.org/10.1016/j.chaos.2013.07.013

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., & Kraut, R. E. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, *10*(4).

Bell, W. J. (1991). *Searching behaviour: the behavioural ecology of finding resources*. New York: Chapman & Hall.

Benhamou, S. (1992). Efficiency of area-concentrated searching behaviour in a continuous patchy environment. *Journal of Theoretical Biology*, *159*(1), 67–81.

Berenzweig, A., Logan, B., Ellis, D. P., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, *28*(2), 63–76.

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE*, *9*(4), e95693.

Berg, N., & Gigerenzer, G. (2010). As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas*, *18*(1), 133–165.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives*, *12*(3), 151–170.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: classification, clustering, and applications*. Chapman & Hall.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

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.

Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

Carneiro, M. J. T. (2012). *Towards the discovery of temporal patterns in music listening using Last. fm profiles*. Retrieved from http://repositorio-aberto.up.pt/handle/10216/61584

Cattuto, C., Baldassarri, A., Servedio, V. D., & Loreto, V. (2007). Vocabulary growth in collaborative tagging systems. *arXiv Preprint*. Retrieved from http://arxiv.org/abs/0704.3316

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

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Christian, B., & Griffiths, T. (2016). *Algorithms to Live By: The Computer Science of Human Decisions*. New York: Henry Holt and Co.

Clark, C. W., & Mangel, M. (1986). The evolutionary advantages of group foraging. *Theoretical Population Biology*, *30*(1), 45–75.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 318–329). ACM.

Davis, I. (2005). Why Tagging is Expensive. Retrieved from http://blogs.capita-libraries.co.uk/panlibus/2005/09/07/why_tagging_is_/

Davison, B. D. (2000). Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 272–279). ACM.

Dias, R., & Fonseca, M. J. (2013). Improving music recommendation in session-based collaborative filtering by using temporal context. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on* (pp. 783–788). IEEE.

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.

Eiron, N., & McCurley, K. S. (2003). Locality, hierarchy, and bidirectionality in the web. In *Workshop on Algorithms and Models for the Web Graph*. Budapest, Hungary.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th Edition). Chichester, West Sussex, U.K: Wiley.

Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., & Giles, L. (2007). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 351–360). ACM.

Figueiredo, F., Pinto, H., Belém, F., Almeida, J., Gonçalves, M., Fernandes, D., & Moura, E. (2013). Assessing the quality of textual features in social media. *Information Processing & Management*, *49*(1), 222–247.

Firan, C. S., Nejdl, W., & Paiu, R. (2007). The benefit of using tag-based profiles. In *Proceedings of the 2007 Latin American Web Conference* (pp. 32–41). IEEE.

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 T. J. Bastiaens, U. Baumöl, & B. J. Krämer (Eds.), *On collective intelligence* (pp. 75–91). Springer.

Fu, W.-T. (2012). From Plato to the World Wide Web: Information Foraging on the Internet. In P. M. Todd, T. T. Hills, & T. W. Robbins (Eds.), *Cognitive search: Evolution, algorithms, and the brain.* Cambridge, MA: MIT Press.

Fu, W.-T., & Dong, W. (2010). Facilitating knowledge exploration in folksonomies: expertise ranking by link and semantic structures. In *Proceedings of the IEEE Second International Conference on Social Computing* (pp. 459–464). IEEE.

Fu, W.-T., Kannampallil, T. G., & Kang, R. (2009). A semantic imitation model of social tag choices. In *International Conference on Computational Science and Engineering* (Vol. 4, pp. 66–73). IEEE.

Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human–Computer Interaction*, *22*(4), 355–412.

Garcia-Retamero, R., & Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin & Review*, *16*(1), 163–169.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482.

Gigerenzer, G., & Todd, P. M. (2000). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Giraldeau, L. A., & Beauchamp, G. (1999). Food exploitation: searching for the optimal joining policy. *Trends in Ecology & Evolution*, *14*(3), 102–106.

Giraldeau, L. A., Valone, T. J., & Templeton, J. J. (2002). Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*(1427), 1559–1566.

Glushko, R. J. (Ed.). (2013). *The Discipline of Organizing*. Cambridge, MA: The MIT Press.

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.

Goldstone, R. L., Jones, A., & Roberts, M. E. (2006). Group path formation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(3), 611–620.

Goodwin, J. C., Cohen, T., & Rindflesch, T. (2012). Discovery by scent: Discovery browsing system based on the Information Foraging Theory. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on* (pp. 232–239). IEEE.

Görlitz, O., Sizov, S., & Staab, S. (2008). PINTS: Peer-to-Peer Infrastructure for Tagging Systems. In *Proceedings of the Seventh International Workshop on Peer-to-Peer Systems (IPTPS)* (p. 19). Tampa Bay, USA.

Gould, S. J., & Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society B: Biological Sciences*, *205*(1161), 581–598.

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., Heilemann, M., & Wolff, C. (2009). Personal Information Management vs. Resource Sharing: Towards a Model of Information Behavior in Social Tagging Systems. In *ICWSM*.

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

Helbing, D., Schweitzer, F., Keltsch, J., & Molnár, P. (1997). Active walker model for the formation of human and animal trail systems. *Physical Review E*, *56*(3), 2527.

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

Hills, T. T. (2006). Animal Foraging and the Evolution of Goal-Directed Cognition. *Cognitive Science*, *30*(1), 3–41.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, *19*(8), 802–808.

Holling, C. S. (1959). Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*, *91*(07), 385–398.

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

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

Howell, D. J. (1979). Flock foraging in nectar-feeding bats: advantages to the bats and to the host plants. *American Naturalist*, 23–49.

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

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.

Hurley, S., & Chater, N. (Eds.). (2005a). *Perspectives on Imitation: From Neuroscience to Social Science - Volume 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book.

Hurley, S., & Chater, N. (Eds.). (2005b). *Perspectives on Imitation: From Neuroscience to Social Science - Volume 2: Imitation, Human Development, and Culture*. Cambridge, Mass: A Bradford Book.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 263–272). Ieee.

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.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

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.

Johnson, N. L., & Kotz, S. (1977). *Urn models and their application: An approach to modern discrete probability theory*. John Wiley.

