

# Social Information Foraging and Collaborative Search

Ed H. Chi, Peter Pirolli  
Palo Alto Research Center  
3333 Coyote Hill Road, Palo Alto, CA 94304  
{echi,pirolli}@parc.com

## INTRODUCTION

*There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers - conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.*

-- V. Bush [Bush45]

V. Bush envisioned a device he called the Memex that would allow scholars to forage through personal stores of multimedia documents, and to save traces of paths through content that could then be shared with other scholars as a way of communicating new findings. The Memex was envisioned as a tool that would increase the capacity of individuals to attend to greater spans of emerging knowledge, and would increase the cooperative information sharing that Bush viewed as necessary to improvements in scientific discovery, which he expected to result in increased benefits to society. Bush's vision was not only to improve the foraging ability of the individual user, but to also improve communication and collaboration.

Despite there being a large amount of CSCW research and literature in this area, significant progress still can be made. For example, what has been amazing is the amount of new social mechanisms invented on the Internet for social foraging. To some extent, the Web, blogs, email, internet groups, and other mundane technologies are all aimed at supporting cooperative information sharing and their success implies their effectiveness.

The major question confronting the field, therefore, is to not only understand how these mechanisms are different or the same as previous generation of tools, but also build models of how social foraging appears to be benefiting individuals, and how these models might in turn tell us how to build new collaboration tools.

In this article, we outline several explorations on models of social information foraging, and how some of these ideas might impact social search engines. We will touch upon general results concerning the costs and benefits of cooperative foraging, the effects of group diversity, and patterns of social structuring that are correlated with innovative discovery.

## COSTS AND BENEFITS OF COOPERATIVE FORAGING

### Problems of Specialization, Bias Barriers, and Tools

Specialization is a natural consequence of too much public knowledge for the individual mind to comprehend. The worry is that knowledge specialization leads to situations in which all the information required to make an important discovery is in the available record somewhere, but it is distributed across specialization boundaries with no single set of eyes in a position to see it all and make sense of it. In the information retrieval and library sciences, this is known as the *undiscovered public knowledge problem*. These problems may include hidden refutations, hidden cumulative strength of individually weak studies, or other hidden links in the logic of discovery.

Indeed, in the intelligence analysis community, since 9/11, there has been increasing focus on the need to reduce individual bias in analysis. Some barriers to information sharing has been systematically studied and publicly debated, including computer security issues, political turf concerns, and multiple intelligence agencies with differing sources of information.

While much of the problem remain at organizational levels such as policy issues, analysts also lack the most basic tools for information sharing, such as the ability to identify the experts in a particular domain, search in a database that has been tagged with the most promising leads, or automatically download and attend to the most promising information sources. The tools are truly quite impoverished currently, but could be greatly improved.

### Studies of Existing Solutions

Existing research showed that the benefits of social information foraging can be systematically studied and understood. Here we introduce some of this past research.

Cooperation may yield more benefits than simply making search more parallel and making it less prone to failure. Membership in a group provides actual or potential resources that can be utilized or mobilized to achieve individual goals. This is known as *social capital* (Bourdieu, 1986; Putnam, 2000).

Pirolli and Card (1999) describe a business intelligence agency whose analysts were tasked to write monthly newsletters about core areas such as computer science or

materials science. In addition to culling material for their own newsletters, analysts would also notice articles pertinent to the specialties of other analysts, and would have such articles copied and routed to the appropriate specialist. An analyst would typically receive about 6–12 relevant articles per month from other analysts, at very little cost. The general belief of the analysts was that such cooperation enhanced the individuals' search capabilities, and reduced the risk of missing something relevant to a specialty area that had emerged in a non-specialty publication (Pirolli and Card, 1999). The social network amongst the analysts is precisely their social capital for solving complex problems of their own.

Interestingly, recent recommender systems use an implicit form of these social networks to provide social capital to their users. Recommender systems exploit social information to make recommendations (documents, movies, music) to individuals. These include collaborative filtering systems (e.g., Herlocker *et al.*, 2004) in which people typically indicate their preference for items in some way (e.g., by rating things such as books), and they receive recommendations based on the preferences of others with similar tastes.

