Towards Improved Web Acceleration: Leveraging the Personal Web

Azarias Reda
University of Michigan
azarias@umich.edu

Edward Cutrell
Microsoft Research
cutrell@microsoft.com

Brian Noble
University of Michigan
bnoble@umich.edu

ABSTRACT

Web acceleration mechanisms play an important role in challenged network environments where connectivity is limited or expensive. However, as web usage gets increasingly personal and fragmented, traditional web acceleration systems that leverage redundancy in user requests to optimize performance find it difficult to perform well. This is unfortunate because personalization is an otherwise important trend that allows users to focus on content that is relevant to them. To start tackling this growing problem, this paper makes three contributions. First, we provide the first personalized, large scale web usage data in a developing country context. This allows researchers to get a nuanced understanding of access behavior that is not offered by aggregate data. Second, we present some analysis on this dataset, which provides tangible evidence for describing the increasingly fragmented and personal nature of web access even in developing countries. Finally, based on lessons learned from the analysis, we provide some recommendations for building effective web acceleration mechanisms in the face of an increasingly personal web. We believe the next generation of web acceleration systems for challenged networks need to have a strong personal component.

Categories and Subject Descriptors

C.2.0 [**Computer Communication Networks**]: Local and Wide-Area Networks

General Terms

Measurment

Keywords

web acceleration, challenged networks

1. INTRODUCTION

Web usage is growing largely personal. The most popular internet destinations increasingly seem to be those that provide individualized experiences to their users. On the other hand, even traditionally 'static' content such as news is increasingly localized and personalized. As the amount of information available on the web grows exponentially [2, 8], personalization is a welcome trend that allows users to focus on what is relevant to them. Just as important, the sheer volume of content available online enables users to choose from a diverse set of services that cater to their personal preferences and interests.

Unfortunately, this creates challenges in designing and building web acceleration mechanisms. These mechanisms are especially important in many developing country environments where connectivity is poor and the network infrastructure is generally challenged. Even when good connectivity is available, it's often very expensive, and comes with costly bandwidth caps [7, 17]. As a result, effective web acceleration systems that work well within these constraints are still very useful. However, many existing systems are built on basic mechanisms that predate the increasingly personal nature of web usage.

As a result, traditional approaches towards web acceleration find it difficult to perform well. For example, recent studies of web acceleration in developing country contexts have reported that overall cache hit rate from prefetching content for users is very low, often given in single digit percentage points [4, 13]. When considering the cost of prefetching unused content in already-constrained environments, those gains dwindle to almost none. This is perhaps to be expected as many systems are built with an aggregate view of their users, while web experience has been getting more personal. This lack of a nuanced understanding of web usage in developing countries limits the effectiveness of many existing systems.

In order to start tackling this growing problem, this paper presents three concrete steps. Our first contribution is collecting the first personalized, large scale web usage dataset in developing countries. While several datasets with aggregate information about web usage patterns in developing countries exist [4, 6, 11, 15], none of them provides an individualized look into personal web usage. Our data set was collected at two sites over a period of one month, and contains web usage information for about 470 users segmented across several sessions. This data has been anonymized to remove personally identifying information, and is available for researchers to access. We believe this dataset will provide a much needed insight for web system developers targeting developing countries.

We have also done some analysis on this individualized dataset in order to quantify the personal nature of web ac-

NSDR'11, June 28, 2011, Bethesda, Maryland, USA.

cess in developing regions. While there has always been anecdotal evidence to suggest web usage is getting increasingly personal and dynamic even in developing regions, our analysis provides some tangible evidence to describe the phenomenon. For example, we found that a majority of our users spent more than 40% of their browsing time on private web destinations such as email and social networking sites. Likewise, nearly 60% of data transferred over the month was initiated from domains that were requested by less than 3% of the user base. On the other hand, we find high similarity between different sessions of the same user, even more than what has been reported for users in well connected environments [16].

