Traffic Characterization and Internet Usage in Rural Africa

David L. Johnson, Veljko Pejovic, Elizabeth M. Belding
University of California, Santa Barbara
davidj,veljko,ebelding@cs.ucsb.edu

Gertjan van Stam
Linknet, Macha, Zambia
gertjan.vanstam@machaworks.org

ABSTRACT

While Internet connectivity has reached a significant part of the world's population, those living in rural areas of the developing world are still largely disconnected. Recent efforts have provided Internet connectivity to a growing number of remote locations, yet Internet traffic demands cause many of these networks to fail to deliver basic quality of service needed for simple applications. For an in-depth investigation of the problem, we gather and analyze network traces from a rural wireless network in Macha, Zambia. We supplement our analysis with on-site interviews from Macha, Zambia and Dwesa, South Africa, another rural community that hosts a local wireless network. The results reveal that Internet traffic in rural Africa differs significantly from the developed world. We observe dominance of web-based traffic, as opposed to peer-to-peer traffic common in urban areas. Application-wise, online social networks are the most popular, while the majority of bandwidth is consumed by large operating system updates. Our analysis also uncovers numerous network anomalies, such as significant malware traffic. Finally, we find a strong feedback loop between network performance and user behavior. Based on our findings, we conclude with a discussion of new directions in network design that take into account both technical and social factors.

Categories and Subject Descriptors

C.2.2 [**Computer-Communications Networks**]: Network Protocols-*Applications*; C.4 [**Performance of Systems**]: Measurement Techniques

General Terms

Measurement, Performance, Human Factors

1. INTRODUCTION

Internet connectivity represents a key factor in the overall development of any nation. It enables competitive participation in the global economy, represents a valuable source of education and facilitates democratization of a society [23, 17]. However, connectivity opportunities are largely unequal among different regions. Poor infrastructure, lack of economic interest from telecommunication companies and insufficient governmental support result in a complete connectivity blackout in a large part of the developing world.

Fortunately, through efforts of university research groups and non-government organizations (NGOs), a number of isolated islands of connectivity have appeared in developing regions in the second half of the last decade. The usual connectivity deployment pattern involves satellite Internet gateways and long distance WiFi- or WiMAX-based networks that connect a gateway with key institutions in the target regions, such as schools, hospitals and community centers. In this model, called the "kiosk model", users access terminals in Internet cafés and other public locations [8, 18]. In some scenarios, Internet connectivity is also available in homes [11], though this is far less common.

Bringing reliable, usable Internet connectivity to remote regions is typically plagued with problems. Satellite connections are slow, often with bandwidth of only a few hundreds of kbps or 1 Mbps [16]; power sources are unreliable and devices are frequently unavailable [22]; networks are managed remotely or by poorly trained local staff [1]; and public Internet cafés have limited availability and high per minute usage costs [26]. However, despite these problems, Internet access has already revolutionized the lives of rural residents. Access to food production and health care information, distance learning programs, and global and local business opportunities has led to vast improvements in health care, quality of life, and economic earning potential. The better the quality of Internet access, in terms of availability, reliability and performance, the more residents stand to gain from their online activity.

Understanding the impact of technology in a developing rural region is a complex undertaking; it requires a multifaceted approach that consists of both detailed traffic analysis and social engagement. Local social customs strongly influence Internet viewpoints and cannot be ignored when trying to interpret observed usage. Unfortunately, there are multiple challenges in technical and social analysis. Remote management and lack of local skilled technical staff render these networks extremely hard to monitor. Partnership with local technical organizations that help deploy and maintain infrastructure is essential. From the social science point of view, on-site data gathering is a very tedious process in remote areas. Villages are often dispersed over a large area with poor roads and lack of communication infrastructure. Language barriers pose additional challenges; often translators are required. Again, partnership with local organizations is critical for obtaining meaningful data.

In this work, we seek to understand the availability and reliability of Internet access, as well as Internet usage and performance, in rural Africa. To do so, we conduct a holistic study that encompasses two remote villages and both technical and social investigation of network usage. Through partnership with Macha Works, we deployed a lightweight

traffic monitoring system in Macha, a village in rural Zambia, which hosts an extensive, long-running wireless network. We monitored the network for two weeks in February 2010, resulting in approximately 50 GB of collected data. As described in section 2, our monitoring system collected data at two measurement points: traffic was collected at the Internet gateway, and squid proxy access logs were archived. We complement our data collection with a series of interviews in Macha, Zambia in order to gain in-depth understanding of the usage patterns of local constituents. We performed the same survey in Dwesa, a remote village in South Africa, that recently obtained Internet access via a satellite link and a local WiMAX network. In total, we conducted 37 interviews.

The analysis of our two-week trace data reveals mostly web based traffic (highest percentage of visits to Facebook), a small fraction of long-lived flows consuming the majority of the bandwidth, a constant stream of malware traffic and poor network performance with large round trip times. We observe a clear influence of the network performance on the user behaviour. In addition, Internet usage habits change once access is available at home, as opposed to being limited to public terminals with restricted usage hours, such as schools and Internet cafés. These findings suggest that rural area Internet penetration should not be evaluated through simple binary "have" and "have nots". We show that only through a comprehensive socio-technical study of a network can we obtain a full picture of the Internet usage in rural areas of the developing world. Such an outcome is crucial for a finer synergy between those who design computer networks and those who use them.

This paper is structured as follows. Section 2 provides necessary background information, including an overview of the population distribution and network structure of Macha and Dwesa. We also describe our approach to network monitoring and social surveying. In section 3 we analyze the traffic characteristics in the monitored network in Macha, while in section 4 we discuss Internet usage trends with an emphasis on time variation. Section 5 examines the presence of virus and other malware traffic in the network. Deeper socio-economic analysis of the interview results is presented in section 6. Based on our analysis, in section 7 we recommend a set of improvements for the Macha network in particular, and possibly rural area wireless networks in general. We complete the paper with the overview of related work in section 8 and our conclusions in section 9.

2. METHODOLOGY

2.1 Macha and Dwesa

Macha, Zambia, highlighted in figure 1, is a typical poor rural area in Africa with scattered homesteads, very little infrastructure, and people living a subsistence lifestyle; the primary livelihood is maize farming. Like many sub-Saharan rural communities, Macha has a concentrated central area, and a large, geographically dispersed rural community with a sparse population. Clusters of homes house members of a single family, and are likely separated from another cluster of homes by one or more kilometers. Macha contains approximately 135,000 residents, spread out over a 35 km radius around the village center [24]. The overall population density is 25 per km^2. In the middle of the community center

Figure 1: Map of Southern Africa with locations of Macha and Dwesa highlighted.

is a mission hospital, a medical research institute, schools, and the Macha Works organization.

