# File Popularity Characterisation.


Chris Roadknight
BT Research Laboratories,
Martlesham Heath, Ipswich,
Suffolk, UK IP5 7RE
+44 (0)1473 648457
roadknic@drake.bt.co.uk

Ian Marshall
BT Research Laboratories,
Martlesham Heath, Ipswich,
Suffolk, UK IP5 7RE
+44 (0)1473 646951
marshall@drake.bt.co.uk

Debbie Vearer
University of East Anglia
Norwich
D.A.Vearer@uea.ac.uk



## ABSTRACT

A key determinant of the effectiveness of a web cache is the locality of the files requested. In the past this has been difficult to model, as locality appears to be cache specific. We show that locality can be characterised with a single parameter, which primarily varies with the topological position of the cache, and is largely independent of the culture of the cache users. Accurate cache models can therefore be built without any need to consider cultural effects that are hard to predict.



## Keywords

File Popularity, web caches.


## 1. INTRODUCTION

WWW caching has proved a valuable technique for scaling up the internet [2, 3]. Caches can bring files nearer the client (with a possible reduction in latency), reduce load on servers and add missing robustness to a distributed system such as the web. A cache's usefulness is directly related to the degree of locality shown in the files it serves, where locality refers to the tendency of users to request access to the same files. The locality is best illustrated using a popularity curve, which plots the number of requests for each file against the file's popularity ranking. It is often said that this popularity curve follows Zipf's law, Popularity = K* ranking$^{-a}$, with a being close to 1 (e.g. [9]); others argue that the curve does not follow Zipf's law [1]. Zipf's law has been observed in several environments where human choice is involved, including linguistic word selection [17] and choice of habitat [12], so there is an expectation that some measures of file popularity should follow Zipf's law too. This would be useful because previous observations of Zipf's law have been largely culture independent, and if some culture independent cache metrics could be established cache models would not need to take account of cultural effects. In the absence of a culture independent metric, modellers are forced to use parameters with embedded culture dependence. The only technique we are aware of to
determine the culture of a cache community is to semantically analyse the files contained in the cache. This is extremely time consuming and does not enable useful predictions about other caches.

It is not at all clear that cache logs reflect human choices, since not all of a user's web requests reach the network cache. Some of the user's requests are intercepted on the user's client, by the cache maintained by the browser. In addition it is hard to establish whether logged requests are user initiated or are the result of embedded object links. The 'Zipf / not Zipf' argument is not helped by the notion that a curve follows Zipf's law if the exponent is close to unity, with the precise meaning of 'close' being vague. In fact (e.g. fig. 1) the observed popularity curves vary significantly [11,14]. Figure 1 gives a basic visual example of popularity curves for the 20 most popular files at 7 WWW caches around the world At first sight it may not be obvious that the variation of an exponent from 0.5 to 0.6 is important. However, a moments thought about the mathematics involved reveals that this small change alters the popularity of the most popular files by a 100% and changes the number of files in the cache by 15%. Therefore, in order to use a fit to Zipf's law in a predictive model, it is necessary to understand the cause of the variations and demonstrate that the value for a cache can be easily predicted from known features of the cache.

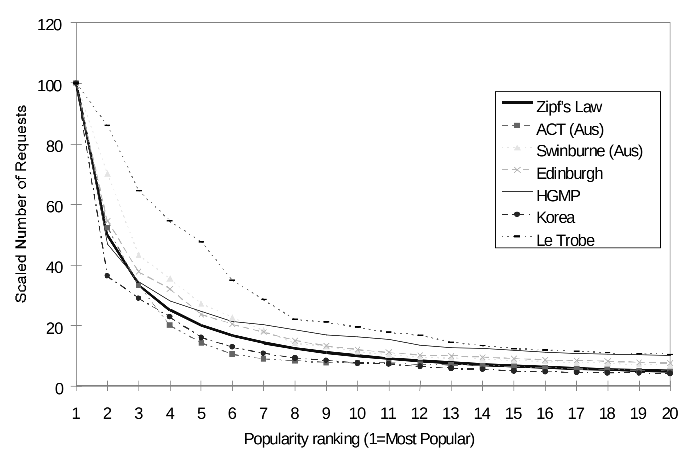


**Figure 1. Scaled popularity curves at 6 caches.**

In this paper we show for the first time that cache locality curves are indeed invariant to the culture of the cache users. We present a set of possible explanations of the observed variance in locality curves, derived from the literature and our own imagination, and propose tests of the explanations. We have performed the tests by analysing a wide variety of caches, and have thereby eliminated all but one of the theories.