Jones, W. (2007). Personal information management. *Annual Review of Information Science and Technology*, *41*(1), 453–504.

Katsikopoulos, K. V., & King, A. J. (2010). Swarm Intelligence in Animal Groups: When Can a Collective Out-Perform an Expert? *PLoS ONE*, *5*(11), e15505.

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

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

Koh, J., Kim, Y.-G., Butler, B., & Bock, G.-W. (2007). Encouraging participation in virtual communities. *Communications of the ACM*, *50*(2), 68–73.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8), 30–37.

Körner, C. (2009). *Understanding the Motivation behind Tagging*. Presented at the Hyptertext 2009.

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.

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, *25*(1), 28–34.

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, *59*(2), 105–117.

Kubek, M., Nützel, J., & Zimmermann, F. (2010). Automatic Taxonomy Extraction through Mining Social Networks. In *Proceedings of the 8th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorporating the 6th International ODRL Workshop*. Namur, Belgium.

Kumar, R., & Tomkins, A. (2010). A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web* (pp. 561–570). ACM.

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, *32*(1), 4–14.

Lambiotte, R., & Ausloos, M. (2005). Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E*, *72*(6), 066107.

Liebman, E., Saar-Tsechansky, M., & Stone, P. (2015). Dj-mc: A reinforcement-learning agent for music playlist recommendation. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (pp. 591–599). International Foundation for Autonomous Agents and Multiagent Systems.

Liu, H., Mulholland, P., Song, D., Uren, V., & Rüger, S. (2010). Applying information foraging theory to understand user interaction with content-based image retrieval. In *Proceedings of the third symposium on Information interaction in context* (pp. 135–144). ACM Press.

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.

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., & Todd, P. M. (2016). Music Tagging and Listening: Testing the Memory Cue Hypothesis in a Collaborative Tagging System. In M. N. Jones (Ed.), *Big Data in Cognitive Science: From Methods to Insights*. New York, NY: Taylor & Francis.

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. *Journal of Web Science*, *1*(1), 16–32.

Luce, R. D. (1959). *Individual Choice Behavior a Theoretical Analysis*. John Wiley and sons.

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

Mahajan, A., & Teneketzis, D. (2008). Multi-armed bandit problems. In *Foundations and Applications of Sensor Management* (pp. 121–151). Springer.

Mandel, M. I., Pascanu, R., Eck, D., Bengio, Y., Aiello, L. M., Schifanella, R., & Menczer, F. (2011). Contextual tag inference. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *7*(1), 32.

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.

Mayr, E. (1983). How to carry out the adaptationist program? *American Naturalist*, 324–334.

McCart, J. A., Padmanabhan, B., & Berndt, D. J. (2013). Goal attainment on long tail web sites: An information foraging approach. *Decision Support Systems*, *55*(1), 235–246.

McFee, B., & Lanckriet, G. R. (2011). The Natural Language of Playlists. In *ISMIR* (pp. 537–542).

McKay, C., & Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? In *ISMIR* (pp. 101–106).

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, *21*(2), 269–288.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures.

Meiss, M. R., Duncan, J., Gonçalves, B., Ramasco, J. J., & Menczer, F. (2009). What's in a session: tracking individual behavior on the web. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia* (pp. 173–182). ACM.

Meiss, M. R., Menczer, F., Fortunato, S., Flammini, A., & Vespignani, A. (2008). Ranking web sites with real user traffic. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 65–76). ACM.

Menczer, F. (2002). Growing and navigating the small world Web by local content. *Proceedings of the National Academy of Sciences*, *99*(22), 14014–14019. http://doi.org/10.1073/pnas.212348399

Menczer, F. (2004). Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*, *55*(14), 1261–1269.

Menczer, F., Fortunato, S., Flammini, A., & Vespignani, A. (2006, February 1). Googlearchy or Googlocracy? *IEEE Spectrum*.

Meo, P. de, Ferrara, E., Abel, F., Aroyo, L., & Houben, G.-J. (2013). Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 14.

Merholtz, P. (2005). Clay Shirky's Viewpoints are Overrated. Retrieved April 21, 2016, from http://peterme.com/archives/000558.html

Mitzlaff, F., Benz, D., Stumme, G., & Hotho, A. (2010). Visit me, click me, be my friend: an analysis of evidence networks of user relationships in BibSonomy. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (pp. 265–270). 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.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Noll, M. G., Au Yeung, C., 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.

Nov, O., & Ye, C. (2010). Why do people tag?: motivations for photo tagging. *Communications of the ACM*, *53*(7), 128–131.

Pachet, F., Cazaly, D., & others. (2000). A taxonomy of musical genres. In *RIAO* (pp. 1238–1245).

Panagakis, I., Benetos, E., & Kotropoulos, C. (2008). Music genre classification: A multilinear approach. In *ISMIR* (pp. 583–588). Retrieved from http://openaccess.city.ac.uk/2109/

Pierce, G. J., & Ollason, J. G. (1987). Eight Reasons Why Optimal Foraging Theory Is a Complete Waste of Time. *Oikos*, *49*(1), 111.

Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 3–10). ACM.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, *29*(3), 343–373.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford University Press.

Pirolli, P. (2009). An elementary social information foraging model. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 605–614). ACM.

Pirolli, P., & Card, S. (1995). Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 51–58). ACM Press/Addison-Wesley Publishing Co.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643.

Pirolli, P., Fu, W.-T., Reeder, R., & Card, S. K. (2002). A user-tracing architecture for modeling interaction with the World Wide Web. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 75–83). ACM.

Pirolli, P., Pitkow, J., & Rao, R. (1996). Silk from a sow's ear: extracting usable structures from the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 118–125). ACM.

Qiu, F., Liu, Z., & Cho, J. (2005). Analysis of User Web Traffic with A Focus on Search Activities. In *WebDB* (pp. 103–108). Citeseer.

Ragno, R., Burges, C. J. C., & Herley, C. (2005). Inferring similarity between music objects with application to playlist generation (p. 73). ACM Press.

Riefer, P. S., & Love, B. C. (2015, April). *Modern Foraging: Exploration and Exploitation in Supermarkets.* University College, London.