Macha Works, through the LinkNet project, has deployed a network of long distance 802.11 wireless links and mesh networks that provides connectivity to approximately 300 community workers and visitors using satellite-based Internet. The majority of users access the Internet through community terminals, though a few people do have 802.11 connectivity in their homes. The community is connected to the Internet through a VSAT connection.

The Dwesa region, also highlighted in figure 1, is located in Eastern Cape province, one of the poorest regions of South Africa. Similar to Macha, Dwesa is characterized by outdated infrastructure, a weak subsistence economy and a sparse population (pop. 15,000, area 150km^2). The Siyakhula project, led by the University of Fort Hare and Rhodes University in South Africa, has established Internet connectivity among local schools via WiMAX links that are several kilometers in range [15]; this is one of the first WiMAX installations in rural South Africa. One of the schools connects to a VSAT satellite, thus serving as the Internet gateway.

Both Macha and Dwesa can be termed "real Africa" and as such present a relevant ground for general conclusions about Internet usage in rural areas of the developing world. At the same time, they provide a great opportunity for a comparative analysis. Internet connectivity in Macha has penetrated much further than just local schools, which is the case in Dwesa, and the network has been operational for longer. In addition, Macha is located in one of the world's poorest countries, while Dwesa, although itself very impoverished, is a part of the richest country in Africa. The social environments of the two areas are therefore very different as are migration patterns, crime rates and other factors.

2.2 Network Architecture

We briefly summarize the LinkNet network in Macha, Zambia and refer the reader to [16] for more detail. Internet connectivity is provided to the community through a VSAT connection that has a committed download speed of 128 kbps bursting to 1 Mbps and a committed upload speed of 64 kbps bursting to 256 kbps with no monthly maximum.

The total monthly cost of the C-band VSAT connection is $1200 (US dollars). Figure 2 shows a scaled view of the core of the Linknet network with the position of each wireless router represented by a symbol. Two radio masts, one shown in figure 2 at the IT tower, spread the connectivity over a 6 km^2 area using a combination of 802.11 point-to-point links, hotspots and mesh networks. Over the past few years, Internet connectivity has spread to a large portion of the medical research institute and the community center, which provides public access through an Internet café. The gateway server makes use of a squid proxy server, which is configured with standard settings and a cache size of 1 gigabyte. The local wireless network can sustain connectivity speeds that are almost always faster than that of the satellite connection. Only a small percentage of local traffic was observed during our analysis, which consisted of primarily DHCP and ARP traffic. As a consequence, monitoring the network at the satellite gateway captures the usage and quality of service supported in the Macha network.

Our data analysis goals are three-fold: The first goal is to understand usage patterns and gain insight into the needs of users in a rural setting. We use interview data to complement our traffic analysis by gaining deeper insight into user's technical literacy and technological attitudes. The next goal is to understand the performance of the network to pinpoint specific inefficiencies and areas for improvement. Our final goal is to make use of the learning from the first two goals and suggest ways in which the performance can be improved.

To meet these goals, two measurement points were required. The first measurement point was located at the gateway and captured all Internet traffic on the interface to the satellite and to the wireless network. The packets were captured in pcap format and a capture length (snaplen) of 96 bytes was used to minimize the size of the log file. This snaplen size was chosen to capture enough information from all the headers in the network packet. In order to analyse the HTTP traffic, the squid proxy access logs were also archived for analysis. All IP addresses were anonymized in order to protect the privacy of the users.

14 days of traffic were captured from midnight, Sunday 31 January to midnight Sunday 14 February in 2010. Approximately 50 GB of packets were captured, consisting of about 6 million packet flows. Captured traffic was compressed and sent to our server in Santa Barbara, CA during off-peak hours for offline analysis. In our analysis, we present traffic from a representative 10 day consecutive time block (Wednesday February 3 - Friday February 12) for visual clarity of the graphs. User-management software installed in April 2010 established that 10% of the traffic was from international visitors. A similar traffic distribution between the local population and visitors was likely for February.

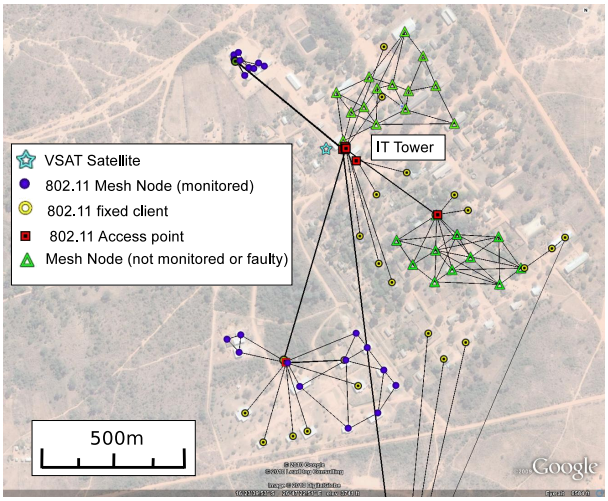

Figure 2: Scaled view of the Macha network.

2.3 Interviews

The interviews were conducted in July/August 2010, onsite in Macha and Dwesa, privately between one interviewee and one interviewer. Interview participation was on a voluntary basis and there were no material awards associated with the participation. The goals of the interview and possible influence on future quality of service within the network were explained to each person in order for them to understand the benefit of participation. In a close-knit African community, residents can be reluctant to openly talk about their habits with a complete stranger. To facilitate the interview process we used our existing ties with local persons of authority for introduction to potential interviewees. Naturally, this method does not result in a completely random sample; however, every effort was taken to ensure that different age, gender and income groups were represented. A total of 37 interviews were conducted: 23 in Macha and 14 in Dwesa. The participants' age ranges from 18 to 57; 15 of them are female, 22 are male, all are literate and have at least some high school education, with income ranging from zero to above US$300 per month.

We opted for a directive, structured questionnaire in the first phase of the conversation, as we wanted to obtain highly quantifiable data that could be correlated with the results of our network trace analysis from Macha. In the second part of the interview the subjects were asked less formal questions and were able to engage in a discussion with the interviewer. In African culture, narrative communication is common, thus these open questions revealed a number of unforeseen issues.

