Exploiting Social Networks for Internet Search

Alan Mislove *Krishna P. Gummadi* *Peter Druschel*

1 Introduction

Over the last decade, the rise of the World Wide Web and the emergence of Web search engines have fundamentally transformed the way people share and find information. Recently, a new form of publishing and locating information, dubbed online social networking, has become very popular. Unlike the Web, whose hyperlinked structure has received significant attention and has been exploited for searching content, the structure of information exchange over these social networks is not as well studied and understood. This paper explores the architectural differences between the Web and the online social networks from the perspective of content sharing systems, and investigates their implications for searching or retrieving useful content.

In the Web, explicit links between content are the primary tool for structuring information. Hyperlinks are used by authors to embed a page in the "web" of related information, they are used by human users to manually "surf" the Web, and they are used by search engines to "crawl" the Web to index content, as well as to "rank" or estimate the relevance of content for a search query.

In contrast to the Web, no explicit links exist between content instances stored in social networks. Instead, explicit links between users, who generate or publish the content, serve as the primary structuring tools. For example, in social networking sites like MySpace [15], Orkut [16], and Flickr [5], a link from user A to user B usually indicates that A finds the information published by B interesting or relevant, or A implicitly endorses B's content due to an established social relationship. Such social links enable users to manually browse for information that is likely of interest to them, and they could be used by search tools to index and locate information.

In the rest of this section, we compare the Web and social networking systems, particularly with respect to their mechanisms for publishing and locating content. Our analysis makes a case for integrating the methods used for locating content in both systems. In later sections, we experimentally evaluate the potential benefits of such a social network based Web search, and outline the research challenges in building integrated search systems for the future Internet.

1.1 Comparing content sharing systems

Content sharing systems, including the web and online social networks, have three distinct mechanisms, which impact how content is exchanged between users:

Publishing content: This is the mechanism by which content creators make information available to other users. It also includes how they relate their content to other content found in the system.

Locating content: Due to the large amount of content these systems offer, the systems must offer ways for users to locate information relevant to them.

Distributing content: Lastly, there must be a distribution mechanism to transfer content between users. In both the web and many current online social networks, the content is transferred using HTTP over TCP, and the users navigate the systems from their Web browser.

Consequently, the differences in the use of the Web and social networks are primarily in the methods used to publish and locate information. Next, we will discuss the mechanims used in the Web and social networks and see how they affect the uses and the type of content exchanged in each system.

1.2 The Web

Publishing: In the Web, users publish by placing content on a Web server. An author places hyperlinks into her content to refer to related content, and she may ask other authors to include links to her content into their content.

Locating: Today, the predominant way of locating information on the Web is via a search engine. Modern Web search engines employ sophisticated information retrieval techniques and impressive systems engineering to achieve high-quality search results at massive scale.

The key idea behind search engines like Google is to exploit the hyperlink structure of the Web to determine both the corpus of information they index and the relevance of a Web page relative to a given query. This approach has proven highly effective, because the incident links to a page are strong indicators of the importance or relevance of the page's content in the eyes of other users.

However, hyperlinks based search has some well known limitations. First, while Web search is very effective for relatively static information, it may under-rate or miss recently published content. The reasons for this are that (*i*) a new page must be discovered and indexed by the search engine, (*ii*) hyperlinks to a new page must be included in subsequently published or edited pages, and (*iii*) all such links must then be discovered by the search engine.

Second, as search engines determine the relevance of a page by its incident hyperlinks, their rating reflects the interests and biases of the Web community at large. For instance, a search for "Michael Jackson" yields mostly pages with information about the pop star. Software engineering researchers, however, may find the Web page of a computer science professor with the same name more relevant. Web search, however, typically ranks highest the candidate search results that are of interest to many people.