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience*, *27*(1), 169–192.

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), 14.

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*(3), 451–469.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

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.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856.

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.

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

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425–440.

Sinha, R. (2005). A cognitive analysis of tagging. Retrieved April 21, 2016, from https://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/

Smith, E. A., & Winterhalder, B. (Eds.). (1992). *Evolutionary Ecology and Human Behavior*. New York: Aldine Transaction.

Stamps, J. A. (1988). Conspecific attraction and aggregation in territorial species. *American Naturalist*, 329–347.

Stearns, S. C., & Schmid-Hempel, P. (1987). Evolutionary Insights Should Not Be Wasted. *Oikos*, *49*(1), 118. http://doi.org/10.2307/3565561

Stephens, D. W. (1990). Foraging theory: up, down, and sideways. *Studies in Avian Biology*, *13*, 444–454.

Stephens, D. W., & Krebs, J. R. (1987). *Foraging Theory* (1 edition). Princeton, N.J: Princeton University Press.

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

Stoilova, L., Holloway, T., Markines, B., Maguitman, A. G., & Menczer, F. (2005). GiveALink: mining a semantic network of bookmarks for web search and recommendation. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 66–73). ACM.

Strohmaier, M., Körner, C., & Kern, R. (2010). Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In *ICWSM*.

Sturm, B. L. (2014a). A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation* (pp. 29–66). Springer.

Sturm, B. L. (2014b). The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research*, *43*(2), 147–172.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (1St Edition edition). Cambridge, Mass: A Bradford Book.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482.

Teevan, J., Jones, W., & Bederson, B. B. (2006). Personal information management. *Communications of the ACM*, *49*(1), 40–43.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological Rationality: Intelligence in the World* (1 edition). Oxford ; New York: Oxford University Press.

Todd, P. M., Hills, T. T., & Robbins, T. W. (Eds.). (2012). *Cognitive Search: Evolution, Algorithms, and the Brain*. Cambridge, MA: The MIT Press.

Tullis, J. G., & Benjamin, A. S. (2014). Cueing others' memories. *Memory & Cognition*, *43*(4), 634–646.

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.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vander Wal, T. (2005). Explaining and Showing Broad and Narrow Folksonomies. Retrieved July 29, 2014, from http://www.vanderwal.net/random/entrysel.php?blog=1635

Vander Wal, T. (2007). Folksonomy Coinage and Definition. Retrieved July 29, 2014, from http://vanderwal.net/folksonomy.html

Veres, C. (2006). The language of folksonomies: What tags reveal about user classification. In *Natural language processing and information systems* (pp. 58–69). Springer.

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 edition). New York: 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.

Wetzker, R., Zimmermann, C., & Bauckhage, C. (2008). Analyzing social bookmarking systems: A del. icio. us cookbook. In *Proceedings of the ECAI 2008 Mining Social Data Workshop* (pp. 26–30).

Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, *1*(2), 20–25.

Wright, A. (2004). Folksonomy. Retrieved April 21, 2016, from http://alexwright.org/blog/archives/000900.html

Yeung, C. A., Man, C., Noll, M., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). On measuring expertise in collaborative tagging systems. In *Proceedings of the 1st Annual ACM Web Science Conference*. Athens, Greece: ACM.

Yeung, C. 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.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *213*, 21–87.

Zheleva, E., Guiver, J., Mendes Rodrigues, E., & Milić-Frayling, N. (2010). Statistical models of music-listening sessions in social media. In *Proceedings of the 19th international conference on World wide web* (pp. 1019–1028). ACM.

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.

Appendices

Details of crawling process

Data collection proceeded in three main phases: In the first, we built our sample of Last.fm users by traversing the site's social network, and collected tagging data for a set of approximately 2 million users. In the second, we collected assorted supplementary data for each user. Finally, we collected complete listening histories for a subset of these users.

Tagging data

Previous work using Last.fm tagging data (Firan, Nejdl, & Paiu, 2007; Schifanella et al., 2010) has crawled content primarily using the site's API, but this proved insufficient for obtaining detailed tagging histories. Though the API does provide accurate summary information of a user's tagging habits (i.e. which tags they have used, and which items they have annotated with each tag), its utility is limited because it provides no temporal tagging information (that is, when a user tagged a particular item with a particular tag). User profile pages, however, contain a timestamped history of a user's tagging activity, with a temporal resolution of one month for all annotations made more than one month prior to when the page is loaded (e.g. user X tagged the artist "Radiohead" with the term "alternative" in March 2011). Though a finer-grained resolution would have been ideal, this limitation could not be avoided.

To crawl the data, we developed a hybrid crawler combining API methods and direct parsing of the HTML content of user profile pages. All data was stored in a MySQL database. We began with a set of arbitrarily selected seed users, whose usernames were used to initialize a crawl queue. The program then proceeded by

1. selecting the next username from the crawl queue;
2. querying the Last.fm API for the user's numeric ID and list of friends on the network, adding each to the crawl queue and recording the friendship relations;
3. extracting a user's list of unique tags utilized from his/her profile; and finally,
4. for each tag, extracting annotation information for each instance of that tag.

Each annotation was stored as a four-element tuple consisting of a user ID, an item URL,[74] the tag assigned, and the month and year in which the item was tagged. The process then repeated, drawing usernames from the crawl queue until a sufficiently large sample of content had been crawled. The number of usernames increased at a faster rate than we could crawl the content of users' profiles, so we did not run into the problem of exhausting the queue. The crawling process is schematized in Figure 46.