We extracted two types of statistics from the data: descriptive and mean/category-comparing. With the former we try to provide a clear picture of the Internet usage in Macha and Dwesa, while with the latter we provide possible explanations for certain observations. Due to the small number of samples we did not perform regression analysis but concentrated on the independent samples t-test for comparing means and χ^2 test for comparing categories. We report test results for which the two-tailed significance was lower than .1. We feel that this slightly looser requirement can be justified for the sample size and the domain in which we are working.[1]

3. TRAFFIC CHARACTERIZATION

In this section we seek to understand the high level usage characteristics of the network, and the network's ability to

[1]We report statistics according to the American Psychological Association standards: χ^2 statistics are reported with degrees of freedom and sample size in parentheses, the Pearson chi-square value, and the significance level; T-tests are reported like χ^2, but only the degrees of freedom are in parentheses; mean values are labeled with M_i.

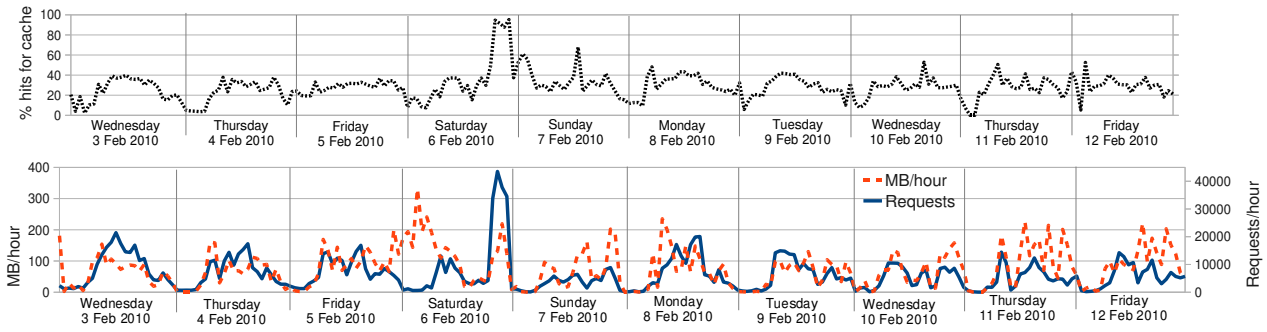

Figure 3: Usage analysis over 10 days.

support the offered load. We are particularly interested in how the bandwidth constrained satellite connection affects network performance. We analyze throughput over time to understand the day/night cycle of usage and then proceed to study the distribution of flow sizes and lifetimes to determine the impact on the user experience. TCP round trip times (RTTs) are used to understand the delays users experience during interactive browsing and real-time activities such as instant messaging and VoIP. The performance of the cache also gives clues into network behaviour.

The proxy had a cache hit rate of 43% with an actual bandwidth saving of 19.59%. This low fraction of bandwidth saved is fairly common in a standard unmodified squid proxy server due to the dynamic nature of the Internet today [5]. The cache size was set to 1 gigabyte; studies have shown there is very little gain from cache sizes beyond 1 GB [5].

Figure 3 shows the traffic load, number of web requests and cache hit rate over the 10 day measurement period. What emerges is a clear, typical diurnal usage pattern, with the exception of Friday night and Saturday evening. The lack of usage during off-peak hours is due to the inaccessibility of public Internet terminals during this time. Not surprisingly, our interview data shows that those who have access at home are more likely to use it after-hours ($\chi^2(1, N = 28) = 5.2, p = .041$).

Friday night and early Saturday morning traffic displayed a sudden increase in aggregate download rate caused by a small number of requests per hour (approximately 700 or 10% of the average request rate). Inspection of the trace files reveals that these large downloads were requested from two single machines. Examination of the proxy logs, during this same period, indicates this was mainly due software updates and some Facebook and sporting web site accesses. A more detailed discussion of web usage behaviour over time is given in section 4. On Saturday evening a satellite failure resulted in anomalous proxy behavior; we discuss this event in section 3.1 when we analyse HTTP response codes.

Three power failures occurred during our monitoring interval: two on Wednesday 10 February and one on Thursday 11 February. These power failures caused corresponding dips in network usage. Power failures in Macha generally last anywhere from an hour to a few days. Our interviews revealed that on the average six such failures happen in a month.

Given the cyclical usage pattern and the bandwidth constraints of the satellite link, it is clear that the available bandwidth could be better utilized. In particular, users and/or administrators should be trained to set up systems that time-shift large downloads to periods when the network is quiet, using cron jobs, for example. This would offload some of the peak-hour traffic, providing more capacity for real-time and interactive traffic. We discuss this further in section 4.

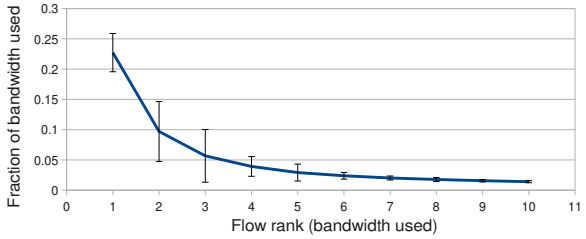

Figure 4: Flow distribution, organized from largest to smallest flow by bytes in one hour bins, over the 10 day measurement window.

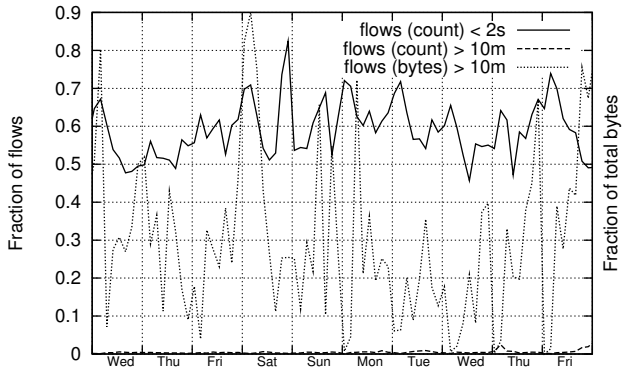

Figure 5: Fraction of short and long-running flows and fraction of bytes used by long-running flows over 10 day measurement window.

3.1 Web object and flow analysis

Studies have shown that traffic in the Internet is typically characterized by many small short flows, such as web requests, and a few large flows, such as file downloads/sharing [2]. Understanding the distribution of flows becomes critical in a network connected over a slow satellite link, as the presence of large flows will likely make interactive activities, such as instant messaging and web browsing, very slow, and possibly even unusable. To determine the traffic composition in the Macha network, we look at flow distribution.

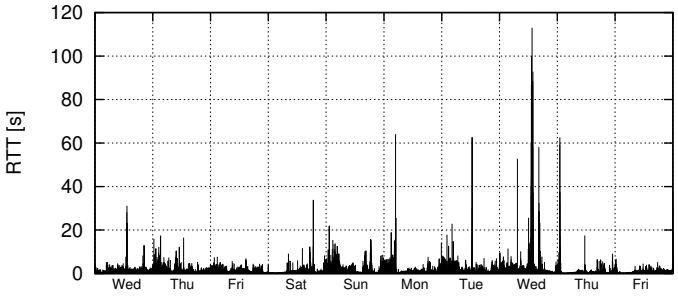

Figure 6: Average RTT measured in one minute bins.

