Navigation Characteristics of Online Social Networks and Search Engines Users

Christopher W. Dunn[†]
chrdunn@cs.indiana.edu

Minaxi Gupta[†]
minaxi@cs.indiana.edu

Alexandre Gerber[*]
gerber@research.att.com

Oliver Spatscheck[*]
spatsch@research.att.com

[†] School of Informatics and Computing, Indiana University [*] AT&T Labs-Research

ABSTRACT

Online social networks (OSNs) represent a significant portion of Web traffic today, comparable with search engines. Even though their primary purpose is different from that of search engines, OSNs are impacting how users navigate the Web and what types of websites they visit. This paper is motivated by the desire to understand the similarities and differences in the websites users visit through OSNs versus through search engines. Using Web traffic logs from 17,000 DSL subscribers of a Tier 1 ISP in the United States, we find that while OSN visitors are less likely to navigate to external websites, when they do, they spend more time at those websites compared to when search engine users visit external websites. Also, OSNs send their users to a narrower subset of the Web than search engines. While websites related to games and video are more commonly visited from OSNs, shopping and reference sites are common for search engines. Finally, OSNs send their visitors to less popular domains more often than search engines. Our findings can be useful to ISPs in network provisioning and traffic engineering.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*

General Terms

Measurement

Keywords

Online Social Networks, Search Engines, Internet Navigation

WOSN'12, August 17, 2012, Helsinki, Finland.

1. INTRODUCTION

Search engines have played an important role in directing users to websites since the mid-nineties. Their role is pivotal today, as finding pertinent information from over a trillion web pages that the current Web boasts [1] presents unique challenges. While search engines continue to act as important conduits between users and websites, a competing phenomenon is now emerging. That phenomenon revolves around the popularity of online social networks (OSNs), such as Facebook and Twitter. Many OSNs have user bases in the hundreds of millions [2] and these users exchange various kinds of objects with each other, including web links. These sharing of links have an interesting side effect, one of making OSNs similar conduits between users and websites as search engines.

While some comparisons between OSNs and search engines are beginning to be made in terms of metrics such as time spent on the website [3], little has been done to understand if OSNs expose or make certain portions of the Web more popular. Given that both search engines and OSNs play an important role in directing users to websites, a comparison of which types of sites each sends its users to is interesting in its own right and is the motivation of this paper. Specifically, our goal is to understand if OSNs popularize different portions of the Web in comparison to search engines. We seek to explore this issue by collecting one day of HTTP traffic logs from 17,000 DSL subscribers of a Tier 1 ISP in the United States and looking for traffic traversing two prominent search engines and two prominent OSNs among the 55 million requests generated by the users.

The key findings of our analysis are the following:

- In comparison with an earlier study [4], we find that OSN visitors are spending more time there compared to what they did in 2009.
- OSN visitors are more likely to stay within the origin website compared to search engine visitors. This finding is unsurprising since users often visit search engines to navigate out. However, we find that when visitors navigate to other domains, OSN visitors are more likely to spend more time at the external domains compared to search engine visitors.
- The variety and number of domains visited from search engines is higher than those visited from OSNs. However, the average number of external domains visited in any given session is less than two for both OSN and search engine

visitors, indicating that they are more similar in this regard than different.
- Search engines lead their visitors to a wider variety of websites. Visitors of OSNs are more likely to visit websites related to games and video while visitors of search engines are somewhat more likely to visit shopping and reference-related websites. News websites, websites offering OSN functionality to improve user experience and advertisement-related websites are visited from both OSNs and search engines.
- OSNs are more likely to send their visitors to less popular domains in comparison with search engines.

2. DATA COLLECTION AND PREPARATION

To collect data for our study, we placed a network monitor on a Broadband Remote Access Server (BRAS). The BRAS we used is an aggregation point for Digital Subscriber Lines (DSLs) for large Tier 1 ISP customers located in the United States and serves approximately 17,000 active broadband subscribers. Our analysis was conducted using aggregated data collected from all HTTP traffic transiting this particular BRAS on 15th February, 2011. We only had access to HTTP headers and no data packets. The privacy of the subscribers was preserved since the dynamic IP addresses were not mapped to individual households and the study focused on the aggregate traffic across all the subscribers. Further, since query strings in URLs can contain sensitive information, we did not examine the query strings in order to protect subscriber privacy.