Finally, using lessons learned from our analysis, we provide some recommendations on how to design better web acceleration mechanisms for developing regions. The personal and dynamic nature of web usage presents both a challenge and an opportunity in building web acceleration systems. The challenge is rethinking traditional mechanisms with an aggregate view of the user base to cope with emergent trends in web usage. However, this presents an opportunity in designing personalized web acceleration mechanism that understand and cater to users in developing countries. We believe the next generation of web acceleration systems for challenged networks need to have a strong personal component.

2. DATA COLLECTION

The collection of personal data usage must be approached with a great deal of sensitivity. Our goal was to collect web usage logs from a diverse set of users, while having some level of repetition among those users. In particular, we would like to attribute each browsing session to a unique user, and do so consistently across sessions. While its often easier to access such data in organizations that require their users to log in for web access, that also limits the cross section of users captured. Instead, we focused on shared access sites that serve a variety of users. We had two sites for data collection—an internet kiosk with 15 Windows workstations, and a vocational computer education center that trains users across several levels. Both of these sites were located in the northern edge of Bangalore in Karnataka, India.

Recent changes in Indian cyber security laws require internet cafe operators to record the identity of their users before every session. Users are required to provide a multitude of information, including their voter ID and telephone number before accessing the web. This law has also given rise to an ecosystem of kiosk management software with personalized information control. However, the internet cafe we worked with was still using a pen and paper form to log users (figure 1). As a result, we had to provide a separate mechanism for associating identity with browsing sessions.

We modified the Event Logger for Firefox [5] plugin to support a login mechanism that asked users for two pieces of credentials, their email and phone number, at startup. The logger collects various pieces of information about web usage, including requests, responses, caching and inactivity. Every event recorded was associated with the individual information that allowed us to establish identity across sessions. Further modifications were made to include information such as the time users spent on various domains and the site of data collection. This data was periodically uploaded to a monitoring and back up server. At the end of one month, there were about 8 million events recorded from around 470 users.

Figure 1: A user signs in to access the web

The second step in the data collection was anonymizing the dataset to remove all personally identifying information. We first group users based on matching credentials, and give each user a globally unique ID. Each session for a user is identified with another globally unique ID, and every event in the session is associated with it. In the current version of the data, we have also removed information from URL parameters that could potentially identify users. However, we plan to include that information in the final release of the data after hashing all parameters in the URL entry for requests and responses. The anonymized dataset will be available as hierarchical record of user-session events.

Understanding the personal nature of web access in developing countries is an obvious use for the dataset. However, we can imagine several other use cases that might be interesting for researchers. For example, it can be used to simulate a realistic service load when evaluating alternatives for web access improvement in developing countries. In addition, it can enable researchers to build and test fine grained access models that represent web usage in these environments. These models are useful in making resource allocation decisions.

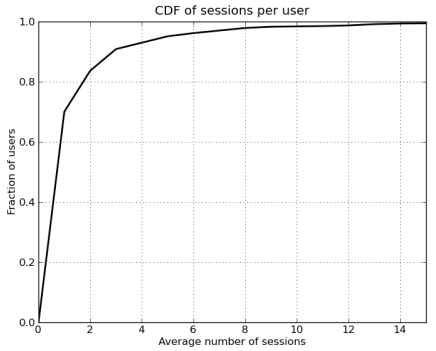
3. ANALYSIS

This section provides our analysis on the personalized dataset. We will start by characterizing the data itself, and then proceed to higher level analysis of user behavior. In the interest of space, we will largely focus on personal patterns in the dataset rather than aggregate analysis.