To understand the general behaviour of all the traffic in the network we plot the flow distribution of the top 10 flows, shown in figure 4, ranked by size of flow. The figure shows the average size of the flow in one hour bins together with standard deviation over the 10 day measurement window. The largest flow, in terms of size, consumed about 22% of the bandwidth while the largest ten flows account for approximately 75% of the bandwidth demand. Thousands of smaller flows form the rest of the long tail distribution not shown in this figure. A flow is always established between two single hosts and thus a small subset of users are consuming the majority of the traffic in this network.

These smaller flows make up many of the interactive applications on the Internet such as web browsing and instant messaging. To understand the balance of small and large flows both in time and in size, the flows were categorized into short flows lasting less than 2 seconds and long flows lasting more than 10 minutes. This is plotted in figure 5.

Typically, about 60% of the flows are short-lived (less than 2 seconds in duration). However, there is also a consistent fraction of long-lived flows (over 10 minutes), about 0.47% of the total. The fraction of long-lived flows increases during the weekend when more large files are downloaded. Though a small fraction of the total number of flows, the long-lived flows often consume a large fraction of the bandwidth, as much as 90% of the traffic on one occasion. A system designed for rural areas needs to cope with this high volume of short flows. Policies could be developed that place a higher priority on short flows to ensure that interactive activities such as instant messaging do not suffer from long delays.

The impact of large flows, as well as latency from the satellite link, is reflected in the RTT of the TCP traffic. Figure 6 shows the RTT of TCP traffic measured in one minute intervals, averaged over all flows within each one minute window measured at the outgoing interface of the gateway. A typical GEO satellite will incur a round trip delay of approximately 560 milliseconds in an uncongested network. Figure 6 indicates that congestion causes delays far beyond those due to the GEO satellite. The 100 second RTT on Wednesday 02/10 was due to a satellite fault; however other large delays of up to 60 seconds represent real delays experienced by users when the network was congested. Many daytime RTT averages are as high as 3 to 10 seconds, making web browsing and real-time applications difficult, if not impossible.

Another effect of large flows and congestion is session time-outs when the proxy is trying to retrieve web content. We analyze HTTP proxy responses to better understand web performance and plot the results in figure 7. We find that during the 10 day period, approximately 86% of the HTTP requests were serviced successfully (found by summing the "OK" and "Not modified" responses). Approximately 4% of the requests resulted in a "Gateway timeout" or "Service unavailable" response, while another 2% received only partial content. To better understand web performance, we analyzed the HTTP responses as a time series (not shown due to space limitations). We find that the "service unavailable" errors occur only when the network is under heavy load, whereas the "gateway timeout" response consistently occurs at any time of the day or night. For example, 90% of HTTP requests received "service unavailable" errors on Saturday night; at the same time the overall number of requests peaked (figure 3). We suspect that it is an automated application, rather than user behaviour that resulted in this phenomenon.

We discuss some ideas from existing and future work to mitigate observed problems in the next steps and related work sections.

4. INTERNET USAGE CHARACTERIZATION

In this section we more closely analyze the offered load of the Macha network to better understand network usage in terms of application breakdown and web traffic classification.

Our previous analysis of our Macha dataset found that Web traffic accounts for 68.45% of total Internet traffic [11]. Web traffic is far more pronounced than in the developed world, where recent studies show only 16% to 34% Web traffic [19]. In our previous analysis, we did not specifically identify any peer-to-peer (P2P) traffic. Since P2P rarely uses known ports we left the possibility that it could be hidden in the 26.47% of traffic that we could not classify. However, our interviews reveal that P2P is indeed not popular in Macha. None of the interview participants reported using this application. Instead, 78% of interviewees transfer large files, such as movies and music downloads, from hand to hand via USB keys. This way of sharing alleviates the problem of the gateway bottleneck in a P2P system. Unfortunately, it has detrimental consequences on network security as it facilitates virus spreading.

The most popular online applications are web browsing and email. Our interviews reveal that 100% of those who use the Internet in Macha use both applications. The next most popular applications are VoIP and instant messaging via Skype and gTalk - 73% of the interviewees use these applications. This suggests that a large part of the unclassified traffic most probably belongs to these applications. A comparison of PC users in Macha and Dwesa yields interesting results. Macha users are more experienced and are more

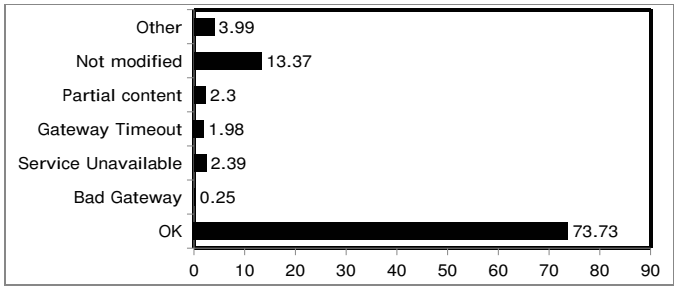

Figure 7: HTTP response codes for proxy server.

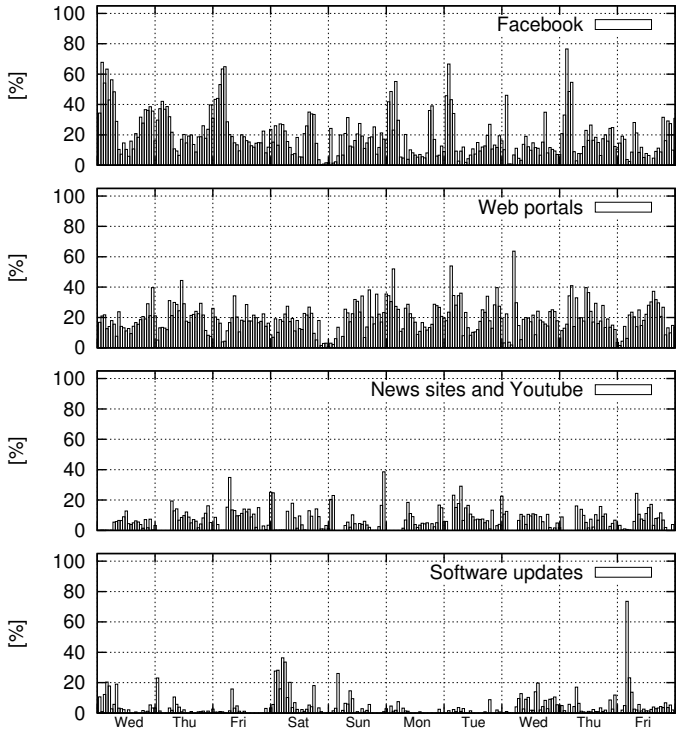

Figure 8: HTTP category requests over time.