Our data contained requests and responses directed to port 80. This generally implies HTTP traffic but there may be other protocols, such as file-sharing traffic that may be using this port. We filter non HTTP traffic out by ensuring that a HTTP header accompanies all request and response traffic. Fields of interest for each HTTP request include the URL objects, its referrer, source and destination IP addresses and port numbers, TCP sequence number, browser user agent and timestamp. The response packets include HTTP status code, MIME type of the returned object, and content length (in bytes) in addition to the source and destination IP addresses and port numbers and TCP sequence numbers. It is noteworthy that the content length field can unfortunately not be used to infer the total number of bytes transferred because any HTTP message that uses any *transfer encoding* should not include a content-length. Even when a message is received with both header fields, clients must ignore the content-length field in that situation [5]. The use of such encodings is very common, especially for data requests and web applications. Though subsequent response packets could be used to gain information on the total bytes, our data filters those packets out for privacy reasons because they contain user data.

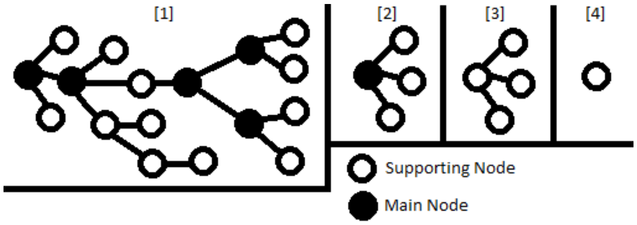

Figure 1: Example session graphs

We begin data processing by pairing HTTP requests with responses. The first step is to match the 5-tuple, {source IP, destination IP, source port, destination port, TCP sequence number}, in each request packet with the corresponding values in response packets (with source and destination IPs and ports reversed). This identifies response packets corresponding to each request packet. Time can be used as a tie breaker in cases where an ambiguity arises during the matching process. In the next step, we assemble the request-response pairs for each {source IP, user agent} pair to identify HTTP transactions for each user behind a subscriber IP. This approach will underestimate users in cases where they use identical browsers. Subsequently, we trace navigation chains of each user by considering each request-response pair to be a node and by linking nodes via referrer fields in HTTP requests. The resulting graphs define *sessions*, as we use them in this paper, and are of the types shown in Figure 1. The first type is an example of a navigation chain where a user has visited a web page and then clicked on resulting links. This graph has two types of nodes: *main nodes* and supporting nodes. The main nodes are the primary URL objects visited by a user, such as `http://www.nytimes.com/index.html` while the supporting nodes are other objects, such as Javascripts and CSS files needed to fully display the visited URL. To identify main nodes in each graph, we start at the beginning of each navigation chain and look for nodes that have multiple outgoing links and a MIME type of HTML or TXT. This approach leverages the fact that almost all modern web pages have multiple objects and that the URLs users use in browser location bars are either HTML or TXT. Since this approach will typically identify advertisements (ads) displayed in an iFrame of a web page to be a main object, we do not consider an object to be a main object if it was downloaded within one second of a previous main object or that object's supporting objects. Work by Ihm et al. used a similar approach [6]. Note that supporting nodes can sometimes bring in other supporting nodes, such as in the case of a Javascript downloading another Javascript. Also note that supporting objects could sometimes be connecting main nodes, as shown in the Figure. A common reason why that happens is *link wrapping* where say, a search engine, wraps the displayed link in order to track the URL followed by a user. Further, note in the first graph that two or more main objects could connect to a main object when a user uses a back button to navigate back to say, the search results, and then clicks on another link. It can also happen when a user uses tabbed browsing to visit the same web page and then clicks on different links from each. The second graph in Figure 1 is a simple version of a navigation chain where a user simply visited a web page and did not click on any resulting links. Note also that our approach may overestimate the number of sessions in cases where navigation chains are broken due to the lack of referrers. A prominent case where that happens is when a website switches a user from HTTP to HTTPS or vice versa. Web-based mail is a common example of such a case. The over-estimation hurts our analysis in the sense that some intermediate visits will be counted as starting points of navigation chains and in turn the lengths of some navigation chains would be underestimated. Finally, the third and fourth graphs in Figure 1 do not contain any main objects. These are fragments of sessions of the first two graph types but the lack of a referrer field, as is often the case for data requests made on a web page, prevented them from being attached to their main session. We could connect such session fragments to their main session by attaching graphs within sub-second granularity from the same HTTP transaction of a user since users cannot navigate across web pages within that time frame. However, we choose not to carry this step since as a final step, we drop sessions without any main objects from further consideration since they do not contain any navigation points. Such sessions are either data requests or ad related.

