Zipf's Law for Web Surfers

Mark Levene and José Borges
University College London
Gower Street
London WC1E 6BT, U.K.
email: {mlevene,j.borges}@cs.ucl.ac.uk

George Loizou
Birkbeck College
Malet Street
London WC1E 7HX, U.K.
email: george@dcs.bbk.ac.uk

August 29, 2000

Abstract

One of the main activities of web users, known as "surfing", is to follow links. Lengthy navigation often leads to disorientation when users lose track of the context in which they are navigating and are unsure how to proceed in terms of the goal of their original query. Studying navigation patterns of web users is thus important, since it can lead us to a better understanding of the problems users face when they are surfing. We derive Zipf's rank frequency law (i.e. an inverse power law) from an absorbing Markov chain model of surfers' behaviour assuming that less probable navigation trails are, on average, longer than more probable ones. In our model the probability of a trail is interpreted as the relevance (or "value") of the trail. We apply our model to two scenarios: in the first the probability of a user terminating the navigation session is independent of the number of links he has followed so far, and in the second the probability of a user terminating the navigation session increases by a constant each time the user follows a link. We analyse these scenarios using two sets of experimental data sets showing that, although the first scenario is only a rough approximation of surfers' behaviour, the data is consistent with the second scenario and can thus provide an explanation of surfers' behaviour.

1 Introduction

The World-Wide-Web (known as the Web) has become a ubiquitous tool, used in day-to-day work, to find information and conduct business, and it is revolutionising the role and availability of information. One of the main activities of users interacting with the Web is that of *navigation* (colloquially known as "surfing") whereby users follow links and browse the destination web pages. Lengthy navigation often leads to the problem of disorientation when users lose track of the context and are unsure how to proceed in terms of satisfying their original goal; this problem is central to web interaction and is known as the *navigation problem* [LL99, LL00]. Understanding user navigation patterns and their underlying distribution is important since it can lead to better web site design with the intention of saving users' effort by directly guiding them to more relevant web pages.

We model the collection of navigation trails, i.e sequence of links, that a user may follow as a stochastic process in terms of an *absorbing Markov chain* [KS60]. In this stochastic process, whenever the user is browsing a web page, he is faced with a choice of continuing the navigation and following one of the available links embedded in the browsed web page,

or terminating the navigation session. Each choice of link has a probability attached to it and the overall trail probability can be interpreted as the relevance (or "value") of following that trail. Hence a trail induces a random walk through the navigation space, and trails with higher probability are more valuable to the user and thus preferable to following trails with lower probability.

Huberman et al. [HPPL98, LH98] propose a random walk model of surfers' navigation behaviour in terms of Brownian motion. In their model at each navigation step the "value" the user obtains from following a link and browsing an additional page is modelled as an independent and identically distributed Gaussian random variable. Thus the optimal stopping rule is to continue navigation if the expected value of following an additional link is greater than zero, otherwise terminate the navigation session. By testing their model on web data Huberman et al. [HPPL98] have demonstrated that the probability of a trail of length t is approximately proportional to $t^{-3/2}$, which is a form of the *Pareto-Zipf rank-frequency law* [Man63], which we simply call *Zipf's law*.

The random walk model proposed in [HPPL98, LH98] does not take into account the topology of the Web and does not provide an analytic derivation of Zipf's law from first principles. This motivates the problem solved herein, which is to derive Zipf's law from a number of reasonable assumptions regarding surfers' navigation behaviour.

For surfers' navigation trails, Zipf's rank-frequency law states that the distribution of trail frequencies obeys an inverse power law when trails are ordered according to their ranks from the most probable to the least probable. The fundamental reason that Zipf's rank-frequency law holds for surfers' navigation trails is that, on average, surfers are much more likely to follow short trails rather than long trails (this assumption is inherent in [HPPL98, LH98]). An elementary explanation of this observation is the fact that the number of short trails is exponentially less than the number of long trails, due to the Web topology, and thus finding a short and relevant trail is much easier than finding a long and relevant trail (if such a trail exists). This explanation does not take into account the overall relevance (or "value") of the trail. Therefore, a deeper explanation for this law of surfers' navigation patterns is that the ratio of "value" gained to "effort" expended is, on average, higher for shorter trails than for longer trails (again this assumption is inherent in [HPPL98, LH98]). Assuming that we measure the relevance of a trail by its probability, we can also give an information-theoretic explanation to this phenomenon in terms of maximising the average information per page browsed (see [Man54]).