We also argue (along with another recent, submitted study [7]) that popularity curves are more accurately modelled by a power law curve with a fitted, negative exponent that is not usually -1. Even for this model to be meaningful, the definitions of what is to be plotted, the sample size, and the fit must be made carefully and precisely. We demonstrate in this paper the degree of care that must be taken, and derive for the first time a quantitative metric for the sample sizes required

## 2. THEORIES

One possible hypothesis (derived from a related proposal by Zipf [17]) is that caches at different levels of the hierarchy have different exponents for best-fit power laws, and caches higher up the hierarchy would have smaller exponents. This is due to a filtering effect of intervening caches. Requests to NLANR, for example, might first go through a browser, local, regional and/or national caches, each one serving some of the requests. Unless there is a strong correlation between the time to live (ttl) allocated to a file and the file's popularity, this 'filtering' will be systematic. This is because requests for more popular files are reduced more than requests for less popular files, since only the first request for a file from a low level cache reaches a high level cache. If the filtering is systematic there should be a reduction in the exponent observed (illustrated in figure 2). Figure 2 also shows that there would be no change in power law exponent if the filtering was in a 'per request' manner (which would be obtained if ttl was inversely proportional to popularity). Seeking a negative correlation between the hierarchical position of caches, and the fitted exponent can test this hypothesis.

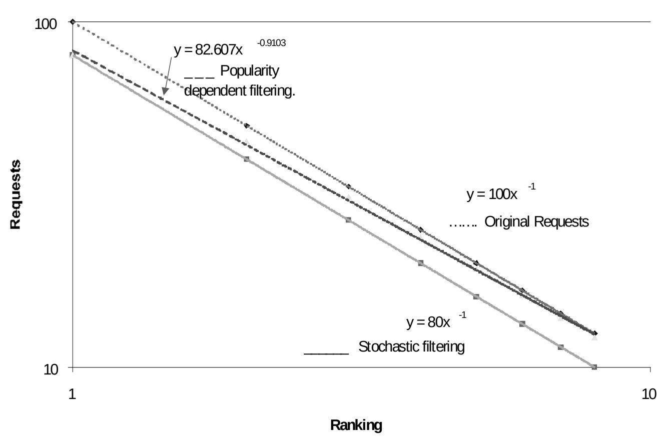


**Figure 2. The possible effects of cache filtering**

While filtering is one possible factor affecting the exponent of the locality curve, other factors possibly influence the exponent. Possible reasons for differences in power law exponent include:

a. **Size of the cache.** It has been proposed [8] that larger caches (i.e. Caches with more requests per day) should have smaller exponents. This can be tested by accurately determining the exponent for a range of caches, at the same position in the hierarchy, and finding a correlation between exponent and size. Taking progressively larger samples from a single cache is not a good test since, as we show below, popularity data is highly bursty and small samples of less than 500000 requests provide unreliable results

b. **The nature of client.** Clients that have large caches will filter requests more than clients with small caches. As the size of the client cache depends on the available disk space, and the disk space is roughly inversely proportional to the age of the computer, areas tending to have newer computers may have lower exponents. So a cache serving an industrial lab should have a lower exponent than a cache serving publicly funded schools.

c. **Number of days that the data is collected over**. It is possible that the popularity curve only approaches stability assymptotically. If the behaviour of individuals is strongly correlated (e.g. by information waves) on a range of timescales with an infinite variance, then the popularity curve exponent will exhibit variation regardless of sample size or timescale. On the other hand if the correlation is only at bounded timescales the exponent will be stable only at timescales larger than the bound. If the behaviour of individual users is only weakly correlated, but has a bounded autocorrelation (e.g. fractional Gaussian statistics), then the exponent should be stable at large sample size regardless of timescale.