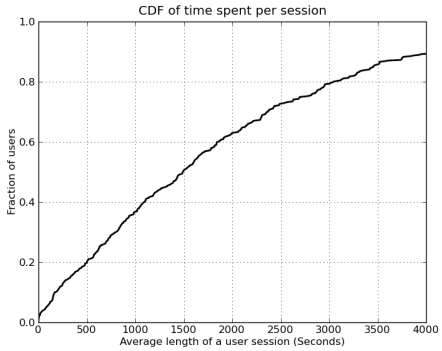
3.1 Data characterization

The usage data was collected for a continuous period of 4 weeks between January and February of 2011. It contains a total of 471 users, with 141 of those users having two or more sessions. Figure 2(a) shows the distribution of the number of sessions. For about 43 users, we have at least 4 separate sessions recorded.

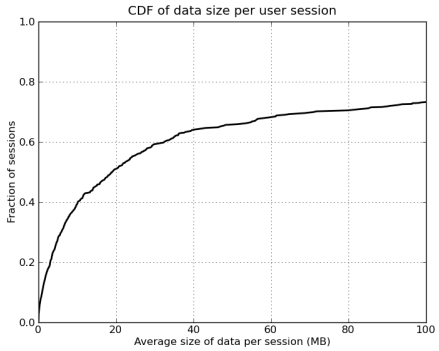
Each session is characterized by the amount of time a user spent on it. Figure 2(b) shows that a majority of the sessions were less than half an hour in length. However, its common to have sessions that are a few hours in length, with 15% of user sessions lasting for at least an hour.



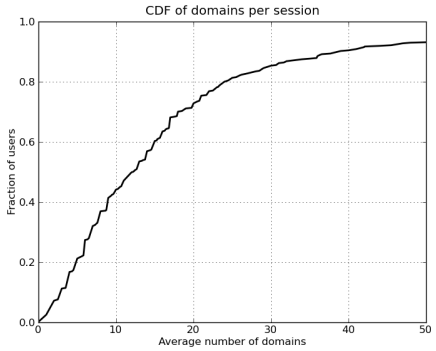
(a) Number of sessions per user



(b) Time spent per session



(c) Data transferred per session



(d) Unique domains requests per session

Figure 2: Characterization of data collected

Figures 2(c) and 2(d) give CDF representations for the amount of data transferred in a session and the number of unique domains that initiated the data transfer. More than 60% of user sessions had less than 15 unique domains responsible for sending data, and transferred less than 30 MBs of data. The size of data transferred varies greatly from one session to other, partly based on the length of the session.

3.2 Time spent on private content

We start out by considering the amount of time users spent on private content. This includes web destinations that require users to log in, such as email and social networking. We have a very conservative approach for identifying private web destinations. We only use well known email and social networking service providers in India, and filter our time records for visits to those domains. Realistically, anything that would require credentials to log in, ranging from financial services to job boards, is private by nature. As a result, our representation underrepresents time spent on private content. Nonetheless, we find that 40% of the users spent at least 60% of their time on private content, while over 30% spent at least 80% of their time on those services. The complementary CDF in figure 3 shows, what fraction users spent *at least* what fraction of their time on private content.

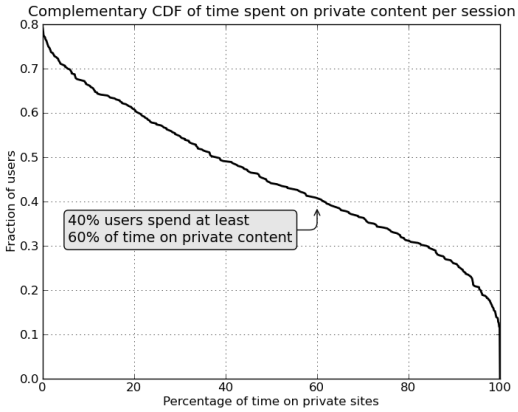


Figure 3: Time spent on private content

3.3 Personally interesting content

Private content, however, is only a small manifestation of the personal nature of web usage. The wide availability of diverse content on the web significantly fragments web usage patterns across users—several destinations are only personally interesting, with a small fraction of users requesting them. We consider this phenomena by analyzing how frequently users make requests to various domains. The complementary CDF in figure 4 plots percentage of users making a request against the fraction of requests. For example, the percentage of requests that were made by at least 10% of the users accounted for less than 2% of the total requests. This severe fragmentation has direct consequences on caching and

prefetching systems that only consider users in aggregate. As shown in the figure, this is even more visible when we weigh each request with the amount of data transfer it initiated.