2 Derivation of Zipf's law

We make the following assumptions concerning users' surfing behaviour:

1. A user's home page acts as a portal for his navigation, i.e. all user navigation trails begin at their home page. When a user is browsing a web page there is a positive probability that he will end their navigation session at that page. In order to model the termination of a navigation session we assume that there is a *global* stopping state and a transition to it from all other states indicating the possibility of stopping the navigation session.

2. Longer trails are less probable, on average, than shorter trails. (As can be seen in

Figure 5 in Section 3 this assumption is backed up by strong empirical evidence; see also [HPPL98].)

3. The *branching factor* of the underlying directed graph representing the topology of the Web is a constant, $b > 1$; the branching factor can be viewed as the average number of outlinks from a page. (We note that the average number of outlinks from a web page has been established as being approximately 8 [KRRT99].)

4. The *average probability* that a user will choose any link out of the b possible ones is given by a function $f(\theta)$, with $0 < f(\theta) < 1$, where θ is an external parameter, such that
$$p_s = 1 - \sum_{i=1}^{b} f(\theta) > 0$$
is the probability of stopping, i.e. of entering the stopping state. (The justification for employing the average probability of choosing a link, say i, rather the actual probability, say p_i, is to allow us to obtain a tractable model that minimises the number of parameters and for which we can deduce general properties of surfing behaviour. The average probability $f(\theta)$ can be estimated from log data using the technique described in [BL00].)

With these assumptions the set of navigation trails a user can follow induces an absorbing Markov chain, $\mathcal{M}$, with a single absorbing state [KS60]. The Markov chain induces a probability distribution on the possible trails a user can follow forming a regular language over the alphabet of web pages. We consider the probability of a trail to correspond to its "value", where the "effort" is taken into account by the nature of the exponential decrease in the probability of a trail as it gets longer.

We now derive Zipf's rank-frequency law from these assumptions using the technique of [Li92]. Let t be the length of a trail in $\mathcal{M}$ and $r(t)$ be the rank of a trail of length t, with $t \geq 0$. Then, when ranking trails from most probable to least probable, we have
$$\frac{b^t - 1}{b - 1} < r(t) \leq \frac{b^{t+1} - 1}{b - 1}, \tag{1}$$
since the number of trails of length t forms a geometric series.

From (1) we conclude that
$$t < \log_b((b-1)r(t) + 1) \leq t + 1$$
thus
$$f(\theta)^t > f(\theta)^{\log_b((b-1)r(t)+1)} \geq f(\theta)^{t+1}$$
implying that
$$g(\theta)f(\theta)^t > g(\theta)\left(\frac{1}{(b-1)r(t)+1}\right)^{\frac{-\log f(\theta)}{\log b}} \geq g(\theta)f(\theta)^{t+1},$$
where $0 < g(\theta) \leq 1$. On setting $\rho = -\log f(\theta)/\log b$, $\alpha = 1/(b-1)$ and $\beta = g(\theta)\alpha^\rho$ we obtain
$$g(\theta)f(\theta)^t > \frac{\beta}{(r(t)+\alpha)^\rho} \geq g(\theta)f(\theta)^{t+1}, \tag{2}$$

which is a generalised form of Zipf's law also known as the Pareto-Zipf-Mandelbrot law [Man54, MN58].

In order to allow a potentially infinite number of ranks we must have $\rho > 1$ (see [Man59], where it is stated that for natural languages one finds, in general, that $\rho > 1$.) It can easily be verified that $\rho > 1$ if and only if

$$f(\theta) < \frac{1}{b}. \tag{3}$$

Using (2) we now demonstrate how to model two scenarios of user navigation. In the first scenario, detailed in Subsection 2.1, the probability of stopping is independent of the number of navigation steps carried out so far, and in the second scenario, detailed in Subsection 2.2, the stopping probability increases linearly with the number of navigation steps.

2.1 Scenario 1

Given that a decision at a given node is to continue surfing, we take the probability of following link i, $1 \leq i \leq b$, to be p_i, where

$$\sum_{i=1}^{b} p_i = 1 - p_s.$$

We can assume that

$$f(\theta) = \frac{1 - p_s}{b} \text{ and } g(\theta) = p_s,$$

whence it can be verified that

$$g(\theta) f(\theta)^t$$

is the average probability of a trail of length t. Moreover, (3) holds in this case so $\rho > 1$. Figure 1 shows a log-log plot of this model with $b = 8$ and p_s varying from 0.05 to 0.9.

