Sources and Characteristics of Web Temporal Locality*

SHUDONG JIN and AZER BESTAVROS
Computer Science Department
Boston University
Boston, MA 02215

{jins,bestavros}@cs.bu.edu

Abstract

Temporal locality of reference in Web request streams emerges from two distinct phenomena: the long-term popularity *of Web documents and the* short-term temporal correlations *of references. In this paper we show that the commonly-used distribution of inter-request times is predominantly determined by the power law governing the long-term popularity of documents. This inherent relationship tends to disguise the existence of short-term temporal correlations. We propose a new and robust metric that enables accurate characterization of that aspect of temporal locality. Using this metric, we characterize the locality of reference in a number of representative proxy cache traces. Our findings show that there are measurable differences between the degrees (and sources) of temporal locality across these traces.*

1. Introduction

Two of the most important properties of Web access patterns are the skewed popularity of Web documents and the temporal locality of reference exhibited in request streams. Studies of Web access streams indicated that long-term popularity follows a Zipf-like distribution [4, 8], whereby the access frequency of a document is proportional to the reverse of its rank. Temporal locality—the property that recently requested documents are more likely to be requested again—was also observed in Web request streams [2, 5, 11]. This property can be captured by the inter-request time distribution [2, 5].

The distinction between popularity and temporal locality properties in Web access patterns is often blurred. This is partially the result of the inherent relationship between these two properties (as we will establish in this paper). However, popularity is not the only determinant of temporal locality of reference. In particular, *temporal correlations* of repeated requests to the same documents is another important contributor to temporal locality. The delineation between these two sources of temporal locality is important due to their implication on Web content caching and replication protocols. An important question is whether these sources can be characterized accuractely with a simple model.

In this paper we provide an affirmative answer to this question. First we show that inter-request time distribution commonly used in characterizing temporal locality [2, 4, 5, 7, 14] is induced through long-term popularity distribution, but is insensitive to temporal correlations of reference. We quantify the asympototical inherent relationship between Zipf-like popularity distribution and inter-request time distribution. Second, we propose a new model that captures both aspects seperately, and together temporal locality. Third, using this method, we characterize the temporal locality of reference in a number of representative proxy cache traces.

2. Related Work

Denning and Schwartz [9] established the fundamental properties that characterize the phenomenon of temporal locality. The presence of temporal locality in Web access streams specifically, has also been widely verified [1, 2, 5, 11, 14]. Such properties are important for the practices in the design of caching and replication mechanisms. Therefore, it is important to characterize the degree of temporal locality. Among others, two methods are well adopted, namely

*This work was partially supported by NSF research grant CCR-9706685.

the stack distance model and inter-request time distribution model.

In [16], Mattson et al introduced the concept of stack distances as a means for analyzing the behavior of demand-paged memory systems and for evaluating the performance of memory management schemes. Stack distance refers to the number of unique documents separating consecutive requests for the same document. In [18], Spirn proposed the use of a Markov stack distance model to capture program behavior. A Markov stack distance model enables the prediction of future inter-request distances based on the most recently generated distances. As such, this model is unable to capture the long-range dependencies among requests and hence is unable to capture the bursty nature of page faults. In [1], Almeida et al used the marginal distribution of stack distance strings to characterize temporal locality. Their analysis of Web request streams, revealed that the marginal distributions of stack distances follow a log-normal distribution. While the stack distance model provides means for characterizing the degree of temporal locality that exists in a request stream, it is not able to delineate the causes of such locality—namely temporal locality due to popularity and due to the temporal correlations of reference [15].

The second method in characterizing temporal locality is the use of the distribution of inter-request times [2, 5, 14]. In [4], Breslau et al found that the distribution of inter-request times is tightly related to document popularity. They argued that the probability of referencing an object t units of time after it has been last requested is *roughly* proportional to $1/t$. However, since their work is based on the Independent Reference Model[1] (IRM) [6] their approach is incapable of capturing the temporal correlations in the reference stream.