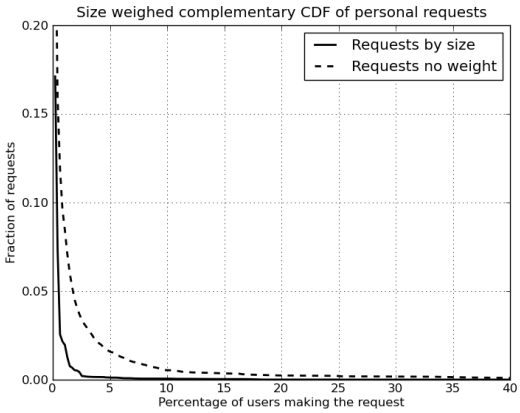


Figure 4: The size and popularity of requests

In figure 5, we look at a 3-way representation of popularity of domains against data size initiated by the domain, and the time users spent on it. Save for a few and well known outliers, we see that most data requests are made by a very small fraction of users. Destinations such as Google and Facebook are requested by a large percentage of people, but account for a small fraction of the total data size, while video content from YouTube accounts for 20% of the total Bytes but represents a small fraction of users.

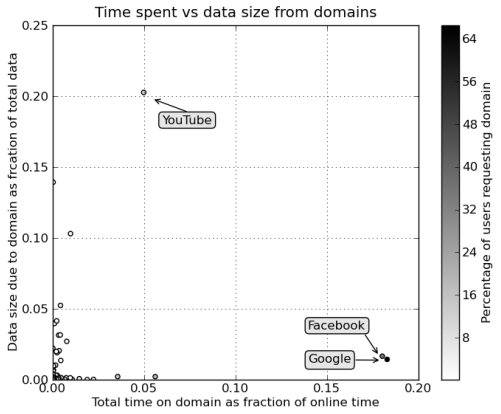


Figure 5: Request rate against data size and time spent

3.4 Understanding individuals

Since web access in developing countries is increasingly personal and fragmented, our next task was analyzing how similar users are to themselves across several sessions. In addition, we would like to compare our results from previous studies of user behavior in well connected environments. Our first analysis looks at pair wise Jaccard indices among domain requests per session. The Jaccard index for two sets gives the ratio of the number of items in the intersection of the sets to the number of items in their union. Therefore, for two disjoint sets, the Jaccard index will be 0, while for two identical sets, the Jaccard index is 1. Jaccard indices for user sessions are interesting because they measure how similar sessions are to each other.

We start out by filtering users that had at least 2 sessions in the dataset, which includes 141 individuals. For each users, we find the average of the pair wise Jaccard indexes among their sessions. The higher the Jaccard index, the more similar sessions are to each other. Figure 6 shows a complementary CDF of Jaccard indexes for two parameters—one weighed by the size of data domains generated, and the other by the amount of time users spent on each domain. These metrics roughly correlate with each other, and 40% of our users had a Jaccard index of at least 0.5. This is significant because it indicates the potential of a personalized web acceleration system to model and understand its users, improving its performance.

