Just a Click Away: Social Search and Metadata in Predicting File Discovery

N. Sadat Shami, Michael Muller, David Millen

Center for Social Software, IBM TJ Watson Research Center
1 Rogers St., Cambridge, MA 02142, USA
{ sadat, michael_muller, david_r_millen } @ us.ibm.com

Abstract

Social search has been claimed to improve content discovery by allowing users to draw on their social network to find relevant content. Thus social network information, complemented with metadata, can enhance the search for new information. We examine the relative contribution of social network information and file metadata in predicting downloads of files by analyzing the file browsing behavior of 5,723 users of a social file sharing service in a large global company. The following factors increase the likelihood of download: (a) if the file author is in the user's social network; (b) if the file has been downloaded by or shared with others in the user's network; and (c) if the file's metadata is a good match to the user's interests. Semi-structured interviews with 18 users provided deeper insight into why these factors are important. Our findings demonstrate the value of the social graph for finding files, with implications for relevant content and people discovery.

Introduction

Finding relevant files in file sharing services has been shown to be difficult (e.g. Jensen et al. 2010; Rader 2009; Voida and Greenberg 2009). Users typically rely on keyword searching or manually navigating among folder hierarchies to find files (Teevan et al. 2004). With the advent of social software, we now have a richer set of features such as social tags, user profiles and collections of files, and a stream of recent events that allow us to discover relevant files through exploration (Shami, Muller and Millen 2011). We refer to social feature enhanced file sharing services as social file sharing services. Examples of such services include slideshare.net and scribd.com.

Social search is an umbrella term to describe search acts that engage social interactions or utilize information from social sources (Chi 2009; Evans and Chi 2008). There has been recent interest in understanding how social search enhances information seeking. Social data is one of many 'signals' used in ranking search results, with some calling it a 'tiny signal' (Stone 2011). Is this indeed the case in file discovery? The main goal of this study is to understand how different social, textual and usage metadata of a file influences the likelihood of file download. When a user arrives at a file page that contains all this metadata and she is "just a click away" from downloading the file, what factors will cause her to move from discovery to download? Using the browsing data of 5,723 users of an enterprise social file sharing service over an 8-month period, we determine the relative weight of factors predicting download. The rest of the paper is organized as follows. First we survey related work and derive hypotheses from it. We then describe the method we used to test our hypotheses and present our results from statistical analysis as well as interviews with users. Finally, we conclude with a discussion and implications for design.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

File Discovery in Social File Sharing Services

The Social Life of Files

Research has looked at how files are discovered through automated methods (e.g. Hardy and Schwartz 1993) or in peer-to-peer networks (e.g. Portmann et al. 2001). Such aspects are out of the scope of this paper. Muller et al. (2010) looked at an enterprise file sharing system and found four patterns of use: a) upload and publicize, b) annotate and watch, c) discover and tell, and d) refind. Muller et al. (2009) also show the importance of grouped files or collections, and the role of curators in maintaining such collections. Additionally, researchers have looked at the design of file sharing systems for improved awareness (Whalen, Toms and Blustein 2008), where file sharing breakdowns occur (Voida et al. 2006), and identifying 'common' files to make storage more efficient and to connect like-minded employees (Tang et al. 2007).

Popular online services such as YouTube and Flickr allow users to share and find videos and pictures. Despite the large audience of such systems, there has been little

systematic research into how users decide which file(s) to download. In this paper, we contribute to that literature by exploring the relative influence of social network information, combined with textual and usage metadata, in the decision to download files. In the sections below, we describe relevant research that allows us to formulate hypotheses regarding how different factors may influence this process.

Social Search

Several researchers have demonstrated the strong influence of social information on individual behavior. Chi (2009) describes two forms of social search systems: social answering systems and social feedback systems. Social answering systems leverage proximity in social network connections to seek out relevant experts for question answering. Social feedback systems use social attention data such as clicks to rank search results or information items. Kammerer et al. (2009) found that the MrTaggy social feedback system, which provides relevance feedback for query refinement, allowed users to successfully perform exploratory search. Shami et al. (2008) found that social network information found in the snippets of search results of an expertise finder predicted whether a user would click on that result to find further information. Shami et al. (2009) also found that users are more likely to contact those that are closer to them in social connectedness. Muller et al. (2009) reported that users' actions related to a file such as recommending, aggregating, and commenting increased the likelihood of downloading. Social networks can thus act as important signals of authenticity and credibility of the file content (Shami et al. 2009). If users recognize that more people from within their social network have acted on a file, that may increase the likelihood that the users themselves will find the file relevant. Based on these findings, we can arrive at the following hypotheses:

H1: *The likelihood of downloading a file increases if the author of the file is in your social network.*

H2a: *The likelihood of downloading a file increases as it is shared with more people in your social network.*

H2b: *The likelihood of downloading a file increases as it is downloaded by more people in your social network.*

File Metadata

Semantic relatedness can act as an important signal of relevance. Matching keywords found on a user's profile with requests for blog posts on a topic was shown to be an effective way to get users to write blog posts on that topic (Dugan, Geyer and Millen 2010). This shows how textual metadata can be used to match users with content.

Rader (2009, 2010) investigated the effect of file labeling, organizing and audience awareness by producers of files on how easily those files can by found by readers. It was easier for readers to find files when producers imagine their audience to be similar to them. While these studies were conducted using a hierarchical file repository, they may have implications for social file sharing services, which have less of a hierarchical structure but more hyperlinked access points for discovery. In particular, textual metadata of a file such as its title, description and tags may predict downloading of files by users, as long as there is semantic similarity with their own interests.

H3: *The likelihood of downloading a file increases the more the textual metadata of the file is similar to the user's interests.*

Research has found a 'rich get richer' effect with internet content. The more attention the content receives, the more popular it becomes. A reason behind this 'rich get richer' phenomenon is provided through the ideas of the 'self fulfilling prophecy' (Merton 1968) and conformity (Asch 1955). These ideas are sometimes used to explain how we think we think, based on what others are doing. Does someone like something because it's good, or since everyone around him likes it, it must be good? In an online experiment about rating and downloading music, Salganik et al. found that participants in the experimental condition were significantly influenced by what others liked and rated highly compared to controls (Salganik, Dodds and Watts 2006). These findings could be applied to file download behavior. The more popular a file is, as represented through being highly downloaded, collected, shared, and commented on, the more likely it is to be downloaded. This leads us to our final hypothesis.

H4: *The likelihood of download increases the more a file is downloaded, collected, shared, or commented on.*

Method

In order to test these hypotheses, we analyzed the usage logs of a social file sharing service called Cattail that was deployed within a large multinational company. We also conducted interviews with Cattail users so we could triangulate our statistical analysis with qualitative data about motivation behind usage. This is not an evaluation of Cattail. Rather, it was a convenient platform to test our hypotheses. The data described here is from the 8 month period starting December 1, 2008 through July 31, 2009.

System Description

Cattail was designed to support social file sharing among workers in an enterprise. A detailed description of Cattail can be found in (Shami et al. 2011). In the interest of space, we focus on the Cattail UI most relevant to our study, which is the 'file' page. The 'file' page provides a consolidated summary of the file and all actions performed on it. This information includes the file author name, those it was shared with and downloaded by, and textual and usage metadata such as its title, tag(s), description, and the number of times it was downloaded, added to named collection(s) of files, and commented on. Figure 1 displays how all this information is organized on the 'file' page. From the file page, users can decide whether to download the file or not.