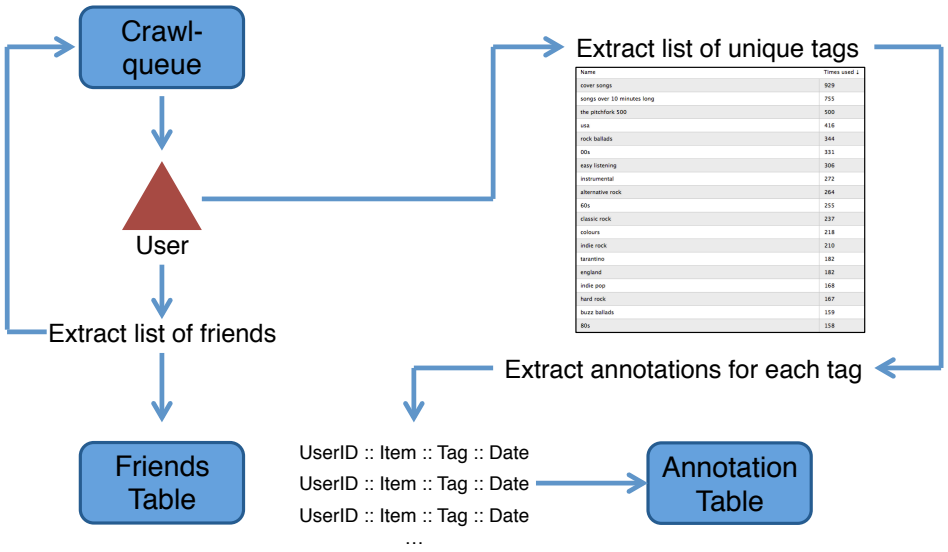


Figure 46: Schematic of the crawling process.

[74] i.e. the relative URL on Last.fm for the artist, album, or song tagged, e.g. "/music/Mogwai/_/Auto+Rock"

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

A more notable problem with this decision is that it necessarily limits our data to Last.fm users who have at least one friend on the network. Previous work, however, indicates that tagging behavior is correlated with having more friends on the social network (Schifanella et al., 2010), and this was confirmed within our own data. To further verify this, we crawled a small truly random sample of users by generating random numbers and crawling the annotation data for users whose numeric IDs matched those values. As we would expect, this resulted in far less annotation data (approximately 479,000 total annotations from 200,000 users, an average of 2.4 annotations per user) than we collected using the friend-based crawler (approximately 33 annotations per user). Given our goal of analyzing tagging behavior, we thus concluded that selectively

collecting data from users with friends was more likely to give us greater amounts of usable tagging data.

Supplemental data

As tagging data was collected, we simultaneously stored profile data (age, gender, registration date, etc.) for each user, as well as the full graph of friendship relationships. We also maintained an item data table, with data on each song, artist, and album encountered. Once tagging data for a given user was complete, we additionally collected his or her complete list of loved and banned tracks, as well as any group memberships.

Listening data

Because Last.fm stores complete listening histories, the amount of listening data per user is quite large (the median number of listens per user in our sample is ~16,000). The time and storage requirements for collecting listening data for all the users encountered in the first phase of crawling were prohibitive, so we were forced to use a sampling procedure.

As one of our interests is how tagging and listening behavior interact (see Chapter 4), and because most users tag very little or not at all, we could not simply sample users randomly. Thus we utilized a supervised sampling procedure that disproportionately favored users with more tagging activity. Users were first binned by their number of annotations (0-10, 10-100, 100-1000, or 1000+), and the crawler repeatedly iterated over the bins, selecting one user at random from each and collecting his or her data on every loop. All listening data was collected by direct querying of the Last.fm API.

Generation of artificial listening sequences

When generating artificial listening sequences for use as a null model, the goal was to maintain high-level statistical properties of the global data while randomizing local

"decisions" about artists listened. Thus we first utilized global (i.e. across all users) distributions of (a) artist space distance between subsequent listens and (b) temporal gaps between listens. Each artificial sequence was then generated with the following procedure:

1. Determine the first artist listened, A_1 by randomly selecting an artist with probability proportional to its global popularity.
2. Determine a time gap (in seconds) between this listen and the next by probabilistically drawing from the overall distribution of inter-listen times.
3. Determine a jump distance, d (in artist space) between this listen and the next by probabilistically drawing from the overall distribution of inter-listen jump distances (i.e. the empirical data in Figure 6).
4. Randomly select an artist A_2 such that $cosine\_distance(A_1, A_2) \approx d$.
5. Repeat steps 2-4 until a listening sequence of length 20,000 has been generated (this value was chosen as it is approximately equal to the median number of listens per user in our data, 19,347).

We repeated this process until we had a sample of 10,000 artificial listening sequences. Due to the computational demands of maintaining the full distance matrix in memory during this procedure, we limited the set of possible artist to the 50,000 most popular (rather than the full set of 112,312).

A second approximation was necessary for step 4, because the empirical distribution of jump distances is necessarily computed over discrete bins, and it is unlikely to find an artist that is a particular distance from another. Thus we first rounded all values of the pairwise distance matrix to two decimal places. Once we had

probabilistically drawn a distance from this reduced precision matrix in step 3, we then randomly selected (with uniform probability) from the set of artists that were approximately that distance from the reference artist. For example, assuming we drew *d=0.34* in step 3 as the distance between A_1 and A_2, artist A_2 would be drawn with uniform random probability from the set of artists whose *rounded* distance from A1 was equal to 0.34. If this set of artists was null (i.e. there exited no other artist distance d from A_1), another distance was draw from the overall distance distribution.

Formalizing patch segmentation

With the decision made to ignore inter-listen times, we clearly will not move forward with the session-based patch definition, and here I detail the patch segmentation process for each of other three methods.

- "Block" method: When treating each artist as a self-contained patch, we assign a unique block index to each contiguous sequence of listening to a given artist. Given the example sequence of artists shown in Table 10, the column "Block_idx" shows the corresponding assignment block index assignment under this method (we assume this is the start of a listener's history, and thus start all indexing at zero).
- "Simple patch" method: This method captures the notion of localities in the artist feature space by drawing a patch boundary any time the distance between two artists exceeds a given threshold. This requires calculating the distance between each adjacent pair of artists (the "Distance" column in Table 10). The "Simple_idx" column shows the segmentation achieved with a distance threshold of 0.2.

- "Shuffle patch" method: This method requires two parameters, a distance threshold as in the simple patch method, and a minimum patch length (*MPL*) parameter. Beginning with the patch boundaries defined with the previous method, this proceeds by checking the number of listens between each patch boundary, and eliminating any boundaries that do not begin or end a patch of length $\geq MPL$. This results in "shuffle patches" constituted by single listens (or short chunks of listening) to various artists that are not long enough to form an independent patch. This also means that, with *MPL=1*, this method is equivalent to the simple patch method. The "Shuffle_idx" column of Table 10 shows an example segmentation of listening with a distance threshold of 0.2 and MPL of 5. Note, for example, that the length-two patch made up of Beck and Radiohead under the simple patch definition becomes part of a short shuffle patch here.