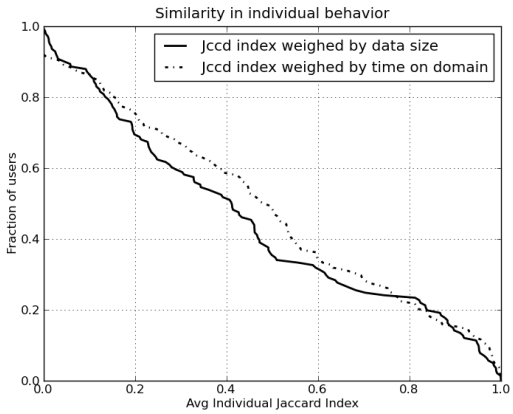


Figure 6: Jaccard indexes among user sessions

Another model that has been commonly applied in studying web access patterns is a Markov model. A first order Markov model assumes the probability of a user getting to a particular state depends only on the user's last state, and this probability is the same anytime this state is observed. In the case of web access patterns, a Markov model tries to predict the user's next request based on the last request. We have used a slight modification of a first order Markov model that continuously updates the probability distribution of states as more information is obtained about the user. Given a user session with an ordered list of requests, our model predicts the top 3 next requests. Afterwards, the model is provided the actual next request, this is used to slightly modify its probability distribution.

Our results in figure 7 show significant success for a Markov models to predict user behavior. On average, the model was able to predict the next request accurately 53% of the time, and the request was in the top 2 and top 3 of the predictions 68% and 75% of the time respectively. We find this to be better than previous studies of web access in developed regions [16], where prediction rates were generally under 40%. However, this is not surprising. Constraints in network access encourage people to mostly focus on important items, and spend their online time on these items. As this reduces

the branching factor of browsing, it helps the model predict requests better. Once again, this presents an opportunity for building personalized web acceleration systems that better understand user behavior in challenged network environments.

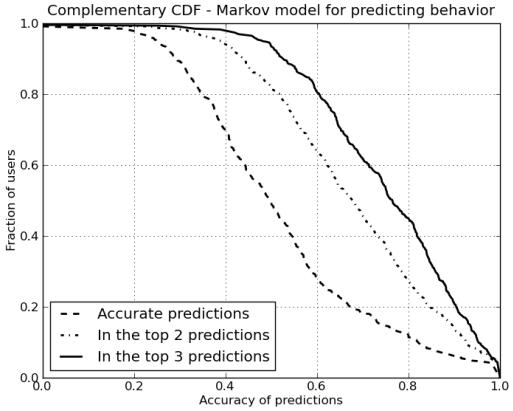


Figure 7: First order Markov model prediction accuracy

3.5 Associations and clustering

Finally, we ran clustering and association algorithms on the global data set to see if we can identify trends that can be exploited in web acceleration. We applied an unsupervised learning algorithm (leader clustering [10]) on a vector representation of the data, where each user had an entry for every domain that is accessed, weighed by the number of times it was accessed. Using an Euclidian metric to calculate distance, we attempt to discover clusters of sessions that show similar interests. However, unlike early studies in web access patterns [21], we were able to cluster only a small fraction of the total sessions, with a significant majority of clusters including only a single session.

We then ran the standard APRIORI algorithm [1] on the dataset to discover association rules between accesses. For example, if users who access site A usually access site B as well, a prefetcher might use this information for intelligently deciding what content to prefetch. While there are some clear associations with high confidence, these tend to be limited to very popular destinations (such as facebook.com ⇒ google.com), or private content(gmail.com ⇒ facebook.com). Our results from associations and clustering generally point to the globally divergent nature of access behavior. This is to be expected from the increasingly personal web.

4. DISCUSSION

With some of the lessons learned from analyzing a personalized web access dataset, we return to the design of web acceleration systems for developing regions. When usage is fragmented and personalized, traditional and aggregate mechanisms such as generic caching and prefetching will increasingly find it difficult to perform well. If the last few years are any example, this trend towards more individualized and diversified experience for web users is going to continue. As a result, we believe web acceleration mechanisms need also to be built with personalization as an important component.

While building personalized web acceleration mechanisms will involve various tradeoffs with storage, computation and privacy, the potential for improving the end user experience is high. This section provides just a few examples on how to take advantage of this trend to improve web experience in developing countries. We imagine system designers can incorporate personalization in web acceleration mechanisms at different levels and in various capacities.