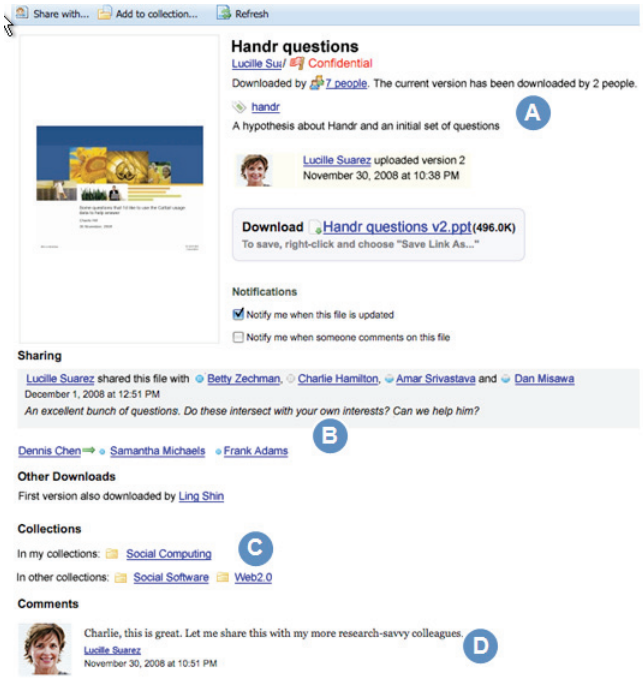

Figure 1. A 'file' page. Names changed to protect privacy. (A) Various file metadata such as file title, tag(s), description, and the number of times it has been downloaded. (B) Social network data: List of people the file has been shared with and downloaded by. Full blue circles indicate the file has been downloaded, half blue circles indicate a previous version has been downloaded, grey circles indicate the file has been shared, but not downloaded. (C) List of collections the file is located in. (D) Comments the file has received from other users.

User Population

During the 8 month study period 154,488 users logged into Cattail at least once. Since Cattail is a social file sharing service, users can easily share files with others. When a user shares a file with others, an email is sent to the share recipient(s). In this paper, we are interested in the discovery of files. Therefore, we focus on users who have downloaded files that were not shared with them. Since they would have no prior knowledge of the file's existence, they would need to discover it. Of the Cattail users, 58,034 downloaded at least one file that was not shared with them. In order to have a manageable dataset to perform inferential statistics, we reduced the study population to those who have downloaded at least 10 files that were not shared with them. This left us with a sample of 5,723 users. As with most social services, the distribution of users who had downloaded files that were not shared with them followed a power law distribution. Focusing on users who had downloaded at least 10 files not explicitly shared with them allowed us to analyze the browsing behavior of the more active users, as opposed to those that only downloaded a few files.

We also conducted 18 semi-structured interviews with Cattail users in order to gain better insight into their behavior. In deciding whom to invite for interviews we took into account the geographic location, business unit and gender of interviewees. We sent invitations to 20 employees. The 18 that accepted our interview requests had an average of 17.03 years of full time work experience (*SD*=9.1, *Min*=2, *Max*=32), were from 9 different countries and 6 different business units, and 7 were female. They had diverse job titles including strategy consultant, HR professional, application architect, learning consultant, sales specialist, and systems engineering professional. Interview questions focused on usage of Cattail, with special emphasis on how interviewees went about downloading files. Interviews were conducted over the phone and recorded with the permission of interviewees, and transcribed. A single researcher performed analysis of interview data by coding transcripts using a Grounded Theory (Strauss and Corbin 1998) approach.

File Browsing Model

Figure 2 shows the six different navigation pathways users can follow to discover a file, as well as how much they were used during our study period. After discovering a file, they evaluate it and decide whether to download it or not.

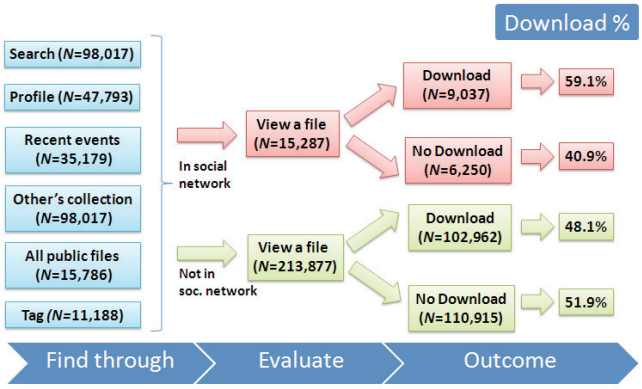

Figure 2. File browsing model. *N* = Total number of events in that category.

Measures

Dependent Variable

Our dependent measure was a dichotomous variable measuring whether or not a user downloaded a file after discovering it (i.e., after viewing its 'file' page). If the user downloaded the file from that page, this variable was given a value of 1, otherwise it was 0.

Independent Variables

Social Navigation Choice: This variable captures whether the user followed a navigation pathway based on someone

in their network. Since the file pathways in Figure 2 contain information about the file author, we can determine if the file author is in the user's social network. In order to determine a user's social network, we utilized an API for gathering social network data that was available within the company. The API aggregates data from several public sources available within the company such as blogs, social tags and bookmarks, people tags, organizational chart etc. in order to determine social connections. The method of calculating social connections is described in Guy et al. (2008) and Ronen et al. (2009). It is noteworthy that Cattail does not provide any feedback to users about who is in their social network. This allowed us to examine how social network information acts as an *implicit* filter in file download decisions.