Finally, the hyperlink structure also influences whether a page is included in the search engine's index. Clearly, unlinked pages and not publically accessible pages (the so-called dark web) are not indexed. In addition, many other pages are not indexed (the so-called deep web) because, based on their location in the hyperlink structure, the search engine deems them insufficiently relevant to be included. As a result, obscure, special-interest content is less likely to be accessible via Web search.

1.3 Social Networks

Online social networking web sites have recently exploded in popularity, with sites built for finding friends like MySpace [15], Orkut [16], and Friendster [7], sharing photos like Flickr [5], sharing videos like YouTube [22] and Google Video [9], and writing blogs like LiveJournal [13] and BlogSpot [4]. These sites are extremely popular with users: MySpace claims to have almost 100 million users, while Flickr boasts 2.5 million, and Orkut claims 13 million. In fact, MySpace recently has been observed to receive more page hits than Google.

Simple uses of the social networking approach have existed for much longer. For instance, the common practice of placing content on the Web and sending its URL to friends or colleagues is essentially an instance of social networking. Typically, the author has no intention of linking the content; thus, the content remains invisible to users other than the explicit recipients of the URL. The content is advertised not via hyperlinks, but via links between users.

Publishing: Users publish content by posting it on a social network site or on their own node. There is no structure between the various pieces of published content and it can be of any type. Often, the content is temporal in nature, such as blog postings or video clips of news stories, and may be of interest only to a small audience.

Independent of the content they store, users maintain links to other users, which imply trust or shared interest. Links can be directed (indicating that the source trusts and is interested in the content of the target) or undirected (indicating mutual trust and interest in each other's content). Some systems maintain groups of users associated with different topics or interests; users then join groups rather than specifying links to individual other users. In some systems, the full social network graph is public; in others, only immediate neighbors can view a node's other neighbors.

Locating: In social networks, users find information by searching content stored by adjacent users in the network. This can be done manually by browsing the content while navigating through the network, or by evaluating search queries over the content stored on adjacent nodes. Content is often rated according to how often users have accessed it, or based on explicit feedback provided by users.

Social networks are far more effective at publishing and locating end user generated content that is primarily of interest to members of a certain social network, or that is of short-term value. Unlike Web search, social networks can rate content rapidly, as they can use both the implicit and explicit feedback of a large community of content consumers rather than a small number of content generators (Hyperlinks in the Web have to be established by content generators.) Because users search adjacent regions of their social network, the content can be rated relative to a community of users with shared interests. For example, blog posts, videos, and comments are all generally relevant for a much smaller period of time than web content, and the audience for these is correspondingly smaller. Thus, social networks enable users to find timely, relevant and reliable information that is hard to find by the way of Web search.

1.4 Integrating Web search and social networks

Today, the information stored in different social networks and in the Web is mostly disjoint, with each system having its own method of searching information. While search companies have started to address the issue with specialized search tools for RSS-based news feeds and for blogs, there is no unified search tool that locates information across different systems. Social network based search methods are also not generally used in the Web, though services like Google scholar support search facilities tailored to a specific community. Given that end users access both the Web and the social networks from the same web browsers, it seems natural to unify the methods to find information on them as well.

In this paper, we explore the idea of integrating Web search with search in social networks. We believe that such an approach could combine the strenghts of both types of systems: simultaneously exploiting the information contained in hyperlinks, and information from implicit and explicit user feedback; leveraging the huge investment in conventional Web search, while also ranking search results relative to the interests of a social network; locating timely, short-lived, or special-interest information alongside the vast amounts of long-lived information on the Web.

In the remainder of this paper, we present results of some preliminary measurements to explore the potential of our approach in Section 2, and follow with a discussion of research challenges that need to be addressed towards an integrated search system in Section 3. Section 4 presents related work and Section 5 concludes.

2 Evaluating Social Networks based Internet Search

Based on the discussion above, we conclude that (i) a growing body of Internet content cannot be retrieved by traditional Web search as it is not well-connected to the hyperlinked Web, and that (ii) social network links can be leveraged to improve the quality of search results. In this section, we describe a social networking experiment we conducted to validate and quantify our analysis.