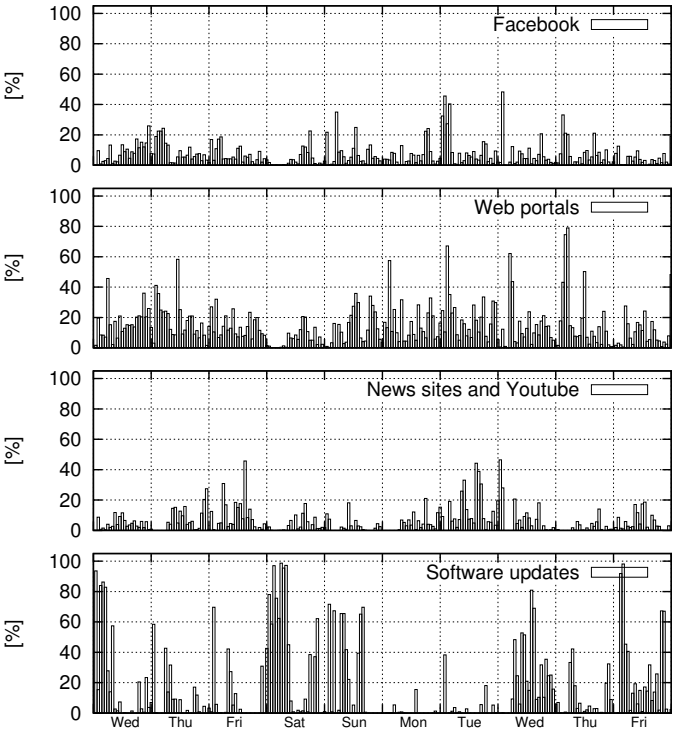

Figure 9: HTTP category traffic load over time.

likely to use web browsing ($\chi^2(1, N = 37) = 8.07, p = .014$) and email ($\chi^2(1, N = 37) = 5.99, p = .035$).

Examination of web proxy logs reveals that web site accesses can be roughly grouped in the following categories: "Facebook"; "Web portals" - including Yahoo, Google and Live.com; "News sites and Youtube" - including Post Zambia and Lusaka Times[2] as well as youtube.com; and "Software updates" - including OS and anti-virus updates. Note that the proxy log entries obscure the difference between the user intended behavior (e.g. accessing a news site), and the automatic hits (e.g. scheduled software updates). We extend our analysis by considering Web access patterns over time. In figures 8 and 9, we break down the total web requests and total web traffic load (in bytes) of each domain type over time, respectively. The values are shown averaged over hour intervals.

From figure 8 we see that, at almost any time, "Facebook" and "Web portals" account for the largest number of HTTP requests. This is not surprising as 77% of interviewees in Macha reported using Facebook and 96% report "Googles" for work or school related material and practical problem solutions. However, the pattern by which the web accesses occur is quite different for the two categories. Web portal access frequency does not change much based on the time of day - the percentage of traffic to portals is roughly constant. Facebook, on the other hand, shows distinctive peaks late at the night. Since our interview data reveals that the late night users are most likely accessing the Internet from their homes, we deduct that restricting connectivity to public locations within a rural area, such as Internet cafés and schools, severely limits the usage of OSNs and leisure applications.

[2]www.postzambia.com, www.lusakatimes.com

The traffic load distribution in figure 9 shows that the majority of bytes transferred often belongs to a small number of flows, which is a similar finding to prior studies [2]. These are typically due to software updates, which can frequently consume the majority of the available bandwidth. Burstiness of these heavy-weight requests leads us to believe that they represent automatic, scheduled OS and anti-virus software updates.

4.1 Media access

Access to news sources is often limited in rural areas. There are only three TV stations in Dwesa and one in Macha. Newspapers from larger cities arrive highly irregularly. Interviewees praised the Internet for its provision of multiple news sources. All of the interviewees in Macha and 79% in Dwesa consider the Internet to be the preferred source of information over television and newspapers. Prior work has shown that Internet usage impacts legacy communication sources in the developed world [12]. We see the same trend in the two African villages. People who extensively use the Internet (five or more hours per day) devote drastically less time to television (M_1 =11 hours, 45 minutes per week; M_2 =3 hours 15 minutes per week; $t(28) = -3.08, p = .005$) and printed media (M_1 =2 hours 30 minutes per week; M_2 =0 hours 0 minutes per week; $t(27) = -5.52, p = .000$).

That the Internet is heavily used as a news source can be observed in figure 8. We see that news site access quickly peaks in the mornings when people arrive to work. The usage then diminishes until late in the evening, when almost no accesses are observed. We extend our analysis over a three week proxy log that overlaps with our ten day measurement period and plot the news site hit distribution over that period in figure 10. We see that the news sites are up to

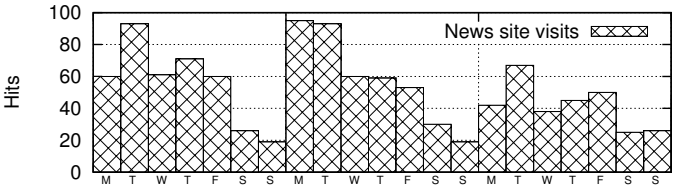

Figure 10: News site visits per day.

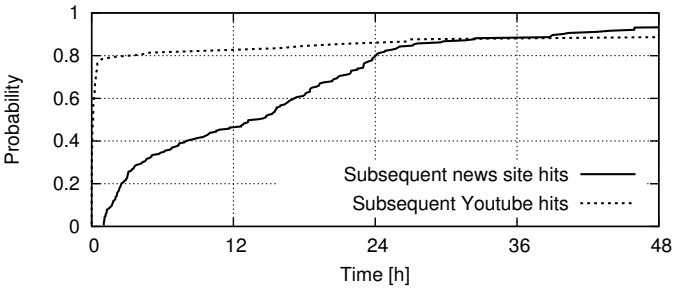

Figure 11: CDF of media site visit interarrival times.

four times more popular during week days than weekends. Interestingly, we observe that the number of news site visits decreases as the week progresses, a phenomenon that is not uncommon for the developed world either[3].