#### **THE EFFECTS OF GROUP DIVERSITY**

Existing research studied the effects of group diversity on cooperative information foraging, as well as the theory that people who provide brokerage of ideas across social clusters are often in position to make valuable novel discoveries.

Organization and management studies (Cummings, 2004) suggest that effective work groups are ones that share information and know-how with external members, and that effectiveness is improved by *structural diversity* of the group. *Structural diversity* is variability in features of the group that expose members to different sources of task information, know-how, and feedback. Such features include geographic locations, functional assignments, number of managers to whom members report, and number of business units associated with the group.

The findings of Cummings (2004) are consistent with the theory of social structural holes (*structural holes theory*) proposed by Burt (in press). Typically, such social networks of information flow will contain densely connected clusters. The sparse linkages between such clusters constitute *structural holes*. People who bridge such structural holes have an advantage of exposure to greater diversity of information and know-how, and *brokerage* across structural holes becomes a form of *social capital* that translates into the discovery of greater amounts of useful, productive knowledge.

One of the exciting prospects for the study of social information foraging is improved ability to find and measure social networks using on-line resources. For instance, it appears that e-mail flow and Web links among

personal home pages provide data that can be used to accurately construct social networks (Huberman & Adamic, 2004) and to study information flow.

#### **SOCIAL INFORMATION FORAGING MODEL**

Pirolli (in press) derived a basic model of social information foraging based on the Information Foraging Theory. The model seeks to understand the effect of the diversity of information foragers and their hints to each other, and how this diversity directly affects the size of cooperating groups.

This model assumes that there are some processing steps that are required to find useful patches of information and this can be modeled as a Poisson process. Cooperating information hints from other foragers can be modeled as the number of search steps that are eliminated due to a distinct hint.

The model suggests that so long as the diversity of agents increases with group size, then the size of a group increases the overall power of cooperative discovery. As individual foragers increase the diversity of their cooperating contacts they will improve in performance.

The model also showed that, assuming there are interference costs that grow with the number of members in a group, these interference costs may reduce the social capital of group foraging down to the level of individual foraging, at which point one expects an equilibrium group size.

We surmise that new social foraging tools could mediate some of these interference costs, focusing users' interactions to enriched information patches, thus enabling larger groups of users to collaborate with each other. Web technologies such as Wiki, Blogs, and Collaborative Tagging might change the scaling effects, enabling larger groups to work more effectively on the web.

#### **Epidemiological Model**

Recent researchers have tried to model the information flow in a social network of information foragers. One set of models for such flow is based on epidemiological models (Huberman & Adamic, 2004; Wu *et al.*, 2004). The model of Wu *et al.* (2004) is based on the application of random graph theory (Bollobas, 1985; Newman, 2002) to the epidemiological models of the spread of diseases. The key idea is that information spreads from person to person in ways analogous to the spread of an infection from person to person. Wu *et al.* (2004) assume a social network whose nodes have an outdegree distribution that follows a scale-free power law. This means that there are a few members of the social network who have a very large number of social connections, but the typical number of connections is small (and less than the arithmetic mean).

Information tends to be selectively passed on to people that the host believes will find the information useful. Individuals tend to form social associations based on the