d. **Cultural differences between user communities.** Popularity curves are a reflection in user behavior, so differences in this behavior should be reflected in the data [1]. From consideration of the work of Zipf on word use in different cultures, it seems likely that cultural differences will often be expressed through differences in the shape of the popularity curve rather than the exponent. If the exponent is significantly affected by cultural factors then the variation should not be explicable by any obvious cache metrics. This can be tested by using caches which are similar in size and topological position, and demonstrating inexplicable variation in the exponent of the popularity curve

## 3. TECHNIQUES

To analyse file popularity, cache logs are usually needed, the only alternative being the correctly processed output from such a cache log. We are indebted to several sources for making their logs available, and hope this is fully shown in the acknowledgements. We have analysed cache logs from several sources including:

**NLANR-lj,** a high-level cache serving other caches worldwide
**RMPLC,** a cache serving schools in the UK

**FIN,** a cache serving Finnish Universities and academic institutions
**SPAIN**, a cache serving most of the universities and polytechnics of Spain
**PISA,** a cache serving the computer science department of Pisa University, Italy

Processed statistics are also available via web pages. We have used published logs from:

**HGMP** (Human Genome Mapping Project) used by scientists working on the HGMP project in the U.K.
**ACT, Swinburne, Letrobe,** Caches serving academic communities in Australia

The range of logs we have looked contain different proportions of academic and home usage. This is of importance because one possible reason for the variation between caches could be the various usage styles at the caches.

Cache logs can be extremely comprehensive, detailing time of request, bytes transferred, file name and other useful metrics [e.g. ftp://ircache.nlanr.net/Traces/]. It is inevitable though, that they cannot contain every variable, that every researcher requires. At the moment cache logs do not contain the means to discriminate between the physical request made by the client and files that are requested by linkage (linked image file, redirections etc) to the requested files. Some heuristic proposals have been made for filtering out linked requests (e.g. only looking at HTML files [10], filtering out file requests with very close time dependencies [11]), but these inevitably introduce some error into the analysis. Another analysis irregularity is that some researchers look at the popularity of generic hosts and not files. We believe that the best approach is to accept that some pages have embedded links and analyse all requests going through a cache, unlinked or otherwise. The popularity curves in this paper were generated using all the logged requests for files in the analysis period.

A simple, least squares method [16] was used to fit power law curves to the data. The quality of the fit was checked using the standard $R^2$ test. The least squares algorithm did not initially fit the upper (most popular) part of the curve very well. The $R^2$ was between 0.7 and 0.9 and the visual fit was poor (figure 3). In an effort to rectify this, a fit on modified data was used [17]. This involved taking all the files that were requested an identical number of times and averaging their ranking, in effect giving them all the same ranking (which seems fairer). For example, if three files are requested 10 times each and are ranked 100, 101 and 102, then one point would appear on the graph at ranking = 101, popularity = 10. As can be seen in figure 3 this makes for a much tighter visual fit. The improvement is confirmed by much higher $R^2$ values (table 1). The least squares calculation could use a weighting for these averaged points, in proportion to the number of files they represent, but with good $R^2$ values this seemed unnecessary.

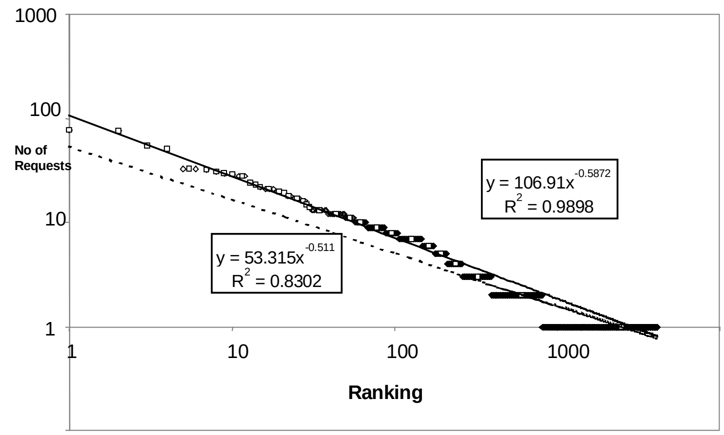


**Figure 3. Illustration of fit calculated by least squares algorithms (dashed line = fit to all points ◆, solid line = fit to average ranking of files requested the same number of times □)**