2.2 Scenario 2

In this scenario we incrementally discount the probability of continuing at each navigation step by a constant γ, with $0 < \gamma < 1$, such that the probability of stopping after navigating for t steps is

$$\min(1, p_s + t\gamma).$$

Thus in this scenario the longer the trail is the more probable it is to stop. In this case, assuming that $p_s + t\gamma \leq 1$, i.e.

$$t \leq \frac{1 - p_s}{\gamma}, \tag{4}$$

the average probability of a trail of length t is given by

$$(p_s + t\gamma) \prod_{i=1}^{t} \left(\frac{1 - p_s - (i-1)\gamma}{b} \right). \tag{5}$$

After some arithmetic manipulation we can transform (5) to

$$(p_s + t\gamma) \left(\frac{\gamma}{b}\right)^t \prod_{i=0}^{t-1} \left\{ \left(\frac{1 - p_s}{\gamma} - t + 1 \right) + i \right\} \quad = \quad (p_s + t\gamma) \left(\frac{\gamma}{b}\right)^t \frac{\Gamma\left(\frac{1-p_s}{\gamma} + 1\right)}{\Gamma\left(\frac{1-p_s}{\gamma} - t + 1\right)}, \tag{6}$$

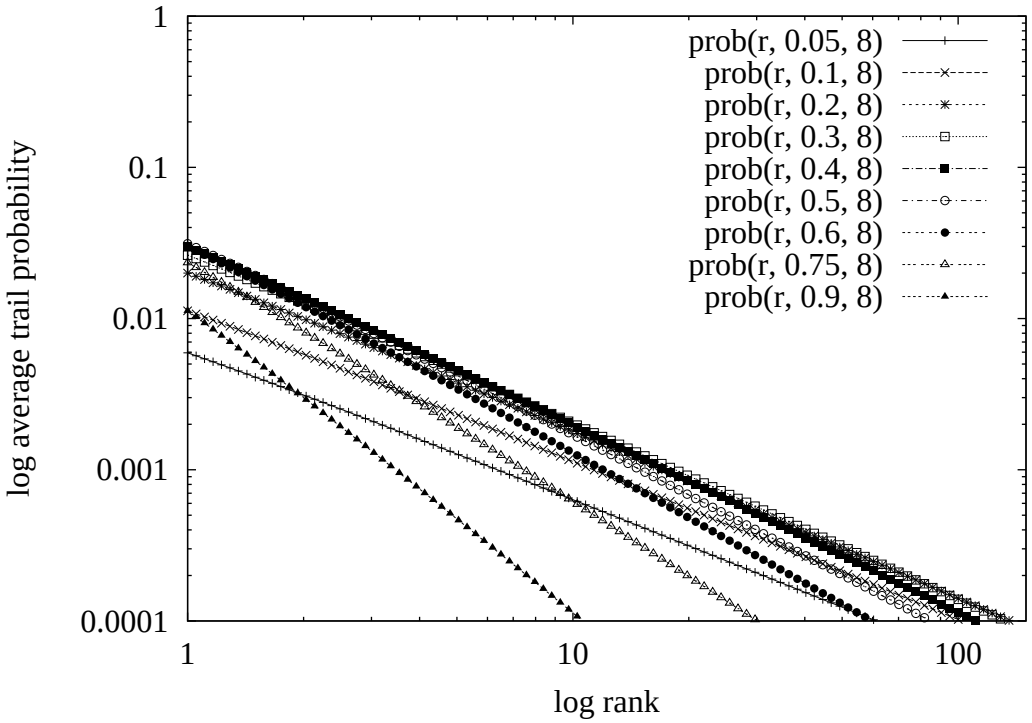

Figure 1: A log-log plot for the first model

where Γ is the gamma function [AS72].