USERS	3x3eyes	biotech	creap	dinopixel	ed	elindbloom	fabio	ffg	gbuser6	guruzilla	jiim	jugglebird	lowellk	rainer	terrorism
3x3eyes	-	0.00%	2.13%	6.80%	0.02%	2.02%	0.06%	0.46%	0.01%	0.52%	1.07%	1.77%	1.71%	0.61%	0.00%
biotech	0.00%	-	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
creap	2.13%	0.00%	-	4.66%	0.01%	2.10%	0.00%	0.08%	0.00%	0.04%	0.54%	1.60%	0.93%	0.40%	0.00%
dinopixel	6.80%	0.00%	4.66%	-	0.01%	6.13%	0.02%	0.59%	0.00%	0.13%	1.18%	2.41%	2.63%	0.34%	0.00%
ed	0.02%	0.00%	0.01%	0.01%	-	0.01%	0.72%	0.04%	0.03%	0.00%	0.03%	0.03%	0.01%	0.00%	0.00%
elindbloom	2.02%	0.00%	2.10%	6.13%	0.01%	-	0.00%	0.20%	0.00%	0.00%	0.60%	1.90%	0.78%	0.16%	0.00%
fabio	0.06%	0.00%	0.00%	0.02%	0.72%	0.00%	-	0.00%	0.08%	0.02%	0.02%	0.00%	0.02%	0.00%	0.00%
ffg	0.46%	0.00%	0.08%	0.59%	0.04%	0.20%	0.00%	-	0.00%	0.28%	0.35%	0.55%	0.18%	0.27%	0.00%
gbuser6	0.01%	0.00%	0.00%	0.00%	0.03%	0.00%	0.08%	0.00%	-	0.00%	0.03%	0.00%	0.06%	0.00%	0.00%
guruzilla	0.52%	0.00%	0.04%	0.13%	0.00%	0.00%	0.02%	0.28%	0.00%	-	0.14%	0.19%	0.03%	0.09%	0.00%
jiim	1.07%	0.00%	0.54%	1.18%	0.03%	0.60%	0.02%	0.35%	0.03%	0.14%	-	0.60%	0.45%	0.21%	0.00%
jugglebird	1.77%	0.00%	1.60%	2.41%	0.03%	1.90%	0.00%	0.55%	0.00%	0.19%	0.60%	-	0.41%	0.37%	0.00%
lowellk	1.71%	0.00%	0.93%	2.63%	0.01%	0.78%	0.02%	0.18%	0.06%	0.03%	0.45%	0.41%	-	0.23%	0.00%
rainer	0.61%	0.00%	0.40%	0.34%	0.00%	0.16%	0.00%	0.27%	0.00%	0.09%	0.21%	0.37%	0.23%	-	0.00%
terrorism	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-

similarity of their characteristics (known as *homophily*), and Wu et al. assume that the similarity of two people diminishes as a function of their distance in a social network (a common observation in sociometrics). One is lead to infer that transmission probabilities for information should decay as a function of network distance from the host source.

The sum of all the predicted infected individuals across all distances from the host provides a measure of the average total size of an “**information outbreak**” (the number of individuals “infected” by the information). Interestingly, Wu et al. showed that the assumption of decay in transmissibility with network distance typically leads to outbreaks of finite size. Empirical studies of samples of email, and numerical simulations using parameters based on observation and a social network graph based on email patterns suggest that outbreaks in an organization of about 7000 people is typically limited to under 50 people.

### IMPLICATIONS FROM THE MODELS

The effects of group diversity and the social foraging models tell us that the structure, composition, and the size of the social network are extremely important.

In terms of **composition**, Cummings’ research show that the diversity of hints is important in bringing fresh information into a group.

In terms of **structure**, the findings from Burt suggest that if we can locate structural holes in the social network, we will be able to bridge across these holes by enabling information to directly flow across more efficiently. One way this can be operationalized is by identifying two workgroups that has tight information sharing, and finding the highest connected members from each group to inform each other.

In terms of **size**, the results from Pirolli and Wu et al. suggest that there is an optimal size of the information outbreak for a given situation. Too large of an outbreak will generate interference, while too small of an outbreak causes information to be disseminated too inefficiently.

There are still other open questions, such as issues surrounding trust and privacy. An information sharing tool should enable further enhancement of trust between users or at least not diminish it. Much more difficult to design for, the privacy of the user must be protected. These are open future research issues.

### COLLABORATIVE SEARCH

We have been interested in applying these findings and implications to the design of a collaborative search engine. In considering the implications of composition and structure, we first wanted to understand the variety of information that different people look at and if there is enough structure and composition in these groups to cover many topics but with enough commonalities to infer possible structural holes to bridge.

We wanted to do this analysis to understand how well people can collaborate in their search process. We wanted to do this by first comparing the history and bookmarks of some real users. We performed some analysis of user browsing histories and their bookmarks, trying to understand to what extent their information sources are diverse but yet overlap. Given the vastness of the web, it turns out that there is a lot more overlap than we might otherwise have expected.

The figure above depicts our analysis results. The user profiles “ed”, “fabio”, and “gbuser6” (highlighted in yellow and orange) are history traces we obtained by directly

examining the browser histories. They all contain 3-6+ months of data. “Biotech” and “terrorism” (highlighted in green) are bookmarks from the Open Directory Project curated listings. Finally, we also examined ten bookmark histories from a social bookmarking site called del.icio.us (users with no highlights). The profiles ranged from 831 URLs to over 9211 (Mean=2243).

