# Modeling Traffic on the Web Graph

Mark R. Meiss[1,3], Bruno Gonçalves[1,2,3], José J. Ramasco[4], Alessandro Flammini[1,2], and Filippo Menczer[1,2,3,4]

[1] School of Informatics and Computing, Indiana University, Bloomington, USA
[2] Center for Complex Networks and Systems Research, Indiana University, Bloomington, USA
[3] Pervasive Technology Institute, Indiana University, Bloomington, USA
[4] Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Turin, Italy

**Abstract.** Analysis of aggregate and individual Web requests shows that PageRank is a poor predictor of traffic. We use empirical data to characterize properties of Web traffic not reproduced by Markovian models, including both aggregate statistics such as page and link traffic, and individual statistics such as entropy and session size. As no current model reconciles all of these observations, we present an agent-based model that explains them through realistic browsing behaviors: (1) revisiting bookmarked pages; (2) backtracking; and (3) seeking out novel pages of topical interest. The resulting model can reproduce the behaviors we observe in empirical data, especially heterogeneous session lengths, reconciling the narrowly focused browsing patterns of individual users with the extreme variance in aggregate traffic measurements. We can thereby identify a few salient features that are necessary and sufficient to interpret Web traffic data. Beyond the descriptive and explanatory power of our model, these results may lead to improvements in Web applications such as search and crawling.

## 1 Introduction

PageRank [6] has been a remarkably influential model of Web browsing, framing it as random surfing activity. The measurement of large volumes of Web traffic enables systematic testing of PageRank's underlying assumptions [22]. Traffic patterns aggregated across users reveal that some of its key assumptions—uniform random distributions for walk and teleportation—are widely violated, making PageRank a poor predictor of traffic, despite its standard interpretation as a stationary visit frequency. This raises the question of how to design a more accurate navigation model. We expand on our previous empirical analysis [22,20] by considering also *individual* traffic patterns [15]. Our results provide further evidence for the limits of Markovian traffic models such as PageRank and suggest the need for an *agent-based* model with features such as memory and topicality that can account for both individual and aggregate traffic patterns.

Models of user browsing have important practical applications. Traffic clearly has a direct impact on the financial success of companies and institutions. Indirectly, understanding traffic patterns aids in advertising, both to predict revenues and establish rates [12]. Second, realistic models of Web navigation can guide the behavior of crawling algorithms, improving search engines' coverage of important sites [9,24]. Finally, improved traffic models may lead to enhanced ranking algorithms [6,28,18].

R. Kumar and D. Sivakumar (Eds.): WAW 2010, LNCS 6516, pp. 50–61, 2010.
© Springer-Verlag Berlin Heidelberg 2010

After background material, we describe in § 3 a data set collected through a field study of over 1,000 users at Indiana University. In § 4 we introduce an agent-based navigation model, *ABC*, with three key, realistic ingredients: (1) *bookmarks* used as teleportation targets, defining boundaries between sessions and capturing the diversity of starting pages; (2) a *back button* is used to model branching observed in empirical traffic; and (3) *topical interests* drive an agent's decision to continue browsing, leading to diverse session sizes. The model also considers the *topical locality* of the Web, so that interesting pages tend to link to other such pages. In § 5 we compare the traffic generated by our model with measurements of both *aggregate* and *individual* Web traffic data. These results allow us to identify features that are necessary and sufficient to interpret Web traffic data.

## 2 Background

Empirical studies of Web traffic have most often involved the analysis of server logs, with scales ranging from small samples of users from a few selected servers [17] to large groups from the logs of large organizations [15]. This approach has the advantage of distinguishing users by IP address (even if they may be anonymized), thus capturing *individual* traffic patterns [15]. However, the choice of target server will bias both the user sample and the part of the Web graph observed. Browser toolbars offer another source of traffic data; these gather information based on the activity of many users. While toolbars involve a larger population, their data are still biased toward users who opt to install such software. Moreover, their data are not generally available to researchers. Adar *et al.* [1] used this approach to study patterns of page revisitation. A related approach is to have a select panel of users install tracking software, which can eliminate many biases but incur experimental costs. Such an approach has been used to describe exploratory browsing [4]. These studies did not propose models to explain observed traffic patterns.