On using the right-hand side of (4) as an upper bound for t, we can compute the average probability of stopping, denoted by $\mathrm{avg}(p_s)$, to obtain

$$\mathrm{avg}(p_s) = \frac{1 + p_s}{2}.$$

Moreover, on using the asymptotic formula [AS72]

$$\Gamma(z + b) \approx \sqrt{2\pi}\ e^{-z}\ z^{z+b-\frac{1}{2}},$$

the quotient of the two gamma functions on the right-hand side of (6) reduces to

$$\left(\frac{1 - p_s + \gamma}{\gamma}\right)^t$$

allowing us to approximate the right-hand side of (6) by

$$\mathrm{avg}(p_s) \left(\frac{1 - p_s + \gamma}{b}\right)^t,$$

which by (2) is of the form required for Zipf's law. Moreover, by (3) we have that $\rho > 1$ if and only if $\gamma < p_s$. Figure 2 shows a log-log plot of this model with $b = 8$, $\gamma = 0.05$ and p_s varying from 0.05 to 0.9.

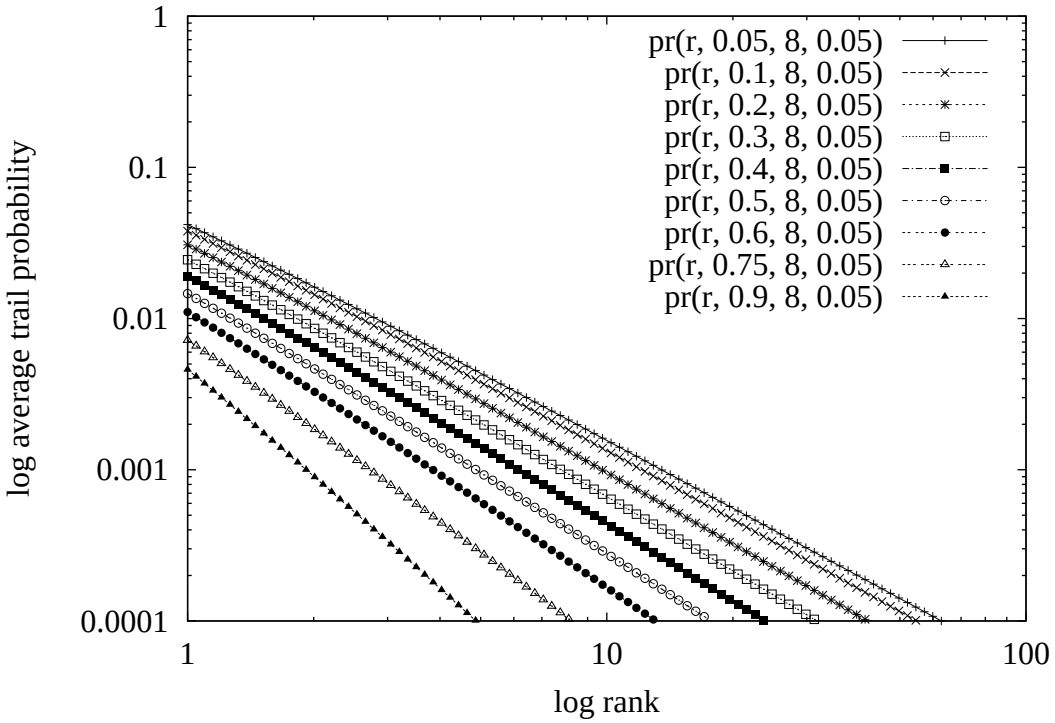

Figure 2: A log-log plot for the second model with averaging

Under this scenario, we can approximate $p_s + t\gamma$ by

$$p_s \exp\left(\frac{\gamma t}{p_s}\right),$$

provided the second term in the polynomial expansion of the exponential, which is of the order of $(\gamma t/p_s)^2$, is negligible. This obtains, for example, if p_s is not overly small and t is not too large, as can be seen in Figure 3, where $\gamma = 0.05$ and t varies between 1 and 10.

We thus obtain an approximation of (6) given by

$$p_s \left(\exp\left(\frac{\gamma}{p_s}\right)\left(\frac{1-p_s+\gamma}{b}\right)\right)^t$$

which by (2) is again of the form required for Zipf's law. Moreover, by (3) we have that $\rho > 1$ if and only if

$$\frac{\gamma}{p_s} < -\log(1-p_s+\gamma). \tag{7}$$

Now assuming that $\gamma < p_s$, then for (7) to be true it is sufficient that

$$p_s - \gamma \geq 1 - \frac{1}{e} \approx 0.632$$

and, for example, if $\gamma/p_s < 1/2$ then it is sufficient that

$$p_s - \gamma \geq 1 - \sqrt{e} \approx 0.393.$$

Figure 4 shows a log-log plot of this model with $b = 8$, $\gamma = 0.05$ and p_s varying from 0.2 to 0.95.