Table 10: Example segmentation of listening excerpt under the three patch definitions. Patch boundaries emphasized by double-lined cell borders. "Distance" indicates cosine distance between the corresponding artist and the preceding one.

Artist	**Distance**	**Block_idx**	**Simple_idx**	**Shuffle_idx**
Sufjan Stevens	NA	0	0	0
The Offspring	0.974887	1	1	0
Beck	0.764417	2	2	0
Radiohead	0.176106	3	2	0
Sparta	0.746763	4	3	1
At the Drive-In	0.088306	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1
At the Drive-In	0.000000	5	3	1

Artist	Distance	Block_idx	Simple_idx	Shuffle_idx
Hollywood Undead	0.916317	6	4	2
Placebo	0.854387	7	5	2
At the Drive-In	0.834327	8	6	3
At the Drive-In	0.000000	8	6	3
At the Drive-In	0.000000	8	6	3
At the Drive-In	0.000000	8	6	3
At the Drive-In	0.000000	8	6	3
At the Drive-In	0.000000	8	6	3
Caetano Veloso	0.979813	9	7	4
Caetano Veloso	0.000000	9	7	4
Caetano Veloso	0.000000	9	7	4
Caetano Veloso	0.000000	9	7	4
Caetano Veloso	0.000000	9	7	4
Antonio Carlos Jobim & Elis Regina	0.066404	10	7	4
Antonio Carlos Jobim & Elis Regina	0.000000	10	7	4

Selecting the two parameters for these algorithms in a principled way is challenging, as they modulate what constitutes (a) a locality in artist space and (b) a patch of listening long enough to permit meaningful analysis, two constructs that lack clear criteria for selection. To explore the effect of changing these parameters, however, we examined the mean distribution of patch lengths generated under different parameter values in Figure 47 across a random sample of ~3,000 users.

In each subplot, line color corresponds to the distance threshold for subsequent listens to be grouped into a patch (we tested values of 0.1, 0.2, …, 0.9), and there is one subplot per value of *MPL* tested. Recall that with *MPL=1*, the simple and shuffle patch methods are equivalent. There is a clear preponderance of length 1 patches, regardless of parameter choice, and a secondary peak at whatever value of *MPL* was chosen. We also observe a "bump" in the distribution for patches of 10-15 listens, which is likely driven by the fact that this roughly corresponds to the length of an album. The effect of the

distance threshold is clearly monotonic and does not substantively change the form of the distributions (though a sufficiently high value does suppress the "album effect").

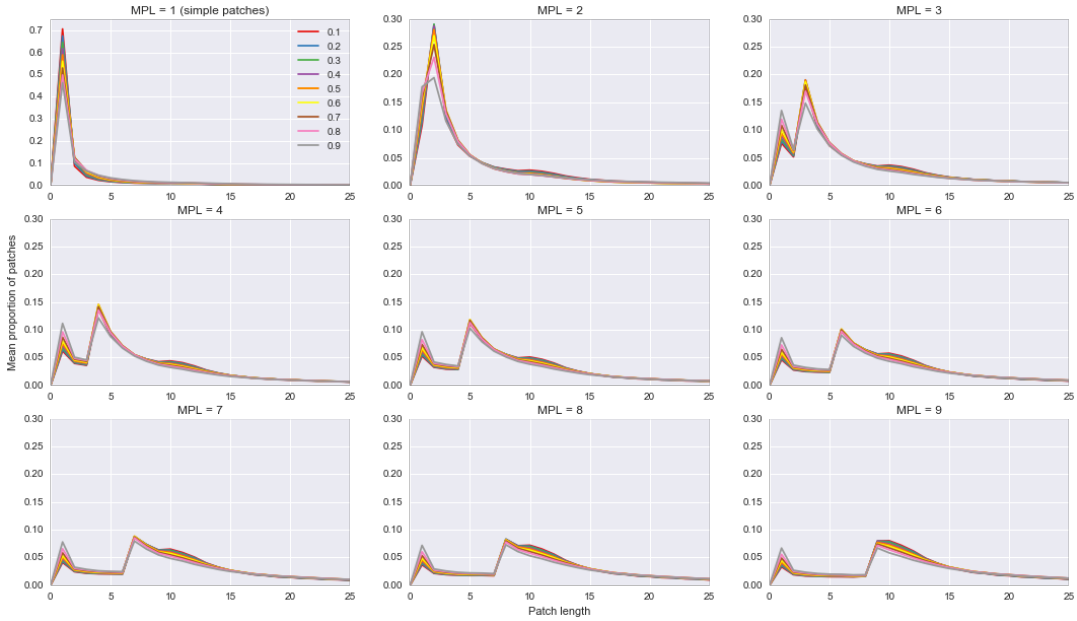


Figure 47: Mean distributions of patch lengths under different distance thresholds (indicated by line color) and values of MPL (indicated by each subplot) and distance. Note different scale for the first subplot.

We extended the analysis by also calculating the probability distributions of listening across patch lengths (i.e. instead of the proportion of patches that were of a given length, we ask what proportion of users' listening fell within patches of a given length), shown in Figure 48.

This makes it clear that length-one patches, though common (Figure 47), do not make up a large proportion of users' listening (Figure 48). The qualitative trends of the "album effect" and peak around patches of length matching *MPL* remains, however. We also observe that the distance threshold does not make a large difference in the qualitative form of the distributions, particularly as we increase *MPL*. This surprisingly low level of sensitivity is likely due to the high proportion of jumps between artists with high

distances (covered in more detail in the next section). These analyses do not suggest a clear cut choice of parameter values, but for now we have selected a distance threshold of 0.2 and a *MPL* of 5. The former represents an intuitively reasonable notion of two artists being "close" on the [0-1] cosine distance scale, and is partly based on qualitative examination of artist pairs of varying distances. The latter was chosen to be long enough such that summary measures describing a patch (e.g. diversity of listening within a patch) can be meaningfully defined, but short enough that we can observe fine-grained segments of listening activity. Furthermore, we can observe in Figure 48 that the proportion of length-five patches remains relatively constant (around 4%) as *MPL* is changed (except for *MPL* values greater than five, of course).