Our own study measures traffic directly through packet capture, an approach adopted by Qiu *et al.* [27], who used captured HTTP packet traces from the UCLA CS department to study the influence of search engines on browsing. We use a larger sample of residential users, reducing the biases attendant to a workplace study of technical users. We focus strongly on the analysis of browsing sessions. A common assumption is that long pauses correspond to breaks between sessions, leading many to rely on timeouts as a way of defining sessions. Flaws in this technique motivated our definition of time-independent *logical sessions,* based on the reconstruction of session trees rooted at pages requested without a referrer [20]. One goal of our model is to explain the broad distributions of size and depth for these logical sessions. The role of page content in driving users' browsing patterns has received relatively little attention, with the notable exception of a study of the correlation between changes in page content and revisit patterns [2].

Under a basic model of Web navigation, users perform a random walk through pages in the Web graph. PageRank [6] is a random walk modified by teleportation, which uses uniformly random starting points to model how users start new sessions. This Markovian process has no memory or backtracking and no notion of user interests or page content. The stationary distribution of visitation frequency generated by PageRank constitutes a prediction of actual click rates, which can then be compared with

empirical traffic data. We have shown that the assumptions underlying PageRank—uniform link selection, uniform teleportation sources and targets—are violated by actual user behavior, making it a poor predictor of actual traffic [22]. Our goal here is to present a more predictive model, using PageRank as a null model for evaluation.

Other researchers have also introduced more realistic models to capture features of real browsing behavior, such as the back button and tabbed browsing [19,5,8]. There have also been attempts to model the interplay between user interests and page content; Huberman *et al.* proposed a model in which visited pages have interest values described by a random walk that continues as long as the current page has a value above a threshold [16]. Such an approach relates to algorithms for improved topical crawlers [24].

We previously proposed a model in which users maintain a list of bookmarks from which new sessions begin, providing memory of past navigation [3]. While it is able to reproduce empirical page and link traffic distributions, it fails to account for patterns exhibited by individual users, such as entropy and session characteristics. The ABC model builds upon this previous model; some initial results were reported in [21]. Here we extend this effort to encompass both individual and aggregate measures of Web traffic, offering a comprehensive comparison among ABC, empirical measurements and a baseline model. We also discuss the key role of the topology of the Web graph.

## 3 Empirical Traffic Data

We gathered our HTTP request data from an undergraduate dormitory at Indiana University under methodology described in detail in our previous work [20]. The requests are gathered directly from a mirror of the building's network connection and reflect only unencrypted traffic. We use some basic heuristics to filter the data to include only requests made from browsers for actual page fetches, retaining a series of (user, referrer URL, target URL) triples. We also strip query parameters from the URLs, which affects roughly one-third of the requests. While this helps in the common case that parameters affect content within a static framework, it is less accurate when embedded CGI parameters select a page. Our analysis indicates that this effect is greatly mitigated by search-engine friendly design principles. The resulting data set contains 29.5 million page requests that come from 967 distinct users. They visited 2.5 million unique URLs, of which 2.1 million appeared as targets and 0.86 million appeared as referrers.

We organize each user's clicks into tree-based *logical* sessions using an algorithm described in our previous work [20]. The basic notions are that new sessions start with empty-referrer requests; that each request represents a directed edge from a referring URL to a target URL; and that requests belong to the session in which their referring URL was most recently requested. These session trees mimic users' multitasking behavior of by permitting several active sessions at once. The properties of these session trees, such as size and depth, are relatively insensitive to an additional timeout constraint [20]. We impose such a timeout as we form the sessions: a click cannot be associated with a session tree that has been dormant for thirty minutes. This process yields 11.1 million logical sessions in all, with a mean of over 11 thousand per user.

The structure of these trees allows us to infer how users backtrack as they browse. Modern caching mechanisms mean that a browser generally does not issue a request for

a recently accessed page, preventing direct observation of multiple links pointing to the same page, within a single session. While we have no *direct* way of detecting when the user presses the back button, session trees allow us to *infer* "backwards" traffic: if the next request in a tree comes from a URL other than the most recently visited, the user must have navigated to that page or opened it in a separate tab.

Any statistical description involves a compromise between summarizing the data and describing it accurately. For many human activities, including Web traffic, the data do are not normally distributed, but rather fit into heavy-tailed distributions best approximated by power laws [7,22]. The mean and median often describe the data poorly, as shown by a large and diverging variance and strong skew. The next best description is a histogram; we thus present these distributions in terms of their probability density functions rather than measures of central tendency. To characterize properties of traffic data and evaluate models of navigation, we focus on six quantities, several of which are discussed in preliminary work [20,21]:

**Page traffic.** The total number of visits to each page. Because of caching mechanisms, the majority of revisits to a page by a single user beyond the first visit within each session will not be represented in the data.