There are 107 million request-response pairs in our data belonging to 22 million sessions destined to 147K domains. However, only 636K of the user sessions contain a main object, with the majority of the rest being singleton data requests similar to the fourth graph in Figure 1. A notable minority are similar to the third graph in Figure 1.

3. CHARACTERIZATION OF SESSIONS

Recall that we use the term *session* to denote a connected graph of web pages visited by a user. Thus, visits to different web pages, in the same browser tab or separate tabs, will be counted as separate sessions unless they are connected by the referrer field. Our data contains 636,427 sessions, of which 75.1% had only a single main node (see Table 1). This could be a result of users choosing certain websites as their home pages or typing in or using a bookmark to choose the next web page they visit. Yet another reason for single main node sessions could be broken referrer chains, which often result from ad-related iFrames embedded in web pages. While not all ads break the referrer chains, they typically have their own HTML URL which qualifies their sessions to contain a main object. Table 1 also shows that most ad-related sessions only have one main node. A small number, 1.5%, have multiple main nodes that belong to the same ad domain and 2.2% have main nodes that traverse multiple domains where say, an ad retrieves objects from other domains. We rule out ad-related sessions from further consideration because users do not start web navigation at advertising networks and these sessions are simply an artifact of the lack of referrer fields to connect these graphs to their original sessions.

	Single node	Single domain	Multiple domains	Total (100%)
All sessions	75.1%	12.9%	12%	636,427
Ad-related	96.3%	1.5%	2.2%	76,040
OSN 1	78.3%	16.2%	5.5%	68,070
OSN 2	64.8%	21.5%	13.7%	1,652
Search engine 1	59.8%	19.3%	20.9%	63,643
Search engine 2	61.2%	5.1%	33.7%	45,092
Search engine 3	73%	8%	19%	21,812

Table 1: Main nodes in various sessions

Of the non-ad-related sessions, only four starting point websites[1], OSN 1, Search engine 1, Search engine 2 and Search engine 3, contribute more than 10K sessions each. These are shown in Table 1. Note that of the four websites that meet this criterion, three are search engines. The fourth is a prominent online social network (OSN). In order to draw meaningful comparison across OSNs and search engines, we add one more OSN to the picture. The second OSN, OSN 2, is the most popular OSN in terms of the websites that acted as starting points for sessions and ranked 27th among the 36 websites that acted as starting points for more than 1K sessions. Specifically, OSN 2 acted as a starting point for 1,652K sessions. It is noteworthy that sessions with a single main node dominate, accounting for 2/3rd to 3/4th of sessions for each OSN and search engine. In fact, all but one search engine have a small percentage of sessions that have multiple main nodes but no navigation to other domains (second column of Table 1). This is expected because users typically visit search engines only to navigate to other domains via searches. The exception, Search engine 1 is also justifiable because the front end of this search engine is a popular news portal, which serves news stories from its own domain. In fact, its search functionality cannot be accessed without accessing the portal. This causes the users to stay at this search engine more often. Search engine 3 also has a news portal but its search functionality can be accessed without going through the portal. This appears to mask its characteristics as a search engine. In contrast to search engines, OSNs have a larger fraction of sessions that have multiple main nodes belonging to their respective domains, indicating that their users tend to stay at the OSN more often. Further, 1/5th to 1/3rd of search engine sessions have main nodes in multiple domains, confirming their navigation to other domains. The corresponding numbers for OSNs are 5.5-13.7%. Finally, note that the total number of sessions for OSN 1 and Search engine 1 are comparable, making a comparison of their composition interesting.