To isolate individual user behavior we calculate inter-visit times for individual IPs and plot the CDF of news websites and Youtube visits for all the unique IPs in figure 11. We observe that most visits to the news sites are made with an interarrival time of less than a day and that the distribution is rather uniform, indicating periodic checking for new information every few hours or each working day. Youtube visits, on the other hand are bursty - almost 79% of the visits by the same user happen within the same hour. We intuit that this can be explained by the Youtube website organization where a visitor is often presented with multiple related videos. In addition, a number of interview participants complained about the unpredictability of Youtube performance. We suspect that in this case the network limitations dictate the user behavior: once the Internet connectivity provides a usable Youtube quality of service, users seize the opportunity to watch multiple videos they were unable to download earlier.

5. MALWARE TRAFFIC

Keeping a computer free from malware in a developing country is a challenge due to the need for up-to-date virus signatures and the increasing rate of new forms of attacks. These signatures can be made available as often as twice per day due to the speed at which new attacks appear. The challenge in a rural area is particularly great because virus signature downloads consume a large amount of the scarce satellite bandwidth, particularly when each computer must individually download its own patches. There is also a lack of specialist knowledge to analyse intrusion detection system logs, enforce timely software patching, and assist with removing malware when machines become infected. In addition, users have a perception that network administrators are solely responsible for protection of the network.

[3]http://www.alexa.com/siteinfo/nytimes.com

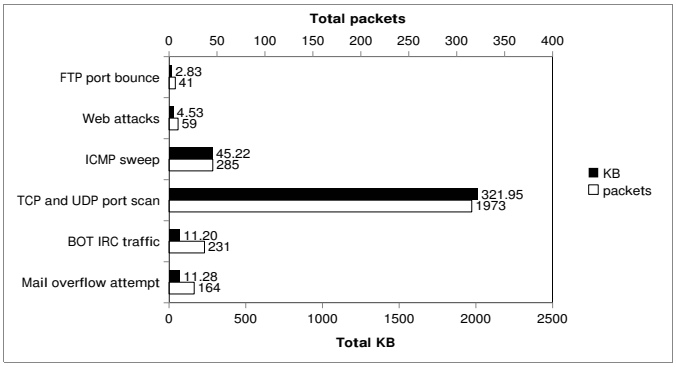

Figure 12: Types of attacks on the Macha network.

Approximately 60% of machines in the Macha network run Microsoft Windows; the remainder of the machines run Linux or MacOS. Windows machines are particularly vulnerable to Malware attacks, especially when the latest security patches have not been installed.

5.1 Detection method

We utilized the *snort* intrusion detection tool for detecting malware. Snort makes use of a set of rules to detect the presence of malware on the network. It is capable of detecting a virus payload on a network as well as the presence of suspicious traffic once the payload has infected a machine and malware traffic is being generated. Our analysis was carried out in non-real-time on truncated packets; it was therefore not possible to detect virus payloads, which require deep packet inspection. We looked for the presence of all known malware type attacks. These are:

- ICMP scans: malware searches for a host to attack
- UDP and TCP port scans: malware has found a host and is looking for open ports on this host to attack
- Bot traffic to IRC server: malware contacts a public IRC server to send messages and receive commands
- Web attacks: malware tries known web site attacks such as buffer overflows

5.2 Analysis

Figure 12 shows the type of attacks present in the network, broken down by total bytes consumed per attack. This is typical behaviour of a bot, in which an IRC server is contacted as a command and control center and the bot scans for potential machines to attack. Out of the 201 unique IP addresses seen over the 10 day measurement window, 9 machines were infected by bots trying to contact an IRC server. 118 machines were carrying out some form of port scanning, however this may have some false positives due to other software that carries out port scanning such as *nmap*. Figure 13 shows the regularity of bot and port scan attack traffic over time. There is a clear pattern of the virus contacting the bot IRC command and control center at certain times of the day. The port scanning traffic is fairly consistent over the the full measurement window except for some brief periods late at night and early in the morning where activity decreases due to some machines being powered down.

The total impact on the satellite network cannot be accurately quantified as many of these malware bots transmit

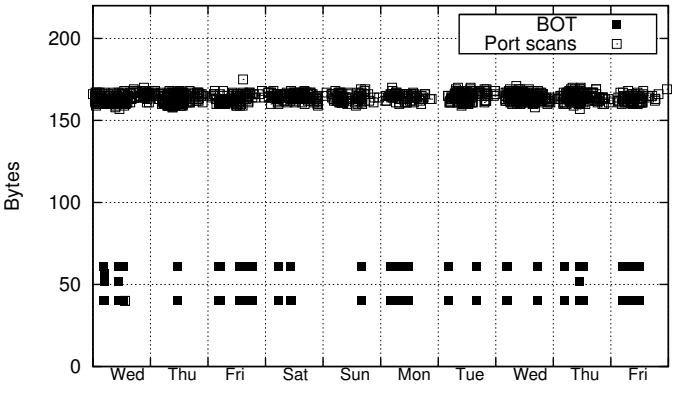

Figure 13: Time series of virus traffic.

large amounts of traffic, such as mail spam, on well-known Internet ports. However, with such a high percentage of infected machines generating unwanted malware traffic in a network, we surmise that the traffic volume due to malware will be significant, negatively impacting an already strained satellite link.

6. SOCIAL CONSIDERATIONS

Internet connectivity has great potential for transforming a society. On one hand it can create multiple opportunities, while on another it can polarize the society and create a digital divide between those who do and those who do not have access. In this section we examine the societal impact of Internet provisioning in Macha and Dwesa. Our analysis is based on the interviews we performed in both villages. Unless otherwise specified, the aggregate results are presented.

6.1 Economic impact

Benefits of Internet access have been quickly realized by the rural Africans. In Macha, for example, local farmers have used the Internet to gain expertise on crop rotations, a move which completely revitalized the local agriculture [16]. Similarly, one of the interviewees in Dwesa trades local arts and crafts online. Unfortunately, a lot of the opportunities are missed because of the lack of large scale plans and business infrastructure. In many rural areas, such as Macha and Dwesa, banks are not present. In fact, obtaining a credit card is often impossible. While micro-financing solutions help with the local economy, they cannot be used to compete in the global economy. Online trading has been attempted by 28% of the interviewees, often with unsatisfactory results. The lack of nation-wide Internet adoption plans severely hampers many of its useful aspects. For example, e-government services are not available in Zambia.

The Internet can create new local jobs, but it can also promote migrations as the local population can apply for distant jobs online. We observe a striking difference between Macha and Dwesa in that sense. In Dwesa, people are much more likely to use the Internet for job searches ($\chi^2(1, N = 30) = 11.32; p = .001$). South Africa has the second highest GINI coefficient in the world[4] and migrations to affluent areas are very common. Thus, our results show a particular case of the technology adaptation to the local context, not a case of technological determinism.