## 4. VARIABILITY OF LOCALITY

In order to compare data from different caches reliably it is necessary to ensure that differences are real and not due to insufficiently large samples. In order to establish the variability of the fitted exponent we examined the popularity curve of one cache over a long period of time. The cache we chose was the Human Genome Research Project (HGMP) cache in the U.K [http://wwwcache.hgmp.mrc.ac.uk/]. This cache receives about 10000 requests per day from a research community. They publish an access count histogram that gives the number of objects accessed N times, this can be easily converted in to a ranking vs. popularity graph. The least squares procedure can then be used to find the slope of the line with best fit. This was carried out for six months of data from January 1998 to July 1998, inclusive. Over these six months the fitted exponent ranged from -0.23 to -1.34 with a mean of -0.5958 and a variance of 0.03 (figure 4), using the 'averaged' ranking method mentioned above.

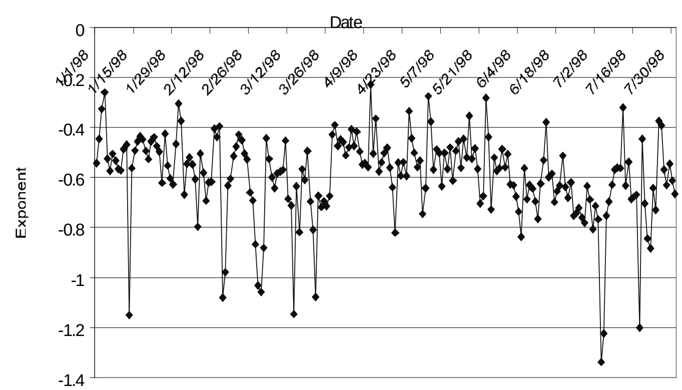


**Figure 4. Variation of best-fit exponent over time**

Figure 4 shows the large amount of variation from the mean. If the data shows long-range dependence the sample size required to get a reliable estimate of the slope of the popularity curve will be considerably larger than might be expected for normal Poisson statistics. A simple aggregated variance graph [15] was plotted using the data in figure 4, to test for long range dependency (fig 5).

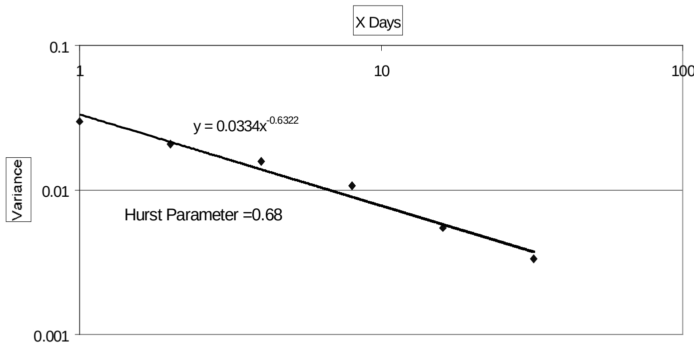


**Figure 5. Aggregate variance of fitted exponents over time.**

The Hurst parameter was calculated using the formula

**Slope = 2H - 2**

The H of 0.68 indicates that there is some self-similarity in 'best fit' exponents over time [5], this implies that large sample sizes will be required for statistics to be 'safe'. It is worth noting that the self-similarity revealed in figure 5 is over 3 orders of magnitude of timescale from hours to weeks. It is thus quite different from the long-range dependency reported elsewhere [13] at timescales of up to a few minutes which is attributable to memory effects in network buffers. It is clear evidence for a second long-range dependency at the timescale of human memory, which system designers will also need to take into account.

In figure 6 we show the effect of increasing the number of requests used to calculate the exponent, at three of the caches we have used. The exponent converges to a stable value for samples of 300 000 or more requests, for all the caches we have analysed. Error estimates are shown at two request ranges, pointing to a need for over 300 000 requests. We have accordingly used samples of 500 000 requests in the analysis reported in the next section. Interestingly the convergence of the exponent does not appear to depend on time – the time taken to collect 500 000 requests varied from 1 day to 1 week. This means that the number of days taken to collect the sample does not significantly affect the slope of the popularity curve. It is worth noting that the apparent dependence of the value of the exponent at small sample size is not real. The trend is simply due to the fact that the exponent has a heavy tailed distribution with most periods giving low exponents and some periods giving very high exponents (as shown in figure 4), thus the most likely result of picking a random sample is to underestimate the exponent. This effect simply diminishes as the sample size increases.