**Link traffic.** The number of times each hyperlink has been traversed by a user, as identified by the referrer and destination URLs in each request. We typically observe only the first visit to a destination page within a session.

**Empty referrer traffic.** The number of times each page initiates a session. We assume that a request without a referrer corresponds to using a bookmark, opening a link from another application, or entering an address manually.

**Entropy.** Shannon information entropy. For an individual user $j$, the entropy is defined as $S_j = -\sum_i \rho_{ij} \log_2 \rho_{ij}$ where $\rho_{ij}$ is the fraction of visits of user $j$ to site $i$ aggregated across sessions.

**Session size.** The number of unique pages visited in a logical session tree.

**Session depth.** The maximum tree distance between the starting page of a session and any page within that session. (Recall that sessions have a tree-like structure because backtracking requests are usually served from the browser cache.)

## 4 ABC Model

We now introduce the models for interpreting the empirical data. As a baseline, we consider a PageRank-like reference model with teleportation probability $p_t = 0.15$. This value is standard in the literature and best approximates the empirical data. We simulate a population of random walkers equal in number to our users. Each walker browses for as many sessions as corresponding real-world user. These sessions are terminated by the jumps, so the total number of pages visited by a walker differs from the corresponding user. Teleportation jumps lead to session-starting pages selected uniformly at random.

We call our own model *ABC* for its main ingredients: agents, bookmarks and clicks, as illustrated in Fig. 1. Agents possess some amount of *energy*, which represents their attention; it is spent by navigating and acquired by visiting interesting pages. Agents also have a *back button* and a *bookmark list* that is updated during navigation. Each agent starts at a random page with initial energy $E_0$. Then, for each time step:

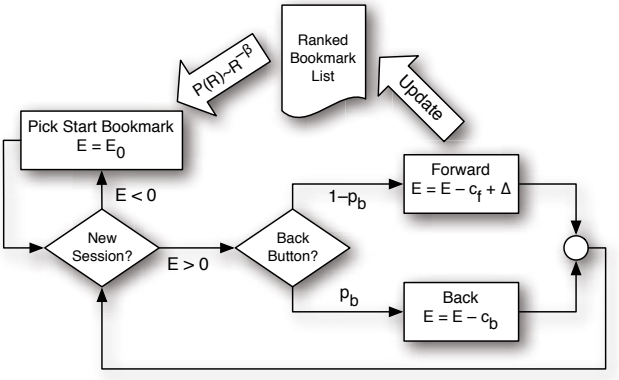



**Fig. 1.** Schematic illustration of the ABC model

1. If $E \leq 0$, the agent starts a new session.
2. Otherwise, if $E > 0$, the user continues the current session, following a link from the present node. There are two alternatives:
   (a) With probability $p_b$, the back button is used, leading back to the previous page. The agent's energy is decreased by a fixed cost $c_b$.
   (b) Otherwise, with probability $1-p_b$, a forward link is clicked. The agent's energy is updated to $E - c_f + \Delta$ where $c_f$ is a fixed cost and $\Delta$ is a stochastic value representing the new page's relevance to the user. The visitation is recorded in the bookmark list, which is kept ranked from most to least visited.

To initiate a new session, the bookmark with rank $R$ is chosen with probability $P(R) \propto R^{-\beta}$. This selection mechanism mimics the use of frequency ranking in various features of modern browsers, such as URL completion. The functional form is motivated by data on selection among a ranked list of search results [14].

The back button is our basic mechanism for producing branching behavior. The data indicate that the incoming and outgoing traffic of a site are seldom equal, with a ratio distributed over many orders of magnitude [22]. This violation of flow conservation cannot be explained by teleportation alone; the sessions of real users have many branches. Our prior results show an average node-to-depth ratio among session trees of almost two. These observations are consistent with the use of tabs and the back button, behavior confirmed by other studies [10,30].

The role of energy is critical. The duration of a real-world session depends on a user's individual goals and interests: visiting relevant pages leads to more clicks and longer sessions. ABC therefore incorporates agents with distinct *interests* and page *topicality*, relying on the intuition that an agent spends energy when navigating and gains it by discovering pages that match its interests. Moving forward costs more than using the back button. Known pages yield no energy, while novel pages increase energy by a random amount representing their relevance. Agents browse until they run out of energy, then start another session.