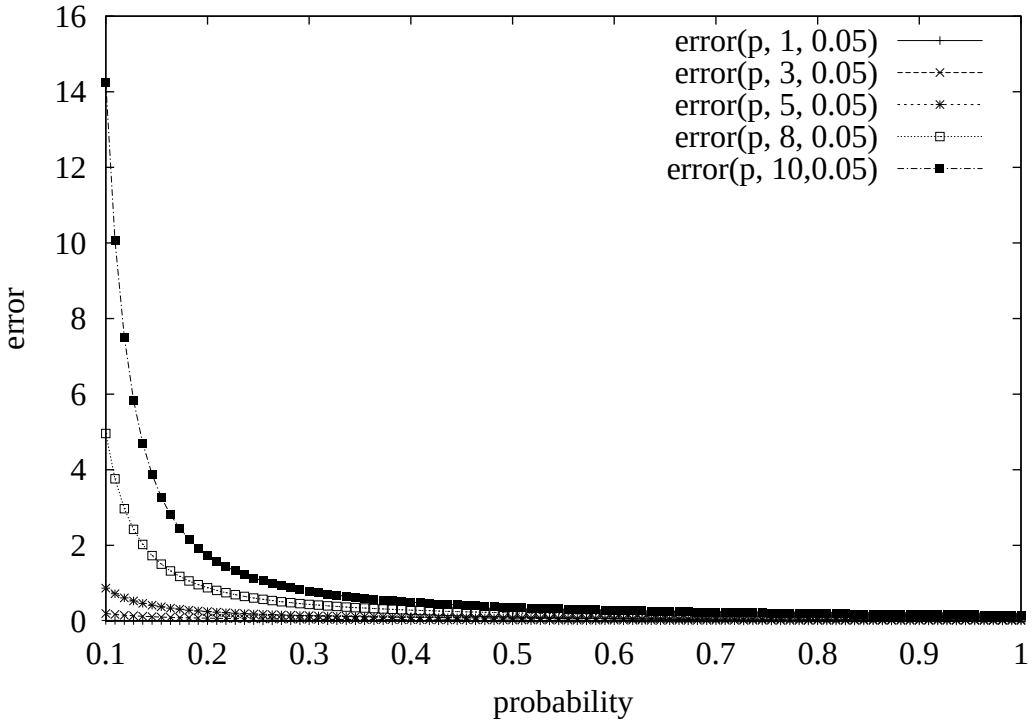

Figure 3: The error in the approximation

3 Concluding Remarks and Experimental Evaluation

We have shown that Zipf's law can be derived by modelling surfers' navigation behaviour as an absorbing Markov chain assuming that longer trails are, on average, less probable than shorter trails. Using our model we considered two scenarios that are consistent with our derivation of Zipf's law. In the first scenario the probability that a surfer will terminate his navigation session remains the same, independently of the number of links that he has followed so far. In the second scenario the probability of stopping, p_s, increases by γ after each link is followed, so that $(1 - p_s)/\gamma$ is an upper bound on the number of navigation steps in a session. We derived two approximations for this scenario one by considering the average probability of stopping and the other by approximating the probability of stopping.

We conclude by examining two sets of empirical data of user web log data, in order to verify our model. The first data set was downloaded from

`www.cs.washington.edu/homes/map/adaptive/download.html`

and contains a week of web log data from 1997 of all accesses from the site:

`machines.hyperreal.org/`

The second data set was downloaded from

`www.cs.berkeley.edu/logs/`

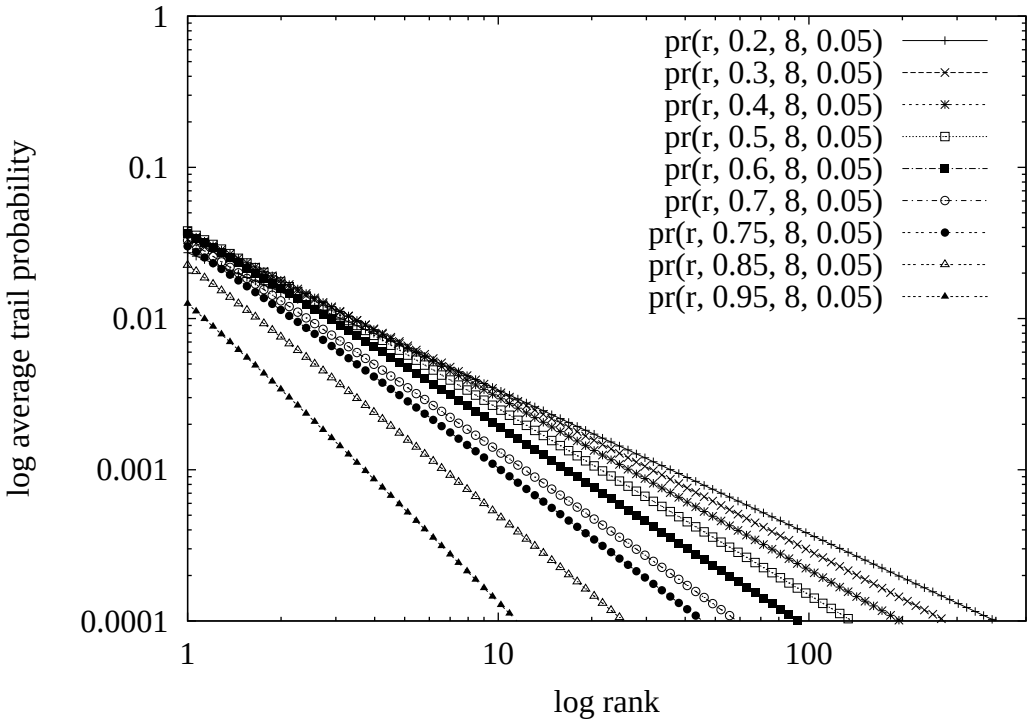

Figure 4: A log-log plot for the second model with approximation

and contains a daily log of all accesses for their site from 1999. In both data sets we considered a trail to be a sequence of URLs with up to 15 minutes between two consecutive requests from the same IP address, and the data was processed using algorithms described in [BL00].

A log-log plot of the two data sets is shown in Figure 5, where we plotted the trail length versus its probability for trails of up to a maximal length of 20, and the trail rank versus its probability for trails ranked up to a maximal rank of 150. It can be seen from the plots of length versus probability that our crucial assumption that longer trails are, on average, less probable is borne out in practice.

We performed regression analysis on both data sets to ascertain whether trail rank versus trail probability data is consistent with our scenarios. We found that scenario 1, where $g(\theta) = p_s$ and $f(\theta) = (1 - p_s)/b$, is inconsistent with the data although it can still provide a rough approximation. On the other hand, the data is seen to be consistent with scenario 2. We next consider our two approximations of Subsection 2.2. The first, which averages $p_s + t\gamma$, is referred to as scenario 2 with averaging; the results are summarised in Table 1. The second, which approximates $p_s + t\gamma$ by an exponential, is referred to as scenario 2 with approximation; the results are summarised in Table 2. (In the tables b stands for branching factor, *cc* stands for correlation coefficient and *se* stands for standard error.)

This work is the first stage of further research looking into the practical applications of results concerning the distribution of surfers' behaviour in the context of Web data mining. One application of our work is that of improving Web site design [PE00]. For example, it is possible to cluster Web site users according to the parameters of our model in order to distinguish between different user communities. Such knowledge can be used to improve the