There are some interesting results:

- Ed and Fabio’s profiles only overlapped with each other strongly, but not with other profiles. This is interesting, especially since they were co-workers.
- Gbuser6’s profile is a real intelligence analyst profile that mainly consists of terrorism and biological weapon URLs. This profile did not overlap with the curated Biotech and Terrorism profiles, and in fact showed very little overlap with any other profiles. This indicates that he used very different information sources. Biotech and Terrorism also had no overlap with other profiles.
- The del.icio.us users overlap strongly with each other, forming their own little community. This is true despite the fact that we have no evidence that these users know each other at all. In particular, dinopixel overlapped with 3x3eyes by 6.8%, which means roughly every 1 in 15 URLs were viewed by both of them! One possible explanation is that they are all uber-geek technologists who tend to use similar information sources on the net such as Slashdot.

What these preliminary results seem to suggest are:

- Users with similar interests tend to skim the same surfaces on the web, if they might be using similar news sources in their everyday browsing activities. Co-workers and communities are most likely to skim the same surfaces, but they are not always getting to the same pages. Given the vastness and diversity of the web, people actually get pretty close to each other, and could easily inform each other of sources that they should not have missed.
- People in related fields but not in the same community do not necessarily skim the same surfaces on the web, and are prime candidates for more information sharing. These results seem to confirm that structural holes do exist. Gbuser6 is well poised to benefit from both biotech and terrorism profiles, and serve as a bridge between the two communities.
- The similarity between profiles could be used as input to a collaborative search engine. It could be used to find experts in particular areas, share new information sources, and bridge across structural holes. We could use the overlap between users to help find related users, related queries, and result pages.

## CONCLUDING REMARKS

In this short paper, we summarized recent work in the understanding of social information foraging, and models of how users socially foraging with diverse hints.

In communities of practice that depend on foraging in overly rich information environments, there appears to be pressure to self-organize into a balance of some division of labor, plus some degree of cooperation. It appears that the division of labor is necessary because of the limits of human attention, but some investment in cooperation can lead to increased returns and less risk of missing something important. The power of cooperation is related to the amount of diversity of the information foragers. Greater diversity leads to greater returns for the group and the individual. This is related to the notion that brokerage (diverse social contacts) provides social capital, and there is evidence that brokers in the flow of information are more likely to be sources of innovative discoveries. Although there are benefits to cooperation, those benefits trade against interference effects.

Recent emerging tools such as blogs, podcasts, and other internet groups serve to bridge across structural holes, but they must at the same time eliminate their potential for interference effects and overwhelm the users.

One way to bridge between different communities of users is to diversify their information sources. We have been examining the possibility of using a collaborative search engine to achieve this effect. Initial studies suggest that a social search engine build with the understanding of the implications of composition, structure, and size of communities should be able to inform users of new potential information sources without creating much additional interference overhead for the users.

## REFERENCES

1. Bollobas, B. (1985). Random graphs. London: Academic Press.
2. Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), Handbook of theory and research in the sociology of education. New York: Greenwood Press.
3. Burt, R. S. (in press). Structural holes and good ideas. American Journal of Sociology, 110.
4. Bush, V. (1945). As we may think. Atlantic Monthly, 176, 101-108.
5. Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing in a global organization. Management Science, 50(3), 352-364.
6. Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5-53.
7. Huberman, B. A., & Adamic, L. A. (2004). Information dynamics in a networked world. In E. Ben-Naim, H.

- Frauenfelder & Z. Toroczkai (Eds.), *Complex networks*. Berlin: Springer-Verlag.
8. Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 66(1), 016128.
  9. Pirolli, P. (in press). Chapter 8: Social Information Foraging. *Information Foraging*.
  10. Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, 106, 643-675.
  11. Putnam, R. (2000). *Bowling alone: The collapse and revival of american community*. New York: Simon and Schuster.
  12. Sandstrom, P. E. (2001). Scholarly communication as a socioecological systems. *Scientometrics*, 51(3), 573-605.
  13. Swanson, D. R. (1986b). Undiscovered public knowledge. *The Library Quarterly*, 56(2), 103-118.
  14. Wu, F., Huberman, B. A., Adamic, L. A., & Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2), 327.