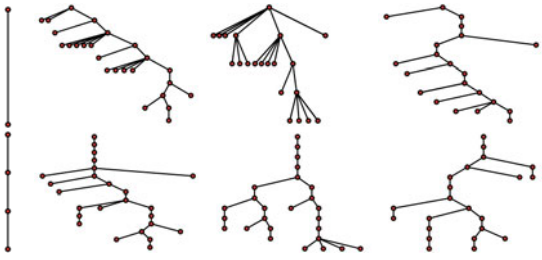


**Fig. 2.** Representation of a few typical and representative session trees from the empirical data (top) and from the ABC model (bottom). Animations are available at `cnets.indiana.edu/groups/nan/webtraffic`.

The dynamic variable $\Delta$ reflects a page's relevance to an agent. If $\Delta$ values are independent, identically distributed random variables, the amount of stored energy will behave as a random walk. The session duration $\ell$ (number of clicks until $E = 0$) will have a power-law tail $P(\ell) \sim \ell^{-\frac{3}{2}}$ [16]. However, empirical results suggest a larger exponent [20]. Moreover, studies show that content similarity between pages is correlated with their link distance, as is a page's relevance to a given topic [11,23]. Neighboring pages are topically similar, and the relevance of page $t$ to a user is close to that of page $r$ linking to $t$. To capture such *topical locality*, we correlate the $\Delta$ values of adjacent pages. We initially use $\Delta_0 = 1$; then, when a page $t$ is first visited *in a given session,* $\Delta_t$ is given by $\Delta_t = \Delta_r(1 + \epsilon)$, where $r$ is the referrer page, $\epsilon$ is uniformly randomly distributed in $[-\eta, \eta]$, and $\eta$ controls the degree of topical locality. A visited page can again become interesting in a later session and provide the agent with energy. However, it will yield different energy in different sessions, modeling drift in user interests.

## 5 Model Evaluation

Our simulations take place on a scale-free network with $N$ nodes and degree distribution $P(k) \sim k^{-\gamma}$, generated according to the Molloy-Reed algorithm [25], which we call G1. This graph has $N = 10^7$ nodes, more than observed in the data, to ensure adequate room for navigation. We also set $\gamma = 2.1$ to match our data set. To prevent dangling links, we construct G1 with symmetric edges. We also ran simulations of ABC on a second graph (G2) derived from an independent, empirical data set obtained by extracting the largest strongly connected component from the Web traffic of the entire university population (about 100,000 people) [22]. G2 is thus an actual subset of the Web graph with no dangling links. Based on three weeks of traffic as measured in November 2009, G2 has $N = 8.14 \times 10^6$ nodes and the same degree distribution, with $\gamma \approx 2.1$. Within each session we simulate caching by recording traffic only when the target page is novel to the session. This lets us count the unique pages visited, which mirrors the empirical session size. These cached pages are reset between sessions.

We must now set the parameters of ABC. The distribution of empty-referrer traffic will depend on the parameter $\beta$ and is well-approximated by $P(T_0) \sim T_0^{-\alpha}$, where $\alpha = 1 + 1/\beta$ [29]. Empirically, this exponent is $\alpha \approx 1.75$ [22]; to match it, we set the parameter $\beta = 1/(\alpha - 1) = 1.33$. We can also fit the back button probability $p_b = 0.5$

