Ranking Web Sites with Real User Traffic

Mark R. Meiss[1,2]* Filippo Menczer[1,3] Santo Fortunato[3]
Alessandro Flammini[1] Alessandro Vespignani[1,3]

[1]School of Informatics, Indiana University, Bloomington, IN, USA
[2]Advanced Network Management Lab, Indiana University, Bloomington, IN, USA
[3]Complex Networks Lagrange Laboratory, ISI Foundation, Torino, Italy

ABSTRACT

We analyze the traffic-weighted Web host graph obtained from a large sample of real Web users over about seven months. A number of interesting structural properties are revealed by this complex dynamic network, some in line with the well-studied boolean link host graph and others pointing to important differences. We find that while search is directly involved in a surprisingly small fraction of user clicks, it leads to a much larger fraction of all sites visited. The temporal traffic patterns display strong regularities, with a large portion of future requests being statistically predictable by past ones. Given the importance of topological measures such as PageRank in modeling user navigation, as well as their role in ranking sites for Web search, we use the traffic data to validate the PageRank random surfing model. The ranking obtained by the actual frequency with which a site is visited by users differs significantly from that approximated by the uniform surfing/teleportation behavior modeled by PageRank, especially for the most important sites. To interpret this finding, we consider each of the fundamental assumptions underlying PageRank and show how each is violated by actual user behavior.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation, user issues*

General Terms

Measurement

Keywords

Weighted host graph, Web traffic, ranking, PageRank, search, navigation, teleportation

1. INTRODUCTION

We report on our analysis of Web traffic from a large and representative sample of real users over an extended period of time. To our knowledge this is by far the largest effort to date to study in depth the structure and dynamics of the weighted Web graph, i.e. the network where links are weighted by actual requests of Web users. A first set of contributions of this work concerns a number of intriguing structural properties revealed by the "dynamic" (traffic-weighted) Web graph, and how they compare to those of the "static" Web graph based on unweighted hyperlinks. We further show that temporal traffic patterns show strong regularities, with a significant portion of traffic that is highly predictable — with implications for Web caching schemes.

A second set of contributions concerns applications of our findings to Web search. In particular, ranking Web pages and sites is one of, if not *the* most critical task of any search engine. The last ten years have brought terrific advances in Web search technology, owing in large part to the development of sophisticated ranking techniques. Here we focus on content-independent algorithms, which rank all pages or sites irrespective of the match between their content and user queries. These ranking algorithms are crucial for distilling the most important pages or sites from the potentially large number that match a user query. PageRank has been the most influential such ranking measure, paving the way for major commercial applications such as Google. While modern search engines have likely refined and improved on PageRank, in addition to combining it with many other criteria, it remains a reference tool for the study of the Web as a complex dynamic network, as well as for the engineering of improved ranking functions. Aside from practical advantages such as efficient computation, the strength of PageRank lies in its intuitive interpretation as the stationary distribution of visitation frequency by a modified random walk on the Web link graph — in other words, PageRank is a simple model of Web traffic generated by user navigation. Our Web traffic data makes it possible to explore how well PageRank models user browsing behavior. In particular, we quantify the degree to which the critical assumptions underlying PageRank are invalid, and discuss how these assumptions affect the resulting ranking of Web sites.

Contributions and Outline

In the remainder of this paper, after some background and related work, we describe the source and collection procedures of our Web traffic data; with 1.3×10^{10} requests from about 10^5 users, this data set provides the most accurate

*Corresponding author. Email: mmeiss@indiana.edu

WSDM '08 Stanford, CA

picture to date of human browsing behavior.

Our main findings are organized into three sections, dealing respectively with:

- general and structural properties of the weighted traffic network (§ 4),
- behavioral and temporal patterns uncovered by the observed user dynamics (§ 5), and
- comparative analysis of ranking based on user traffic versus topological PageRank (§ 6).

We conclude with a discussion of the limitations of our data, implications of this work for search applications, and a look to future work.

2. BACKGROUND

Many studies have used Web crawlers to reveal important insights on the large-scale *structure* of the Web graph, such as the "bow-tie" model, the presence of self-similar structures and scale-free distributions, and its small-world topology [3, 11, 1, 17, 16, 35]. While these insights have informed the design of a variety of applications such as crawlers and caching proxy servers, structural analysis has seen its greatest application in ranking pages returned by search engines. In particular, the well-known PageRank [10] and HITS [27] algorithms are able to use the pattern of links connecting pages to rank them without needing to process their contents; these algorithms have inspired a vast amount of research into ranking algorithms based on link structure. The structural properties of the link graph extend to the host graph, which considers the connectivity of entire Web servers rather than individual pages [7].

Researchers have been quick to recognize that structural analysis of the Web can become far more useful when combined with *behavioral* data. Some paths through the Web are used far more heavily than others, and a variety of behavioral data sources exist that can allow researchers to identify these paths and improve Web models accordingly. The earliest efforts have used browser logs to characterize user navigation patterns [12], time spent on pages, bookmark usage, page revisit frequencies, and overlap among user paths [15]. The most direct source of behavioral data comes from the logs of Web servers, which have been used for applications such as personalization [30] and improving caching behavior [37]. Because search engines serve a central role in users' navigation, their log data is particularly useful in improving results based on user behavior [2, 28].

Other researchers have turned to the Internet itself as a source of data on Web behavior. Network flow data generated by routers, which incorporates high-level details of Internet connections without revealing the contents of individual packets, has been used to identify statistical properties of Web user behavior and discriminate peer-to-peer traffic from genuine Web activity [29, 18].

The most detailed source of behavioral data consists of actual Web traffic captured from a running network, as we do here. The present study most closely relates to the work of Qiu *et al.* [33], who used captured HTTP packet traces to investigate a variety of statistical properties of users' browsing behavior, especially the extent on which they appear to rely on search engines in their navigation of the Web. The study presented here involves a much larger user population over a longer period of time, but our deletion of identifying client information prevents us from associating a series of clicks with any particular user or drawing any conclusions as to the duration of individual browsing sessions. We also focus on host-level activity rather than individual URLs in this first phase of our effort.

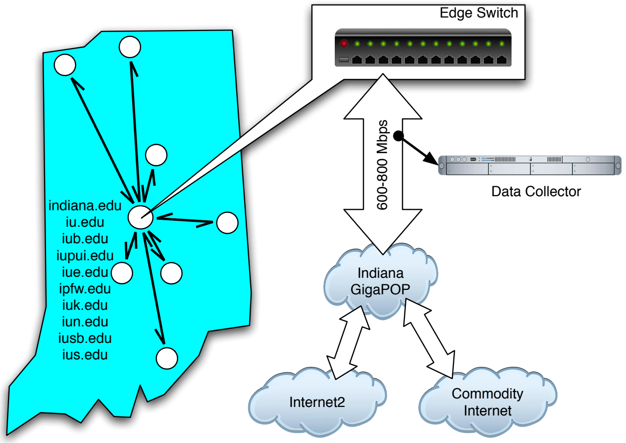

Figure 1: Sketch of Indiana University's Internet connectivity and our experimental setup.

3. DATA DESCRIPTION

3.1 Data Source

The click data we use in this study is gathered by a dedicated FreeBSD server positioned at the edge of the Indiana University network. One of its 1 Gbps Ethernet ports is attached to a switch monitoring port that mirrors all traffic passing between the eight campuses of Indiana University and both Internet2 and the commodity Internet, representing the combined Internet traffic of about 100,000 users. Under normal conditions, we observe a sustained data rate of about 600–800 Mbps on this interface. Fig. 1 illustrates our data collection framework.