3. Trace Characteristics

In this paper we use traces from DEC [10] and NLANR [17]. Some characteristics of these traces are shown in Table 1, including the total/unique number of requests and their aggregated sizes. Our preprocessing of the DEC traces excluded non-cache-able requests, including cgi-bin requests and queries. Our preprocessing of the NLANR traces was more elaborate. The NLANR traces include many IMS (If-Modified-Since) and REFRESH requests with reply code 304 (Not Modified). In order to include such requests in the workload, we had to find the sizes of the documents of such requests. We do so through a 2-pass scanning of the entire trace. In addition to this preprocessing, we have also excluded non-cacheable requests, including cgi-bin requests and queries.

[1]IRM assumes that a request stream consists of a sequence of independent, identically-distributed random variables.

Table 1. Traces used in this paper

Trace	Period	All reqs/GB	Unique reqs/GB
DEC	'96:29/8-4/9	3,543,968 / 45	1,354,996 / 22
NLANR.RTP	'99:4/6-17/6	9,113,027 / 91	3,249,549 / 45
NLANR.SD	'99:4/6-17/6	9,082,461 / 129	3,549,609 / 62
NLANR.UC	'99:4/6-17/6	8,983,585 / 113	2,459,366 / 47

Popularity Distribution: Numerous studies [3, 4, 8] have shown that a Zipf-like distribution, i.e. a power law, can model the relationship between the long-term popularity of a document and its popularity rank. This relationship can be expressed as: $P \sim \rho^{-\alpha}$, $0 < \alpha < 1$, where P is the document's long-term popularity (number of references) and ρ is the rank of its popularity. Thus the value of α could be used to characterize the Zipf-like popularity of Web documents in a request stream. Table 2 shows a least-square fit of the values of α for our trace set.

Table 2. Values of α found in our trace set.

Traces	DEC trace	RTP trace	SD trace	UC trace
α	0.77	0.71	0.72	0.66

Distribution of Inter-Request Times: Cao et al [5] observed that the probability of a future request is quantitatively related to the time elapsed since the last request. Breslau et al [4] indicated such property is not an artifice. They found that the probability of referencing a document at t time after it has been last requested is *roughly* proportional to $1/t$. To examine this model, in Figure 1(left) we plot the fraction of requests as a function of the distance[2] between two consecutive requests for the same documents.

For the DEC trace, the log-log scale plots in Figure 1(left) are nearly straight lines with slope quite close to -1.0 except for the existence of diurnal spikes [5, 11]. This confirms the applicability of the inter-request time distribution. For the NLANR traces, the plots display similar properties, except that the slopes of the bodies are in the $-0.65 \sim -0.73$ range. The

[2]We measure the distance using the number of intervening requests as opposed to the absolute time to mask diurnal effects. Characteristics of request interarrivals measured in absolute time are available from [12].

variations in the slope across the NLANR traces can explained by noting that the traces reflect different workloads. These variations aside, the slopes for the NLANR traces are clearly less pronounced than that measured for the DEC trace.

4. From Popularity to Temporal Locality

The skewed distributions evident in Figure 1(left) may well be a reflection of the skewed popularity distribution characterized in 2. Highly popular documents tend to be requested frequently, and thus will exhibit shorter inter-request times; less popular documents, on the other hand, tend to be requested infrequently, and thus will exhibit longer inter-request times. We establish this relationship quantitatively next.

We consider the changes in the inter-request time distribution when the request streams are subjected to a random permutation. The right-hand-side plots in Figure 1 are obtained by applying a random permutation to the request streams, thus eliminating temporal correlations while preserving the document popularity distribution in the request streams. Comparing the left-hand-side and right-hand-side plots in Figure 1 reveals no significant changes in the distributions except for being smoothed as a result of trace scrambling. Table 3 shows the slopes of these plots using a least-square fit for $0 < x < 1,000,000$. The fact that there is little change as a result of scrambling suggests that the inter-request time distribution is predominantly determined by the document popularity distribution in the trace, and thus cannot effectively quantify the degree of temporal correlations in the request streams.