Figure 2 shows the CDF of total session durations. We rule out sessions consisting only of one main object because many of these are likely to be a result of users setting an OSN or a search engine as their home page and do not indicate navigation. A key observation is that the graph for OSN 1 stands out in that it has more longer-duration sessions compared to the rest. In fact, the sessions are longer than observed by Schneider et al. [4], who observed that 12.5% sessions were longer than an hour. In comparison, 28.4% of sessions are longer than an hour in our data, which is newer. *The trend suggests that users are interacting with OSN 1 for longer durations now.* Further, the session durations for OSN 2 are closer to that of Search engine 1, whose new portal effect appears to make its sessions longer compared to those of Search engine 2 and Search engine 3.

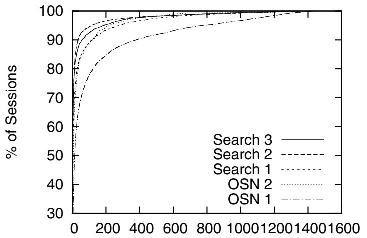

Figure 2: CDF of total session durations (for sessions with multiple main nodes)

Figure 3 shows a CDF of percentage of time spent at the starting domain versus total time. We focus only on sessions consisting of more than one main object. Overall, we find that *sessions starting at search engines spend lesser time*

[1] We merge support domains, such as those serving images on behalf of the primary domain, for each of the four websites.

there compared to those starting at OSNs. Specifically, 2/3rd of OSN sessions with multiple main nodes spend at least 87-91% time there. In contrast, 2/3rd of search engine sessions of the same type spend at least 27-55% there. This is unsurprising because a larger percentage of OSN sessions do not visit other domains.

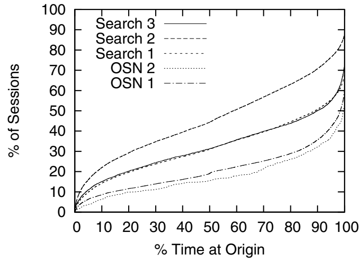

Figure 3: CDF of the percentage of time at the origin domain versus other domains (for sessions with multiple main nodes)

4. NAVIGATION TO OTHER DOMAINS

Most users visit search engines to be directed to other websites based on the keywords of their queries. However, in the case of OSNs, a user can choose to stay at the OSN site or could click on a link and be directed to another website. Irrespective, both search engines and OSNs send users to other websites. Here, we draw comparisons among search engines and OSNs by looking at the websites they send traffic to.

A total of 18,307 unique domains were visited from the OSNs and search engine under consideration. More domains were visited from search engines (6,764 from `Search engine 1` 11,531 from `Search engine 2` and 2,095 from `Search engine 3`) compared to those from OSNs (858 from `OSN 1` and 177 from `OSN 2`). Figure 4 shows the CDF of unique domains visited in a session. We only focus on sessions with multiple domains. This rules out sessions with one or more main nodes that stay at the origin website. First we note that 62-80% of the sessions with multiple domains traverse exactly two domains, including the origin domain. The effect is more pronounced for `OSN 1` and `OSN 2` where 80% and 62% of sessions respectively have exactly two domains. Also, even when the number of domains in a session is larger, most sessions have less than 10 domains, which is why we truncate the x-axis at 10 domains. Overall, we find that *the average number of domains visited in a single typical OSN session with multiple domains is similar to those in a single search engine session*, with the averages being between 2.43 and 2.82 domains. At the tail end, a few sessions have more than 10 domains. Specifically, the maximum domains visited in any session is 12 for `OSN 2` and 21 for `OSN 1`. In contrast, the corresponding number domains are 49 and 45 for `Search engine 2` and `Search engine 1` and 32 for `Search engine 3`.