[4]GINI coefficient is a measure of income inequality.

6.2 Social capital

Online social networks (OSNs) are highly popular in rural Africa, with 69% of interviewees being active OSN users. While more preferred among the younger population[5], they are much more than a leisure activity. In Macha 29% of the interviewees use Facebook for some sort of business correspondence. For example, a local pastor uses Facebook to send spiritual messages to his church followers. OSNs can serve as a great cultural bridge - 82% of the interviewees have remote online friends whom they had not yet met in the real world.

Social ties are established over OSNs with both local and remote friends and relatives (77% have OSN friends within their village and 91% have remote OSN friends). To a lesser extent email is also used for local communication (47% use it for local and 91% for remote correspondence). Of all the means of online communication instant messaging has the highest level of usage locality (80% use it for local and 75% for remote correspondence). The results are not surprising as asynchronous communication provided by email and Facebook messages is more suitable for cases where personal contact does not happen often. However, the insights should be carefully considered from the network systems point of view as the locality of interaction can be used to save the satellite bandwidth [25].

6.3 Gender roles

When the Internet revolution started in the developed world, a digital divide appeared as the first appropriators of the new technology were mostly young white males [9]. Since the relative uniformity of the demographic in Macha and Dwesa and our sample size do not allow race and age bisection, we concentrate on differences in Internet usage among men and women.

While both men and women use computers, women we interviewed are less likely to own a computer ($\chi^2(1, 37) = 8.29, p = .007$). Women spend significantly less time using the computer: on the average men spend 22 hours per week using a computer, while women spend 8 hours ($t(34) = 3.21, p = .003$). When online, both genders are equally likely to use email, OSNs and Youtube.

In Macha, men are more likely to perform computer maintenance themselves than women ($\chi^2(1, 22) = 6.14, p = .023$). They are also more likely to do virus scans ($\chi^2(1, 22) = 4.62, p = .054$). In Dwesa we did not observe such a discrepancy. The reason is, we believe, that one of the project leaders in Dwesa is a woman and a number of local champions are women. As a consequence, the outreach is stronger and women are not likely to see such tasks as "men only".

7. NEXT STEPS

It is clear from both our traffic analysis and our interview data that the network in its current configuration undergoes periods of unusability; extremely slow response times, web page request timeouts, and instant message losses are some of the key problems observed. At the core of these problems is over-saturation of the slow satellite link due to the large number of users. In Macha, there are IT support staff who have achieved an impressive skill set in network engineering, considering that many of the administra-

[5]The average age of an OSN user is 26, while the non-user average age is 36 years ($t(30) = -2.63, p = .024$).

tors who have participated in the network deployment and maintenance do not have a college education. However, the problems experienced in this network require advanced skills in networking engineering. Due to this limited skill set, a pre-configured gateway, *ClarkConnect*, was used. Such a preconfigured gateway is not optimized for rural bandwidth constrained links; there are some fairly simple improvements that can be made using known solutions. There is also scope for appropriate technology research to mitigate these problems. For example on the client side, a different abstraction of web search with underlying intelligent proxies could be provided. Some good examples are the TEK search engine [14] that delivers a low-bandwidth copy of a web page to a user using email and Ruralcafe [4] that provides an expanded search query interface, allowing users to enter additional search terms and maximize the utility of their search results.

We focus on interventions in the network and from our analysis, we identify the need for four key areas:

- The caching behaviour of squid must be changed to identify when the same content is being served by different URLs, as in the case of CDNs.
- Content servers, to mirror services such as Wikipedia, operating system updates and virus updates, need to be installed in Macha with traffic redirected to these servers.
- Traffic destined for local recipients, such as shared pictures and local voice calls, needs to avoid traversal of the satellite link.
- Internet access should be extended to more homes so that off-peak hours of the satellite link can be more optimally used.

There are additional actions that, if enforced, would notably improve the performance of the network. For instance, when a machine with malware is detected, it should be disconnected from the network and only be reconnected once it has been disinfected.

The first two interventions have some known solutions [6], however these need to be constantly adapted as the web evolves. The third intervention is an area that will require new solutions; determining which traffic is of local interest is not a trivial problem. This is especially true of web-based social interaction [25]. The fourth intervention requires solutions to more widely cover rural houses. This is a challenge because homes are often many kilometers away from the community center where Internet is available. There are some new options, using white spaces spectrum, which can achieve longer distances and non-line-of-sight links that go beyond the current gambit of WiFi solutions for rural areas.

8. RELATED WORK

The performance of rural area wireless networks has been investigated since their inception in the early 2000s. Work by Chebrolu et al. [3] and by Sheth et al. [21] deals with wireless propagation problems over long distance links. A comprehensive study of rural area network problems is presented in [22]. In this paper, technical issues were juxtaposed with social obstacles of deploying networks in the developing world. Our work is similar in a sense that it also considers both perspectives. However, Surana et al. concentrate on system troubleshooting with an accent on the energy problems. In addition they provide only anecdotal evidence of the social problems. Our study is geared towards understanding of user behavior and provides a quantifiable description of social occurrences. Specifics of network usage in the developing world is also the subject of [20] and [13]. The former investigates a specific application (VoIP) and its economic feasibility in rural areas, while the latter includes interviews of telecenter users in rural India. In this paper, we do not focus on a single application and supplement our on-site interview data with a comprehensive network trace.

From a large body of work that analyses the behavior of the Internet traffic, [5] is the closest to ours; it examines web traffic usage in Internet Cafés and community centers in Cambodia and Ghana. While it serves as a good starting point, this paper only studied HTTP traffic, and there was not a wireless network aggregating traffic to the Internet connection. In our analysis, we analyzed full TCP dumps and provide insights on a number of problems, including virus and bot presence. Moreover, the profile of Web traffic changed drastically. The popularity of Web 2.0 sites, such as Facebook and Youtube, that we observe in rural Africa could not be envisioned even in the developed world up to a few years ago.

Estimates, in 2007, of the 600 million machines connected to the Internet infected with botnets range from 150 million which are vulnerable to infection to a conservative estimate of 12 million currently infected by bots [6]. A study at Aachen University in Germany in 2007 showed that in 8 weeks, 13.4 million successful exploits were discovered due to 2034 unique malware binaries circulating amongst 16,000 unique IP addresses [6]. Botnests, which are the key platform for most Internet attacks, are surveyed in [7]. This work highlights that botnet command and control has moved beyond the common methods such as contacting an IRC server. Methods used today make use of HTTP and P2P traffic and are thus harder to detect with common intrusion detection systems, such as snort. This emphasizes the fact that the malware we detected may only be the tip of the iceberg. While we are not aware of any previous work studying the problem of viruses in rural network deployments, the importance of user education in rural networks is highlighted in [10].