Table 3. Slopes of the curves in Figure 1.

Traces	DEC	RTP	SD	UC
Original	0.95	0.71	0.65	0.73
Scrambled	0.84	0.68	0.62	0.69

The strong relationship between popularity and inter-request time distributions is evident in the relationship between the slopes identified in Table 2 and Table 3. This relationship is formalized by the following Theorem.[3]

[3]For $\alpha = 1$, Breslau et al [4] have also related popularity and inter-request time distributions. Our theorem is more general as it applies to any value of α, suggesting that a trace with less skewed popularity distribution tends to have weaker temporal locality and worse cacheability.

Theorem 1 *If the distribution of document popularity in a request stream asymptotically follows a power law with parameter α, where $0.5 \ll \alpha \leq 1$, then the distribution of inter-request time in a random permutation of this request stream can be characterized asymptotically using a power law with parameter $(2 - \frac{1}{\alpha})$.*

Proof: Let k denote the number of requests for a page and $N(k)$ be the number of pages requested k times. Let $F(t)$ denote the number of instances whereby two requests for the same document are separated by t units of time, where $0 < t < 1$ is the normalized time spanning the randomly-permuted request stream. Notice $F(t)$ reflects the probability distribution of inter-request time. Let $P(t,k)$ denote the number of instances whereby two requests for the same document with total k requests are separated by t units of time, where $k \geq 1$ (i.e. we are only interested in documents requested at least twice). We make the following two observations:

1. $N(k) \sim k^{-1-\frac{1}{\alpha}}$, where α is the parameter of the power law governing the Zipf-like distribution of document popularity. This observation follows by noting that $N(K)$ can be obtained by transposing the X-axis and Y-axis of the popularity distribution and taking the derivative of the resulted function. Let $N(k) = Ck^{-1-\frac{1}{\alpha}}$.

2. $P(t,k) \sim ke^{-kt}$. The inter-arrival time of randomly scattered requests for an object follows an exponential distribution with a mean close to $1/k$. Normalizing it, we get $P(t,k) = \frac{ke^{-kt}}{1-e^{-k}}$. Since $1 - e^{-k}$ approaches unity so fast when k increases, it follows that we can take $P(t,k) \approx ke^{-kt}$, without affecting the asymptotic property we are attempting to establish.

For a document requested k times, the number of inter-request time appearing in the request stream is $(k-1)$. From the definition of $P(t,k)$, we get that the total occurrences of inter-request time t is given by $(k-1)P(t,k) = k(k-1)e^{-kt}$. Thus, the total number of occurrences for all documents in the request stream is $N(k)(k-1)P(t,k) = Ck^{-\frac{1}{\alpha}}(k-1)e^{kt}$. Finally, we compute $F(t)$ by summing up the value of $P(t,k)$ for all k's:

$$\begin{aligned} F(t) &= \sum_k N(k)(k-1)P(t,k) \\ &= \sum_k Ck^{-\frac{1}{\alpha}}(k-1)e^{kt}, k \geq 2 \qquad (1) \end{aligned}$$