The API was used to generate the social network of each of our 5,723 users. It was able to provide us with social network information for 98% of them. The mean network size was 124.84 (*SD*=20.41). The social navigation variable was thus a dichotomous variable coded as 1 if the file author was in the social network of the user viewing the file, and 0 if not.

Social network overlap: As shown in Figure 1, Cattail displays the list of people a file was shared with and downloaded by. Using these lists, we can determine the overlap between the user's social network and those that the file was shared with and downloaded by. We calculate this using the Overlap Co-efficient – by dividing the users in common by the smaller of the number of users in each, as described in (Chapman), and used in (Muller 2007). The Overlap Co-efficient results in a number between 0 and 1. We have one variable for the overlap between a user's social network and those the file was shared with, and another one for the overlap between a user's social network and those that have downloaded the file.

Textual metadata: Textual metadata captures the similarity of a file's textual metadata with the user's interests. Figure 1 shows that a file has three forms of textual metadata: title, description, and tag(s). We treat these as three separate variables. All three of these were converted to term vectors by removing duplicates and common stop words.

It is not possible for us to precisely know each of our users' interests. However, we can utilize 'people tags' from the same internal company API as an approximation of interests. Similarly to the way it calculates a person's social network, the API assigns 'people tags' to users. These tags are based on those used by others to tag the user, and tags the user used to tag their own files and bookmarks. While not perfect, we felt these 'people tags' associated with a user is a reasonable approximation of their interests.

We again use the Overlap Co-efficient to determine the similarity between a file's textual metadata, and a user's interests, as expressed through the combination of tags they have authored and tags associated with them. This gives us three variables: title overlap, description overlap, and tag overlap, all between 0 and 1.

Usage metadata: As shown in Figure 1, a file includes information about how many times it was downloaded, shared with others, added to collections, and comments received. These are the usage metadata of a file and reflect the activity on the file when it was viewed by a user. Additionally, we included a dichotomous variable that was 1 if a file had been added to a collection and 0 if it had not.

Control Variable

As a control measure, we entered into our model the total number of times each user had downloaded a file. On average, our participants downloaded 639.01 files (*SD*=2525.23). Table 1 shows descriptive statistics for all the variables in the model.

		Mean (SD)
Dependent variable		
	Downloaded or not [a]	51% downloaded
Independent variables		
Social navigation	File author in social network or not [a]	6.7% in social network
Social network overlap	Share overlap [b]	0.12 (0.26)
	Download overlap [b]	0.06 (0.15)
Textual metadata	Title overlap [b]	0.14 (0.19)
	Tag overlap [b]	0.18 (0.31)
	Description overlap [b]	0.07 (0.12)
Usage metadata	Download count of file [c]	48.83 (137.35)
	Num. of times file shared [c]	6.08 (32.7)
	Num. of collections file is in [c]	2.17 (4.39)
	File is in a collection or not [a]	60.4% files in a collection
	Num. of comments on file [c]	0.69 (3.38)
Control variable		
	Total files downloaded by each participant [c]	639.01 (2525.23)

[a.] dichotomous variable, [b.] values between 0 and 1, [c.] continuous variable

Table 1. Descriptive statistics. *N* = 229,164 observations

Results

One of the major assumptions of many parametric statistical tests is independence of observations. Because our users contribute multiple observations each, these observations are theoretically inter-correlated and therefore not independent of each other. When using a regression framework, this within-cluster correlation typically leads to underestimated, or biased, standard errors and increased likelihood of Type I errors. General Estimating Equations (GEE) – an extension of generalized linear models – was thus used to account for correlated repeated measurements within individuals, resulting in unbiased regression parameters (Liang and Zeger 1986). Furthermore, GEE models are robust to violations of normality and homogeneity of variance (Garson).

GEE was run with a random factor representing intra-individual correlation per user and without sensitivity to the entry order of predictors. It was ensured that predictors produced variance inflation factors <3 to mitigate against multicollinearity. Because of the large sample size, the study alpha was set to 0.01 to reduce the risk of Type 1 errors. Table 2 summarizes our analysis.