9. CONCLUSION

The results from our analysis, as well as from previous studies [1], highlight the fact that rural networks need to be treated as a special class of network due to their unique set of challenges. We find that while traffic is primarily web based, a large portion of cache-able traffic from CDNs is not cached. Further, there are many long-lived, non-real-time flows during peak usage times of the network that negatively affect the interactivity of web browsing and often cause instant messaging to fail. These large flows are typically due to automated file downloads, for example when an operating system requests an update. Very high TCP round trip times during the day, sometimes over 10 seconds, with high probability lead to a change in user behaviour; a highly interactive multi-search web experience is likely to become far more singular and deliberate in purpose, almost approaching the environment of a DTN network.

As we described in section 7, there are a number of avenues for improvement of rural network performance. We

[6]http://drupal.airjaldi.com/node/264

believe that more attention should be given to building prepackaged networking solutions for rural wireless networks that are cognizant of the characteristics that have been highlighted in this paper. Some of the problems were addressed in the early days of the Internet when last-mile access was similar to what is currently experienced in rural networks, but many of the problems are new as the Internet has become more dynamic and media-rich. As the average web page size continues to grow, and as Internet applications become increasingly more interactive and demand real- or near-real-time performance, the digital divide will widen unless innovative networking techniques that mitigate these increasing bandwidth demands are employed.

10. ACKNOWLEDGEMENTS

The authors would like to thank the following: Linknet in Macha, Zambia for their cooperation on this project, specifically Fred Mweetwa, CEO Macha Works, for facilitating the interviews explicitly; The Meraka Institute and University of Fort Hare in South Africa for organising the visit to Dwesa and facilitating the interviews; Professor Jennifer Earl for assisting with the interview design and analysis.

11. REFERENCES

[1] E. Brewer, M. Demmer, M. Ho, R. J. Honicky, . J. Pal, M. Plauche, and S. Surana. The challenges of technology research for developing regions. *IEEE Computer*, 38(6):15–23, June 2005.

[2] N. Brownlee and K. Claffy. Understanding Internet traffic streams: Dragonflies and tortoises. *IEEE Communications Magazine*, 40(10):110–117, 2002.

[3] K. Chebrolu, B. Raman, and S. Sen. Long-Distance 802.11b Links: Performance Measurements and Experience. In *MobiCom'06*, Los Angeles, CA, September 2006.

[4] J. Chen, L. Subramanian, and J. Li. RuralCafe: web search in the rural developing world. In *WWW'09*, Madrid, Spain, April 2009.

[5] B. Du, M. Demmer, and E. Brewer. Analysis of WWW Traffic in Cambodia and Ghana. In *WWW'06*, Edinburgh, UK, May 2006.

[6] J. Goebel, T. Holz, and C. Willems. Measurement and analysis of autonomous spreading malware in a university environment. *Detection of Intrusions and Malware, and Vulnerability Assessment*, 4579:109–128, 2007.

[7] G. Gu, R. Perdisci, J. Zhang, and W. Lee. BotMiner: Clustering analysis of network traffic for protocol-and structure-independent botnet detection. In *USENIX Security Symposium (Security'08)*, San Jose, CA, July 2008.

[8] S. Guo, M. H. Falaki, E. A. Oliver, S. U. Rahman, A. Seth, M. A. Zaharia, U. Ismail, and S. Keshav. Design and Implementation of the KioskNet System. In *ICTD'07*, Bangalore, India, December 2007.

[9] E. Hargittai and S. Shafer. Differences in Actual and Perceived Online Skills: The Role of Gender. *Social Science Quarterly*, 87:432–448, June 2006.

[10] J. Ishmael, S. Bury, D. Pezaros, and N. Race. Deploying Rural Community Wireless Mesh Networks. *IEEE Internet Computing*, 12(4):22–29, 2008.

[11] 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'10*, San Francisco, CA, June 2010.

[12] T. L. Kennedy, S. Aaron, A. T. Wells, and B. Wellman. Networked Families. Technical report, Pew Internet and American Life Project, October 2008.

[13] R. Kumar and M. L. Best. Social Impact and Diffusion of Telecenter Use: A Study from the Sustainable Access in Rural India Project. *The Journal of Community Informatics*, 2:1–22, 2006.

[14] L. Levison, W. Thies, and S. Amarasinghe. Providing Web search capability for low-connectivity communities. In *International Symposium on Technology and Society*, Raleigh, NC, June 2002.

[15] M. Mandioma. Rural Internet Connectivity: A Deployment in Dwesa-Cwebe, Eastern Cape, South Africa. Master's thesis, University of Fort Hare, November 2007.

[16] 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*, Bangalore, India, December 2007.

[17] H. V. Milner. The Digital Divide: The Role of Political Institutions in Technology Diffusion. *Comparative Political Studies*, 39:176 – 199, March 2006.

[18] A. Pentland, R. Fletcher, and A. Hasson. Daknet: Rethinking connectivity in developing nations. *IEEE Computer*, 37(1):78–83, January 2004.

[19] H. Schulze and K. Mochalski. ipoque Internet Study 2008/2009, 2009.

[20] S. Sen, S. Kole, and B. Raman. Rural Telephony: A Socio-Economic Case Study. In *ICTD'06*, Berkeley, CA, May 2006.

[21] A. Sheth, S. Nedevschi, R. Patra, S. Surana, L. Subramanian, and E. Brewer. Packet Loss Characterization in WiFi-based Long Distance Networks. In *INFOCOM'07*, Anchorage, AK, May 2007.

[22] S. Surana, R. Patra, S. Nedevschi, M. Ramos, L. Subramanian, Y. Ben-David, and E. Brewer. Beyond Pilots: Keeping Rural Wireless Networks Alive. In *NSDI'08*, San Francisco, CA, April 2008.

[23] L. Waverman, M. Meschi, and M. Fuss. The Impact of Telecoms on Economic Growth in Developing Countries. *The Vodafone Policy Paper Series*, 2:10–23, 2005.

[24] Wikipedia. Macha, Zambia. http://en.wikipedia.org/wiki/Macha,_Zambia.

[25] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao. Exploiting Locality of Interest in Online Social Network. In *CoNEXT'10*, Philadelphia, PA, December 2010.

[26] S. P. Wyche, T. N. Smyth, M. Chetty, P. M. Aoki, and R. E. Grinter. Deliberate Interactions: Characterizing Technology Use in Nairobi, Kenya. In *CHI'10*, Atlanta, GA, April 2010.