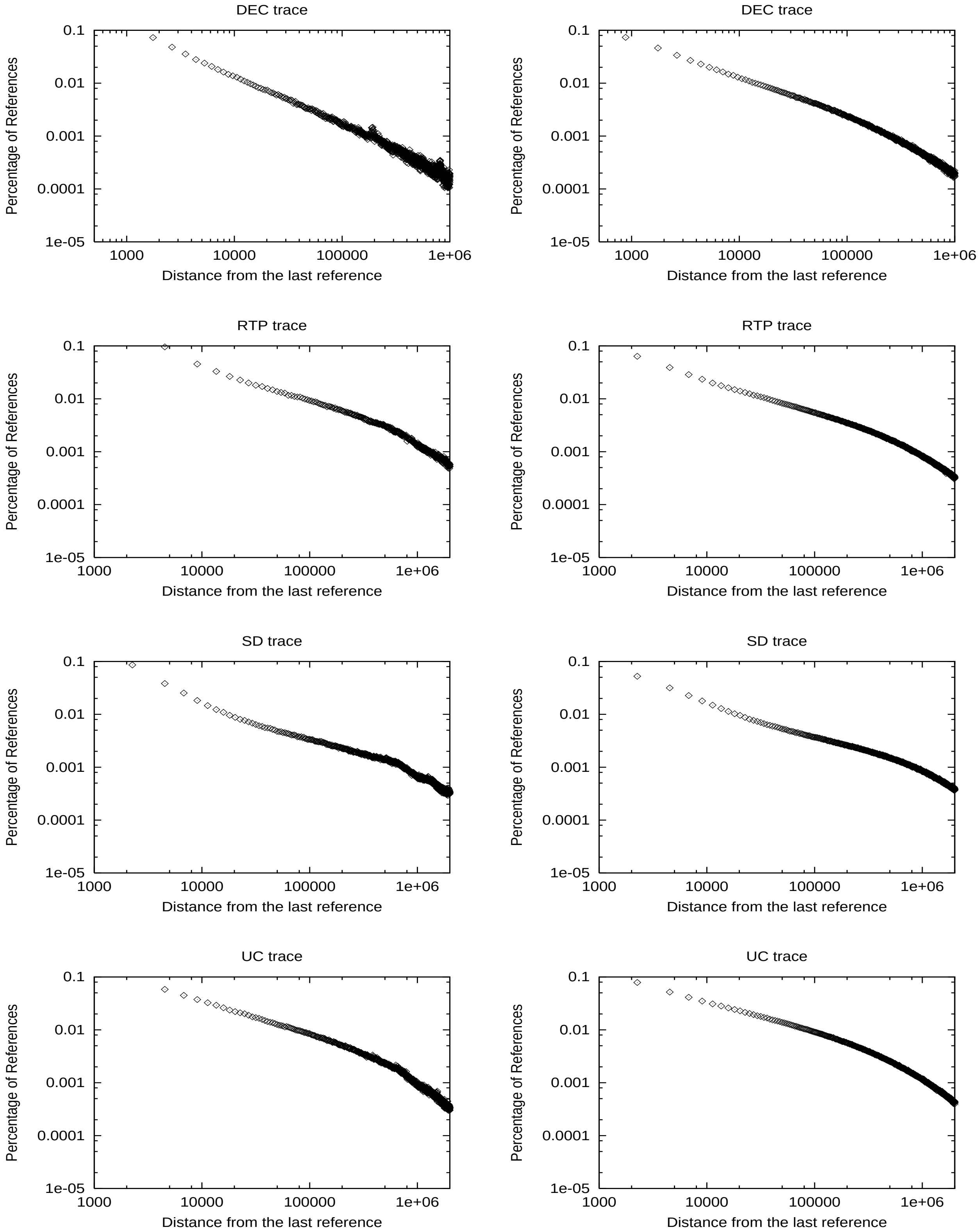

Figure 1. Probability distribution of inter-request time of the original request streams (left) and a random permutation thereof (right).

Replacing the summation in the equation with an integral, we obtain:

$$F(t) \approx \int_2^{\infty} Ck^{-\frac{1}{\alpha}}(k-1)e^{kt}dk = C\frac{\Gamma(2-\frac{1}{\alpha},2t) - t\Gamma(1-\frac{1}{\alpha},2t)}{t^{2-1/\alpha}} \quad (2)$$

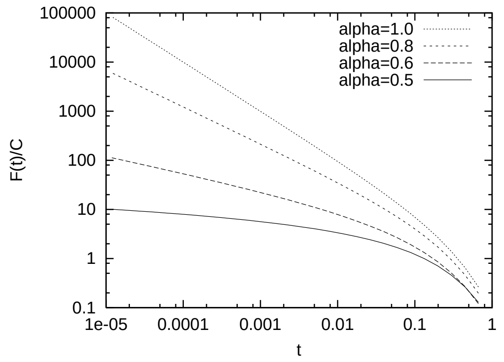

Figure 2. $F(t)/C$ with parameter α.