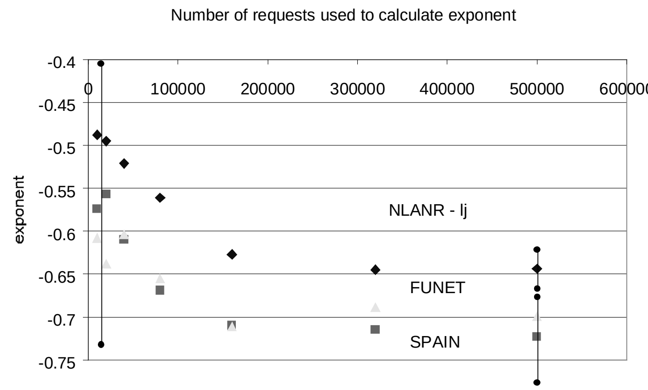


**Figure 6. Effects on exponent of increasing number of requests sampled.**

## 5. ANALYSIS

We have been able to obtain samples in excess of 500000 file requests for 5 very different caches. We show in figures 7 and 8 the popularity curves for these caches, and the curves fitted to the data using the techniques outlined in section 3. In table 1 we show the estimated value of the exponent in the power law, together with the error interval and the confidence limit established by the $R^2$ test.

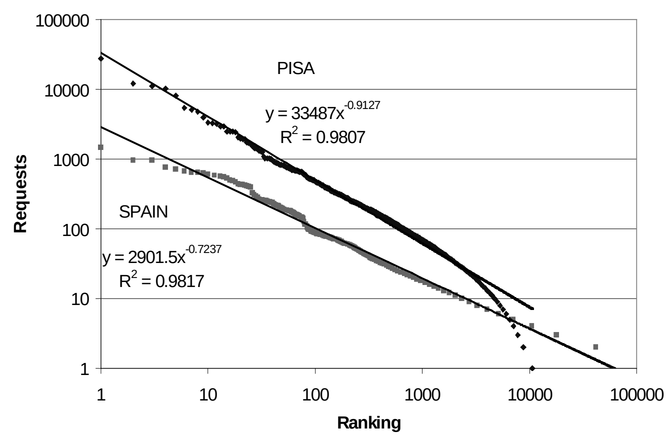


**Figure 7. Least square fit for 2 caches**

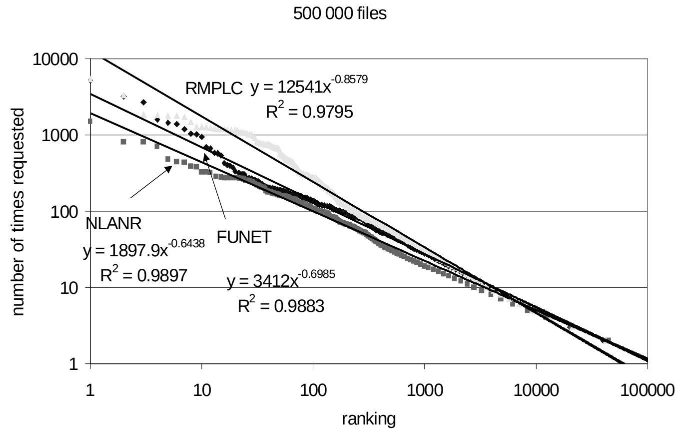


**Figure 8. Least square fits for 3 large caches.**

Figures 7 & 8 clearly show that even when a suitably large number of requests is used, very different best-fit curves are generated for each cache. The NLANR cache is used as a parent by some national caches. FUNET and Spain are national caches, RMPLC and PISA are local caches serving very different communities. The best-fit exponents in table 1 follow this order, so our results are a strong indication that the variation is at least partly due to hierarchy filtering as proposed in section 2.

| Cache | Position | Exponent | R squared | Error estimate |
|---|---|---|---|---|
| NLANR - lj | Highest | -0.644 | 0.9897 | ±0.024 |
| PISA | Local | -0.913 | 0.9807 | ±0.038 |
| FUNET | National | -0.699 | 0.9883 | ±0.046 |
| SPAIN | National | -0.724 | 0.9817 | ±0.045 |
| RMPLC | Local | -0.858 | 0.9795 | ±0.109 |

Table 1. Cache factors of interest.