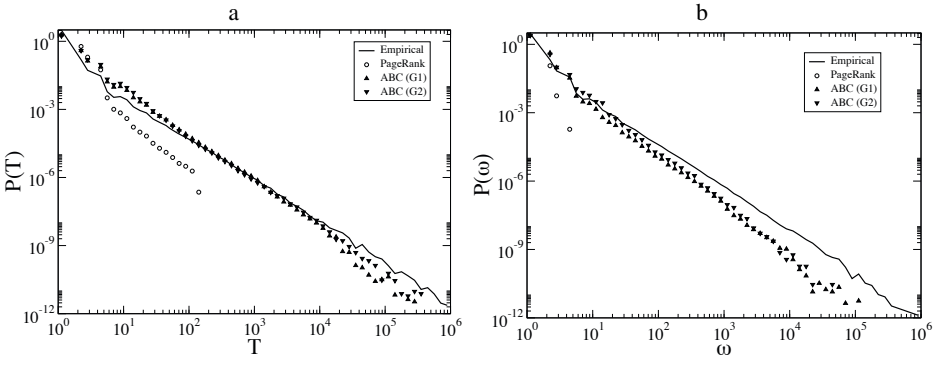



**Fig. 3.** Distribution of (a) page traffic and (b) link traffic generated by ABC model versus data and baseline

from the data. The initial energy $E_0$, the forward and backward costs $c_f$ and $c_b$, and the topical locality parameter $\eta$ control session duration. We thus set $E_0 = 0.5$ arbitrarily and estimate the costs as follows. Empirically, the average session size is roughly two pages. The net energy loss per click is $-\delta E = p_b c_b + (1 - p_b)(c_f - \langle \Delta \rangle)$, where $\langle \Delta \rangle = 1$ is the expected energy value of a new page. By setting $c_f = 1$ and $c_b = 0.5$, we obtain an expected session size $1 - (1 - p_b)E_0/\delta E = 2$ (counting the initial page). In general, higher costs lead to shorter sessions and lower entropy. We explored the effects of $\eta$ by simulation, settling on $\eta = 0.15$. Small values mean that all pages have similar relevance, making the session distributions too narrow. Large values erase topical locality, making the distributions too broad. Our results refer to this combination of parameters, with the numbers of users and sessions per user being drawn from the empirical data. We use the same parameters for both G1 and G2, without any further tuning to mach the properties of these networks.

The ABC agents generate session trees similar to those in the empirical data, as shown in Fig. 2. For a quantitative evaluation, we compare ABC with the empirical distributions described in § 3 and the reference PageRank model as simulated on the artificial G1 network.

We first consider the aggregate distributions, starting with traffic received by individual pages, as shown in Fig. 3(a). The empirical data show a broad power-law distribution for page traffic, $P(T) \sim T^{-\alpha}$, with exponent $\alpha \approx 1.75$, consistent with prior results for host-level traffic [22,20]. Theoretical arguments [26] suggest that PageRank should behave similarly. In general, a node will be visited if a neighbor has just been visited, making its traffic proportional to its degree in the absence of assortativity. This idea and prior results [22] lead us to expect PageRank's distribution of page traffic to fit a power law $P(T) \sim T^{-\alpha}$ where $\alpha \approx 2.1$ matches the exponent of the in-degree [7,13], as shown in Fig. 3(a). In contrast, traffic from ABC is biased toward previously visited pages (bookmarks), yielding a broader distribution and matching empirical measurements.

In Fig. 3(a), we compare the distributions of traffic per link $\omega$ from the models with the empirical data, revealing a power law for $P(\omega)$ with degree 1.9, agreeing with prior measurements of host-level traffic [22]. The comparison with PageRank illustrates the diversity of links with respect to their probability of being clicked. A rough argument

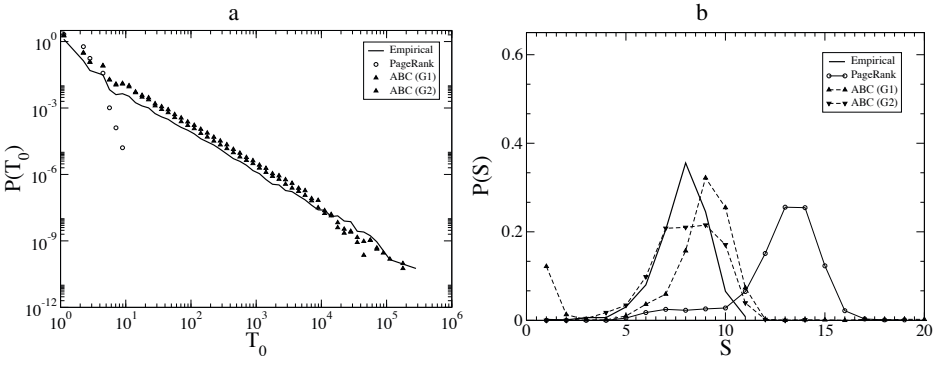



**Fig. 4.** Distribution of (a) traffic originating from jumps (page requests with empty referrer) and (b) user entropy generated by ABC model versus data and baseline

can explain the reference model's poor performance at reproducing the data. Recall that, disregarding teleportation, page traffic is roughly proportional to in-degree. The traffic expected on a link would thus be *proportional* to the traffic of the source page and *inversely proportional* to its out-degree, assuming that links are chosen uniformly at random. In-degree and out-degree are equal in our simulated graphs, leading to link traffic that is independent of degree and nearly constant for all links, as shown by the decaying distribution for PageRank. For ABC, the stronger heterogeneity in the probability of visiting pages is reflected in a heterogeneous choice of links, resulting in a broad distribution better fitting the empirical data, as shown in Fig. 3(b).