Figure 2 displays $F(t)/C$ for different values of α. In this equation, $\Gamma(a,x) \equiv \int_x^{\infty} t^{a-1}e^{-t}dt$ is the incomplete Gamma function. Comparing it to $\Gamma(2-1/\alpha, 2t)$ when t is small, $t\Gamma(1-1/\alpha, 2t)$ is insignificant for any $0.5 \ll \alpha \leq 1$. Moreover, the complete Gamma function $\Gamma(2-\frac{1}{\alpha}) \equiv \Gamma(2-\frac{1}{\alpha}, 0)$ is analytical since $2-\frac{1}{\alpha} > 0$. When t is small, $\Gamma(2-\frac{1}{\alpha}, 2t)$ is close to $\Gamma(2-\frac{1}{\alpha})$, a constant. Therefore in this case:

$$F(t) \sim \frac{1}{t^{2-1/\alpha}} \quad (3)$$

When t is not very small, the second item of the numerator is non-negligible, so there will be a visible "dip" in the log-log scale plot. ■

It is important to note the difference between the above asymptotic relationship and the distributions obtained from realistic request streams. In particular, when α is close to 0.5, theoretically, the slope of inter-request time distribution approaches 0. However, since the length of a realistic request stream is limited, this asymptotic slope can not be reached. Moreover, for any α generally, the derived inter-request time distribution does not strictly follow power law (as shown in our proof). This is also indicated by the shapes of the lines in Figure 1(right) which are not perfectly straight. Even though, we are still able to measure slopes closer to their asymptotic values $(2-\frac{1}{\alpha})$, e.g., by applying the least-square fit over a shorter range $0 < x < 100,000$ instead of $0 < x < 1,000,000$ in order to avoid the corruption attributed to the limited trace length. Table 4 shows the closeness of the measured slopes in this way and the theoretical slopes.

Table 4. Derived & estimated slopes in Figure 1.

Traces	DEC trace	RTP trace	SD trace	UC trace
Derived	0.70	0.59	0.61	0.49
Estimated	0.71	0.62	0.64	0.55

5. Characterizing Temporal Correlation

The strong relationship between popularity and temporal locality, as evidenced by the closeness of the slopes in Table 3, often disguises an important aspect of temporal locality, i.e., the temporal correlations of repeated requests for the same documents. This has lead to (for example) the inadequate conclusion in [4] that a Zipf-like popularity distribution, together with an independent reference model, is enough to explain/characterize temporal locality. In this section we propose a new model to accurately measure the degree of short-term temporal correlations of reference.

To characterize only the temporal correlations in a request streams, we need to eliminate the effect of popularity. The new model we are proposing here is *the probability distribution of inter-request time for equally popular documents.*

To demonstrate the expressiveness of this model, we consider the distribution of inter-request time for documents after they appear k times in the request stream. We assume there is a warm-up period, so that long-term popularity is rather accurate. Then we get the samples from the second half of the request stream. Figure 3 (left) shows the plots when k=1, 2, 4, and 8 for the RTP trace. Because we restrict our characterization of request interarrivals to equally-popular documents (those requested k times so far), the effect of popularity is eliminated but the effect of temporal correlations is preserved. Figure 3 (right) shows the inter-request time distribution for documents requested k times when the trace is scrambled. The results indicate that the slope has all but disappeared, which is expected due to the elimination of both popularity and temporal correlations effects.