To obtain information on individual HTTP requests passing over this interface, we first use a Berkeley Packet Filter to capture only packets destined for TCP port 80. While this eliminates from consideration all Web traffic running on non-standard ports, it does give us access to the largest body of it. We make no attempt to capture or analyze encrypted traffic using TCP port 443. Once we have obtained a packet destined for port 80, we immediately remove all identifying information about the client from the IP and TCP headers, making it impossible to associate the payload data with any particular client system. We then use a regular expression search against the packet payload to determine whether it contains an HTTP GET request.

If we do find an HTTP request, we analyze the packet further to determine the identity of the virtual host contacted, the path requested, the referring host, the advertised identity of the user agent, and whether the request is inbound to or outbound from the university. We then write a record to our raw data files that contains a timestamp, the virtual host, the path requested, the referring host, a flag indicating whether the user agent matches a mainstream browser

(Internet Explorer, Mozilla/Firefox, Safari, or Opera), and a direction flag. We reduce the user agent field to a single flag in order to save disk space: most agent strings are quite long, and we observe well over 10,000 unique agent strings over the course of a day. Most of this analysis is done using a small set of regular expressions; coupled with careful optimization of our network settings, this allows us to record about 30% of all HTTP requests directed to TCP port 80 during peak traffic hours. On a typical weekday, we log around 60 million HTTP requests, the raw records for which require about 6–7 GB of storage.

The most directly comparable source for HTTP request data is that of Alexa, which gathers traffic information based on the surfing activity of several million users of its browser toolbar. However, this traffic information includes only the destinations of HTTP requests, not the identity of the Web server from which a link was followed. Other Internet companies that provide browser toolbars may have more detailed traffic data, but this information is not generally available to researchers and, as with Alexa, includes only users who have opted to install a particular piece of software.

While our collection method allows us to gather a substantial volume of click information from a large and diverse user population, we do recognize several potential disadvantages of our data source. First, the academic community whose traffic we monitor is a biased sample of the population of Web users at large. This is inevitable when collecting traffic data from any public Internet service provider (ISP). The fact that we cannot log clicks at line rate during peak usage hours means that our sampling rate is not uniform throughout the day: we miss many requests during the afternoon and very few in the early morning hours. Because we do not perform stream assembly, we can only analyze HTTP requests that fit in a single 1,500 byte Ethernet frame. While over 98% of all HTTP requests do so, some Web services generate extremely long URLs. Finally, the HTTP referrer field can be spoofed; we make the assumption that the few users at Indiana University who do so generate a small portion of the overall traffic.

3.2 Generation of Host Graph

In principle it is possible to capture the entire URLs of the referring and requested pages with our experimental setup, and to build a weighted link graph with pages as nodes. This is indeed our goal. In this paper, however, we report on an initial stage of the project in which we focus on the host graph. One reason is that this is more feasible with our current storage and computing resources, and indeed necessary to tune our collection and analysis algorithms; another is that the host graph already reveals several interesting insights about Web traffic.

To derive a weighted version of the Web host graph from the raw click data, we first reduce the data into "click lists" containing only the indices of the referring and target servers for each observed HTTP request. These indices are pointers into an external database that contains the fully qualified domain names of the Web servers involved. We generate two sets of these click lists for the raw data: FULL, which includes every HTTP request detected on port 80; and HUMAN, which is a subset of the FULL data set that includes only those requests that are (1) made by a common browser and (2) for URLs that are likely to be actual Web pages (instead of media files, style sheets, etc.).

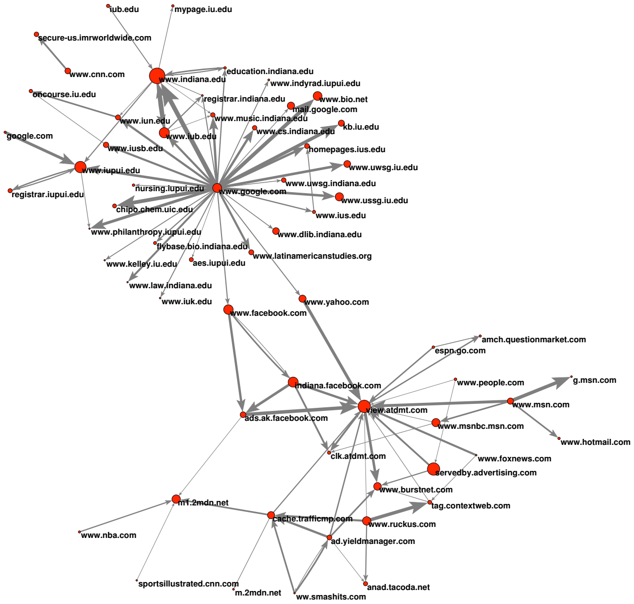

Figure 2: Visualization of the most requested hosts and the most clicked links between them. Node size is proportional to the log of the traffic to each site, and edge thickness is proportional to the log of the number of clicks on links between two sites.

The zero index in our scheme refers to an illusory Web server we call the "empty referrer." This server is identified as the referring host for every HTTP request that does not include referrer information; sources of such requests include bookmarks, browser start pages, mail systems, office applications, clients with privacy extensions, and so forth. It is also identified as the destination host for the small portion of clicks for which we could not identify a virtual host, usually because of old or primitive client software that generates HTTP/0.9 requests.

The click lists represent lists of directed edges in the Web host graph. When we merge a set of these edges, we obtain a subset of the actual Web host graph, weighted according to observed user traffic over a period of time. We are thus able to apply various levels of aggregation to the click lists to generate hourly, daily, monthly, and cumulative versions of the observed host graph. These graphs are stored as sparse connectivity matrices for analysis in Matlab.

4. STRUCTURAL PROPERTIES

The click data was collected over a period of about seven months, from 26 September 2006 to 19 May 2007, with no data collected from 15 to 28 January 2007 and from 1 to 8 April 2007. Fig. 2 offers a view of a small portion of the resulting weighted host graph, consisting of the most popular destination sites and the most clicked links between them. We first report on general properties of this data and on the structure of the weighted host graph.

Table 1 summarizes the dimensions of the click data and host graphs analyzed in this paper. Each human page click involves an average of 14.2 HTTP requests for embedded media files, style sheets, script files, and so on. One notable observation is that a majority of human-generated clicks do

Table 1: Summary statistics of the FULL and HUMAN host graphs.

		FULL		HUMAN	
		Number	Percent	Number	Percent
requests	with empty referrer	2,632,399,381	20.4%	490,290,850	54.0%
	to unknown destination	232,147,862	1.8%	2,078,725	0.2%
	total	12,884,043,440		907,196,059	
hosts	referring	5,151,634	67.8%	2,199,307	54.5%
	destination	7,026,699	92.5%	3,743,074	92.8%
	total	7,595,907		4,031,842	
edges		37,537,685		10,790,759	

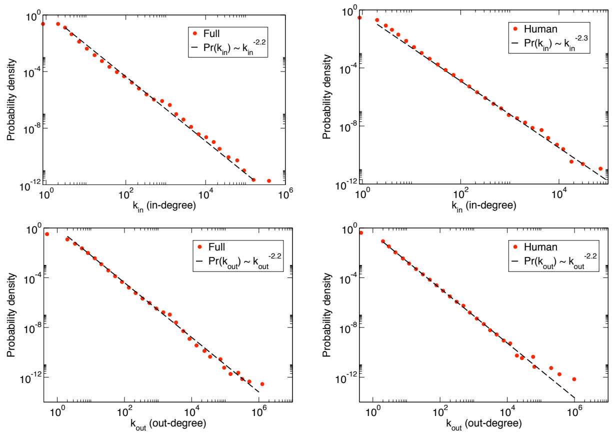

Figure 3: Distributions of in-degree (top) and out-degree (bottom) for the FULL (left) and HUMAN (right) host graphs. In these and the following plots in this paper, power-law distributions are fitted by least-squares regression on log values with log-bin-averaging, and also verified with the maximum likelihood methods and Kolmogorov-Smirnov statistic as proposed by Clauset *et al.* [14].

