Influences on Tag Choices in del.icio.us

Emilee Rader
School of Information
University of Michigan
ejrader@umich.edu

Rick Wash
School of Information
University of Michigan
rwash@umich.edu

ABSTRACT
Collaborative tagging systems have the potential to produce socially constructed information organization schemes. The effectiveness of tags for finding and re-finding information depends upon how individual users choose tags; however, influences on users' tag choices are poorly understood. We quantitatively test competing hypotheses from the literature concerning these choices, using data from del.icio.us (a collaborative tagging system for organizing web bookmarks) and a computer model of possible tag choice strategies. We find evidence that users choose tags in a pattern consistent with personal information management goals, rather than as a result of social influence.

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces – web-based interaction, collaborative computing

Author Keywords
collaborative tagging, information management, logistic regression, computer model

INTRODUCTION

User-contributed metadata, also known as *tagging*, provides a means for users to associate personally salient keywords or labels with content items [8, 23], enabling them to find the content later via information they are predisposed to recognize or recall [12]. Tagging helps users "package" information for future information seeking and reuse [13]. Tagging has not only been applied to personal information management; many *collaborative tagging* systems have appeared in recent years. Collaborative tagging systems such as del.icio.us and citeulike.org publicly expose individual users' associations between content items and tags, thereby providing visibility into words others have used to tag similar items. Grudin [9] suggests that collaborative tagging can be a low-effort solution for shared or group information management, because it does not require that users try to conform to a controlled vocabulary or organization scheme. However, in other shared information management contexts, the effort required to "package" information is necessary for effective information reuse [13].

In a collaborative tagging system, users interested in viewing content tagged a certain way by others can browse the system by clicking on tags. Tags provide the "information scent" [21] that connects users with information; they are the infrastructure upon which information organization and finding takes place, allowing users to navigate by recognition rather than recalling terms by which to search [25]. This has interesting consequences when one considers the potential utility of tags for information management, finding, and re-finding. If a given tag is applied in an inconsistent manner among many users, more variability exists in the content items displayed when a user browses to a particular tag. For example, users tend to assign high-level tags like "technology" and personal tags like "to read", as well as words like "apple" that can refer to a computer or a fruit, and "photos" or "pictures" which are synonyms. Influences on how users choose words as tags could affect not only their own use of a particular tagging system for personal information management, but also impact how the system supports the information finding of others [26].

We focus on the social bookmarking website del.icio.us as a case study of a collaborative tagging system supporting both personal and shared information management. del.icio.us is an online application that allows users to save and tag their own web bookmarks so they are accessible from any networked computer. It is an interesting case for several reasons. The bookmark and tag histories for over one million users are public and can be viewed (and analyzed) by anyone. del.icio.us has received attention in the research literature as the canonical example of a collaborative tagging system for information management [8] (in contrast with the photo sharing website flickr.com, which incorporates tagging but has a different overall purpose). Finally, researchers suggest [8, 10] that a socially constructed shared vocabulary might emerge from individual users' tag choices on del.icio.us.

Our objective in this research was to look for a pattern of evidence indicating that a social process could affect tag choices. Golder and Huberman [8] speculate that users might imitate each others' tag choices; in other words, tag choices might be influenced by tags that had been previously applied to the same web page by other users. However, it is reasonable to assume that there might be other sources of influ-

CSCW'08, November 8–12, 2008, San Diego, California, USA.

ence on users' tag choices having to do with personal information management goals. For example, Wash and Rader [26] found that users of del.icio.us chose tags for organizing and re-finding their own bookmarks according to mental rules and definitions they had established, striving for consistency within their own personal "controlled vocabulary". Or, users might desire to expend as little effort as possible when choosing tags, and simply select tags suggested in the del.icio.us posting interface when they create a new bookmark.

We conducted a multiple-method investigation that teases apart these competing explanations. In Study One, we used a logistic regression analysis of a large sample from del.icio.us, in which we evaluated the influence of several predictors on users' tag choices. In Study Two we developed a computer model in which we assume a number of different tag choice strategies one at a time, and compare aggregate patterns in model results against the same measures in the data from del.icio.us[1]. We found evidence that users' tag choices are not a result of imitation of others' tags; instead, they follow an individual, idiosyncratic pattern. This suggests that personal information management goals, rather than social processes, have a greater influence on tag choices in del.icio.us.

BACKGROUND AND RELATED WORK

By default, bookmarks and tags in del.icio.us are public information. Each new bookmark has the following metadata associated with it: the username of the person saving the bookmark, the tags selected by that user, and the date and time the bookmark was created. Users browsing del.icio.us view subsets of bookmarks delimited by metadata such as a particular username, tag, or user-tag combination. For example, clicking the tag *library* in the list of popular tags on del.icio.us displays all web pages bookmarked by any del.icio.us user having the tag *library* associated with them. Clicking on a username displays web pages bookmarked by a particular person. The metadata for a given web page can also be displayed including the usernames of all the users who bookmarked it, and all the tags ever associated with it. The Library of Congress home page has been bookmarked in del.icio.us by 3060 different users, and tagged "library" by 1337[2]. When a user creates a new bookmark, the interface (Figure 1) displays *recommended tags* selected automatically by the system, *your tags* which are all tags chosen in the past by that user, and *popular tags* for that particular web page.

Furnas et al. [7] began the study of tagging with their paper on the vocabulary problem, in which they reported that when two random people create a label for the same document, they choose identical words less than 20% of the time. Tagging has been studied in a mobile context [2] and for photos [14]. It has been applied to personal information management [5, 24], and in a corporate environment [16, 17]. Researchers want to better understand tagging patterns [10, 8, 26] and make recommendations for how users might produce better tags [6, 22, 23]. We focus here in particular on the findings of Golder and Huberman [8] and Sen et al. [23], because they motivated and guided our investigation most directly.