The results for the other traces and for other values of k are similar. When we used other traces, again we plotted curves of different shapes for the original request streams and their scrambled versions. For larger values of k, we need only to state the following finding: the corresponding curves of the original

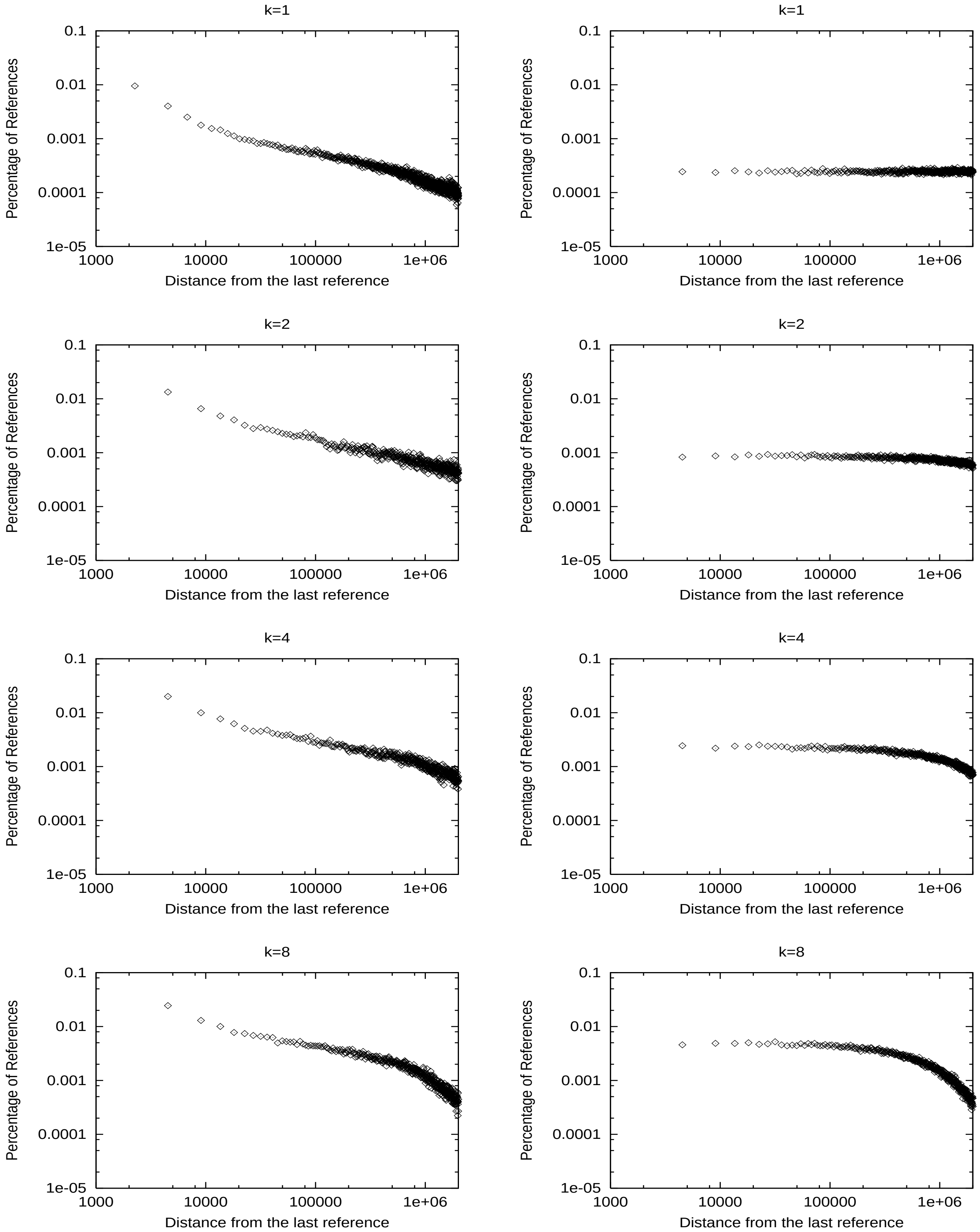

Figure 3. The inter-request time probability distribution for the equally popular documents in the RTP trace for the original request streams (left) and a random permutation thereof (right).

and scrambled request streams have different shapes reflecting contrary short-term temporal correlations properties.