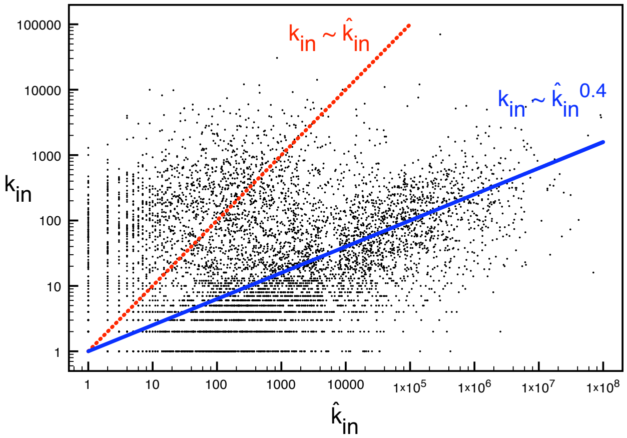

Figure 4: Scatter plot of k_{in} values estimated from the HUMAN host graph versus $\hat{k}_{in}$ values obtained from Yahoo. We show the proportionality line as a reference along with the best power-scaling fit, although a power relationship may not be the best model.

not have a referrer page, meaning that users type the URL directly, click on a bookmark, or click on a link in an email.

The first question about the host graph reconstructed from our sample of traffic is whether it recovers the well-known topological features of the link graphs built from large-scale crawls [3, 11, 17, 35]. The most stable signature of the Web graph is its scale-free in-degree distribution, which many studies consistently report as being well fitted by a power law $\Pr(k_{in}) \sim k_{in}^{-\gamma}$ with exponent $\gamma \approx 2.1$. As shown in Fig. 3, we indeed recover this behavior from the FULL host graph ($\gamma = 2.2 \pm 0.1$); although Web traffic may not follow on every link, it produces a picture of the Web that is topologically consistent with those obtained from large-scale crawls. The power-law in-degree distribution in the HUMAN host graph has a slightly larger exponent $\gamma = 2.3 \pm 0.1$. This hints at an important caveat. While the structure of the traffic-induced and crawler-induced networks may be similar, they are based on very different sampling procedures, each with its own biases. One cannot compare the two networks directly on a node-by-node basis. To illustrate this point, we sampled nodes from the HUMAN graph and compared their in-degree with that given by a search engine (via the Yahoo API). As evident from the scatter plot in Fig. 4, the correlation is weak (Pearson's $\rho = 0.26$ on the log-values), and we cannot assume proportionality. If one conjectures a power-law scaling $k_{in} \sim \hat{k}_{in}^{\eta}$ where $\hat{k}_{in}$ is the in-degree obtained from crawl data, we see that a sublinear bias $\eta < 1$ fits the data better than proportionality $\eta = 1$. While we cannot say that such a power-law scaling is the most appropriate model of the relationship, this does highlight a sample bias whereby the in-degree of popular nodes is underestimated by a greater amount than that of low-degree nodes. The lack of proportionality explains the higher exponent in the power-law distribution of in-degree. Assuming again that k_{in} and $\hat{k}_{in}$ are deterministically related by the power formula conjectured above, it follows immediately that $\Pr(k_{in})dk_{in} = \Pr(\hat{k}_{in})d\hat{k}_{in}$. Therefore

$$\begin{aligned}\Pr(k_{in})dk_{in} &\sim k_{in}^{-\gamma}dk_{in} \sim \hat{k}_{in}^{-\eta\gamma}d(\hat{k}_{in}^{\eta}) \\ &\sim \hat{k}_{in}^{-\eta\gamma+\eta-1}d\hat{k}_{in} \sim \hat{k}_{in}^{-\hat{\gamma}}d\hat{k}_{in}\end{aligned}$$

and thus the k_{in} exponent changes to $\gamma = (\hat{\gamma}-1)/\eta + 1 > \hat{\gamma}$ if $\eta < 1$.

The literature is less consistent about the characterization of the Web's out-degree distribution, for reasons outside the scope of this paper. Our data (Fig. 3) is consistent with a power law distribution $\Pr(k_{out}) \sim k_{in}^{-\gamma}$ with exponent $\gamma \approx 2.2$.

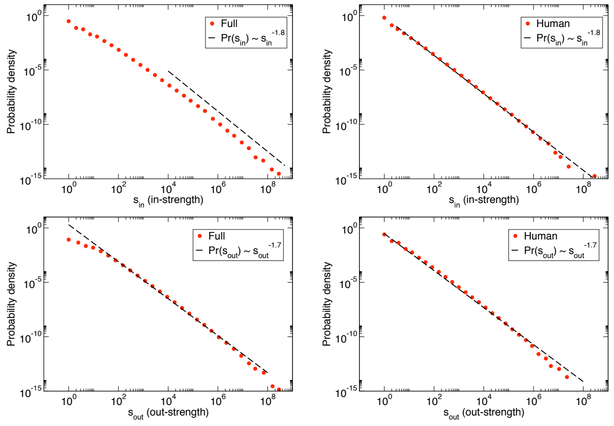

Figure 5: Distributions of in-strength (top) and out-strength (bottom) for the FULL (left) and HUMAN (right) host graphs.

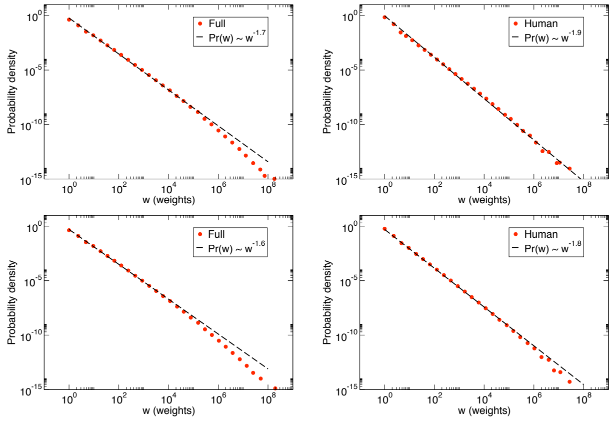

Figure 6: Distributions of weights excluding (top) and including (bottom) requests with empty referrer for the FULL (left) and HUMAN (right) host graphs. Requests with non-empty referrer correspond to clicks from one page to another, whereas an empty referrer may originate from a bookmark or a directly typed URL.

Table 2: Requests by source. The percentages of edges are shown under k_{out} (total out-degree) and the percentages of traffic are shown under s_{out} (total out-strength). For requests with empty source, the percentage of edges is computed by representing these requests as originating from the special "empty referrer" host.