	β (SE)	95% CI for Exp β		
		Lower	Exp β	Upper
Intercept	- 1.04 *** (0.05)	0.32	0.35	0.39
Social navigation choice				
File author in social network or not	0.65*** (0.05)	1.75	1.9	2.1
Social network overlap				
Share overlap	0.76*** (0.1)	1.75	2.13	2.6
Download overlap	1.41*** (0.17)	2.95	4.1	5.75
Textual metadata				
Title overlap	0.49*** (0.11)	1.31	1.65	2.06
Tag overlap	1.17*** (0.08)	2.77	3.24	3.79
Description overlap	0.55** (0.19)	1.19	1.73	2.5
Usage metadata				
Num. of collections file is in	- 0.01*** (0.003)	0.98	0.98	0.99
File is in a collection or not	0.21*** (0.03)	1.15	1.23	1.31
Control variable				
Total files downloaded by each participant	5.4E-5*** (4.2E-6)	1.0	1.0	1.0

Table 2. GEE predicting whether a user will download a file. Only significant predictors shown.
Note: *N*=229,164 observations, ***$p < 0.001$, **$p < 0.01$

Social Navigation Choice

The odds of downloading a file increase roughly 2 times when the file author is in the social network of the user ($Exp\ \beta = 1.9, p < 0.001$). H1 is thus supported.

Our interviewees mentioned that being able to recognize authors of files influenced them to download.

"Depending on who wrote the document increases the likelihood that it would be relevant." [S3, Application Architect, Male, USA]

Social Network Overlap

The odds of downloading a file increase 2.13 times as the overlap between the user's social network and those that the file was shared with increases ($Exp\ \beta = 2.13$, $p < 0.001$). H2a was thus supported.

One might assume that if a user follows the path where the file author is in her social network, the likelihood of having social network overlap with those that the file was shared with and downloaded by might also be high. In order to test this, we ran a GEE analysis with only the observations where users followed a path where the file author was not in their social network ($N = 213{,}877$). In the interest of space, suffice it to say that share overlap and download overlap were still statistically significant. In fact, their odds ratios were higher than that reported in Table 2.

The list of people the file had been shared with allowed users to do a cursory check to determine if the file had been shared with people they knew. If it was shared with people they knew that were working in a related area, users thought the file would be relevant to them as well.

"I wouldn't look around too much except to note if it's a big [share] list or a short list, and if there was somebody in particular I knew on the list." [S8, Learning Consultant, Male, USA]

The odds of downloading a file increase 4.1 times as the overlap between a user's social network and those that the file was downloaded by increases ($Exp\ \beta = 4.1, p < 0.001$). H2b was thus supported. Download overlap is the strongest predictor among the independent variables. This was confirmed by quotes from interviewees who spoke about the strong signal of knowing someone you know has downloaded a file is.

"If you know that there are some experts for a product or project and you see that these experts have downloaded some files, and you work in the same area, you normally can be sure that this can be of interest to you as well." [S18, Expertise Management Specialist, Female, USA]

Download by a trusted other conveyed trust for some:

"Again if I see a file that has been downloaded by people that I think that I trust or I know are good, then I may be more likely to download it… because if they think it is worth [downloading] that makes a lot of difference in whether I'll go into that file or not." [S12, Learning Consultant, Male, New Zealand]

Textual metadata

The odds of downloading a file increase 1.65 times as the overlap between the title of the file and the tags associated with a user increase ($Exp\ \beta = 1.65, p < 0.001$). Similarly, the odds increase 3.24 times as the overlap between the tags associated with the file and the tags associated with the user increase ($Exp\ \beta = 3.24, p < 0.001$). For the third textual metadata variable, we see a similar result. The odds of downloading a file increase 1.73 times as the overlap between the description of the file and the tags associated with the user increase ($Exp\ \beta = 1.73, p < 0.01$). H3 is thus supported. Regarding the value of tagging, one participant mentioned:

"The notion of tagging is important for me… when I find the document on Cattail, I always try to read the tags potentially to see if I can find something else on the document that I didn't find for example. For example, if I tag the name 'collaboration' on Cattail to find documents relevant for that topic, if I find a document with the tag 'collaboration' and for example 'communication' I have additional information regarding which type of

information I will find in that document." [S2, Client Technical Specialist, Male, France]

Another participants mentioned the value of having a file description:

"The only metadata I normally use is the description of the file. Because that gives me the most accurate overview if there's a value for me in the file or not." [S16, Application Architect, Male, Ireland]