4.1 Personalized prefetching

Predictive prefetching algorithms are an important component of web acceleration [5, 14]. Prefetching, however, has an important constraint in developing regions that the cost of mispredictions is quite high—limited and expensive bandwidth makes it difficult to justify meager gains from aggressive prefetching that consumes a lot of resources. Personalized web usage makes it especially hard to have good, general purpose prefetching algorithms that work well for all users.

Fortunately, however, at the individual level, web usage is not as diverse as it seems. As some of our analysis has shown, this is even more so in developing countries. This suggests an important consideration for web acceleration mechanisms should be incorporating user identity in determining what to prefetch. Such a mechanism can be implemented at a client or a proxy level depending on the private or shared nature of access. Identity does not have to be personally identifiable, and a personal prefetcher will have responsibility in providing some privacy guarantees to its users. Some existing systems provide a framework for personal prefetching [9, 20], and system designers can incorporate them in building web acceleration mechanisms.

4.2 Weighted personal caching

The success of caching relies on redundancy in data requests. If a user requests an object, and this object is requested again by the same or a different user, then the cache can serve it locally. However, for systems dealing with many users, redundancy in data requests can be quite low. As we saw in our analysis, most content people access tends to be either private or only personally interesting. Therefore, another approach towards better caching might be weighting the cache based on what the system knows about its users.

There could be several ways to implement a weighted, personal caching system, and the decision to do so is highly influenced by the fraction of users that repeatedly visit a shared access system. One option is reserving a portion of the caching space for users that are known to frequently visit the system. This will reduce the general caching space available for new and one time visitors, but it could be a worthwhile tradeoff given the average overall improvement. Another alternative is to incorporate personal usage patterns in the cache eviction policy, allowing the cache to adapt to individual patterns.

4.3 Personalization as a service (PaaS)

Another dimension to incorporate personalization is to build cloud-based tools that guide local clients through web acceleration. This might turn out to be more convenient as it does not tie users to particular access sites. This also gives designers more flexibility to incorporate large scale data in improving individual performance. For example, as a user

starts a browsing session, she might be offered to sign in to her PaaS account. Without a local administrator having to keep track of individual behavior, the PaaS can provide the required information to the browser that would help it improve web performance to the current user. At the same time, this allows the user to get improved performance across various service providers and platforms.

5. RELATED WORK

There have been several studies that consider web usage in developing world scenarios [4, 5, 6, 11, 13, 15]. The datasets for these studies are generally collected from client or proxy level loggers in various contexts and can be characterized as aggregate views into web access in developing countries. Du et all looked at HTTP traffic captured from internet kiosks in two developing countries [6]. Their results point out various features of web usage in developing countries. Work by Ihm et al. [11] analyzes a large amount of data collected at a global scale and tries to observe differences in developed and developing world traffic. Their dataset comes from a worldwide proxy server that has access to the content of requests in addition to access logs. More specific web access analysis in schools and universities [4, 5] has considered traces obtained from educational environments in developing countries. Our dataset differs from these projects because it captures personalized web usage of individuals in a developing country context, allowing us to perform finer grained analysis of web access behavior. As web experience gets increasingly personalized, a deeper understanding of user behavior is essential in designing appropriate systems for challenged network environments. Our dataset is available for researchers to further investigate individual behavior in web access for developing country contexts.

Web usage acceleration is one important class of services for challenged networks. There has been a lot of interest in this area of research [3, 5, 12, 13, 14, 18, 19, 20]. In environments where network connectivity is limited and expensive, web acceleration techniques play an important role in improving end user experience. Caching and prefetching are some of the widely used techniques for accelerating web access, with several flavors of implementations and various levels of success. We believe our work provides tangible evidence for considering personalization as an important design factor for such systems. As web usage gets more and more personal, acceleration techniques built atop an aggregate analysis of their user base will be hard pressed to cope with diverse behavior and content.