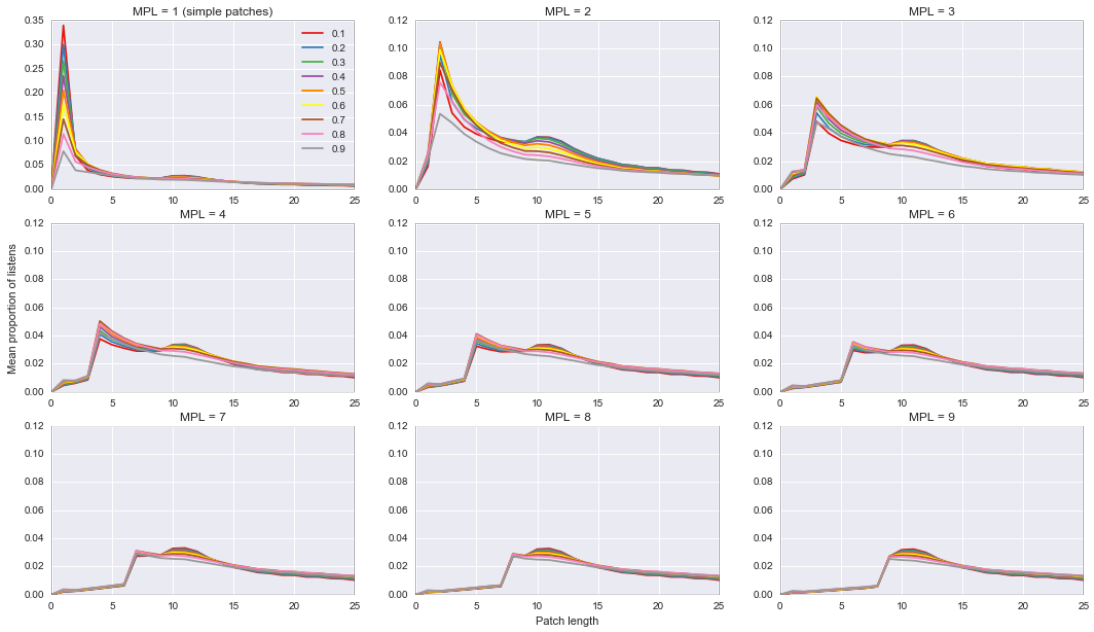


Figure 48: Mean distributions of listening within patches of a given length. Values of MPL and distance threshold follow the conventions of Figure 47. Note differing scale for the first subplot.

Patch clustering

The "shuffle patch" method described above, on which we focus in the main text, in fact only describes half of the patch identification process. After segmentation, it is necessary match disparate listening segments so we can determine when a user has return to the "same" patch. Our patch segmentation method identifies contiguous regions of listens wherein the distance in artist space from one artist to the next is below our specified cosine distance threshold of 0.2, so of course there is variability in the precise makeup (in terms of length and the set of artists listened) of identified listening segments, so we cannot expect any one-to-one to correspondence between different segments and utilize a clustering algorithm to match segments.

The first step is to separate exploratory versus exploitative behavior. We do so by imposing a conservative threshold on the diversity of listening within listening segments, where the diversity is defined as the mean of pairwise distances between all listens (not between unique artists, such that repeated listens to the same artist thus *do* affect the diversity measure). A reasonable choice for this threshold is to use the same value of 0.2 from our segmentation algorithm. This is equivalent to defining an exploitative segment as any in which the artist space distance between any two listens is, on average, less than 0.2. This division emerges naturally from the segmentation algorithm, as is clear in Figure 49, where we examine the distribution of listening segment diversities across all users. Although over half of all segments have zero diversity (i.e. constitute listening to a single artist, left), excluding zero-diversity segments reveals a bimodal distribution divided around a diversity value of approximately 0.2.

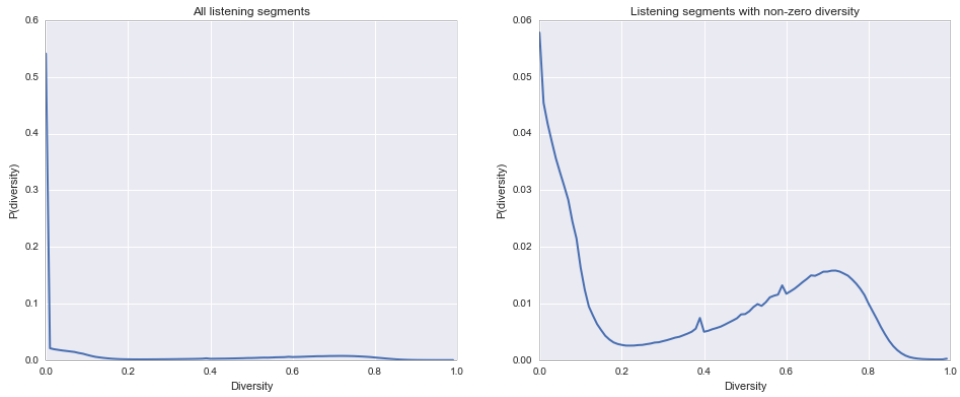


Figure 49: Distributions of listening segment diversities across all users and segments (left) and across only those segments with non-zero diversity (right).

There exist a broad variety of possible clustering approaches (Jain, 2010), but in the interest of simplicity and clarity, we apply a basic agglomerative hierarchical lustering technique *to exploitative segments only*. Though there is variety in exploratory behavior, we save study of such variation for future work, and instead only identify patches from among a user's exploitative listening, essentially lumping together all other segments as homogenous exploratory behavior.