Figure 5 shows the total session time spent at domains other than the origin domain. We focus only on sessions containing multiple domains for this analysis. While Figure 3 showed that OSN visitors spend more time there than search engine visitors do at search engines, this Figure shows that *OSN visitors spend more time at external domains compared to search engine visitors when they navigate out of the origin website.*

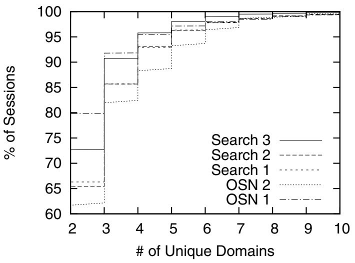

Figure 4: CDF of the unique domains visited in a session (for sessions with multiple domains)

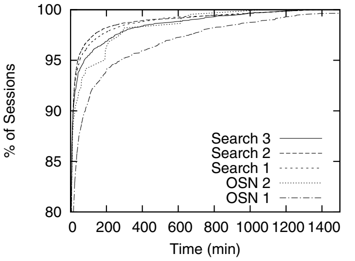

Figure 5: CDF of the total session time spent at external domains (for sessions with multiple domains)

Finally, we found that different external domains are visited from each starting point. Specifically, 63% of the domains in sessions starting at `OSN 1` occur only in sessions that start with `OSN 1`. Similarly, 44% of `OSN 2`, 72% of `Search engine 1`, 82% of `Search engine 2`, and 57% of `Search engine 3` domains are unique. Further, even when an external domain occurs in sessions with different starting points, it is typically dominated by a single starting point. Only 7% of all domains receiving traffic from one of our OSNs or search engines receive 0.1% of their traffic from two or more of the five starting points in question. Overall, external domains are strongly correlated with the starting point of their sessions, and the five starting points being analyzed represent large close knit neighborhoods of traffic that are only weakly connected to each other or the rest of the web.

4.1 Rankings and Categories

Next, we look at the categories of domains users navigate to. We initially used the DMOZ [7] open directory project to categorize, which is the largest human edited directory of websites around the world. However, since it left a large number of domains uncategorized, we switched to manual categorization. Since it was not possible to manually categorize each of the 18,307 domains, we focused on categorizing the top-50 most-visited domains for each OSN and search engine. Table 2 lists unique domains found for each OSN and search engine in each category and the percentage of sessions that navigate to them. We focus only on sessions that navigate to domains other than the starting website. (Table 1 shows the percentage of such sessions in the second last column from the right.) Note that since one session can navigate to domains in multiple categories, the percentage of sessions may be over 100.

Category	OSN 1	OSN 2	SE 1	SE 2	SE 3
Adult	0	0	3	5	1
Apps	6	2	1	2	1
Blog	2	4	1	4	2
Communications	0	0	1	0	1
Employment	0	0	2	0	1
Games	20	0	0	0	0
Maps	0	0	0	0	1
Music	0	1	0	1	0
News	2	10	9	6	11
Organization	1	0	2	3	2
OSN	7	11	7	8	9
Reference	0	1	5	9	9
Search	4	2	5	4	4
Shopping	1	5	9	7	4
Sports	0	5	3	0	2
Unknown	3	0	0	0	0
Video	2	8	2	1	1
Web Services	1	1	0	0	1

Table 2: Categorization of top-50 external domains found in OSN and search engine sessions (SE denotes search engine)

We note a few interesting facts about external domains from Table 2. First, domains related to games figure heavily for OSN 1. No other OSN or search engine has any. Advertisement networks also contribute to main nodes for both OSN and search engine sessions. In fact, they are second only to game-related domains for OSN 1. Websites offering functionality for improving a user's OSN experiment figure prominently for all. News-related websites are more popular for OSN 2 and the three search engines but not OSN 1. Reference and shopping sites dominate search engines and are less common for sessions originating at OSNs. Video websites are more commonly visited from OSN 2.

Table 3 shows the popularity of domains visited from OSNs versus search engines. *We find that OSNs send a significantly smaller portion of their traffic to popular Alexa-100 domains relative to search engines.* Search engines send a greater percentage of their sessions to popular websites than OSNs, but still less than the overall percentage of sessions in our data that visit popular websites.