Error estimates were calculated using several methods, the ones shown were the largest calculated. These were derived by sampling the points on the curve, fitting least squares curves and then looking at the variances of the exponents fitted to the subsets of points to derive standard errors.

As a second approach, it is possible to look at daily locality curves and their fitted exponents if these exponents are averaged over a long enough period (see section 4). This approach was carried out on 4 caches, with results following those of the previous section (table 2). This result is useful since some caches do not publish individual file requests, but do publish daily popularity curves. An example is the HGMP cache, which is shown in the table although the result shown is for too small a sample to be meaningful.

| Cache | Request Based Analysis | Averages of Daily Analysis |
|---|---|---|
| NLANR - lj | -0.644 | -0.6211 |
| PISA | -0.913 | |
| FUNET | -0.699 | -0.701 |
| SPAIN | -0.724 | |
| RMPLC | -0.858 | -0.824 |
| HGMP | | -0.596 |

Table 2. Exponent calculated using 500000 requests compared with averaging daily exponent

## 6. DISCUSSION

The data in section 5 supported the notion that the variation in cache popularity curves is simply due to the hierarchical position of the cache. Our data shows no evidence supporting any other of the possibilities we considered in section 2. We showed in section 3 that the variation was not a temporal effect, and that size was probably not significant. As it is commonly supposed that size is a major factor it is worth presenting futher support for our contention that it is not here. Figure 9 shows cache size plotted against exponent for 6 caches. We cannot fit a line to this data with a meaningful degree of confidence, but any line would show no relationship between size (requests/day) and exponent.

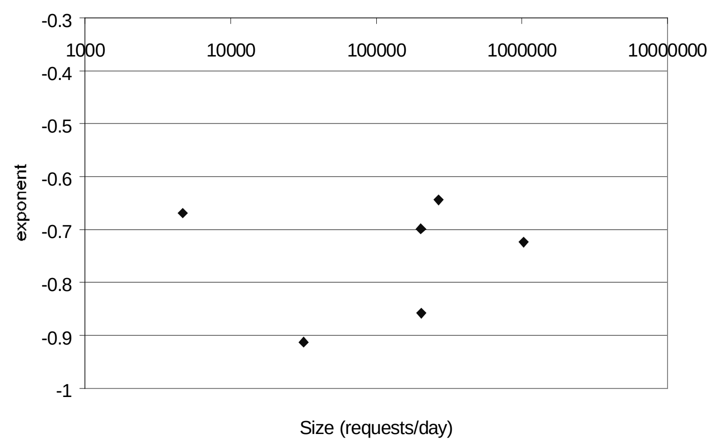


**Figure 9. Relationship between requests/day and exponent.**

As for culture dependence, consider the low level caches RMPLC and PISA. RMPLC serves UK secondary schools and has a user community which is very different from a ‘normal’ web user community. The users are severely constrained, strongly guided by teachers and syllabi and very young. It is hard to imagine a user community more different from the undergraduates, lecturers and researchers at Pisa university. Yet both sites show similar exponents. There is thus no evidence in our data for cultural dependency in the value of the exponent. Of course this does not mean that other cache metrics will not show culture dependencies.

The lack of significant differences between caches at similar apparent levels in the hierarchy means that client effect are not significant either. However, we believe that increased cache capability at clients may be one reason why old cache logs tend to have higher exponents than current logs. This supposition deserves further analysis.

The discovery of a relatively straightforward cache metric is proving very beneficial in a cache model we have built which is reported elsewhere [6], that has been proved to make accurate predictions about the performance of caching systems. Previous versions of the model, which used hit rates, were not accurate since the hit rate depends strongly on the (unknown) cultural interests of the users. The model has already been instrumental in improving the resilience and scalability of BT caching architectures, without incurring the cost of deploying an experimental cache network. The model is currently being further developed at BT, along with other simulation programs (e.g. 4) that evaluate cache performance by generating a representative stream of requests. These models require an accurate description of real cache behavior so that system performance can be accurately assessed. Any simulator must express the same behavior that has been demonstrated in this paper if the results are to be accurate.

The accurate determination of the parameter requires large samples. This is due to a large timescale, long range dependency in the user requests, which we discuss more fully elsewhere. It is worth emphasising here that any deductions about cache properties based on small sample sizes will be misleading and all inputs to models should be based on large sample sizes.