Our empirical data in Fig. 4(a) show that pages are not equally likely to start a browsing session. Their popularity as starting points is roughly distributed as a power law with exponent of about -1.8 (consistent with results for host-level traffic [22]), implying diverging variance and mean as the number of sessions increases. While not unexpected, this demonstrates a serious flaw in the hypothesis of uniform teleportation. Because PageRank assumes uniform probability among starting pages, its failure to reproduce the empirical data is evident in Fig. 4(a). In contrast, ABC's bookmarking mechanism captures the non-uniform probability of starting pages, yielding a distribution similar to the empirical data, as shown in Fig. 4(a), supporting the idea that rank-based bookmark selection is a sound cognitive mechanism for initiating sessions.

When it comes to individual users, the simplest hypothesis is that the broad distributions for aggregate behavior reflect extreme variability within the traffic of each user, suggesting that there is no "typical" user as described by their overall traffic. To examine users' diversity of behavior, we adopt Shannon's information entropy as defined in § 3. Entropy measures the focus of a user's interests, offering a better description of a single user than, e.g., the number of distinct pages visited; two users who have visited the same number of pages can have very different measures of entropy.

Given a number of visits $N_v$, the entropy is maximum ($S = N_v \log(N_v)$) when $N_v$ pages are visited once, and minimum ($S = 0$) when all visits are to a single page. The distribution of entropy across users is shown in Fig. 4(b). We observe that the PageRank model produces higher entropy than observed in the data: a PageRank walker picks starting pages with uniform probability, while a real user most often starts from

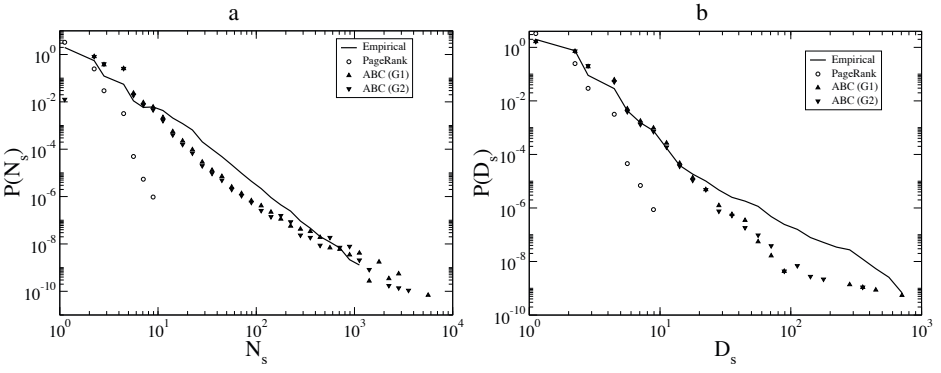



**Fig. 5.** Distribution of (a) session size (unique pages per session) and (b) session depth generated by ABC model versus data and baseline

a previously visited page, leading them to revisit neighboring pages. The ABC model yields entropy distributions that are influenced by the underlying network but fit empirical entropy data better than PageRank, suggesting that bookmarks, the back button, and topicality help to explain the focused habits of real users.

Finally, we consider two distributions that describe logical sessions: size (number of unique pages) and depth (distance from the starting page), both of which affect entropy. Figs. 5(a) and (b) show that the empirical distributions are broad, spanning three orders of magnitude, with a large proportion of long sessions. The brief sessions seen for the PageRank model originate from its teleportation mechanism, which cannot capture broadly distributed session sizes. The jump probability $p_t$ bounds the length $\ell$ (number of clicks) of a session, with a narrow, exponential distribution $P(\ell) \sim (1 - p_t)^\ell$. These exponentially short sessions do not conflict with the high entropy of PageRank walkers (Fig. 4(b)), which arises from jumps to random targets rather than browsing itself. In contrast, user interest and topical locality in ABC yield broad distributions of both session size and depth, as seen in Fig. 5(a) and (b). Agents visiting relevant pages tend to keep browsing, and relevant pages lead to more relevant pages, creating longer and deeper sessions. We believe the diversity shown in aggregate measures of traffic is a consequence of this diversity of interests rather than the behavior of extremely eclectic users—as shown by the narrow distribution of entropy.