Exploitation segments, which following the parameter choice above are of length ≥5, are first represented by their mean feature vectors, defined as the mean of the feature vector for each artist in the segment, weighted by the number of occurrences of the artist in the segment. We then apply complete-linkage agglomerative clustering. In general, agglomerative clustering methods function by initially assigning each observation to its own cluster. At each iteration of the algorithm, the two most similar clusters are combined into a new cluster until all observations eventually fall into a single cluster. The history of these meres between clusters results in a dendrogram which can be used to visualize hierarchical relationships in the data, or cut a specified threshold to achieve a

flat clustering of observations. Agglomerative clustering methods differ in how the determine which clusters are most similar (i.e. the linkage method), and we opt for complete linkage, wherein the distance between clusters is defined as the maximum distance between any two observations across the clusters. This conservative method leads to relatively small clusters of a fixed diameter (Everitt, Landau, Leese, & Stahl, 2011) that align with our conception of patches as localities in artist space.

After passing exploitative listening segments through the clustering algorithm, we cut the dendrogram at a distance threshold of 0.2 to generate a flat clustering (see Figure 50). All listening segments assigned to the same cluster can then be characterized as belonging to the same patch. By using the same threshold as above (0.2), we can describe the patch clustering process as first identifying segments where the mean pairwise distance between listens is below our listening threshold, then identifying all listening segments within a given diameter in artist space as being visits to the same patch. For our current purposes, we perform clustering independently for each user and as such do not identify patches across users. Generating global patch definitions is a promising direction for future work.

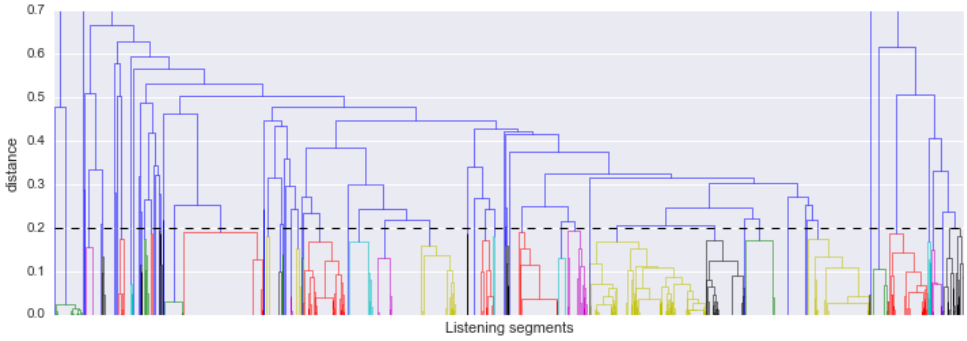


Figure 50: Example patch clustering for one user. The dendrogram is cut at a distance threshold of 0.2 (black dashed line), so all branches below that point are collapsed into a single segment cluster (patch), indicated by contiguous regions of the same color. The ordering of clusters is arbitrary.

Jared Lorince
The Northwestern Institute on Complex Systems
Chambers Hall, 600 Foster Street
Evanston, Illinois 60208

Email: jared.lorince@kellogg.northwestern.edu
Homepage: https://jlorince.github.io

Education

Joint Ph.D. Program in Cognitive Science and Cognitive Psychology
Indiana University, Bloomington, September 2016

B.A. in Cognitive Science
University of California, Berkeley, May 2009

Research & Work Experience

Northwestern University, Kellogg School of Management
Postdoctoral fellow, The Northwestern Institute on Complex Systems (NICO), September 2016 - Present

StumbleUpon.com, San Francisco, CA
Data scientist, July 2015 – August 2016
Data science intern, April 2015 - July 2015

Indiana University, Department of Psychological & Brain Sciences, Cognitive Science Program
Graduate student researcher, Adaptive Behavior and Cognition Lab, August 2010 - Present
Graduate student researcher, "Heuristica" project, August 2011 - March 2015

Yahoo! Labs, Sunnyvale, CA
Research scientist student, User Intent Analysis Group, July 2011 - March 2012

Sonoma State University, Department of Psychology
Research assistant, August 2009 - February 2010

UC Berkeley, Department of Psychology
Research assistant, Computational Cognitive Science Lab, August 2008 - May 2009
Research assistant, Concepts & Cognition Lab, August 2008 - May 2009

Universidad Complutense de Madrid, Facultad de Psicología
Research assistant, EQUIAL ("Language acquisition research team"), March 2008 - July 2008

Peer-Reviewed Publications

Lorince, J. & Todd, P. M. (in press). Music Tagging and Listening: Testing the Memory Cue Hypothesis in a Collaborative Tagging System. In Jones, M. N. (Ed.), Big Data in Cognitive Science: From Methods to Insights (pp. xxx-xxx). New York, NY: Psychology Press (Taylor & Francis).

Lorince, J., Zorowitz, S., Murdock, J., & Todd, P. M. (2015). The Wisdom of the Few? "Supertaggers" in Collaborative Tagging Systems. The Journal of Web Science, 1(1), pp. 16-32.

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) (pp. 141-152). Springer International Publishing.

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 (WebSci 2014) (pp. 129-138). ACM.

Lorince, J., Donato, D., & Todd, P. M. (2014). Path Following in Social Web Search. In Proceedings of the 7th Annual Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP 2014) (pp. 119-127). Springer 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 (WebSci 2013) (pp. 215-224). ACM.

Mullinix, G., Gray, O., Colado, J., Veinott, E., Leonard, J., Papautsky, E. L., . . . , Lorince, J., et al. (2013). Heuristica: Designing a serious game for improving decision making. In Proceedings of the 2013 IEEE Games Innovation Conference (IGIC) (pp. 250-255). IEEE.

Veinott, E. S., Leonard, J., Papautsky, E. L., Perelman, B., Stankovic, A., Lorince, J., et al. (2013). The effect of camera perspective and session duration on training decision making in a serious video game. In Proceedings of the 2013 IEEE Games Innovation Conference (IGIC) (pp. 256-262). IEEE.

Talks/Presentations

Lorince, J., Joseph, K., & Todd, P. M. Do tags really function as retrieval aids? International Conference on Computational Social Science (ICCSS2015). Helsinki, Finland. 10 June 2015.

Lorince, J., Joseph, K., & Todd, P. M. Analysis of music tagging and listening patterns: Do tags really function as retrieval aids? Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP 2015). Washington, D.C. 3 April 2015.