	FULL		HUMAN	
Source	k_{out}	s_{out}	k_{out}	s_{out}
Empty	10.2%	20.4%	14.5%	54.0%
Search	8.2%	2.9%	21.2%	4.9%
WebMail	3.1%	2.0%	1.6%	0.6%
Other	78.5%	74.6%	62.8%	40.4%

The difference between our network representation of the Web host graph and that obtained from crawls, of course, is that we have weighted edges telling how many times links between hosts are clicked. For weighted networks, the notion of degree is generalized to that of *strength,* defined as the sum of the weights over incoming or outgoing links:

$$s_{in}(j) = \sum_i w_{ij} \quad s_{out}(i) = \sum_j w_{ij}$$

where w_{ij} is the weight of edge (i, j), i.e. the number of clicks on the link from host i to host j. Note that because $s_{in}(j)$ represents the total number of times that site j is visited, this is what we refer to by the less formal term *traffic.* Fig. 5 plots the distributions of strength for the host graphs. Not all the curves are best fit by power laws; nevertheless, all distributions are extremely broad, spanning eight orders of magnitude. The portions of the distributions fitted by power laws $\Pr(s) \sim s^{-\gamma}$ yield γ values between 1.7 and 1.8. These exponents $\gamma < 2$ imply that the average strengths diverge as the networks grow, being bounded only by the finite size of the data. Such broad distributions of traffic suggest that the static link graph captures only a portion of the actual heterogeneity of popularity among Web sites.

Finally, in Fig. 6 we plot the distribution of the weights w_{ij} (*link traffic*) across all edges. These, too, are broad distributions over many orders of magnitude, that we can fit to power laws $\Pr(w) \sim w^{-\gamma}$ with exponents γ between 1.6 and 1.9. Such extreme heterogeneity tells us that not all links are created equal: a few carry a disproportionate amount of traffic while most carry very little traffic. This, of course, could simply result from a trivial correlation with the traffic of the originating hosts. In § 6 we discuss the local heterogeneity of traffic across links from individual hosts.

5. TRAFFIC PROPERTIES

5.1 Behavioral Traffic Patterns

The traffic data allows us to address some basic questions about how users navigate the Web. First, to what extent do people wander through pages (surf by following links) versus visiting pages directly (teleport using bookmarks, typing URLs, or by other means)? Table 2 shows the percentages of requests originating from different types of sources. Let us focus on the HUMAN host graph. As already noted in § 4, the majority (54%) of requests have an empty referrer, corresponding to pages visited without clicking on a link. Such a high number suggests that traditional browser bookmarks are still widely used, in spite of the growing popularity of online bookmark managers. On the other hand, all this traffic corresponds to only 14.5% of the edges. This small k_{out}/s_{out} ratio indicates that each additional empty-referrer request is less likely to lead to a destination host that has not been seen before. This is reasonable for sites that are bookmarked or have easily remembered URLs. Fig. 7 shows that while the volume of requests is affected by both the academic calendar and the time of day, the relative ratios of clicks from different sources are fairly stable.

Second, we see from Table 2 that less than 5% of traffic originates from search hosts. This analysis was carried out by matching the DNS names of referring hosts against a list of common search engines, including Google, Yahoo!, MSN, Altavista, and Ask. Such a low percentage is some-

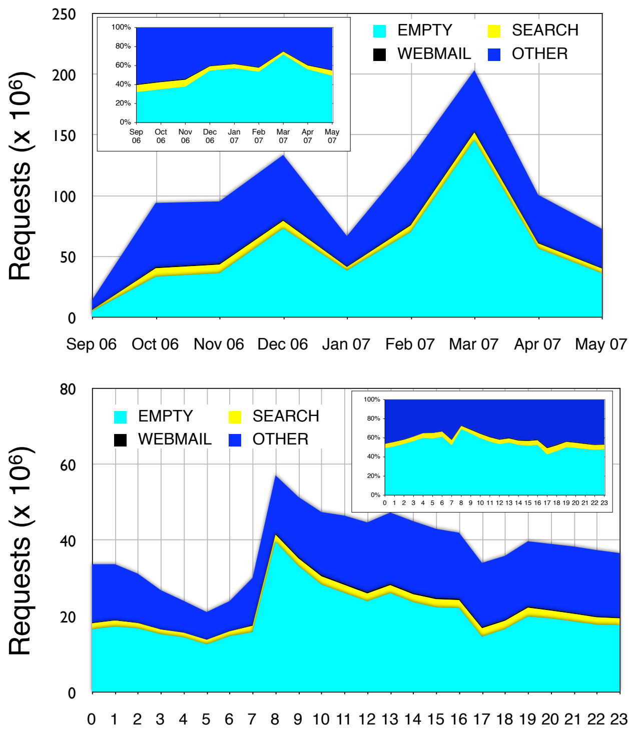

Figure 7: Volume of requests from various sources in the HUMAN host graph. Top: seasonal variations, aggregated by month. Bottom: daily variations, aggregated by hour. The insets plot percentages of request sources (total strength). The monthly and hourly ratios of total degree by source (not shown) are even more stable.

what surprising when one considers the wide impact generally attributed to search engines in steering Web traffic. Of course, our measure is a lower bound of the influence of search engines, since it only monitors requests that are *directly* generated by search; successive clicks appear as regular navigation, even if the path was initiated by a search. (Our disposal of identifying client information makes it impossible to recover chains of clicks.) Nevertheless, we expected a higher percentage of clicks originating from search engines. Another notable statistic is the much higher fraction (21.2%) of edges corresponding to these clicks. The high k_{out}/s_{out} ratio suggests that each search click is more likely than others to lead to a new host. In other words, search engines promote the exploration of unvisited sites, an "egalitarian" effect that is in agreement with earlier findings [33, 21].

Another way to use our data to measure the impact of search on Web navigation is to inspect how traffic scales with in-degree. Earlier literature conjectured that due to the use of PageRank by search engines, established sites would attract a disproportionate fraction of traffic [5, 25, 31, 13], corresponding to a superlinear scaling of search traffic with in-degree:

$$s_{in} \sim k_{in}^{\beta}$$

with $\beta > 1$. In contrast, a random walk would yield a linear scaling ($\beta = 1$). Preliminary analysis of our data yields a trend that is slightly above (though not statistically different from) the linear model of a random walk ($\beta \gtrsim 1$). This is consistent with the observed distributions of s_{in} and k_{in}; since both are power laws with exponents γ_s and γ_k respectively, if the two are related by a power relationship, we must have $\Pr(s)ds = \Pr(k_{in})dk_{in}$, and this leads to

$$\begin{aligned} s_{in}^{-\gamma_s} ds &\sim (k_{in}^{\beta})^{-\gamma_s} d(k_{in}^{\beta}) \sim k_{in}^{-\beta\gamma_s+\beta-1} dk_{in} \sim k_{in}^{-\gamma_k} \\ \beta &\approx (\gamma_k - 1)/(\gamma_s - 1). \end{aligned}$$

Considering the errors on the fitted parameter, the empirical values of β, γ_s and γ_k are consistent. However, this finding $\beta \gtrsim 1$ would appear to disagree with our own prior measurements based on traffic data from Alexa, where a sublinear scaling fitted the data better, suggesting that search engines would *mitigate* the "rich-get-richer" dynamics of the Web [21]. In fact, there is no contradiction when one considers the in-degree sampling bias discussed in § 4; this bias affects the relationship between traffic and in-degree. For example, conjecturing again the power scaling $k_{in} \sim \hat{k}_{in}^{\eta}$, one would find $s_{in} \sim k_{in}^{\beta} \sim \hat{k}_{in}^{\eta\beta}$. If $\eta < 1/\beta$ (as in Fig. 4), then $\eta\beta < 1$, i.e. one recovers a sublinear scaling of traffic with the crawl-based in-degree $\hat{k}_{in}$. The traffic data is therefore consistent with our prior empirical finding, yet we cannot say much about the impact of search engines on this trend, since search directs such a small percentage of the overall traffic in our data.