6. CONCLUSION

This paper considered the issue of personalization in web access behavior in developing countries, and its implications for web acceleration mechanisms. Web acceleration mechanisms are especially important in challenged network environments where connectivity is limited or expensive. However, traditional solutions that rely on the redundancy of user requests for improving performance find it increasingly difficult to be effective as web usage gets personalized and fragmented. We have collected a large scale personalized web usage data in a developing country context, and will make it available for researchers. Our analysis on this dataset provides some tangible evidence for describing the personal nature of web access in developing countries. Finally, using lessons learned from this analysis, we provide some initial recommendations for incorporating personal design into web acceleration mechanisms.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[2] N. Azzouna and F. Guillemin. Analysis of ADSL traffic on an IP backbone link. In *Global Telecommunications Conference,IEEE*, pages 3742–3746 vol.7, Dec. 2003.

[3] A. Badam, K. Park, V. S. Pai, and L. L. Peterson. HashCache: cache storage for the next billion. In *NSDI'09: Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, pages 123–136, Berkeley, CA, USA, 2009. USENIX Association.

[4] J. Chen, S. Amershi, A. Dhananjay, and L. Subramanian. Comparing web interaction models in developing regions. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 6:1–6:9, New York, NY, USA, 2010. ACM.

[5] J. Chen, D. Hutchful, W. Thies, and L. Subramanian. Analyzing and Accelerating Web Access in a Shool in Peri-Urban India. In *WWW 2011*. ACM, 2011.

[6] B. Du, M. Demmer, and E. Brewer. Analysis of www traffic in cambodia and ghana. In *WWW 2006*, pages 771–780.

[7] Ethiopia Tel. Co. Services, 2009. `www.telecom.net.et/services`.

[8] J. Gantz. The diverse and exploding digital universe. Technical Report White paper, IDC, 2008.

[9] Google Gears, 2011. `www.gears.google.com/`.

[10] J. Hartigan. *Clustering Algorithms*. John Wiley and Sons, New York, 1975.

[11] S. Ihm, K. Park, and V. S. Pai. Towards understanding developing world traffic. In *NSDR 2010*, pages 8:1–8:6.

[12] S. Ihm, K. Park, and V. S. Pai. Wide-area network acceleration for the developing world. In *Proceedings of the 2010 USENIX conference on USENIX annual technical conference*, pages 18–18.

[13] S. Isaacman and M. Martonosi. Potential for collaborative caching and prefetching in largely-disconnected villages. In *WiNS-DR 2008*, pages 23–30.

[14] S. Isaacman and M. Martonosi. The C-LINK System for Collaborative Web Usage: A Real-World Deployment in Rural Nicaragua. In *NSDR 2009*. ACM, 2009.

[15] D. L. Johnson, E. M. Belding, K. Almeroth, and G. van Stam. Internet usage and performance analysis of a rural wireless network in macha, zambia. In *NSDR 2010*, pages 7:1–7:6, 2010.

[16] B. Lambert and O. Fatemieh. Generating intelligent links to web pages by mining access patterns of individuals and the community, 2005.

[17] K. W. Matthee, G. Mweemba, A. V. Pais, G. van Stam, and M. Rijken. Bringing Internet connectivity to rural Zambia using a collaborative approach. In *ICTD '07*, pages 47–58, 2007.

[18] S. Michel, K. Nguyen, A. Rosenstein, L. Zhang, S. Floyd, and V. Jacobson. Adaptive web caching: towards a new global caching architecture. *Comput. Netw. ISDN Syst.*, 30:2169–2177, November 1998.

[19] M. Rabinovich and O. Spatscheck. *Web Caching and Replication*. Addison-Wesley, 2001.

[20] A. Reda, B. Noble, and Y. Haile. Distributing private data in challenged network environments. In *WWW 2010*. ACM, 2010.

[21] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, pages 1007–1014, 1996.