To study the dependence of ABC on network topology, we ran the model on additional artificial networks. We eliminated any limitation of network size by simulating an infinite graph that generates new nodes as agents navigate. In this limit case, an agent's energy level is a random walk, with session size obeying a power law with exponent $-3/2$ [16]. However, the constant availability of new content leads to too many large sessions, as shown in Fig. 6. We then considered a Barabasi-Albert (BA) tree with the same node count as G1. The large number of leaf nodes affects the distribution of session size dramatically. Agents that begin a session in a leaf seldom to backtrack sufficiently high up the tree to discover new nodes; they quickly run out of energy, yielding a narrow distribution (Fig. 6). If we lift the constraint that the network contain no cycles, agents can escape these cul-de-sacs. Using an Erdös-Renyi (ER) network broadens the

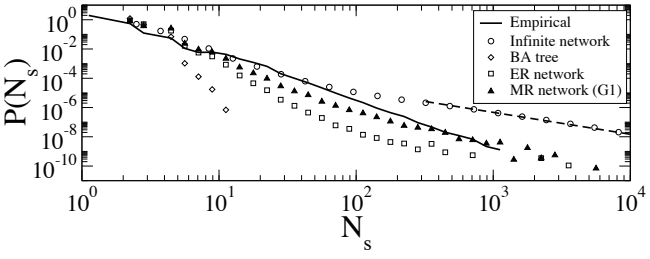



**Fig. 6.** Effect of network topology on session size. The curves correspond to simulations of the ABC model on different artificial networks. The resulting session size distributions are compared with the empirical one. The dashed line is a guide to the eye for $P(N_S) \sim N_S^{-3/2}$.

distribution of session size (Fig. 6), bringing it closer to the empirical data while still underestimating the number of large sessions due to the lack of hubs.

For comparison, Fig. 6 also shows the distribution obtained with G1, a network with cycles and broadly distributed degree. As already seen (Fig. 5(a)), this network gives excellent results, showing that both hubs and cycles are needed for the exploration of distant regions of the network. If either element is missing, agents can reach only limited content, leading to shortened sessions.

## 6 Conclusions

Previous studies have shown that Markovian processes such as PageRank cannot explain many patterns observed in real measurements of Web activity, especially the diversity of starting points, the global diversity of link traffic, and the heterogeneity of session sizes. Furthermore, individual behaviors are quite focused in spite of such diverse aggregate measurements. These observations call for a stateful, agent-based model that can help explain the empirical data through more realistic browsing behavior. We have proposed three key ingredients for such a model. First, agents maintain individual lists of bookmarks (a memory mechanism) for use as teleportation targets. Second, agents have a back button (a branching mechanism) that can also simulate tabbed browsing. Finally, agents have topical interests that matched by page content, modulating the probability of an agent starting a new session and leading to heterogeneous session sizes.

We have shown that the resulting ABC model is capable of reproducing with remarkable accuracy the aggregate traffic patterns we observe in our empirical measurements. More importantly, our model offers the first account of a mechanism that can generate key properties of logical sessions. This allows us to argue that the diversity apparent in page, link, and bookmark traffic is a consequence of the diversity of individual interests rather than the behavior of very eclectic users. Our model is able to capture, for the first time, the extreme heterogeneity of aggregate traffic measurements while explaining the narrowly focused browsing patterns of individual users. While ABC is more complex than prior models, its greater predictive power suggests that bookmarks, tabbed browsing, and topicality are salient features of how we browse the Web. We believe that ABC may lead the way to more sophisticated, realistic, and hence more effective ranking and crawling algorithms.

The model does rely on several key parameters. While we have attempted to make reasonable and realistic choices for most of these parameters and explored the sensitivity of our model with respect to the rest, further work is needed to understand the combined effect of these parameters in a principled way. For example, we already know that parameters such as network size, costs, and topical locality play a key role in modulating the balance between individual diversity (entropy) and session size. In the future, we hope to analyze the model from a more theoretical perspective.

Finally, while the ABC model is a clear step in the right direction, it shares some limitations of existing efforts, most notably the uniform choice among outgoing links from a page, which may cause the imperfect match between the individual entropy values of our agents and those of actual users.

*Acknowledgements.* We thank CNetS and PTI at Indiana University and L. J. Camp of the IU School of Informatics and Computing, for support and infrastructure. We also thank IU's network engineers for support in data collection. This work was supported in part by the I3P research program, managed by Dartmouth College and supported under Award 2003-TK-TX-0003 from the Science and Technology Directorate of the U.S. DHS. BG was supported in part by grant NIH-1R21DA024259 from the NIH. JJR is funded by the project 233847-Dynanets of the EUC. This material is based upon work supported by NSF award 0705676. This work was supported in part by a gift from Google. Opinions, findings, conclusions, recommendations or points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. DHS, Science and Technology Directorate, I3P, NSF, IU, Google, or Dartmouth College.