5.2 Temporal Traffic Patterns

With the timestamp information in our Web request data, we can look at the predictability of traffic over time. Using the host graph to predict future traffic has potential applications for Web cache refreshing algorithms, capacity allocation, and site design. For example, if an ISP knows that a certain news page is regularly accessed every morning at 10am, it can preload it into a proxy server. Likewise, knowledge about regular spikes in traffic can guide provisioning decisions. Finally, site owners can adapt sites so that content can be made most easily accessible based on its predicted demand at different times or days.

As a first, crude analysis on the predictability of Web request patterns, let us use the simple precision and recall measures, which are well-established in information retrieval and machine learning. The goal is to predict the host graph for time interval t using a snapshot of the host graph at time interval $t - \delta$, as a function of the delay δ. Given a weighted network representation of the host graph, where an edge $w_{ij}(t)$ stands for the number of clicks from host i to host j in time interval t, we define *generalized temporal precision* and *generalized temporal recall* based on true positive click predictions:

$$\begin{aligned} P(\delta) &= \left\langle \frac{\sum_{ij} \min[w_{ij}(t), w_{ij}(t-\delta)]}{\sum_{ij} w_{ij}(t-\delta)} \right\rangle_{t\in[\delta,T]} \\ R(\delta) &= \left\langle \frac{\sum_{ij} \min[w_{ij}(t), w_{ij}(t-\delta)]}{\sum_{ij} w_{ij}(t)} \right\rangle_{t\in[\delta,T]} \end{aligned}$$

where the averages $\langle\cdot\rangle$ run over hourly snapshots. Our analysis is based on $\delta_M T$ comparisons of hourly snapshots of the network, where $\delta_M = 168$ hours is the maximum delay considered (one week) and $T = 4996$ is the total number of hourly snapshots. The baseline for these measures is sta-

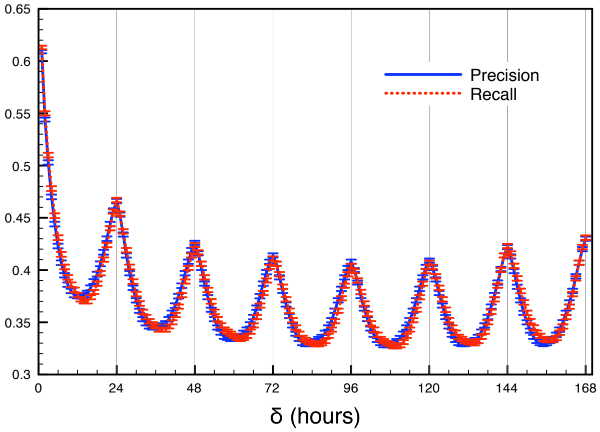

Figure 8: Average temporal precision and recall as a function of delay δ for the HUMAN host graph. Error bars correspond to ±1 standard error.

tionary traffic; if the host graph does not change, perfect predictability is obtained and $P = R = 1$ for any δ.

Fig. 8 plots generalized temporal precision and recall versus delay for HUMAN clicks. As one would expect, predictability decays rapidly; however, both precision and recall are quite high (above 50%) for $\delta \leq 3$ hours. We observe very strong daily and weekly cycles; after more than 4 hours, the requests from the prior day at the same time yield higher precision and recall. Even after going back two or more days, one can predict more than 40% of the clicks, which yields better performance than using data from 10-12 hours earlier. We also observe a large volume of stationary data, as suggested by the fact that P and R never fall below 32%. Precision and recall track each other closely, which serves as further evidence of a large volume of stationary traffic. For example, 47% of clicks at any given time are predicted by the clicks from the previous day at the same time, and the same percentage are repeated the next day at the same time.

The FULL host graph has almost identical trends, except that both precision and recall are about 10% higher. This may be due to the higher predictability of crawler traffic, as well as to commonly embedded files such as style sheets and images.

6. REAL VS. RANDOM SURFING

In this section we address the question, *How good is PageRank as a model of Web navigation?* Or, more specifically, *How well does the ranking of Web sites produced by PageRank approximate that obtained from actual Web user traffic?* Content-independent ranking is of course critical for search engines, so that the most important pages that match a query can be brought to the user's attention. Yet PageRank's importance goes beyond its search applications; this topological network measure remains a key tool in studying the structure of large information networks — the Web being, of course, the premier example — as well as the reference model for the dynamic behavior of the many people who forage for information in these networks.

PageRank has several interpretations, ranging from linear algebra to spectral theory, and many implementation issues. Here we focus solely on the intuitive interpretation of PageRank as the stationary distribution of visit frequency by a modified random walk on the Web link graph, i.e., as a simple model of the traffic flow produced by Web navigation. Formally, PageRank is the solution of a set of linear equations:

$$PR(j) = \frac{\alpha}{N} + (1-\alpha) \sum_{i:w_{ij}\neq 0} \frac{PR(i)}{k_{out}(i)}$$

where N is the number of nodes (Web pages, or, in the current context, sites) and α is a so-called *teleportation* factor (a.k.a. *damping* or *jumping* factor). The first term describes the process by which a user stops browsing at some random node and jumps (teleports) to some other random node. The second term describes a uniform random walk (surfing) across links, with the sum running over the incoming links of node j. The parameter α models the relative probabilities of surfing versus teleporting. We have already discussed in § 4 and 5 that our empirical data would support a higher teleportation probability than the customary value $\alpha = 0.15$ [10]: $\alpha \approx 0.54$ for human browsers, or $\alpha \approx 0.2$ even including crawlers. Other studies have addressed the role of α in PageRank [9, 19]; we use the customary value $\alpha = 0.15$ in the present PageRank calculations.

Aside from the teleportation factor, the interpretation of PageRank as a graph navigation model is based on three fundamental assumptions that are implicit in the above definition:

1. Equal probability of following each link from any given node: $\forall i, j : \Pr(i \to j|\text{click}) = \begin{cases} 1/k_{out}(i) & w_{ij} > 0 \\ 0 & \text{otherwise;} \end{cases}$
2. Equal probability of teleporting *to* each of the nodes: $\sum_i \Pr(i \curvearrowright j|\text{jump}) = 1/N$.
3. Equal probability of teleporting *from* each of the nodes: $\sum_j \Pr(i \curvearrowright j|\text{jump}) = 1/N$;

We can now compare the traffic predicted by the PageRank model on the host graph with the actual traffic flow generated by our sample of users, and captured by the in-strength $s_{in}(j)$ for each host j. This provides indirect validation of the above assumptions.