2.1 Experimental methodology

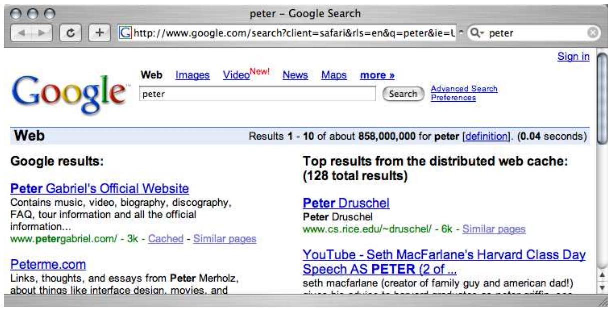
In our integrated social networks based Internet search experiment, users not only exploit the hyperlinked structure of the content, but also leverage the content previously viewed or rated as relevant by the neighbors in their social network. We recruited a small group of 10 graduate students and researchers in our distributed and networked systems group to share all Internet content downloaded or viewed by them with one another.

To estimate the limits of hyperlink-based search, we check what fraction of the content or URLs actually visited by the users are not indexed by a state-of-the-art search engine, Google. Examples of such URLs could vary from recently created blog postings to pages in the "deep Web". Such URLs represent information that is of relevance to the users (as they have been viewed by them), but which is not sufficiently well connected to the Web to be indexed by Google.

To estimate the benefits of social network based search, we allow users to search over content previously viewed by others. Each user runs a lightweight web proxy, which transparently indexes all visited URLs. When a Google search is performed, the proxy transparently forwards the query to both Google as well as peer proxies of other users in the social network. Each proxy executes the query on the local index and returns the result to the sender. The results are then collated and presented alongside the Google results as shown in Figure 1. We measure and compare how often users click on results from the social network and Google.

Our experimental prototype combines the Lucene [2] text search engine with the FreePastry [6] peer-to-peer overlay. To rank the results obtained from the social network, we (a) configured Lucene to follow Google's query language, using only boolean AND queries and respecting quotations and 'site:' directives, (b) multiplied the Lucene score of a search result by the Google PageRank of that result, and (c) added the scores from all users who previously viewed a given result. Thus, our ranking takes advantage of both the hyperlinks of the Web and the social links of the user community.

Given that our user base is small, includes the authors, and represents only a single community with highly specialized interests, we cannot claim that our results would be representative of a deployment with a larger, diverse user base. Our results should be viewed as just indicators for the potential of social network based Web search. A more comprehensive study based on web access traces collected at the gateway router of a major university is currently in progress.



Web Results **1** - **10** of about **858,000,000** for **peter** [definition]. (**0.04** seconds)

Google results:

Peter Gabriel's Official Website
Contains music, video, biography, discography, FAQ, tour information and all the official information...
www.**peter**gabriel.com/ - 3k - Cached - Similar pages

Peterme.com
Links, thoughts, and essays from **Peter** Merholz, about things like interface design, movies, and

Top results from the distributed web cache: (128 total results)

Peter Druschel
Peter Druschel
www.cs.rice.edu/~druschel/ - 6k - Similar pages

YouTube - Seth MacFarlane's Harvard Class Day Speech AS **PETER** (2 of ...
seth macfarlane (creator of family guy and american dad!)

Figure 1: **Screenshot of our search interface**. Results from the distributed cache appear alongside the normal Google results.

2.2 Results

We present measurements and experiences from a month long usage of our experimental deployment. During this time, the 10 users have downloaded 296,831 URLs. Of these, only 26.3% could be indexed, i.e., their content type was either 'text/html' or 'application/pdf'.

2.2.1 Limits of Hyperlink-based search

Even the best Web search engines cannot index content that is not well-linked to the Web or not publicly available to the entire world. So our first goal is to understand the Internet content that is viewed by users, but is not captured by the search engines. We would also like to know if this content is already indexed by our social network.