	% sessions to Alexa top-100
OSN 1	20%
OSN 2	25%
Search engine 1	39%
Search engine 2	34%
Search engine 3	42%
All multi-domain sessions	50%

Table 3: Percentage of sessions navigating to top-100 Alexa domains

5. RELATED WORK

Various aspects of OSNs have received attention from the research community. Traffic characterization, privacy, security, OSN social graph analysis and data collection techniques have been investigated in the recently. In the interest of brevity, we focus on works falling in the area of OSN traffic characterization only since it is the most pertinent for this paper. In the context of search engines, relevance of search results, search engine bias and optimization and security issues have been examined. We focus on works where search engines lead their visitors due to its relevance for our paper.

Works closest in spirit to ours have examined how users interact with OSNs. In a 2009 study, Schneider et al. [4] examined traffic characteristics and dynamics of user interaction with popular OSNs using from real-world traffic logs containing user interactions. They found that 12.5% of total sessions on Facebook lasted longer than an hour. We find that users now interact with OSNs for longer durations. They also found that 7% of user sessions visited other OSN websites. We find this number to be 5.5-13% for our OSNs, indicating that this aspect of OSNs has changed little. While this study has a few similarities with ours, there are key differences. Our paper focuses in depth of the composition of external domains visited during OSN sessions. In [8], the authors examined ways to group user interactions with OSNs into classes of similar behaviors, particularly the production and consumption of content. This work focused on YouTube, which specializes on user interaction primarily via video upload and viewing activities.

Facebook applications have been analyzed from various perspectives. In [9], Nazer et al. created Facebook applications and examined the network costs of these applications. In their previous work [10], they examined the changes in popularity of Facebook applications over time, and how this popularity was distributed among users. In [11], Gjoka et al. used a public crawl of Facebook to produce high-level coarse statistics and usage patterns for third party applications launched from Facebook. In contrast with these works, we focus on domains instead of specific applications.

In [12], Cho et al. examined the impact of search engines on the popularity of web pages. In [13], Fortunato et al. found that search engines avoided a theoretical vicious cycle of popularity, and instead had a tendency to send users to less popular pages than they would have otherwise found. We find that OSNs send their users more often to websites falling in limited number of categories compared to search engines. Also, our data indicates that search engines send their visitors to more popular websites compared to OSNs.

6. CONCLUSION

This paper takes a first look at comparing where users navigate from OSNs versus from search engines. We find that not only does the outgoing traffic for a major OSN compares with that of a major search engine but also that modern OSNs are competing with search engines in their ability to act as conduits between users and websites. However, there are differences in the types of sites OSNs expose to their users, suggesting that these two types of conduits are somewhat complementary. The limited availability of real-world HTTP logs limited our ability to see if our findings hold for other vantage points from other parts of the world. We leave that investigation to future work.

7. REFERENCES

[1] J. Alper and N. Hajaj, "The official Google blog: We knew the Web was big," 2011, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[2] Facebook, "Statistics | Facebook," 2011, http://www.facebook.com/press/info.php?statistics.

[3] L. Rao, “Comscore: Facebook keeps gobbling people’s time,” 2011, http://techcrunch.com/2011/02/07/comscore-facebook-keeps-gobbling-peopl%es-time/.

[4] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, “Understanding online social network usage from a network perspective,” in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2009.

[5] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, “Hypertext transfer protocol – HTTP/1.1,” IETF RFC 2616, 1999.

[6] S. Ihm and V. S. Pai, “Towards understanding modern Web traffic,” in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011.

[7] DMOZ, “Open directory project,” http://www.dmoz.org/.

[8] M. Maia, J. Almeida, and V. Almeida, “Identifying user behavior in online social networks,” in *Workshop on Social Network Systems (SocialNets)*, 2008.

[9] A. Nazer, S. Raza, D. Gupta, C.-N. Chuah, and B. Krishnamurthy, “Network level footprints of facebook applications,” in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2009.

[10] A. Nazir, S. Raza, and C.-N. Chuah, “Unveiling facebook: A measurement study of social network based applications,” in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2008.

[11] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang, “Poking facebook: Characterization of osn applications,” in *ACM Workshop on Online Social Networks (WOSN)*, 2008.

[12] J. Cho and S. Roy, “Impact of search engines on page popularity,” in *International World Wide Web Conference (WWW)*, 2004.

[13] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, “The egalitarian effect of search engines,” in *International World Wide Web Conference (WWW)*, 2006.