## 7. CONCLUSION

The analysis of cache popularity curves requires careful definition of what is to be analysed and, since the data displays significant long range dependency, very large sample sizes. With appropriate care it is possible to fit an inverse power law curve to cache popularity curves, with an exponent of between -0.9 and -0.5, and with a high degree of confidence. The exponent does not appear to depend on cache size, on time, or on the culture of the cache users, but only depends on the topological position of the cache in the network. It is an extremely useful metric for modelling purposes since cache position is easy to determine. Other metrics which have cultural dependencies are extremely hard to use since the culture of a cache user community is near to impossible to determine without time consuming semantic analysis of the requested files. Further data should be analysed to fully confirm the relative independence of the metric.

## 8. ACKNOWLEDGEMENTS

We would like to thank Pekka Järveläinen (Funet), Simon Rainey (RMPLC), Javier Puche (Centro Superior de Investigaciones Cientificas) and Luigi Rizzo (Pisa University) for supplying us with anonymised logs. We would also like to thank Luis Alberto Velasco for his assistance with writing the log analysis scripts. We would also like to thank the 3 anonymous reviewers for their useful and informative comments.

## 9. REFERENCES

[1] V Almeida, M Cesario, R Fonseca, W Meira Jr., C Murta. Analyzing the Behavior of a Proxy Server in Light of Regional and Cultural Issues. 3rd International WWW Caching Workshop. http://wwwcache.ja.net/events/workshop/21/

[2] M Abrams, C Standridge, G Abdulla, S Williams and E Fox. Caching Proxies: Limitations and Potentials. Proc. 4th Inter. World-Wide Web Conference, Boston, MA, Dec. 1995.

[3] M Baentsch, L Baum, G Molter, S Rothkugel and P Sturm. Enhancing the web's infrastructure: From caching to replication. IEEE Internet Computing. March 1997. P. 18-27.

[4] P Barford and M Crovella. 'Generating representative Web workloads for network and server performance evaluation.' in Proceedings of Performance '98/ACM SIGMETRICS '98, pp. 151-160, Madison WI.

[5] J Beran. 'Statistical methods for data with long-range dependence'. Statistical Sciences 1992, Vol. 7, No. 4, 404-427.

[6] George Bilchev, Sverrir Olafsson, Ian Marshall and Chris Roadknight
'Modelling HTTP Traffic Generated by a Community of Users.' HPCN99 12-13-14 April.
http://www.wins.uva.nl/events/HPCN99/index.html

[7] L Breslau, P Cao, L Fan, G Phillips and S Shenker. 'Web Caching and Zipf-like Distributions: Evidence and Implications.' IEEE Infocom '99. http://www.cs.wisc.edu/~cao/papers/zipf-implications.html.

[8] L Breslau, P Cao, L Fan, G Phillips and S Shenker. 'On the implications of Zipf's Law for web caching'. in 3W3Cache Workshop, Manchester, June 1998.

[9] C Cunha, A Bestavros, and M Crovella. Characteristics of WWW client-based traces. Technical report TR-95-010, Boston University Department of Computer Science, April 1995.

[10] B Huberman, P Pirollo, J Pitkow, R Lukose, "Strong regularities in world wide web surfing", Science, 280:95-97 (April 3, 1998).

[11] I Marshall, C Roadknight, 'Linking cache performance to user behaviour', 'Computer Networks and ISDN systems' 30 (1998), pp.2123-2130.

[12] M Marsili, Y Zhang, "Interacting individuals leading to Zipf's law", Physical Review Letters, 80(12), 2741-2744 (March 23, 1998)

[13] J Pitcow, 'Summary of WWW characteristics', Computer Networks and ISDN Systems 30 (1998) 551-558

[14] C Roadknight, I Marshall, 'Variations in cache behaviour', in 'Computer Networks and ISDN systems' 30 (1998), pp.733-735.

[15] M Taqqu, V Teverovsky and W Willinger. 'Estimators for long-range dependance: an empirical study.' Fractals. Vol 3, No. 4 (1995) 785-788.

[16] J Topping. Errors of observation and their treatment. Harper and Row Publishers, INC. 1972

[17] G. K Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, Cambridge, MA, 1949.