Usage Metadata

The usage metadata provided the most surprising results in our study. The odds of downloading a file were found to decrease 0.9 times the more collections a file is in (*Exp* β = 0.9, $p < 0.001$). However, the odds of downloading a file increase 1.23 times as long as it is in a single collection (*Exp* β = 1.23, $p < 0.001$). This indicates that being in a collection shows diminishing returns. Another interpretation is that the closer the odds ratio of a predictor is to 1.0, the more it is independent of the dependent variable, with 1.0 representing full statistical independence. We can thus interpret the *Exp* β of 0.9 for the number of collections a file is in as having little effect on the odds of download, even though it is significant.

Surprisingly, the number of times a file was downloaded was not a significant predictor of file download (*Exp* β = 1.00, $p = 0.07$). Neither was the number of times a file was shared (*Exp* β = 1.00, $p = 0.021$), and the number of comments it received (*Exp* β = 1.01, $p = 0.017$). H4 was thus partially supported.

The following quote from a participant may explain the non-significance of the download count of a file.

"I would say that if there had been a lot of downloads, I almost always took that as an indication that it was pretty good content. However, I usually did not have the opposite… like I said, that if it had never been downloaded, I just assumed nobody had discovered it. Or maybe I was one of the first to find it or whatever. The lack of downloads wasn't always a negative." [S3, Application Architect, Male, USA]

Table 3 provides a summary of the hypotheses supported and not supported.

Hypothesis	Supported?
H1: *The likelihood of downloading a file increases when the author of the file is in your social network.*	Yes
H2a: *The likelihood of downloading a file increases as it is shared with more people in your social network.*	Yes
H2b: *The likelihood of downloading a file increases as it is downloaded by more people in your social network.*	Yes
H3: *The likelihood of downloading a file increases the more the textual metadata of the file is similar to the user's interests.*	Yes
H4: *The likelihood of download increases the more a file is downloaded, collected, shared, or commented on.*	Partially

Table 3. Study hypotheses

Discussion

The data from our study confirms three out of four of our hypotheses. Following their social network leads users to download files more compared to not relying on such information. It is interesting that the majority (93.33%) of file views occur through users not following navigation paths where the file author is in their social network. However, when they do follow such paths, it leads to significantly more downloads. Perhaps some of the file views from not following known authors represent users looking to expand their network by viewing files of unknown others; Millen et al. (2007) reported a similar process of discovering new colleagues with social tagging. Alternatively, the high proportion of search events may indicate that users viewed many files that were not relevant but were nonetheless available in the search results.

Having overlap between one's social network and the people that downloaded a file was the strongest predictor of file download. Even though Cattail does not show who is in a user's social network, users downloaded files based on an inspection of the list of those they knew had downloaded the file. This was backed up by comments from our interviewees. The list of file downloaders was intended to help file authors get awareness of activity on their files, but it was also useful to those viewing the files. In fact, textual metadata was intended to help file viewers determine relevance, but the strongest predictor was downloads by others in a user's network. Cattail thus expands the search space, but users use their social network to validate and filter content. In this paper we did not tease out the nature of the social relationship. Future work will involve disentangling different types of social connections (e.g. same job, same business unit) on users' file downloading behavior.

Rader (2009, 2010) demonstrated the value of audience awareness in hierarchical file systems. We extend those findings and show that the same pattern may be prevalent in social file sharing. We were surprised that download count and the number of comments a file received was not significant in our data. This could be due to several reasons. First, our data is based on when a user viewed a file in a particular point in time. When the user viewed the file, it may not have had many downloads or comments (comments are relatively rare, occurring on only 7.3% of all files). Second, the interface of Cattail did not make file's popularity salient, as can be seen in Figure 1. File lists were ordered by default by recency, although users did have the option of sorting by download count. Finally, we note that the multinational corporation in which Cattail was deployed pursues many diverse lines of research, manufacturing, software development, sales, and consulting. Files that are popular with unknown other users (i.e., in a generalized popularity ranking) may not reflect the interests of any particular user.

Limitations

As with any study, our findings suffer from some limitations. The use of an API that infers social connections, as opposed to relying on articulated contacts

likely introduced some noise in the data. Nonetheless, the social network overlap variables were highly significant.