Lorince, J., Zorowitz, S., Murdock, J., & Todd, P. M. "Supertagger" Behavior in Building Folksonomies. ACM Web Science Conference (WebSci 2014). Bloomington, IN. 25 June 2014.

Lorince, J. & Todd, P. M. Identifying Canonical Music Listening Patterns on Last.fm. Computational Approaches to Social Modeling (ChASM 2014) Workshop at Websci 2014. Bloomington, IN. 23 June 2014.

Lorince, J., Donato, D., & Todd, P. M. Path Following in Social Web Search. Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP 2014). Washington, D.C. 3 April 2014.

Lorince, J., Donato, D., & Todd, P. M. From Spatial Search to Information Search: Can Users Benefits from the Web Search Paths of Others? Midwest Cognitive Science Conference (MWCSC 2012). Bloomington, IN. 7 May 2012.

Posters

Lorince, J. & Todd, P. M. (2014, May). Metadata and Memory Cues in Collaborative Tagging: Music Listening and Tagging on Last.fm. Poster Presented at the 4th Annual Midwest Cognitive Science Science Conference, Dayton, Ohio.

Lorince, J. & Todd, P. M. (2013, May). Can simple social copying heuristics explain tag popularity in a collaborative tagging system? Poster Presented at the 5th Annual ACM Web Science Conference, Paris, France.

Lorince, J., Malviya, S., & Todd, P. M. (2012, July). Social information environments of collaborative tagging systems: Individual and group-level cognitive perspectives. Poster presented at the ABC Summer Institute on Bounded Rationality, Berlin, Germany.

McGlasson, C., Lorince, J., Crandall, D., & Todd, P. M. (2012, June). Testing an Evolutionary Account of Color Preferences Using Online Photos. Poster presented at the 24th Annual Meeting of the Human Behavior and Evolution Society, Albuquerque, New Mexico.

Balzarini, R., Goode, C., Lorince, J., Grenier, G., Plouffe, A., Pella A. & Smith, H.J. (2010, August). When Will Social Norms Persuade Group Members to Drink Less Alcohol? Poster presented at the annual meeting of the American Psychological Association, San Diego, California.

Lorince, J., Griffiths, T. & Lombrozo, T. (April, 2009). Exploring a bias towards unknown variables in causal attribution. Poster presented at California Cognitive Science Conference, Berkeley, CA.

Other peer-reviewed work

Lorince, J., Joseph, K., & Todd, P. M. (2015). Do Tags Really Function as Retrieval Aids? (Extended abstract presented at the International Conference on Computational Social Science)

Lorince, J. & Todd, P. M. (2014). Identifying Canonical Music Listening Patterns on Last.fm. (Extended abstract presented at the Computational Approaches to Social Modeling (ChASM) workshop at Web Science 2014)

McGlasson, C., Lorince, J., Crandall, D. J., & Todd, P. M. (2013). Exploring the use of big data in color preference research. Journal of Vision, 13(9), 1167-1167. (Meeting abstract presented at VSS 2013)

Invited Talks

Lorince, J. Adaptive Decision Making, Social Knowledge Generation, & Fun with Big Numbers. Indiana University Student Organization for Cognitive Science (SOCS) Meeting. Bloomington, IN. 19 February 2014.

Lorince, J. Socially mediated decision making in a collaborative tagging system. Indiana University Department of Telecommunications Media Arts & Sciences Speaker Series. Bloomington, IN. 16 November 2012.

Lorince, J. Get out of the lab and into the game: Why there needs to be dialogue between game designers and behavioral scientists. The IU TED-like Salon. Bloomington, IN. 4 December 2011.

Lorince, J. & Ross, T. Play how you want (or not): How the crowd modifies/limits individual behavior in online games. Indiana University Department of Telecommunications Media Arts & Sciences Speaker Series. Bloomington, IN. 18 November 2011.

Skills

Statistical analysis, machine learning, and visualization: Proficient with Python scientific analysis stack (Pandas, Scikit-learn, Scipy, Numpy, Matplotlib, etc.), Graphlab Create, and Apache Spark

Web data mining and databases: Web crawler development; experience with various database systems (MySQL, Hive/HDFS, Redis)

Experimental design: Experience developing and testing hypotheses in various contexts, including data analytics questions using big data tools, in-lab psychological studies and online research using Amazon Mechanical Turk

Other skills: Multi-agent model design; topic modeling; report writing (LATEX); basic web design tools (HTML, CSS, Javascript); other data analysis tools (R, Unix, Bash scripting)

Awards and Honors

Research fellowship, Templeton Foundation "What drives human cognitive evolution?" grant (2016)

National Science Foundation IGERT Fellowship, trainee in dynamics of brain-body-environment interaction in behavior and cognition, IU Bloomington, 2010-2016

Research fellowship, AFRL "Heuristica" grant for development of serious games for mitigating negative decision making biases (2011-2015)

Accepted to the ABC Summer Institute on Bounded Rationality, Berlin, Germany (2012)

Yahoo! Labs Faculty Research and Engagement Program grant recipient, 2011

Cognitive Science departmental citation winner, UC Berkeley 2009

High honors in Cognitive Science, UC Berkeley, 2009

High distinction in general scholarship, UC Berkeley, 2009

Teaching

Teaching Assistant, Introduction to Cognitive Psychology, Department of Psychological & Brain Sciences, Indiana University (Fall 2015)

Associate Instructor, Experimental Methods in Psychology, Department of Psychological & Brain Sciences, Indiana University (Fall 2014)

Service

Reviewer:
Transactions on Computer-Human Interaction, ACM
Topics in Cognitive Science (TopiCS), Cognitive Science Society.
Behavioral Research Methods, The Psychonomic Society.

Program Committee Member:
International Conference on Social Informatics (SocInfo 2016)

International Conference on Computational Social Science (ICCSS 2015-2016)
Computational Social Science Workshop (CSS 2014) at ECCS 2014.
6th Annual ACM Web Science (WebSci 2014).
Computational Approaches to Social Modeling Workshop at WebSci 2014.

Publicity Committee:
6th Annual ACM Web Science (WebSci 2014).