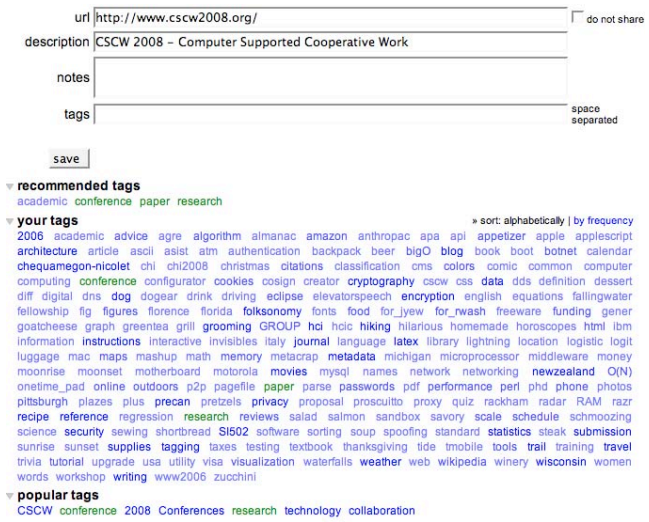

Figure 1. Screen capture of the previous version of the del.icio.us bookmark posting interface (the interface was changed as of Aug. 1, 2008)

Golder and Huberman [8] argue that users' tag choices are not random; instead, consensus emerges for which tags best represent a given web page. They show that web pages bookmarked in del.icio.us demonstrate a stable frequency distribution following a power-law pattern in which the same few tags are chosen by many users, while most other tags are selected by only one or two people. Golder and Huberman hypothesize that when a user bookmarks a web page in del.icio.us, their tag choices are influenced by tags that had been previously applied to that web page by others (p206). They illustrate this imitation hypothesis through a mathematical construct: the stochastic urn of Polya [20]. For users to behave according to Polya's Urn, they must randomly select tags from the tag distribution for a given webpage. This means that if the tag "library" makes up 13.5% of all tags applied to the Library of Congress home page, users must somehow choose "library" 13.5 times out of 100. However, the del.icio.us interface does not provide users with sufficient information about the tag frequency distribution to behave according Polya's Urn. Rather, users are presented a nonrandom, biased sample in the posting interface: the *recommended* and *popular* tags (see Figure 1). Golder and Huberman suggest that imitation occurs via these tags presented in the interface, but do not address the distinction between biased sampling methods and the unbiased random draws of Polya's Urn. In our computer model we implemented several different forms of sampling from the tag distribution, allowing us to clarify the difference.

Sen et al. [23] manually assigned tags from MovieLens, a movie recommendation system, to one of three *classes*: factual, subjective, or personal. Through a field experiment manipulating the information displayed in the MovieLens

[1]Our database schema and code for our computer model and analyses may be downloaded from http://bierdoctor.com/papers/cscw08
[2]As of April 17, 2008

tagging interface, they found that users imitated tag classes when tagging movies, and concluded that "community influence plays an important role in vocabulary" (p186). In our analysis we focus on a more fine-grained dependent variable, individual users' exact tag choices, rather than subjectively assigning tags to classes. This allows us to test competing hypotheses from the literature using a larger dataset containing tag choices made over a longer period of time, albeit without the experimental control afforded by the ability to make changes to the interface. The difference in the unit of analysis (tag classes versus exact tag choices) allows us to potentially reach different conclusions. As we will show below, our findings contradict those of Sen et al.; we found little support for the hypothesis that users imitate one another's exact tag choices.

STUDY ONE: TAG CHOICES IN DEL.ICIO.US

Over two weeks in January 2007, we downloaded the entire bookmark and tag history for approximately 20,000 different web pages in del.icio.us. The web pages were chosen by periodically sampling the "recently posted" and "popular" del.icio.us pages. We randomly chose 30 web pages from our sample that had been bookmarked by at least 100 users. Then, in June 2007 we downloaded the complete public bookmark and tag histories for all of the approximately 12,000 users who had ever bookmarked any of these 30 web pages. In other words, our dataset contains the complete tag histories for 30 web pages bookmarked in del.icio.us, as well as tag histories for all users who ever bookmarked any of those 30 web pages as of June 2007.

Model and Data Setup

We used a logistic mixed model regression[1] to evaluate the influence of three hypotheses on users' tag choices:

1. *Imitation*: Users imitate tags that previous users have applied to a web page
2. *Organizing*: Users re-use tags that they have applied to other web pages
3. *Recommended*: Users choose tags that are suggested via the del.icio.us posting interface[3]

If the *imitation* hypothesis has a strong influence, tags previously associated with a given web page by other users will be correlated with tag choices. We can assume a social process is at work, and a socially constructed vocabulary is truly emerging. If tagging behavior is determined more by *Organizing* than by *Imitation*, then we expect tags a user has applied before to other web pages to be correlated with tag choices. Finally, if the *Recommended* hypothesis is true, users' tag choices are influenced by tags suggested in the del.icio.us posting interface.

We model the dependent variable — the choice of a single tag — as a yes/no choice. Because we lack evidence indicating what tags users have or have not viewed prior to choosing tags, we make a simplifying assumption that the list of observations for each user consists of a yes/no choice for all tags applied to the particular web page at the time our data were collected. We attempt to estimate the probability of saying "yes" to each tag as a function of three different factors included in the model as predictors. First, if *Imitation* is shown to have strong influence on a particular tag choice by a particular user, then the probability that a tag is chosen should be higher if the word has been used previously as a tag. This would be reflected in the model as a large, positive coefficient for the "$\mathrm{used.onsite}$" predictor. Second, if *Organizing* is shown to have strong influence, the probability that a word is chosen should be higher if the word has been previously used by that user as a tag for a different web page. This would be reflected by a large, positive coefficient for "$\mathrm{used.byuser}$". For the *Recommended* hypothesis, the algorithm for selecting tags to display in the posting interface is not publicly known; however, some experimentation with del.icio.us has led us to believe that a tag is much more likely to be recommended if it has both been applied previously to that web page and used previously by the user. Therefore, we approximated the *Recommended* hypothesis by including an interaction term that is 1 when both $\mathrm{used.onsite} = 1$ and $\mathrm{used.byuser} = 1$.

The model also includes several controls for other factors that may influence the probability of choosing a tag. Some tags seem to "fit" the web page better than others (i.e., *library* for the Library of Congress home page), and are more frequently applied. Since the data include repeated measures for each tag, it is important to control for per-tag variability using fixed effects. This is represented in the model by "$\mathrm{tag\_dummys}$". Also, some users tend to assign more tags to their bookmarks than others; we controlled for this within-user variability using random effects. Finally, we account for temporality in the $\mathrm{used.onsite}$ variable. This variable is 0 for early bookmarks and 1 for later bookmarks, switching after a tag is used. We believe $\mathrm{used.onsite}$ controls for any autocorrelation that might result from the previous use of certain tags, and therefore a time series model is not necessary. The model is set up as follows:

$$\mathrm{tag\_chosen} = f(\mathrm{used.onsite}, \mathrm{used.byuser}, \mathrm{interaction}, \mathrm{tag\_dummys}, \mathrm{random\_effect(user)})$$

Logistic Regression Results

We estimated the model using maximum likelihood estimation, separately for each of the 30 web pages in the study. This allowed us to compare web pages and determine whether an overall pattern exists.[4] We summarize the estimates for the model coefficients in Table 1.