6.1 Rank Correlations

Since the primary use of PageRank is to rank sites, we focus on the ranking obtained by PageRank rather than on the actual values of the PageRank vector. To compare rankings of Web sites according to two criteria (e.g. PageRank vs. actual traffic), we use the established Kendall's τ rank correlation coefficient [26], which is intuitively defined from the fraction of pairs whose relative positions are concordant in the two rankings:

$$\tau_b = \frac{4C}{N(N-1)} - 1$$

where C is the number of concordant pairs and the subscript b (dropped henceforth) refers to the method of handling ties. Values range from 1 (perfect agreement) to -1 (perfect inversion), and $\tau = 0$ indicates the absence of correlation. We compute τ efficiently with Knight's $O(N \log N)$ algorithm in the manner implemented by Boldi *et al.* [8].

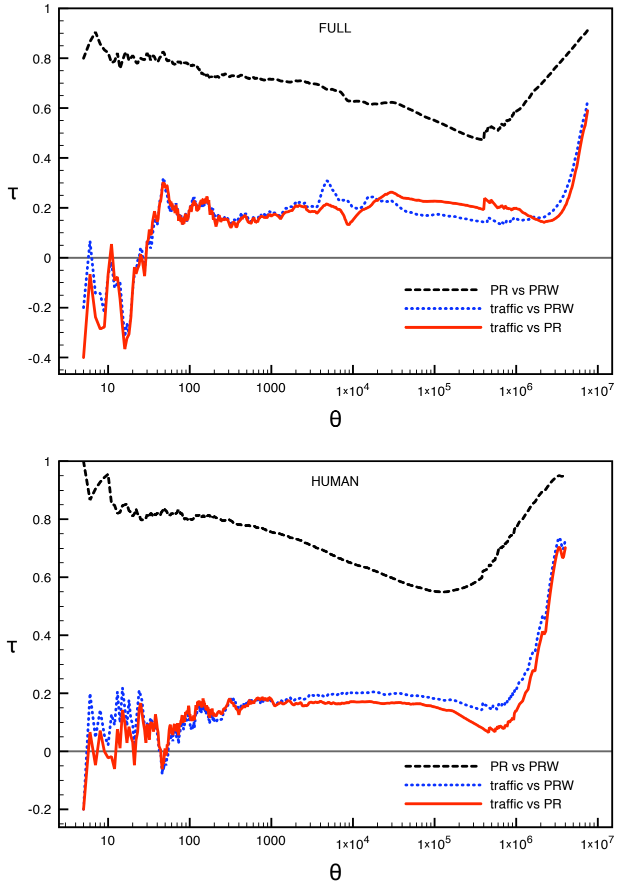

Figure 9: Kendall's τ correlations between different rankings of the sites in the FULL (top) and HUMAN (bottom) host graph, versus traffic rank threshold θ.

Fig. 9 plots rank correlations for subsets of θ top-traffic hosts. Let us focus on the HUMAN host graph; we can see the correlations from small sets of most popular sites, all the way to the entire 4-million-host network. Let us first consider the correlation between the ranking estimated by PageRank (PR) and that obtained from the empirical traffic data (s_{in}). We see that as more low-traffic hosts are added, the correlation increases up to almost 0.7. Low-traffic sites dominate and are mostly tied at the bottom, driving up the correlation. For the top sites, however, the correlation is weak ($\tau < 0.2$ up to a million hosts or so). This is consistent with earlier traffic data from the Polish Web [36]. Surprisingly, PageRank is quite a poor predictor of traffic ranks for the most popular portion of the Web.

To tease out the factors that contribute to the low correlation, we ranked the hosts according to a third, intermediate measure between empirical traffic and PageRank: let us define *weighted PageRank* by plugging the empirical link weights into the PageRank expression:

$$PRW(j) = \frac{\alpha}{N} + (1-\alpha) \sum_{i:w_{ij}\neq 0} \frac{w_{ij}}{s_{out}(i)} PRW(i).$$

As shown in Fig. 9, weighted PageRank is only a slightly better predictor of traffic, and is better correlated with PageRank than with traffic. This suggests that the errors in PageRank ranking are dominated by violations of assumptions 2 and 3 about the random teleportation model.

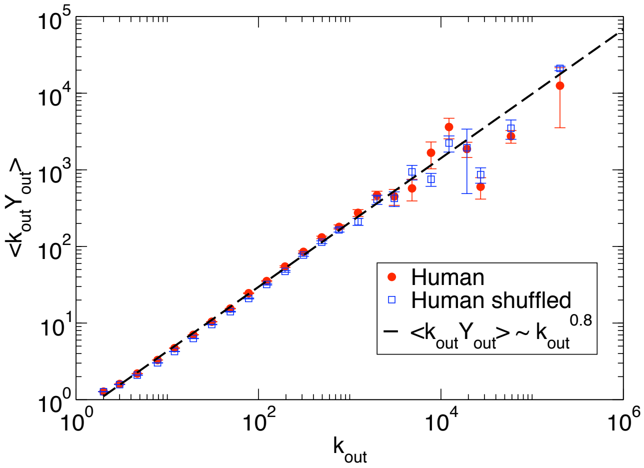

Figure 10: Average behavior of $k_{out}Y(k_{out})$ versus k_{out} from the HUMAN host graph. $\langle Y(k_{out})\rangle$ vales are obtained averaging Y_i across nodes i grouped by out-degree in logarithmic bins. Error bars correspond to ±1 standard error on the bin averages. We also plot the same measure for a shuffled version of the HUMAN host graph, in which the link weights have been randomly reassigned across all edges.

6.2 Non-Uniform Distributions

To better understand how the PageRank model assumptions affect the ranking of Web sites, we can consider each hypothesis directly in an attempt to quantify the degree to which it is supported by the data.

Assumption 1 is about *local* homogeneity of link weights. Note that we have already seen in § 4 that link weights are very *globally* heterogeneous; here we look instead at the links from each individual node, i.e., whether surfers are equally likely to click on any of the links from a given site. A local heterogeneity implies that only a few links carry the biggest proportion of the clicks. Such a heterogeneity would define specific pathways within the host graph that accumulate most of the total traffic. In order to assess the effect of inhomogeneities at the local level, for each host i we calculate

$$Y_i = \sum_j \left(\frac{w_{ij}}{s_{out}(i)} \right)^2 .$$

The function Y_i is known as the Herfindahl-Hirschman index and extensively used in economics as a standard indicator of market concentration [23, 24]; it is also known as a disparity measure in the complex networks literature [6, 4]. Y_i as a function of out-degree $k_{out}(i)$ characterizes the level of local heterogeneity among the links from i. If all weights emanating from a node are of the same magnitude, the quantity $k_{out}Y(k_{out})$ scales as a constant independently of k_{out}, whereas this quantity grows with k_{out} if the local traffic is heterogeneously organized with a few links dominating. Increasing deviations from the constant behavior therefore signal local heterogeneity, where traffic from a site is progressively focused on a small number of links, with the remaining edges carrying just a small fraction of the clicks.

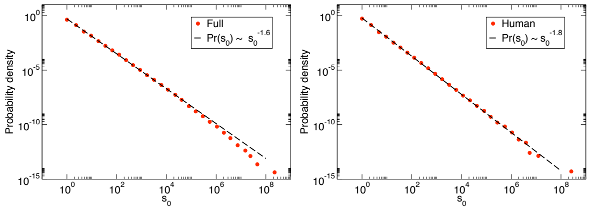

Figure 11: Distribution of requests with empty referrer for FULL (left) and HUMAN (right) host graphs.