Another potential limitation is using download as a dependent variable. It is difficult to tell whether users found the file relevant after they had a chance to open it. However, we feel that this is a better proxy than using file views as it involves slightly more investment of energy. Additionally, in commercial systems, the crucial event – i.e., the purchase event – is the download. So file providers like Amazon or Apple iTunes would want to maximize such activity.

Although we tried our best to obtain a diverse sample of participants for qualitative interviews, it is simply not possible to characterize the motivations of such a large number of users through interviews with so few participants. Findings from our interviews may thus not be representative of the larger population of Cattail users.

Finally, this study was conducted in a large global organization where employees were comfortable with using various forms of social software. The organizational culture of the company likely influenced our results. We hope future studies will replicate and extend our findings in other settings. However, understanding what works in large organizations such as companies, academic departments, governments, NGOs etc. is important. Enterprise phenomena can also be compared with file-sharing in other large-scale, non-enterprise settings for insights about work-oriented vs. recreational file-sharing (e.g., Napster), for centralized vs. distributed file-sharing, and for file-sharing within a trusted environment.

Implications For Design

Several important design implications arise from our findings. Since social network overlap with the list of people that have downloaded a file was the strongest predictor, a recommender system could suggest files to users based on what others in their social network have downloaded. It may be useful to replace the system-wide popularity rankings with a more limited popularity measure based on each user's own social network. If this were done as an anonymous aggregate listing of popularity, it would not violate privacy standards – i.e., the reading behavior of individual users would remain private, even though their aggregate reading behavior could provide a useful index to users in their social networks.

Our results support the idea of combining people tags with social network information to suggest relevant new connections to users based on topics. We may find it useful to compute a similarity score between the user and each other person whose information appears on the file-description page. This could facilitate not only file download, but also the discovery of potential new colleagues (i.e., previously unknown people whose profile is calculated to be similar to the user).

In our study we used the tags assigned by users to their own files as one way to obtain people tags. However, many users do not assign tags to files, and tags in one social software service may sometimes have poor overlap with tags in another social software service (Muller 2007). Another way to express the interests of users through tags is to compute term frequency – inverse document frequency (tf-idf) values from the text of the files they have authored. A third way, focusing on a model of the user's current interest, is to compute an index of words in the files that they have downloaded during the same session, and treat the highest-ranking words as a temporary "interest model" for the remainder of the session.

Another potential design feature is the ability to move directly from browsing an individual's file, to engaging in social relations with that user. In the vernacular of the internet, if you can "like" a file and share/recommend it to others, why not also have a one-click method to "friend" its creator? Or another person who recommended it? These speculations extend the range of serendipitous discovery beyond discovering previously unknown content, to discovering previously unknown colleagues as well (Millen, Feinberg and Kerr 2006; Millen et al. 2007; Ronen et al. 2009; Shami et al. 2009).

Conclusion

In this study, we sought to determine the factors that lead to download of files in a social file sharing service. Users are more likely to download a file when the file author is in their social network, and there is overlap between their network and those that the file was shared with and downloaded by. Our study thus adds to the literature on social search by empirically demonstrating the value of social networks in file discovery. It further contributes to the growing literature on audience design by illustrating the importance of proper tags, title and description on file discovery. Our findings can be used to improve relevant file discovery features by combining file content with social graph information in social file sharing services, and by actively "reaching out" from the file-sharing system to take actions in other, related social software services.

Our study clearly demonstrates the value of one's social network in discovering content expected to be relevant. While traditional keyword search may sometimes be adequate, advances in utilizing the social graph has the potential to improve our ability to uncover content and people we might have not otherwise stumbled upon.

References

Asch, S. E. 1955. Opinions and social pressure. *Scientific American* 193: 31-35.

Chapman, S. Sam's string metrics. Retrieved July 20, 2010, from http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

Chi, E. H. 2009. Information Seeking Can Be Social. *Computer* 42(3): 42-46.

Dugan, C., Geyer, W., and Millen, D. R. 2010. *Lessons learned from blog muse: audience-based inspiration for bloggers*. In *Proc. CHI 2010*. ACM Press.

Evans, B. M., and Chi, E. H. 2008. *Towards a model of understanding social search*. In *Proc. CSCW 2008*.

Garson, G. D. Generalized Linear Models and Generalized Estimating Equations. Retrieved September 18, 2010, from http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm

Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., and Farrell, S. 2008. *Harvesting with SONAR: the value of aggregating social network information*. In *Proc. CHI 2008*. ACM Press.