In logistic regression, the dependent variable is dichotomous, meaning it takes only two possible values. The model is used to estimate the probability of the dependent variable taking on the value 1, given a set of predictors. This probability is represented in the form of *odds*. For example, a probability of 50% can be represented as 1:1 odds, and 2:1 odds trans-

[3]It is difficult to concretely specify this hypothesis because del.icio.us does not reveal its method for selecting tags to suggest, and the method may have changed multiple times.

[4]Combining the data for all 30 web pages into one large dataset proved computationally infeasible.

Table 1. Logistic Regression Results. Coefficients for tag_dummys **and per user random effects are omitted due to space constraints.**

Title	*Users*	*Used.onSite*	*Used.byUser*	*Interaction*	$G_m(df)$	R^2_L	*P*	λ_p
A List Apart: Alternative Style	395	-0.1665	3.764 ***	-0.5780 *	6019 33 ***	0.5218	0.9824	0.2322 ***
London Underground History	369	-0.3365 *	3.226 ***	-0.0501	6638 34 ***	0.4923	0.9778	0.1049 ***
Haiku	161	-0.7128 *	2.368 ***	0.8593 *	2100 20 ***	0.5081	0.9494	0.2086 ***
Spread Firefox	214	-1.0990 ***	2.825 ***	0.5331 *	2277 24 ***	0.4360	0.9799	0.1293 **
PayPalSucks.com	121	-0.4083	3.140 ***	-0.1475	1146 17 ***	0.4174	0.9557	0.0760
OS X Maintenance	282	-0.5596 **	3.106 ***	-0.1510	3686 28 ***	0.4648	0.9744	0.1935 ***
The Library of Congress	552	-0.4079 ***	3.740 ***	-0.3113 *	7882 39 ***	0.4455	0.9921	0.0986 ***
GDI+ FAQ main index	114	-0.1986	3.528 ***	-0.6113	1299 21 ***	0.4602	0.9485	0.1974 ***
MetaGer	174	-0.4910 *	4.776 ***	-1.3510 ***	1318 20 ***	0.3952	0.9736	0.1367 **
eHomeUpgrade	270	-0.1153	3.712 ***	-0.5184 *	3495 35 ***	0.4207	0.9797	0.0809 *
Getting started with SSH	938	-0.0064	3.289 ***	-0.3577 *	18337 43 ***	0.5645	0.9846	0.1625 ***
err.the_blog	456	0.4622 .	3.578 ***	-0.4908 .	7622 31 ***	0.5496	0.9755	0.2847 ***
Beer Advocate - Respect Beer.	489	0.0400	3.222 ***	-0.2351	6899 27 ***	0.5357	0.9846	0.2392 ***
Old Computers	258	-0.2279	4.055 ***	-0.6511 **	3770 28 ***	0.4777	0.9785	0.1637 ***
DotNetNuke	714	-0.1742	3.659 ***	-0.6486 ***	12376 55 ***	0.4761	0.9878	0.0950 ***
BibDesk	303	-0.4937 **	3.859 ***	-0.2865	5941 35 ***	0.5116	0.9800	0.2163 ***
Tiny Icon Factory	819	0.0009	2.916 ***	0.4367 **	15902 58 ***	0.5041	0.9865	0.1337 ***
Mint: A Fresh Look at Your Site	560	-0.1202	3.570 ***	-0.3691 *	10730 46 ***	0.4701	0.9869	0.0430 *
Telegraph newspaper online	447	-0.4688 **	4.350 ***	-0.7939 ***	5221 23 ***	0.5094	0.9890	0.2109 ***
Glimpses? The Uncanny Valley	166	0.1536	2.995 ***	0.0883	2300 36 ***	0.3701	0.9668	0.0364
DVDStyle	157	-0.7305 *	2.656 ***	0.4941	2469 20 ***	0.4974	0.9532	0.2153 ***
digg labs / swarm	499	-0.4685 ***	2.876 ***	0.5907 ***	9877 55 ***	0.4768	0.9867	0.1044 ***
Flickr: The HDR Pool	596	-0.3258 .	3.208 ***	-0.2578	9210 29 ***	0.5472	0.9833	0.2459 ***
Sxip Identity	496	-0.2473 .	3.958 ***	-0.8236 ***	8318 39 ***	0.4833	0.9870	0.1158 ***
Many Eyes	466	0.3220 *	3.032 ***	-0.0818	9276 54 ***	0.4729	0.9820	0.1421 ***
Obscure Sound - Indie Music Blog	116	-0.4727	2.899 ***	0.0480	1136 15 ***	0.5008	0.9488	0.3354 ***
JotSpot Wiki (dojomanual)	218	0.0855	3.82 ***	-1.1510 **	3150 29 ***	0.5009	0.9533	0.2314 ***
BasKet Note Pads	124	-0.7212 **	3.211 ***	-0.3224	2183 24 ***	0.4612	0.9584	0.1339 ***
101 Cookbooks[5]	1000	0.1086	4.297 ***	-1.1060 ***	18231 43 ***	0.6028	0.9932	0.2595 ***
Snipplr - Code 2.0[5]	850	0.4304 ***	3.536 ***	-0.1440	20329 83 ***	0.4934	0.9888	0.1199 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

lates to a 66% probability. The coefficients for the predictors in a logistic regression model are the natural logarithm of odds *ratios*, or the ratio of the odds of one possible outcome divided by the odds of another outcome. In the model, our predictors are dummy variables that can be either 1 or 0. Therefore, the coefficient represents the natural logarithm of the ratio between the odds that a tag will be chosen when the value of the predictor is 1 to the odds when the predictor is 0. If the coefficient is positive, then the probability of a tag being chosen is greater when the value of the predictor is 1 (or true). If the coefficient is negative, the probability of a tag being chosen is greater when the predictor is 0 (or false).