The fit in Fig. 10 shows that the traffic follows the scaling law $k_{out}Y(k_{out}) \sim k_{out}^{\lambda}$ with $\lambda \approx 0.8$. This represents an intermediate behavior between the two extreme cases of perfect homogeneity ($\lambda = 0$) and heterogeneity ($\lambda = 1$ if all traffic from a node goes through a single link). The picture is therefore consistent with the existence of major pathways whereby most traffic enters a site from its major incoming links and leaves it through its major outgoing links (see Fig. 2). However, such local heterogeneity is to be expected given the broad distribution of weights (see Fig. 6). In fact, the same scaling behavior of $k_{out}Y(k_{out})$ is observed when shuffling the weights, as also shown in Fig. 10. This suggests that the local link weight heterogeneity is mainly a reflection of accidental local correlations between highly diverse weights. In this light one can interpret the observed correlation between traditional and weighted PageRank (Fig. 9): local weight diversity does not explain much of the difference between PageRank and traffic.

Assumption 2 is about homogeneity of teleportation destinations, that is, whether all sites are equally likely to be the starting points of surfing paths. For each host i we denote by $s_0(i)$ the number of jumps to i, i.e. the number of requests that have i as the target and an empty referrer. This is a direct measure for the probability that a site is a starting point for surfing. Fig. 11 plots the distributions of s_0, showing a very broad power law distribution with exponent between 1.6 (for the FULL host graph) and 1.8 (for the HUMAN host graph). The exponent below 2 implies that both the variance and the mean of the distribution diverge in the limit of large graphs, and are bounded only by the finite size of the data. This result violates PageRank's homogeneous teleportation assumption, which would manifest itself in a narrow distribution, and helps to explain PageRank's low correlation with traffic. Intuitively, people are much more likely to jump to a few very popular sites than to the great majority of other sites.

Assumption 3 is about homogeneity of teleportation sources, that is, whether all sites are equally likely to be end-points for sequences of surfing clicks and jumping points to new paths. For each host i, $s_{in}(i)$ is the number of arrivals into i (requests having i as the target) and $s_{out}(i)$ is the number of departures from i (requests having i as referrer). The strength differential $s_{in}(i) - s_{out}(i)$ is not the same as the number of paths that have terminated at i, because multiple paths can start from i, for instance when users hit the back button or follow multiple links in different browser tabs; cached pages do not generate new requests. For this reason the very nature of traffic data does not allow us to validate assumption 3 directly. However, we can use $s_{out}(i)/s_{in}(i)$ to measure the likelihood that traffic into i leaves i by clicking on links from i. Note that this "hubness" measure is not a probability; in fact we can have $s_{out}(i)/s_{in}(i) \gg 1$ due to multiple traffic paths from i. Yet, the larger $s_{out}(i)/s_{in}(i)$, the more i is likely to be a starting hub, and the less it is likely to be a teleportation source (or surfing sink). Fig. 12 plots the distributions of s_{out}/s_{in}. We observe a very broad distribution, with an initial plateau in the regime where $s_{out} < s_{in}$ followed by a power law decay for $s_{out} > s_{in}$. The central peak corresponds to sites where traffic is conserved ($s_{out} = s_{in}$). While this result is not a direct check on the validity of assumption 3, it shows that people follow many more links from a few very popular hubs than from the great majority of less popular sites, helping to further explain the low correlation between PageRank and traffic rankings. The two clearly demarcated regimes lead us to speculate on the possibility of using the s_{out}/s_{in} ratio as a topology-independent criterion to identify hubs.

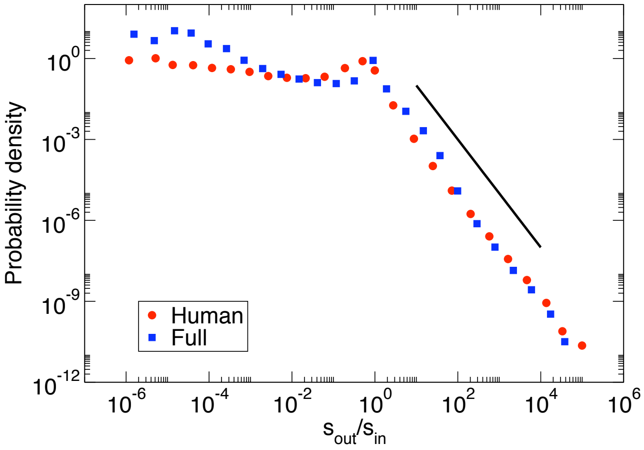

Figure 12: Distribution of the ratio of outgoing to incoming strength for FULL and HUMAN host graphs. The traffic ratio is very large for popular hubs from which users follow many links. A power law trend with exponent 2 is included as a guide to the eye.

6.3 Discussion

Having found that such a large fraction of traffic is driven by teleportation —starting for example from bookmarks or default home pages— rather than hyperlinks, and that this process is not captured well by uniform random jumps, an important question is how to better model teleportation. What are the preferred starting points of our navigation? The analyses in § 6.2 tell us that strong preferences exist leading to scale-free distributions, but do not say anything about what the preferences are.

A first step toward this inquiry is to see if the probability of starting from a site is correlated with the probability of arriving to the same site through navigation. Fig. 13 shows that indeed there is a very strong correlation between traffic to a site through navigation ($s_{in} - s_0$) and traffic to the same site from the empty referrer (s_0). The two are almost linearly related (see fit in Fig. 13). Therefore, the pages from which people start their browsing tend to be the same as those where they are likely to end up — there is a single notion of popularity.

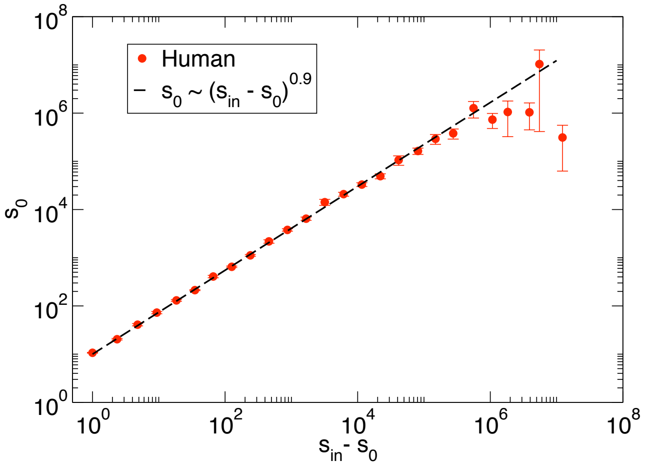

Figure 13: Correlation plot between traffic from empty referrer and traffic from navigation in the HUMAN host graph. For better visualization, we average s_0 values within logarithmic bins in the $s_{in} - s_0$ axis. Error bars correspond to ±1 standard error on the bin averages.

7. CONCLUSIONS

In this paper we have reported on our first analysis of the host graph constructed from a large collection of Web clicks. Our data set provides the most accurate picture to date of human browsing behavior and the largest-scale monitoring effort to date in terms of size of user sample, temporal duration, and amount of Web traffic captured. The data reveals that the dynamic network of Web traffic is even more heterogeneous than the static link graph previously studied through crawl data. Not only in-degree and out-degree, but also site-level incoming and outgoing traffic, as well as link traffic, exhibit scale-free distributions with remarkably broad tails.