## References

1. Adar, E., Teevan, J., Dumais, S.: Large scale analysis of web revisitation patterns. In: Proc. CHI (2008)
2. Adar, E., Teevan, J., Dumais, S.: Resonance on the web: Web dynamics and revisitation patterns. In: Proc. CHI (2009)
3. Gonçalves, B., Meiss, M.R., Ramasco, J.J., Flammini, A., Menczer, F.: Remembering what we like: Toward an agent-based model of Web traffic. Late Breaking Results WSDM (2009)
4. Beauvisage, T.: The dynamics of personal territories on the web. In: Proc. HT (2009)
5. Bouklit, M., Mathieu, F.: BackRank: an alternative for PageRank? In: Proc. WWW Special Interest Tracks and Posters, pp. 1122–1123 (2005)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks 30(1-7), 107–117 (1998)
7. Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the Web. Computer Networks 33(1-6), 309–320 (2000)
8. Chierichetti, F., Kumar, R., Tomkins, A.: Stochastic models for tabbed browsing. In: Proc. WWW, pp. 241–250 (2010)
9. Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. Computer Networks 30(1-7), 161–172 (1998)
10. Cockburn, A., McKenzie, B.: What do web users do? an empirical analysis of web use. Int. J. of Human-Computer Studies 54(6), 903–922 (2001)

11. Davison, B.: Topical locality in the Web. In: Proc. 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 272–279 (2000)
12. Douglis, F.: What's your PageRank? IEEE Internet Computing 11(4), 3–4 (2007)
13. Fortunato, S., Boguna, M., Flammini, A., Menczer, F.: Approximating PageRank from in-degree. In: Aiello, W., Broder, A., Janssen, J., Milios, E.E. (eds.) WAW 2006. LNCS, vol. 4936, pp. 59–71. Springer, Heidelberg (2008)
14. Fortunato, S., Flammini, A., Menczer, F., Vespignani, A.: Topical interests and the mitigation of search engine bias. Proc. Natl. Acad. Sci. USA 103(34), 12684–12689 (2006)
15. Gonçalves, B., Ramasco, J.J.: Human dynamics revealed through web analytics. Phys. Rev. E 78, 026123 (2008)
16. Huberman, B., Pirolli, P., Pitkow, J., Lukose, R.: Strong regularities in World Wide Web surfing. Science 280(5360), 95–97 (1998)
17. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide-Web. Computer Networks and ISDN Systems 27, 1065–1073 (1995)
18. Liu, Y., Gao, B., Liu, T.Y., Zhang, Y., Ma, Z., He, S., Li, H.: BrowseRank: letting Web users vote for page importance. In: Proc. SIGIR, pp. 451–458 (2008)
19. Mathieu, F., Bouklit, M.: The effect of the back button in a random walk: application for PageRank. In: Proc. WWW Alternate Track Papers & Posters, pp. 370–371 (2004)
20. Meiss, M., Duncan, J., Gonçalves, B., Ramasco, J.J., Menczer, F.: What's in a session: tracking individual behavior on the Web. In: Proc. HT (2009)
21. Meiss, M., Gonçalves, B., Ramasco, J.J., Flammini, A., Menczer, F.: Agents, bookmarks and clicks: A topical model of Web navigation. In: Proc. HT (2010)
22. Meiss, M., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Ranking web sites with real user traffic. In: Proc. WSDM, pp. 65–75 (2008)
23. Menczer, F.: Mapping the semantics of web text and links. IEEE Internet Computing 9(3), 27–36 (2005)
24. Menczer, F., Pant, G., Srinivasan, P.: Topical web crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology 4(4), 378–419 (2004)
25. Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. Random Structures and Algorithms 6(2-3), 161–180 (1995)
26. Noh, J.D., Rieger, H.: Random walks on complex networks. Phys. Rev. Lett. 92, 118701 (2004)
27. Qiu, F., Liu, Z., Cho, J.: Analysis of user web traffic with a focus on search activities. In: Proc. 8th International Workshop on the Web and Databases (WebDB), pp. 103–108 (2005)
28. Radlinski, F., Joachims, T.: Active exploration for learning rankings from clickthrough data. In: Proc. KDD (2007)
29. Fortunato, S., Flammini, A., Menczer, F.: Scale-free network growth by ranking. Phys. Rev. Lett. 96, 218701 (2006)
30. Tauscher, L., Greenberg, S.: How people revisit web pages: Empirical findings and implications for the design of history systems. Int. J. of Human-Computer Studies 47(1), 97–137 (1997)