To interpret the results in Table 1, first focus on the columns for $\mathrm{used.onsite}$, $\mathrm{used.byuser}$, and $\mathrm{Interaction}$. The values in these columns are the coefficient estimates for predictors representing our three hypotheses. The size of the coefficient and whether it is positive or negative indicates whether that predictor increases or decreases the probability of a given tag being chosen, and how strong the effect is. From these coefficients, we can calculate the predicted probability of being chosen for each tag applied to a given web page. An example of fitted probabilities for "101 Cookbooks" is presented in Table 2. The remaining columns of Table 1 present the results of statistical tests to evaluate the validity of our model.

[5]These two web pages in our sample had more users bookmark them than listed (1427 and 1137 respectively) but we truncated the dataset for computational reasons.

The three hypotheses are operationalized as follows:

1. *Imitation*: When bookmarking a given web page, users choose tags previously associated with that web page by other users ($\mathrm{Used.onSite} > 0$)

2. *Organizing*: When bookmarking a given web page, users choose tags they had applied before to other web pages ($\mathrm{Used.byUser} > 0$)

3. *Recommended*: When bookmarking a given web page, users choose tags suggested in the del.icio.us posting interface, operationalized in our model as tags that had previously been both applied the web page and used by the user on other web pages ($\mathrm{Interaction} > 0$)

A Wald test[6] can be done on each parameter estimate, similar to the standard t-test used in Ordinary Least Squares (OLS) regression. It compares the Null hypothesis that the true value of the parameter is 0 with the alternative hypothesis that the parameter is not 0. The stars in Table 1 show the statistical significance of these Wald tests.[7]

[6]The likelihood ratio test is more accurate, but requires more time to compute; this can be problematic for very large samples (like ours). Using the Wald statistic can increase the standard error when the estimated coefficient is large, leading to failure to reject the null hypothesis (Type II error) [15].

[7]Multi-collinearity can produce large standard errors, making it impossible to get statistically significant estimates. We frequently rejected the null, indicating that collinearity is not a problem [11].

Table 2. Fitted Probabilities for the top 4 tags on *101 Cookbooks*

used.onsite	used.byuser	food	cooking	recipes	blog
no	no	0.2034	0.2285	0.2183	0.0340
yes	no	0.2216	0.2482	0.2374	0.0377
no	yes	0.9494	0.9561	0.9535	0.7209
yes	yes	0.8737	0.8892	0.8833	0.4879

The *Imitation* hypothesis was supported for one web page, ManyEyes, which has a positive (though small) parameter estimate significant at the 5% level. Although the Wald test for the used.onsite predictor is significant for 13 sites, 12 have a negative parameter estimate, which does not support the *Imitation* hypothesis. From this we reject Hypothesis 1. However, the *Organizing* hypothesis is supported for all 30 web pages at the 0.1% level. The parameter estimates are quite high, indicating a strong effect. From this pattern, we conclude that Hypothesis 2 is supported. Finally, the *Recommended* hypothesis is supported for 4 of the 30 web pages at the 5% level (the other 12 significant estimates are negative). However, because we are uncertain how well we have approximated the recommendation algorithm in del.icio.us, we hesitate to draw conclusions about this hypothesis.

To illustrate the pattern of our results, Table 2 shows the model-estimated probabilities of choosing the 4 most frequently used tags (as of June 2007) on the *101 Cookbooks* web page for an average user. When a tag has been used previously by a user, our analysis shows a much greater probability of it being chosen again than if the user had not used it before. This pattern is consistent across all 30 webpages.

Model Fit and Diagnostics

We conducted two different types of goodness-of-fit tests to ensure that these results actually represent what is in the data. The first test is analogous to the standard goodness-of-fit test for OLS regression. In OLS regression, the F statistic is a statistical test that the model actually fits the data. Technically, it is a hypothesis test that the specified model fits the data better than the simplest possible model – the mean of the data. For logistic regression, the G_m statistic is analogous to the F statistic. It compares the specified model to the mode of the data, which is the simplest explanatory statistic for a binary variable. The G_m test is statistically significant at the 0.1% level for all 30 models. The OLS R^2 statistic represents how much of the variability in the data the model is able to explain. It is a substantive, rather than statistical test of significance. The R^2_L statistic is the logistic equivalent of R^2 [15], and represents the percentage of the likelihood explained by the model.[8] For our models, R^2_L indicates that the models explain about 50% of the likelihood on average. This indicates that there is definite room for improvement in understanding why users choose certain tags, but that our predictors account for a nontrivial portion of the likelihood.

The second test we conducted concerns the predictive efficiency of the model. With a binary independent variable, we can use the model to "predict" our dependent variable. To do this, we calculate the estimated probability (as we did in Table 2) and predict that a user will choose that tag if this probability is greater than 50%. The *P* column in Table 1 shows the percentage of tag choices that our model predicts correctly. For every web page, our model is over 94% accurate. However, this number can be misleading, as always predicting that the user will choose no tags can achieve above 90% correct for many web pages. To measure how much better our model predicts tag choices, we calculated the λ_p statistic [15]. This statistic represents the "proportional reduction in errors" — how many fewer errors does our model make than expected? This statistic ranges from 1 when all errors have been eliminated, to 0 when we make the same number of errors as a simple predict-the-mode model, to potentially negative if we make more errors than expected. In general, our model allows us to predict tag choices approximately 10 to 20 percent better than a simple predictor, and never worse. This improvement is statistically significant at the 0.1% level for all but 6 web pages (and 4 of those 6 are significant at the 5% level).

Interpretation

The results in Table 1 have a clear pattern; of our three explanatory variables, the strongest influence is users' previous tag choices. The coefficients on used.byuser consistently indicate a much larger influence than that of used.onsite or the interaction term. While user variability and individual tag 'fit' (represented by control variables in the model) play an important role in the choice of tags, the data indicate that users' previous tag choices are also important. This analysis also casts doubt on the *Imitating* and *Recommended* hypothesis, as operationalized in our model. We were only able to detect influence of these predictors in 1 and 4 web pages, respectively, and in these instances the influence was small. If there is a social process at work promoting a socially constructed vocabulary, we doubt that it takes the form of direct imitation. We are less sure about the effect of recommended or popular tags because we do not have a compelling measure of this explanatory variable.