The analysis reveals a few surprises. First, much more of the traffic than anticipated (more than half of human requests) is generated not from clicking on links, but from bookmarks, default pages, or direct typing of Web addresses. Second, search engines direct a surprisingly small fraction of traffic (less than 5% of human requests). However, they lead to a larger fraction of the sites visited. Third, the temporal traffic patterns are more predictable than we expected; much less surprising are the very strong cyclic regularities exhibited on daily and weekly bases. The latter findings may have implications for the design of improved proxy and browser caching techniques.

The traffic data has also allowed us to validate PageRank as a model of Web navigation, along with its random walk and random teleportation assumptions. PageRank ranks sites very differently than actual human traffic, especially for the most important hosts. This finding is interpreted in light of our empirical analysis, showing how each of the random behavior assumptions underlying PageRank is violated: not all links from a site are followed equally, but even more importantly, some sites are much more likely than others to be the starting or ending points of surfing sessions. From an application perspective, this suggests that Web traffic data available to an Internet Service Provider (or Autonomous System) could be used to induce a ranking measure over all sites to better reflect their relative importance according to the dynamic behavior of the population of Web users [34]. Search engines could form partnerships with ISPs to explore the potential benefit of integrating traffic data into ranking algorithms. Alternatively, one could consider variations of PageRank in which the teleportation process is modeled according to the empirical traffic data. However, such steps are likely to amplify the search bias toward already popular sites [21].

Aside from the limitations of our data source discussed in § 3, an important bias emerged from our analysis, namely how user traffic samples the Web graph. We have shown that the bias exists and is likely to be strong, although further work is needed to better understand its nature. One consequence is that the topological portion of the traffic-induced host network cannot be directly compared with the link graphs obtained from Web crawlers, which have different types of bias [22, 32]. The bias also affects the measure of how traffic scales with in-degree, and the comparison of such measure with earlier work based on crawler data.

While search traffic could not be separated from surf traffic in the data collected from sources such as Alexa, this is possible with our data. However, at present the small percentage of search traffic precludes a meaningful analysis. We are in the process of collecting more data and will in the future use it to study the possible biases of search in directing user traffic. To this end, we will also need to collect information about search queries from HTTP requests, enabling us to dissect the roles of query generality, search engine ranking algorithms, and user interface issues in shaping search traffic [21]. The study of search bias, which is also critical for Web growth modeling [20], provides further motivation to better understand link sampling bias.

Finally, we plan to extend our analysis from the host graph to the page graph. The increased resolution may be key for a better insight into user browsing behavior, topology-based ranking algorithms, the role of search in Web navigation, and Web evolution modeling.

Acknowledgments

We are grateful to José Ramasco, Tamas Sarlos, Bo Pang, and four anonymous reviewers for their helpful suggestions. MM acknowledges support from Indiana University's Advanced Network Management Lab and University Information Technology Services. SF, FM, and AV acknowledge funding from the Institute for Scientific Interchange Foundation in Torino, Italy. FM, AF, and AV thank the IU School of Informatics for support. Finally FM and AV were funded in part by NSF awards 0348940 and 0513650, respectively.

8. REFERENCES

[1] L. Adamic and B. Huberman. Power-law distribution of the World Wide Web. *Science*, 287:2115, 2000.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving Web search ranking by incorporating user behavior information. In *Proc. 29th ACM SIGIR Conf.*, 2006.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World Wide Web. *Nature*, 401(6749):130–131, 1999.

[4] E. Almaas, B. Kovacs, T. Vicsek, Z. N. Oltvai, and A.-L. Barabasi. Global organization of metabolic

fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843, 2004.

[5] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web structure, dynamics and page quality. In A. H. F. Laender and A. L. Oliveira, editors, *Proc. 9th Intl. Symp. on String Processing and Information Retrieval (SPIRE 2002)*, volume 2476 of *Lecture Notes in Computer Science*, pages 117–130. Springer, 2002.

[6] M. Barthelemy, B. Gondranb, and E. Guichardc. Spatial structure of the internet traffic. *Physica A*, 319:633–642, March 2003.

[7] K. Bharat, B.-W. Chang, M. Kenzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of First IEEE International Conference on Data Mining (ICDM'01)*, 2001.

[8] P. Boldi, M. Santini, and S. Vigna. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. *Internet Mathematics*, 2(3):387–404, 2005.

[9] P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 557–566, New York, NY, USA, 2005. ACM Press.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[11] A. Broder, S. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1–6):309–320, 2000.

[12] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[13] J. Cho and S. Roy. Impact of search engines on page popularity. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, *Proc. 13th intl. conf. on World Wide Web*, pages 20–29. ACM, 2004.

[14] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. Technical report, arXiv:0706.1062v1 [physics.data-an], 2007.

[15] A. Cockburn and B. McKenzie. What do Web users do? An empirical analysis of Web use. *Intl. Journal of Human-Computer Studies*, 54(6):903–922, 2001.

[16] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[17] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *Eur. Phys. J. B*, 38:239–243, 2004.

[18] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson. Identifying and discriminating between web and peer-to-peer traffic in the network core. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 883–892, New York, NY, USA, 2007. ACM Press.

[19] S. Fortunato and A. Flammini. Random walks on directed networks: the case of pagerank. *International Journal of Bifurcation and Chaos*, 2007. Forthcoming.

[20] S. Fortunato, A. Flammini, and F. Menczer. Scale-free network growth by ranking. *Phys. Rev. Lett.*, 96(21):218701, 2006.

[21] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proc. Natl. Acad. Sci. USA*, 103(34):12684–12689, 2006.

[22] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *Proc. 9th International World Wide Web Conference*, 2000.

[23] O. Herfindahl. *Copper Costs and Prices: 1870-1957.* John Hopkins University Press, Baltimore, MD, 1959.

[24] A. Hirschman. The paternity of an index. *American Economic Review*, 54(5):761–762, 1964.

[25] L. Introna and H. Nissenbaum. Defining the web: The politics of search engines. *IEEE Computer*, 33(1):54–62, January 2000.

[26] M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.

[27] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[28] J. Luxenburger and G. Weikum. *Query-Log Based Authority Analysis for Web Information Search*, volume 3306 of *Lecture Notes in Computer Science*, pages 90–101. Springer Berlin / Heidelberg, 2004.

[29] M. Meiss, F. Menczer, and A. Vespignani. On the lack of typical behavior in the global Web traffic network. In *Proc. 14th International World Wide Web Conference*, pages 510–518, 2005.

[30] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):141–151, 2000.

[31] A. Mowshowitz and A. Kawaguchi. Bias on the Web. *Commun. ACM*, 45(9):56–60, 2002.

[32] M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. In *Proc. 10th International World Wide Web Conference*, 2001.

[33] F. Qiu, Z. Liu, and J. Cho. Analysis of user web traffic with a focus on search activities. In A. Doan, F. Neven, R. McCann, and G. J. Bex, editors, *Proc. 8th International Workshop on the Web and Databases (WebDB)*, pages 103–108, 2005.

[34] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: machine learning for static ranking. In *Proc. 15th International World Wide Web Conference*, pages 707–715, New York, NY, USA, 2006. ACM.

[35] M. A. Serrano, A. Maguitman, M. Boguna, S. Fortunato, and A. Vespignani. Decoding the structure of the WWW: A comparative analysis of Web crawls. *ACM Trans. Web*, 1(2):10, 2007.

[36] M. Sydow. Can link analysis tell us about web traffic? In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 954–955, New York, NY, USA, 2005. ACM.

[37] Q. Yang and H. H. Zhang. Web-log mining for predictive web caching. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):1050–1053, 2003.