For each URL viewed, we checked (i) whether Google's index contains the URL, and (ii) if some peer in the social network has previously viewed the URL. Since search engines only index static HTML content, we considered only URLs that end in .html, .htm, .php, .jsp, or .asp. Further, we discarded URLs with an auto-refresh feature (such as the scoreboard sites for sports), as they would artificially bias the results against Google. This left us with 6,679 URL requests for 3,987 URLs.

Our analysis shows that Google's index contains only 62.5% of these URLs viewed, representing 68.1% of the distinct URLs. This implies that one third of all URLs browsed by our users cannot be retrieved by searching Google!

The collective index of users' caches contained only 30.4% of the URLs at the time of their viewing. More interestingly, 13.3% of the URLs viewed were contained in peer caches but not in Google's index. This amounts to a 20% improvement in the documents that could be retrieved by social network based search compared to Google search. It is worth noting that, for our small social network of CS researchers, this improvement comes at the expense of adding just a few thousand URLs to a Google index containing billions of URLs.

Our results naturally raise the question, *what are these documents that are of a lot of interest to users, but are so hard to index for search engines like Google?* We manually analyzed a number of such URLs and show a random sample of them in Figure 2.

The content that is beyond the scope of Google falls under three broad categories. Many URLs refer to recently generated content that is too new or too deep in the Web structure (i.e., of interest to a narrow community) to be indexed by Google. As shown in Figure 2, examples include blog postings to "Live Journal" and "craigslist", news articles at "CNN.com", and discussion forums on "kde.org". Further, a small fraction of the URLs can be accessed by users within our department, but are not publicly accessible. Such URLs belong to the "dark web", and cannot be indexed by search engines. Our sample, shown in Figure 2 contains such URL, which refers to the status page of a departmental wireless router.

2.2.2 Benefits of Social Networks based search

Another challenge facing search engines is ranking all the indexed documents in the order of their relevance to a user's query. Ranking is very crucial for search, as most users rarely go beyond the first few query results [18]. Our goal here is to study how often users click on query results from caches of other peers as opposed to Google. As shown in Figure 1, our users are presented with results from both Google and peer caches for every Google query.

During the course of a month, we observed 1,478 Google searches. While Google's first result page contained an average of 9.465 results, our smaller shared cache index resulted in an average of 5.510 results on the first page. Of the 1,478 queries, 944 (64%) resulted in clicks on search result links. 86.5% of these search result clicks were produced only by Google, 7.7% of the result clicks were obtained only from shared caches, while 5.7% appeared in both sets of results. This amounts to a 9% hit rate increase over the Google results, and a 5% improvement in overall search experience.

It should be kept in mind that this 9% improvement over Google, considered by many to be the gold standard for Web search engineering, is coming from a simple, very small, social-network based system quickly put together by three systems researchers over a period of a few days. Based on our early experience, we feel that these benefits suggest fundamental, inherent advantages of using social links for search, that could be exploited better with more careful engineering.

To better understand the scenarios when search results from peers outperform Google results, we manually analyzed both the queries and result clicks. We show a random sample of the data we analyzed in Firgure 3. We observed that they fall under two categories:

Disambiguation and Relevance: Often search terms have multiple meanings to different people in different contexts. Search engines take the most popular or common term definition, while social networks can take advantage of links between users, who share similar definitions or interpretation of these terms. An example for disambiguation is shown in Figure 3, where a user's query for "bus" yielded our local bus schedule, as it is the page most visited by other users that contains the term. Same with the term "Peter". Another interesting example is the term "coolstreaming". Google search leads to popular sites (such as wikipedia) discussing the "coolstreaming" technique for P2P streaming of multimedia content, while searches over caches lead to the INFOCOMM paper that is of more interest to our users.