STUDY TWO: MODELING TAG CHOICE STRATEGIES

The logistic regression analysis described above allows us to detect patterns in tag choices a posteriori; as such we are only able to speculate about what processes may have caused those patterns to occur. To address this weakness, we developed a computer model to evaluate competing explanations for the aggregate pattern of tags that appears on del.icio.us. We call these competing explanations *tag choice strategies*. In addition, the analysis described in Study One lumped together several forms of what might be considered "imitation" strategies into one explanatory variable. Computer modeling allows us to specify different forms imitation might take, and control what strategy is used to choose tags. It would be nice to instruct collections of real people to use one or more of the strategies suggested by the literature; we could then determine whether those tag choices resulted in tagging patterns similar to those found on del.icio.us. However, this technique would be prohibitively costly. Computer modeling allows us to explore the effects of different strate-

[8]Menard [15] points out that the standard R^2 statistic can be calculated for logistic regressions, but is biased. For this reason, we do not report it here.

gies, and compare them with the real-world data [18]. Such models cannot tell us which strategy or strategies real users of del.icio.us used; they can only tell us which strategies result in patterns of tags that are different from those observed in our large sample downloaded from del.icio.us, henceforth called the *real world* data. In other words, this technique cannot confirm which strategy is prevalent on del.icio.us, but it can be used to rule out possible explanations.

Measures

Axtell et al. [3] described two types of measures for validating computer models against each other or against a real world dataset. *Distributional equivalence* is achieved when the distributions of results being compared are statistically indistinguishable; *numerical identity* exists when samples from different sources are shown to produce results that are numerically equivalent. We selected two measures to compare the tag choice strategies in our computer model against the real world data, one to test for distributional equivalence, and the other to test for numerical identity.

Baseline for Distributional Equivalence

To establish a baseline measure against which to evaluate the distributional equivalence of tag choice strategies implemented in our model, we identified the theoretical distribution that most closely matched the tag frequency distribution in our del.icio.us sample. We fit the data from each web page to seven different discrete probability distribution families (discrete powerlaw, negative binomial, binomial, discrete lognormal, discrete exponential, poisson, and geometric), estimating parameters with maximum likelihood estimation, to discover which distribution fit "best" (a statistical determination [4]). We then used a non-nested Kolmogorov-Smirnov (KS) test to conduct pairwise comparisons between these distributions. The KS test is a common goodness-of-fit test to determine how well a set of data points fits a particular theoretical distribution. We are using it here to fit our data to distribution types other than normal.

The discrete powerlaw distribution fit the empirically observed (real world) tag distributions better than the other seven distributions we tested. The fitted distribution had an average exponent α of 1.92 ± 0.40. This is a low exponent for a powerlaw distribution, and indicates that the "long tail" of tags is very long and heavy. This low exponent also has another important implication. Newman [19] explains that powerlaw distributions with an exponent less than 2 have an infinite (or undefined) mean. Therefore, estimates of a "mean" or average tag are undefined, and any inferential statistics based on the mean of the tag distribution cannot be used.

Baseline for Numerical Identity

To measure the extent of the vocabulary problem [7], we calculated the average inter-user agreement (IUA) for a sample of 200 users from each of the 30 web pages in our sample from del.icio.us described above; this measure became our baseline for establishing numerical identity. On average, users who bookmarked these web pages chose the same tag only $14\% \pm 5\%$ of the time. IUA is sufficiently different than the goodness of fit to a powerlaw distribution of tags, and is a complementary measure for characterizing a set of tag choices.

Modeling Tag Choices

We modeled 120 web pages for each of five tag choice strategies we implemented, described in detail below. Each modeled web page was paired with one of 30 real web pages used in Study One, and the number of users for each web page modeled was chosen to match the real web page. In essence, we are simulating what would happen if the same set of users bookmarked the real web page, but chose their tags according to one of our five hypothesized strategies (and bookmarked it in a random order). To simulate a user choosing tags for a web page, two choices have to be made. First, the computer model chooses how many tags that user will apply to the web page. Second, the model chooses which specific tags will be applied. These parameters are selected by the model for each web page based on the distribution of parameters we found on del.icio.us.

The tags from the matched real web page are ordered from most-frequently used to least-frequently used, with ties broken randomly; each tag is then mapped onto a number according to its rank in the frequency distribution. When the random-number generator produces a 1, this is mapped to the most-frequently-used tag, 2 onto the second most frequently used tag, and so on. Any numbers larger than the number of tags on the matched web page are left as numbers. For each user, the specific tags they choose depends on which tag choice strategy is being modeled. The only difference between these strategies is in specific tag choice; all other decisions (number of users, number of tags per user, etc.) are identical. We implemented five different tag selection strategies in our computer model:

Zipf: Zipf's law states that word frequency in most written works follows a powerlaw distribution. Therefore, del.icio.us users might naturally choose their words from this distribution [19]. This could potentially account for Golder and Huberman's observation that the stable pattern in the tag frequency distribution for web pages bookmarked on del.icio.us is evident even for less common tags not popular enough to be recommended in the del.icio.us posting interface (p. 206) [8]. The model chooses random numbers from the base powerlaw distribution until it has the required number of unique numbers. These numbers are then mapped onto tags as described above.

Organizing: Users might favor tags that they had used previously. This strategy was described by Wash and Rader [26]. Simulated users have a 50% chance of choosing tags according to Zipf's law, and a 50% chance of choosing tags they had used before. When choosing tags they had used before, the model computes the overlap (set intersection) between tags the user had ever used and tags that were ever applied to the matched web page. It then randomly chooses among the tags in this overlap set. If that is not enough tags, then additional tags are chosen randomly from the base powerlaw distribution.

Imitation-Urn: Imitation of other users' tag choices might be achieved using a path-dependent process, as described by Golder and Huberman [8] in the Polya's Urn example. For users to imitate previous users' tag choices, it is necessary for those previous users to exist; the first few users who bookmark a web page have no one to imitate. To handle this, the first 20 simulated users draw as described above for *Zipf* and serve as 'seeders.' All users after the first 20 choose a tag from the current empirical distribution of tags for the simulated web page. This means that if there are two tags, 'A' and 'B', and 'A' has been used twice previously and 'B' only once, then tag 'A' is chosen with probability $\frac{2}{3}$ and tag 'B' is chosen with probability $\frac{1}{3}$. However, to ensure growth of the vocabulary beyond that used by the initial 20 seeders, each tag choice has a 10% probability of choosing a new, previously unused tag. This probability was chosen to match the average empirically observed probability from the del.icio.us data. The average web page in our original sample from del.icio.us has a new tag probability of $10.5\% \pm 8.3\%$.