To elaborate on this point, we should point out that: (1) short-term temporal correlations still exist in the original request streams even when the long-term popularity increases, as we plotted the curves for larger k's. Generally, for any k, the slope of the curves is a little more pronounced when the inter-request time is shorter (i.e, convex), which suggests that short-term temporal correlations exists a little stronger in shorter time scales. (2) However, the same curves for the scrambled request streams exhibit a contrary property: namely that the slope is quite closer to 0 (i.e. no slope) when the inter-request time is shorter (i.e., concave).

Notice that the inter-request time distribution of equally popular documents (of k requests to each) roughly follows an exponential distribution when the trace is scrambled. Let us denote this distribution as $F(t) \sim ke^{-kt}$. Applying logarithm to both sides, and let $Y = \log P$ and $X = \log t$, then we have an equation of form $Y = C_1 - C_2 e^X$, where C_1 and C_2 are constants independent of X. It explains the shapes of the log-log scale curves in Figure 3 (right). When t is small, the curves are close to horizonal. This also explains the drops at the extreme right of both-sides plots in Figure 3.

The above findings indicate that our proposed model, the probability distribution of inter-request time for equally popular documents, is an expressive model for capturing the short-term temporal correlations in a request stream. The degree of short-term temporal correlations can be approximately quantified by a parameter β, the slope of the log-log scaled inter-request time distribution for equally popular documents (though we noticed such curves are slightly convex). Specifically, for a set of documents with the same popularity, the probability that the inter-request time equals t is roughly proportional to $t^{-\beta}$.

Table 5 gives the ranges of β for $0 < x < 100,000$, corresponding to less than about 3 hours. To obtain the values of β in Table 5, plots similar to those in Figure 3 are drawn and the value of β is estimated with a least-square fit for $0 < x < 100,000$. Table 5 indicates that the value of β is rather stable for different values of k's, but that it is quite different from across traces.

Table 5. Values of β for level-k popularities.

Traces	DEC trace	RTP trace	SD trace	UC trace
$k = 1$	0.61	0.51	0.39	0.50
$k = 2$	0.63	0.49	0.41	0.50
$k = 4$	0.63	0.47	0.40	0.46
$k = 8$	0.65	0.46	0.41	0.43

6. Locality Characteristics & Implications

To summarize the findings of the last two sections, temporal locality is induced through two sources: long-term popularity of documents and short-term temporal correlations of reference. Long-term popularity is captured by α, the parameter of the power law that governs the relationship between frequency of access and rank. Short-term temporal correlations are captured by β, the parameter of the power law that governs the relationship between frequency of access and the interarrival of requests to equally-popular documents. Thus, temporal locality is adequately characterized using the pair (α, β).

The values of the pair (α, β) for the different traces are summarized in Table 6, where α values are from Table 2 and β values are averaged from the sets of equally popular documents up to 10 references. We notice that both parameters vary accross the traces. Interestingly, both values for the DEC trace are the largest. The larger α value for DEC indicates that the requests are more likely to be concentrated on fewer very popular documents. This might be is a property of the documents themselves or a reflection of the decrease in popularity skew from 1996 to 1999 (also documented in [3]). The larger β value for the DEC trace shows that repeated requests to the same documents are more likely to be correlated in time (more so than what has been observed for the NLANR traces).

Table 6. Reference locality parameters (α, β)

Traces	DEC trace	RTP trace	SD trace	UC trace
(α, β)	(0.77,0.64)	(0.71,0.47)	(0.72,0.40)	(0.66,0.46)

We argue that the significant differences between the α and β values across traces is likely due to: (1) the different workloads experienced at the NLANR and DEC proxy servers; the NLANR proxy servers deal with requests from more diverse client population, and (2) the improving efficiency of client-side caching in '99 versus '96 [3], which serves as a filter of the requests [19].