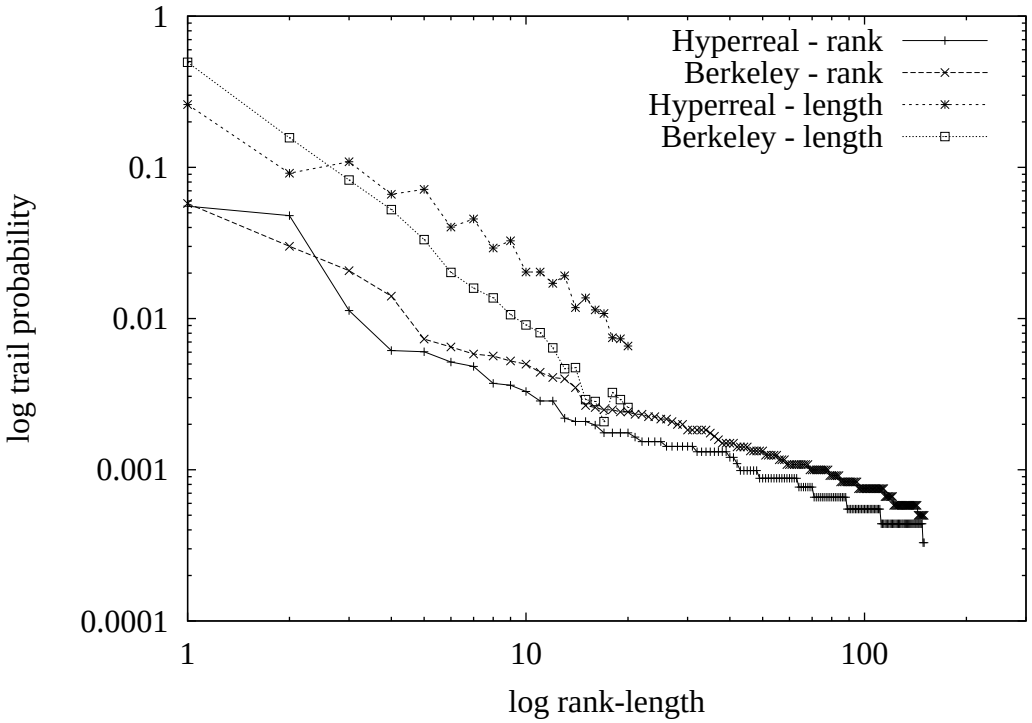

Figure 5: A log-log plot of experimental data

Data set	b	p_s	avg(p_s)	$f(\theta)$	γ	cc	se
Hyperreal	5	0.853	0.926	0.072	0.214	0.930	0.0022
Berkeley	2	0.862	0.931	0.072	0.007	0.864	0.0028

Table 1: Regression analysis of scenario 2 with averaging

Web site in order to target these communities, for example by encouraging users whose trails tend to be shorter to explore other parts of the Web site. We are also augmenting our model with the distribution of the time spent on each page in a trail, which could have an effect on the actual design of Web pages. By examining further data sets we intend to establish whether any universal patterns, such as those reported in [HPPL98], arise in our model.

References

[AS72] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover, New York, NY, 1972.

[BL00] J. Borges and M. Levene. Data mining of user navigation patterns. In B. Masand and M. Spiliopoulou, editors, *Web Usage Mining*, To appear in Lecture Notes in Artificial Intelligence (LNAI 1836). Springer-Verlag, Berlin, 2000.

[HPPL98] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose. Strong regularities in world wide web surfing. *Science*, 280:95–97, 1998.

Data set	b	p_s	$f(\theta)$	γ	cc	se
Hyperreal	5	0.46	0.13	0.048	0.949	0.0019
Berkeley	2	0.10	0.45	0.00003	0.972	0.0013

Table 2: Regression analysis of scenario 2 with approximation

[KRRT99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of International Conference on Very Large Data Bases*, pages 639–650, Edinburgh, 1999.

[KS60] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. D. Van Nostrand, Princeton, NJ, 1960.

[LH98] R.M. Lukose and B.A. Huberman. Surfing as a real option. In *Proceedings of the International Conference on Information and Computation Economies*, pages 45–51, Charleston, SC, 1998.

[Li92] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38:1842–1845, 1992.

[LL99] M. Levene and G. Loizou. Navigation in hypertext is easy only sometimes. *SIAM Journal on Computing*, 29:728–760, 1999.

[LL00] M. Levene and G. Loizou. Web interaction and the navigation problem in hypertext. In A. Kent, J.G. Williams, and C.M. Hall, editors, *Encyclopedia of Microcomputers*. Marcel Dekker, New York, NY, 2000. To appear.

[Man54] B. Mandelbrot. On recurrent noise limit coding. In *Proceedings of the Symposium on Information Networks*, pages 205–221, Brooklyn, NY, 1954.

[Man59] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by H.A. Simon. *Information and Control*, 2:90–99, 1959.

[Man63] B. Mandelbrot. New methods in statistical economics. *The Journal of Political Economy*, 71:421–440, 1963.

[MN58] G.A. Miller and E.B. Newman. Tests of a statistical explanation of the rank-frequency relation for words in written English. *The American Journal of Psychology*, 71:209–218, 1958.

[PE00] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118:245–275, 2000.