Imitation-Popular: Users might prefer to click on the tags that are suggested in the del.icio.us posting interface. This was also hypothesized by Golder and Huberman [8] to be a plausible form of imitation, via biased sampling. Suggested tags in the del.icio.us posting interface come in two forms: *recommended* and *popular*. Del.icio.us has not publicized their algorithm for choosing which tags to display in the interface; however, we implemented a simple approximation in our model. We proposed that the tagging system could simply recommend the N most popular tags for that web page. Then users could randomly choose among those N tags. The model first creates 20 'seeders' in the same way it did for the *Imitation-Urn* strategy. All of the remaining users choose randomly among the $N = 5$ most popular tags at that point. If they need to apply more than 5 tags, then the remaining tags are chosen randomly from the base powerlaw distribution.

Imitation-Random: As a counterpoint to the flavors of imitation described above, we tested one final strategy. Rather than choosing randomly from the 5 most popular tags as in the *Imitation-Popular* strategy, users choose uniformly from among all tags previously used to that point (after the first 20 users have chosen tags according to Zipf's law).

Computer Model Results

One of the benefits of computer modeling was that the development process forced us to be very explicit about what information users would need to follow a hypothesized strategy. Golder and Huberman [8] suggested that the powerlaw distribution of tags for a given web page could arise from path-dependent choices. When trying to replicate these decisions for our simulation, we found that this only works if a user chooses tags from the empirical distribution *at the time of decision*. This is a very high information requirement for users; they must know the exact proportions of existing tags to choose appropriately. Del.icio.us does not present this information in its interface; however we assume this knowledge for our simulations with the *Imitation-Urn* strategy so that it models a truly path-dependent process.

Table 3. Measures of distributional equivalence and numerical identity.

	Mean KS		*St.dev. KS*		*Mean IUA*		*St.dev. IUA*	
Real World	0.069		0.026		0.144		0.057	
Zipf	0.080	*	0.011	++	0.374	***	0.074	+++
Organizing	0.084	***	0.029		0.182	***	0.052	
Im-Urn	0.139	***	0.067	+++	0.184	***	0.056	
Im-Popular	0.223	***	0.149	+++	0.317	***	0.088	+++
Im-Random	0.386	***	0.063	+++	0.070	***	0.042	+++

Wilcoxon test signif. codes: '***' 0.001 '**' 0.01 '*' 0.05
Levene test signif. codes: '+++' 0.001 '++' 0.01 '+' 0.05

Distributional Equivalence Measure

For each simulated web page, we fit the tag distribution produced by the model to a discrete powerlaw distribution using maximum likelihood. We then conducted a Kolmogorov-Smirnov (KS) goodness-of-fit test to see how well the simulated distribution fit a powerlaw. A KS statistic ranges from 0 to 1; 0 means that the distribution is identical to a powerlaw, and higher numbers indicate greater deviation from a powerlaw (0.22 is a bad fit). The second and third columns (KS) of Table 3 show the mean and standard deviation of the KS statistic for each strategy. We used a Wilcoxon matched-pairs rank-sum test with Bonferroni correction to compare the set of KS statistics for each tag choice strategy (one for each simulated web page) against the set of KS statistics for the real world web pages, and found all comparisons to be significant, likely due to our large sample size for both the strategies modeled and our real-world sample. Therefore, it is more instructive in this case to consider practical, rather than statistical significance when interpreting the results of our analyses. In fact, in light of our large sample sizes we can interpret the significance of these results to mean that we can be confident the pattern of results we observed is unlikely to have occurred by chance. We can then focus on the actual differences observed, which for some strategies were large and for others were very small.

The mean KS statistic for the *Imitation-Popular* and *Imitation-Random* strategies indicate that data generated in this way do not fit a powerlaw very well. Figure 2 illustrates the tag distribution (on a log-log plot) for all five strategies on one simulated web page, along with the paired real world distribution. The straight line on each plot represents the theoretical powerlaw distribution that best fit the data. The non-powerlaw nature of the three *Imitation-** strategies is noticeable compared to the nearly linear plots for the *Organizing* and *Zipf* strategies, as well as the *Real World* data. However, the distributions based on the *Zipf* and *Organizing* strategies fit as well as the real data from del.icio.us.

The mean KS statistic for the *Imitation-Urn* strategy is closer to that of the *Real World*, *Zipf*, and *Organizing* tag distributions than the other two *Imitation* strategies' distributions; however, the standard deviation of the *Imitation-Urn* strategy was significantly different from that of the *Real World* data using the Levene homogeneity of variance test. Figure 3 shows the density plot of KS statistics for each strategy; the narrow distributions clustered to the left are visibly different from the wide distributions produced by the *Imitation* strate-