Our characterization of temporal locality indicated that long-term popularity is the predominant contributor to temporal locality, but that temporal correlations exists. It advocates the use of a frequency-based, recency-aware cache replacement policy. In [13], we derived an algorithm—called GreedyDual*—that captures both long-term popularity and short-term temporal correlations in an adaptive fashion. We compared the performance (both hit ratio and byte hit ratio) of this algorithm with other algorithms such as LRU, GDS and LFU-DA. Our findings include: (1) Our proposed algorithm consistently achieves the best performance. The improvement is usually 10-30%. (2) The incorporation of long-term frequency is important for Web proxy caching algorithm, in line with our finding that document popularity is the main contributor to temporal locality. (3) Capturing temporal correlations is more crucial for smaller cache, in line with our finding that temporal correlations exists in short term. (4) Sensitivity analysis indicated that incorporation of β successfully captures the relative strength of long-term popularity and short-term temporal correlations.

7. Conclusion

In this paper we have shown that there are two phenomena that contribute to temporal locality in Web request streams: the long-term popularity of documents and short-term temporal correlations of reference. To capture both, we suggest the use of two power laws—one characterizing popularity distribution and the other characterizing the inter-request time distribution for equally-popular documents. Using this model we characterized both sources of temporal locality in different traces and established that the parameters of these power laws are different across traces, but consistent within each trace.

References

[1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the WWW. In *Proceedings of PDIS*, December 1996.

[2] M. Arlitt and C. Williamson. Web server workload characteristics: The search for invariants. In *Proceedings of ACM SIGMETRICS*, May 1996.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: Characteristics and caching implications. *World Wide Web*, 2(1):15–28, 1999.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *Proceedings of IEEE INFOCOM*, April 1999.

[5] P. Cao and S. Irani. Cost-aware WWW proxy caching algorithms. In *Proceedings of USITS*, December 1997.

[6] E. G. Coffman and P. J. Denning. *Operating systems theory*. Prentice-Hall, 1973.

[7] E. Cohen and H. Kaplan. Exploiting regularity in Web traffic patterns for cache replacement. In *Proceedings of ACM STOC*, May 1999.

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

[9] P. Denning and S. Schwartz. Properties of the working set model. *Communications of the ACM*, 15(3):191–198, 1972.

[10] Digital Equipment Corporation. ftp://ftp.digital.com/pub/DEC/traces/proxy/.

[11] S. D. Gribble and E. A. Brewer. System design issues for Internet middleware services: Deductions from a large client trace. In *Proceedings of USITS*, December 1997.

[12] S. Jin and A. Bestavros. Temporal locality in Web request streams: Sources, characteristics, and caching implication. Technical Report BU-CS-99-014, Computer Science Department, Boston University, October 1999.

[13] S. Jin and A. Bestavros. GreedyDual* Web Caching Algorithm: Exploiting the Two Sources of Temporal Locality in Web Request Streams. In *Proceedings of the 5th International Web Caching and Content Delivery Workshop*, Lisbon, Portugal, May 2000.

[14] P. Lorenzetti, L. Rizzo, and L. Vicisano. Replacement policies for a proxy cache. Technical Report LR-960731, Univ. di Pisa, 1997.

[15] A. Mahanti. Web proxy workload characterization and modelling. Master's thesis, Department of Computer Science, University of Saskatchewan, September 1999.

[16] R. Mattson, J. Gecsei, D. Slutz, and I. Traiger. Evaluation techniques and storage hierarchies. *IBM Systems Journal*, 9:78–117, 1970.

[17] National Laboratory for Applied Network Research. ftp://ircache.nlanr.net/Traces/.

[18] J. Spirn. Distance string models for program behavior. *IEEE Computer*, 13(11), November 1976.

[19] D. Weikle, S. McKee, and W. Wulf. Caches as filters: A new approach to cache analysis. In *Proceedings of IEEE MASCOTS*, July 1998.