Hardy, D. R., and Schwartz, M. F. 1993. *Essence: A resource discovery system based on semantic file indexing*. In *USENIX Winter Conference 1993*. Citeseer.

Jensen, C., Lonsdale, H., Wynn, E., Cao, J., Slater, M., and Dietterich, T. G. 2010. *The life and times of files and information: a study of desktop provenance*. In *Proc. CHI 2010*. ACM Press.

Kammerer, Y., Nairn, R., Pirolli, P., and Chi, E. H. 2009. *Signpost from the masses: learning effects in an exploratory social tag search browser*. In *Proc. CHI 2009*. ACM Press.

Liang, K. Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. . *Biometrika* 73(1): 13-22.

Merton, R. K. 1968. *Social Theory and Social Structure*. New York: Free Press.

Millen, D. R., Feinberg, J., and Kerr, B. 2006. *Dogear: Social bookmarking in the enterprise*. In *Proc. CHI 2006*. ACM Press.

Millen, D. R., Yang, M., Whittaker, S., and Feinberg, J. 2007. *Social bookmarking and exploratory search*. In *Proc. ECSCW 2007*.

Muller, M. 2007. *Comparing tagging vocabularies among four enterprise tag-based services*. In *Proc. GROUP 2007*. ACM Press.

Muller, M., Millen, D. R., and Feinberg, J. 2009. *Information Curators in an Enterprise File-Sharing Service*. In *Proc. ECSCW 2009*.

Muller, M., Millen, D. R., and Feinberg, J. 2010. *Patterns of usage in an enterprise file-sharing service: publicizing, discovering, and telling the news*. In *Proc. CHI 2010*. ACM Press.

Portmann, M., Sookavatana, P., Ardon, S., and Seneviratne, A. 2001. *The Cost of Peer Discovery and Searching in the Gnutella Peer-to-peer File Sharing Protocol*. In *Proc. ICON 2001*.

Rader, E. 2009. *Yours, mine and (not) ours: social influences on group information repositories*. In *Proc. CHI 2009*. ACM Press.

Rader, E. 2010. *The effect of audience design on labeling, organizing, and finding shared files*. In *Proc. CHI 2010*. ACM Press.

Ronen, I., Shahar, E., Ur, S., Uziel, E., Yogev, S., Zwerdling, N., et al. 2009. *Social networks and discovery in the enterprise (SaND)*. In *Proc. SIGIR 2009*. ACM Press.

Salganik, M. J., Dodds, P. S., and Watts, D. J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311(5762): 854-856.

Shami, N. S., Ehrlich, K., Gay, G., and Hancock, J. T. 2009. *Making sense of strangers' expertise from signals in digital artifacts*. In *Proc. CHI 2009*. ACM Press.

Shami, N. S., Ehrlich, K., and Millen, D. R. 2008. *Pick me! Link selection in expertise search results*. In *Proc. CHI 2008*. ACM Press.

Shami, N. S., Muller, M., and Millen, D. R. 2011. *Browse and Discover: Social File Sharing in the Enterprise*. In *Proc. CSCW 2011*. ACM Press.

Stone, B. 2011. Larry Page's Google 3.0. *BusinessWeek*.

Strauss, A. L., and Corbin, J. M. 1998. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks: Sage Publications.

Tang, J. C., Drews, C., Smith, M., Wu, F., Sue, A., and Lau, T. 2007. *Exploring patterns of social commonality among file directories at work*. In *Proc. CHI 2007*. ACM Press.

Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. 2004. *The perfect search engine is not enough: a study of orienteering behavior in directed search*. In *Proc. CHI 2004*. ACM Press.

Voida, S., Edwards, W. K., Newman, M. W., Grinter, R. E., and Ducheneaut, N. 2006. *Share and share alike: Exploring the user interface affordances of file sharing*. In *Proc. CHI 2006*. ACM Press.

Voida, S., and Greenberg, S. 2009. *WikiFolders: augmenting the display of folders to better convey the meaning of files*. In *Proc. CHI 2009*. ACM Press.

Whalen, T., Toms, E. G., and Blustein, J. 2008. *Information displays for managing shared files*. In *Proc. CHIMIT 2008*. ACM Press.