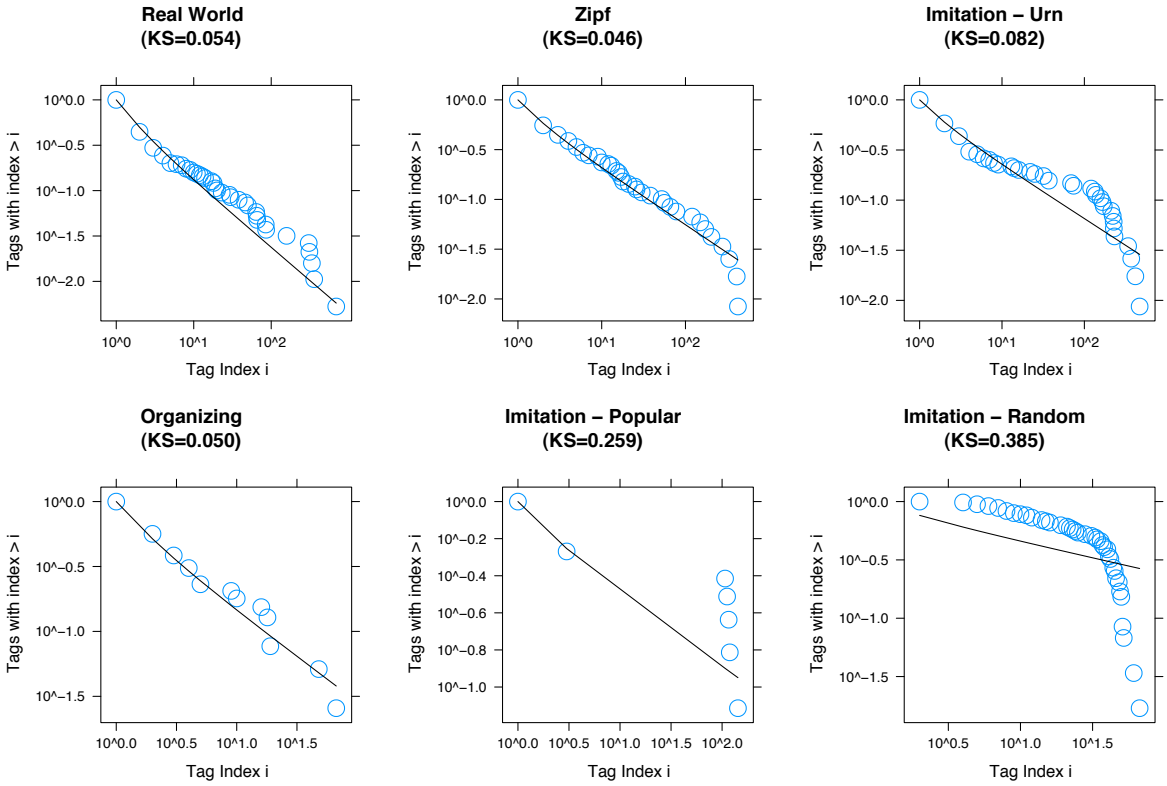

Figure 2. Tag frequency distributions for the real world data and strategies implemented in the computer model on a log-log scale.

gies. We therefore conclude from our distributional equivalence measure that we can rule out the three *Imitation-** strategies as plausible processes that might give rise to the distributional pattern we saw in our sample from del.icio.us.

Numerical Identity Measure

We calculated the average inter-user agreement between simulated users of each modeled web page, for each strategy. Table 3 also provides the inter-user agreement means and standard deviations for each tag choice strategy, and for the real-world data from del.icio.us. IUA ranges from 0 to 1 and represents how often two random users chose the same tag; higher numbers mean greater agreement. We again found that Wilcoxon matched-pairs rank-sum tests with Bonferroni correction were significant for pairwise comparisons between the *Real World* data and all tag choice strategies. Inter-user agreement was much higher for the *Zipf* and *Imitate-Popular* strategies than observed in the sample data from del.icio.us. The *Organizing* and *Imitation-Urn* strategies are negligibly different in terms of practical significance, as they are well within one standard deviation of the mean of the *Real World* data. They are also the only tag choice strategies for which the Levene test for homogeneity of variance was not significant when compared with the *Real World* data. From our numerical identity measure we can therefore rule out rule out the following strategies as plausible processes that might produce inter-user agreement values we saw in our sample from del.icio.us: *Zipf*, *Imitation-Popular*, and *Imitation-Random*.

Interpretation

Based on the two measures described above, we can make the following determinations about the plausibility of each tag choice strategy producing data like that in our sample downloaded from del.icio.us:

1. *Zipf* — rule out based on numerical identity
2. *Organizing* — cannot rule out
3. *Imitation-Urn* — rule out based on distributional equivalence
4. *Imitation-Popular* — rule out based on both numerical identity and distributional equivalence
5. *Imitation-Random* — rule out based on both numerical identity and distributional equivalence

SUMMARY AND IMPLICATIONS

The two studies reported in this paper focus on detecting patterns in tag choices on del.icio.us. We used two different methods, logistic regression performed on sample data collected from del.icio.us and computer modeling of tag choice strategies, to examine competing hypotheses describing processes that might produce observed tag choice patterns.

Our logistic regression showed that users' past tag choices had a large influence on future tag choices, while the fact that a tag had been used before on a web page had very little influence. In addition, we were able to rule out all tag

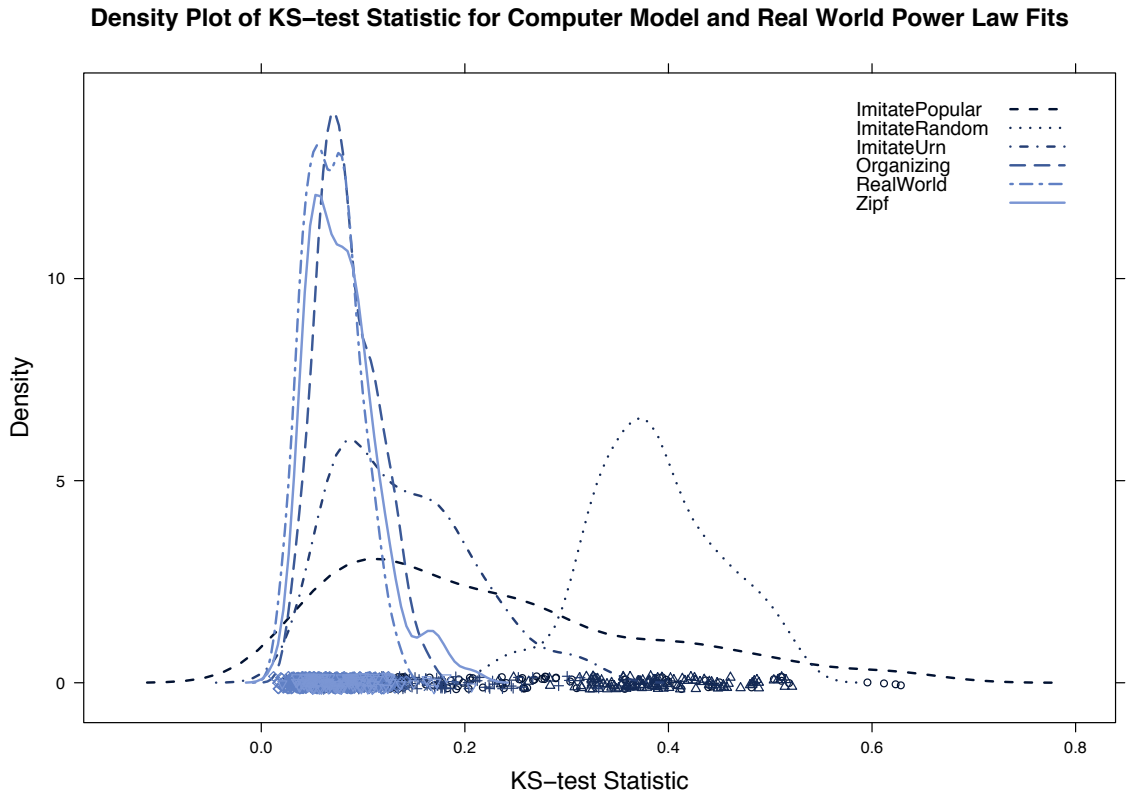

Figure 3. Shows the frequency distribution shape of KS statistics for real world data and modeled tag choice strategies.

choice strategies implemented in our computer model, except for the *Organizing* strategy. In other words, our results indicate that the most plausible hypothesis among those we tested is that tag selection in del.icio.us is governed by individual, idiosyncratic processes rather than a form of direct imitation. These results contradict both the hypotheses presented in Golder and Huberman [8] and the results of Sen et al. [23], and suggest that the potential for emergence of a socially-constructed vocabulary on del.icio.us due to tag imitation is unlikely.