Serendipity: Rather ironically, we discovered the serendipity benefits of our search, serendipitiously. By serendipity, we refer to users discovering relevant information quite by accident, clicking on links/information

URL	Too New	Deep Web	Dark Web
http://jwz.livejournal.com/413222.html	X	X	
http://www.mpi-sws.mpg.de/∼pkouznet/oulia/pres0031.html		X	
http://sandiego.craigslist.org/w4m/179184549.html	X	X	
http://edition.cnn.com/2006/ ... /italy.nesta/index.html	X		
http://72.132.241.163/status.asp			X
http://www.itv.com/news/ ... a8e4b6ea.html	X		
http://www.stat.rice.edu/∼riedi/ ... /target21.html		X	
http://amarok.kde.org/forum/index.php/board,9.20.html	X	X	

Figure 2: **Sample URLs which are not indexed by Google**. The results are also tagged by the reason for not being in Google's index.

Query	Result URL	Disambiguation & Relevance	Serendipity
bus	Saarbrücken bus schedule	X	
peter	Peter Druschel's home page	X	
serbian currency	XE.com exchange rates	X	
coolstreaming	CoolStreaming INFOCOM paper	X	
stefan	FIFA World Cup site		X
münchen	Peter Druschel's homepage		X

Figure 3: **Sample search queries where our social network returned results not in Google**. The results are categorized into different scenarios discussed in Section 2.

that they never intended to query specifically for. Serendipity is an integral part of the Web browsing experience, and results from distributed caches provide ample opportunities for such serendipitious discoveries. For example, one of our users looking for information about "München" (Munich) disovered that another fellow researcher did his schooling in München, thus, finding a convenient source of information about the city.

3 Research opportunities and challenges

Online social networking enables new forms of information exchange in the Internet. First, it makes it very easy and convenient for individuals to publish information, without necessarily linking it to the wider WWW. Second, social networks make it possible to locate and access information that was previously exchanged by "word of mouth", that is, by explicit communication between individuals. Lastly, unlike Google, which organizes the world of information according to popular opinion, social networks can organize the world of information according to the tastes and preferences of groups of individuals.

We see great potential in the integration of the Web and social network search technologies. Such an integration can provide unified access to all of the world's public on-line information, not just the information in the shallow Web. We presented evidence that it can also improve the quality of Web search results, because it can rank results relative to the interests and biases of groups of individuals. In this section, we discuss research opportunities and challenges associated with realizing this vision.

Privacy: Participants in a social network must be willing to disclose which information they find interesting and relevant. This creates a tension between the privacy concerns of individuals and the effectiveness of the social network, which depends on the willingness of individuals to share information. In small social networks of mutually trusting participants (e.g., family members or close friends) the problem reduces to access control. However, in larger social networks (e.g., all researchers in computer networking), a solution that is acceptable to users would require mechanisms to control information flow and anonymity.

Scalable search in social networks: Existing search mechanisms in social networks are often based on a simple breadth-first search of the network, starting from the originator of the query. Given the small-world properties (6 degrees of separation) in social networks, this does not scale to large networks, because a 3-hop neigborhood would include a third of the participants. Scalable index-based search algorithms are needed for social networks.

Membership and clustering of social networks: An individual may generally be a participant in multiple social networks, e.g., networks related to professional interests, networks related to hobbies, and networks related to family and friends. This raises several questions. First, can the graph consisting of all users and their social links be automatically derived and maintained? For instance, by observing which users exchange email, or by considering similarity in content browsed or stored between pairs of users? In the absence of such techniques, users have to explicitly declare and manage their social network memberships. Second, can the social network graph topology be used to automatically identify different clusters of communities associated with certain interests? Third, given such a clustering, how should a search query be resolved with respect to the different clusters that a user participates in?

Content rating and ranking: How should search results over social networks be ranked? There are many alternatives that could be explored: Based on global page-rank, as in Google? Based on an a local page-rank specific to the social network? Based on the number of users who have browsed or stored the content? Based on explicit user rankings? Based on some combination of the above? How should the search results from the social network be displayed or ranked relative to the Google results?