We believe ours is the first quantitative study of how users of del.icio.us choose tags to compare competing hypotheses from the literature. Our logistic regression allows us to control for sources of variability that the cosine similarity measures used by Sen et al. [23] do not. Also, our emphasis on exact tag choice rather than tag class means we are able to consider how the processes shaping the tag vocabulary on del.icio.us might affect its utility as a tool for personal and shared information management. Del.icio.us users do not navigate by tag classes; specific words and the multiple meanings associated with them are important for finding and re-finding. It might also be that tagging is just different on del.icio.us and MovieLens. Del.icio.us has a strong information management component (storing and organizing bookmarks), while it is less clear for what purpose tags might be chosen or used on MovieLens. Finally, our computer model allows us to assume different strategies and look at the tag choice patterns they produce, rather than identifying patterns and speculating on what might have caused them, as Golder and Huberman reported in [8].

Our results suggest that users choose tags for personal information management rather than according to a shared vocabulary; it is possible that the diversity of contexts in which the same terms are applied as tags results in more variability in the content returned when a user searches with that tag. If this is true, users might not find tags to be useful for finding and re-finding. A study by Millen et al. [16] of the Dogear system logs hints that this might be the case: they counted more events associated with keyword search than navigating by personal or shared tags in their sample. As tagging is incorporated into more and more tools for information management (Tang, et al. [24] for example), it is increasingly important to understand how users choose tags, and the implications these choices have for how the system is used.

LIMITATIONS AND FUTURE WORK

It is important to note that through this research we are only able to **rule out** competing hypotheses. Data downloaded from del.icio.us are evidence which may be used to detect potential tagging strategies, but we cannot make assumptions about what information users may have seen and acted upon, or infer what a given user was thinking when making tag choices. Therefore, we are not able to say with absolute certainty that users choose tags according to their own per-

sonal organization scheme; nor can we determine whether popular tags are chosen more often due to imitation, because they are topically relevant, or for some as-yet unknown reason. The major weakness of the methods we have used that we cannot make any claims about users' perceptions, goals, or motivations that might shed more light on tagging strategies. In reality, the same tag choice strategy might not be used by all users, or even apply to all the tag choices of an individual user.

We also lack empirical evidence regarding the usefulness of tags for organizing, finding and re-finding personal and shared information. These limitations leave ample opportunity for future work in this area, including field studies of tagging behavior, measurement of the effectiveness of tags for information management, and experiments which will allow us to infer causal relationships between factors affecting tag choices and characteristics of the resulting tag distributions. The research described in this paper leads us to focus our efforts on more rigorous investigations of the predictions of the *Organizing* hypothesis when tags are used for personal and shared information management.

ACKNOWLEDGMENTS

We would like to thank CSCAR for providing advice on our logistic regression model, and Judy Olson, Lian Jian, Bethany Rader, members of the BlearyTheory lab group, and iConference reviewers for comments on earlier drafts. This material is based upon work supported by the National Science Foundation under Grant No. CNS 0716196.

REFERENCES

1. A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, second edition, 2007.
2. M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *CHI '07*, 971–980, 2007.
3. R. Axtell, R. Axelrod, J. M. Epstein, and M. D. Cohen. Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory*, 1(2):123–141, 1996.
4. A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. http://arxiv.org/abs/0706.1062v1, June 2007.
5. E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with phlat. In *CHI '06*, 261–270, 2006.
6. U. Farooq, T. G. Kannampallil, Y. Song, C. H. Ganoe, J. M. Carroll, and C. L. Giles. Evaluating tagging behavior in social bookmarking systems: Metrics and design heuristics. In *GROUP '07*, 351–360, 2007.
7. G. Furnas, T. Landauer, L. Gomez, and S. Dumais. Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806, 1983.
8. S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
9. J. Grudin. Enterprise knowledge management and emerging technologies. In *HICSS '06*, 2006.
10. H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07*, 2007.
11. G. G. Judge, W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T.-C. Lee. *The Theory and Practice of Econometrics, Ch. 22*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition, 1985.
12. M. Lansdale. The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66, 1988.
13. L. M. Markus. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *J. of MIS*, 18(1):57 – 93, 2001.
14. C. Marlow, M. Naaman, d. boyd, and M. Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *WWW '06 Collaborative Tagging Workshop*, 2006.
15. S. Menard. *Applied Logistic Regression Analysis*. Quantitative Applications in the Social Sciences. Sage University Press, 2002.
16. D. Millen, M. Yang, S. Whittaker, and J. Feinberg. Social bookmarking and exploratory search. In *ECSCW '07*, 21–40, 2007.
17. D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *CHI '06*, 2006.
18. N. Nan, E. W. Johnston, J. S. Olson, and N. Bos. Beyond being in the lab: using multi-agent modeling to isolate competing hypotheses. In *CHI '05*, 1693–1696, 2005.
19. M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.
20. S. Page. Path dependence. *Quarterly Journal of Political Science*, 1(87-115), 2006.
21. P. Pirolli. Rational analyses of information foraging on the web. *Cognitive Science*, 29(3):343–373, 2005.
22. S. Sen, F. M. Harper, A. LaPitz, and J. Riedl. The quest for quality tags. In *GROUP '07*, 2007.
23. S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *CSCW '06*, 181–190, 2006.
24. J. C. Tang, E. Wilcox, J. A. Cerruti, H. Badenes, S. Nusser, and J. Schoudt. Tag-it, snag-it, or bag-it: combining tags, threads, and folders in e-mail. In *CHI '08*, 2179–2194, 2008.
25. J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04*, 415–422, 2004.
26. R. Wash and E. Rader. Public bookmarks and private benefits: An analysis of incentives in social computing. In *ASIS&T '07*, 2007.