System architecture: Should the system be centralized or distributed? A centralized architecture, similar to current Web search engines, may raise concerns about privacy, trust and market dominance. Also, a centralized approach may not scale with the bandwidth requirements of a central data store or the number of different social networks. A decentralized architecture, on the other hand, faces challenges of its own: Building even a conventional Web search engine in a decentralized fashion is a difficult research problem.

4 Related work

Several projects have looked at replacing the functionality of the large centralized web search engines with a decentralized system, built from contributing users' desktops [12]. Both Minerva [3] and YaCy [20] system implement a peer-to-peer web search engine without any points of centralization. Addtionally, other projects [11, 17] have examined replacing the centralized PageRank computation of Google with a decentralize approach. All of these projects, though, are primarily focused on replacing the functionality of existing centralized search engines with a decentralized architecture.

A few systems have looked at query personalization, or taking a user's preferences and interests into account when ranking pages. Most notably, the A9 [1] and Google Personalized Search [8] allow users to create profiles to which search results are tailored. There has also been much research into methods for accurately personalizing search queries [10, 19]. While these projects are concerned with personalization, our work is complementary and examines the ability to use social links to automatically derive users interests.

Lastly, a number of projects have looked at using social networks to aid a variety of applications. Notable distributed systems projects include SPROUT [14], which uses the trust of social links to increase the probability of successful DHT routing, and Maze [21], which allows users to create friends in the file sharing network.

5 Conclusions

We have examined the architectural differences between the Web and social networking systems with respect to publishing and locating content. The resulting analysis indicates that social network based search may be able to find Internet content that Web search engines currently have problems finding. We reported results from our ongoing social network based search experiment that further illustrates the potential benefits. Finally, we outlined research challenges and opportunities in leveraging societal relationships to build search systems for the future Internet.

References

[1] A9 Search. `http://www.a9.com`.

[2] Apache Lucene Search Engine. `http://lucene.apache.org`.

[3] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Minerva: Collaborative p2p search. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05)*, August 2005.

[4] BlogSpot. `http://www.blogspot.com`.

[5] Flickr. `http://www.flickr.com`.

[6] FreePastry Project. `http://www.freepastry.org`.

[7] Friendster. `http://www.friendster.com`.

[8] Google Personalized Search. `http://www.google.com/psearch`.

[9] Google Video. `http://video.google.com`.

[10] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th International World Wide Web Conference (WWW'03)*, May 2003.

[11] K.Sankaralingam, S. Sethumadhavan, and J.C.Browne. Distributed pagerank for p2p systems. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC-12)*, June 2003.

[12] J. Li, B. T. Loo, J. Hellerstein, F. Kaashoek, D. R. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, February 2003.

[13] LiveJournal. `http://www.livejournal.com`.

[14] S. Marti, P. Ganesan, and H. Garcia-Molina. Dht routing using social links. In *Proceedings of the 3rd International Workshop on Peer-to-Peer Systems (IPTPS'04)*, February 2004.

[15] MySpace. `http://www.myspace.com`.

[16] Orkut. `http://www.orkut.com`.

[17] J. X. Parreira, D. Donato, S. Michel, and G. Weikum. Efficient and decentralized pagerank approximation in a peer-to-peer web search network. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06)*, September 2006.

[18] B. Smyth, E. Balfe, O. Boydell, K. Bradley, P. Briggs, M. Coyle, and J. Freyne. A live-user evaluation of collaborative web search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, July 2005.

[19] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'05)*, August 2005.

[20] YaCy Search Engine. `http://www.yacy.net`.

[21] M. Yang, H. Chen, B. Y. Zhao, Y. Dai, and Z. Zhang. Deployment of a large-scale peer-to-peer social network. In *Proceedings of the 1st Workshop on Real, Large Distributed Systems (WORLDS'04)*, December 2004.

[22] YouTube. `http://www.youtube.com`.