Network traffic locality in a rural African village

D.L. Johnson, Elizabeth M. Belding,
University of California, Santa Barbara
davidj,ebelding@cs.ucsb.edu

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

ABSTRACT

The Internet is evolving from a system of connections between humans and machines to a new paradigm of social connection. However, it is still dominated by a hub and spoke architecture with inter-connectivity between users typically requiring connections to a common server on the Internet. This creates a large amount of traffic that must traverse an Internet gateway, even when users communicate with each other in a local network. Nowhere is this inefficiency more pronounced than in rural areas with low-bandwidth connectivity to the Internet. Our previous work in a rural village in Macha, Zambia showed that web traffic, and social networking in particular, are dominant services. In this paper we use a recent network trace, from this same village, to explore the degree of local user-to-user interaction in the village. Extraction of a social graph, using instant message interactions on Facebook, reveals that 54% of the messages are between local users. Traffic analysis highlights that the potential spare capacity of the local network is not utilized for direct local communication between users even though indirect communication between local users is routed through services on the Internet. These findings build a strong motivation for a new rural network architecture that places services that enable user-to-user interaction and file sharing in the village.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance

Keywords

Rural network, Social network, Traffic analysis, Developing regions, Measurement

ICTD2012 2012 Atlanta, GA, USA

1. INTRODUCTION

The Internet has evolved both in terms of size and application since its birth in the early 1990s. Recent studies have shown that the average web page size in 2011 is 48 times larger than the average size in 1995 (14.12K in 1995 [3] and 679K in 2011 [14]). There is an increasing amount of off-PC storage and processing using cloud computing for services such as navigation, photo sharing and file hosting. Many applications that in the past were run on a user's operating system, such as email, word processors and instant message clients, are now run on web browsers. These features have brought users in developed countries closer to the vision of "anywhere any-time" computing, where devices connected to high speed Internet connections delegate computing power and storage capacity to cloud computing services. However, what is not clear are the implications for users in developing regions where Internet access speeds of between 128kbps and 256kbps are common, as was typical of dial-up users of the 1990s.

One of the key consequences of web-based computing is an increase in traffic load. In the rural villages we have studied in Zambia [6], there can be as many as 60 concurrent users sharing a single relatively slow satellite link. As a result, web-applications become increasingly slow, to the point where they are unusable. We have seen that social networking is the most popular web application in the village of Macha, Zambia and users share messages, pictures, music and software with each other, where the sender and recipient are often in the same village. When a user shares an object with another user in the same village using an off-PC storage server such as Facebook or Dropbox, the same file traverses the slow satellite gateway twice: once as it is uploaded to the server, and again as it is downloaded to the recipient. This leads to congestion of an already constrained satellite link.

The most widely used solution to deal with inefficiencies in this centralized model is to use peer-to-peer (P2P) networking. Bittorrent is an example of a popular P2P networking protocol and accounts for between 27% to 55% of Internet traffic, depending on geographic location [13]. Bittorrent works on the principle of "tit for tat" in which users may not download content faster than they can upload content. Peers are always located outside the local network, rendering Bittorrent unusable or extremely slow due to the limited capability of the satellite link. Although we found only traces of P2P traffic (1% of total aggregate traffic passing through gateway) it did account for a surprising portion of traffic with large outbound flows (35% of all outbound flows greater than 100KB). Many users in Macha use Skype, which also uses P2P networking. A large portion of these large P2P outgoing flows may be Skype traffic, which is difficult to distinguish from Bittorrent traffic. We found no evidence of P2P traffic being directly routed between two local users as most local users have no direct network routes to each other and super-peers — well provisioned nodes for routing P2P traffic — are always located non-locally. Clearly traditional P2P networking has not been the panacea to save Internet gateway bandwidth in rural networks.

In order to find a more well suited solution to the inefficiencies of centralized web access, we analyse the locality of interest in the Macha network and its potential to save gateway bandwidth. We carry out two tasks in this study. The first task is to measure the strength of social connections in the village of Macha by extracting a social graph from Facebook instant message chat. This helps us to understand the potential of users to share information and files. The second task is to measure the fraction of local traffic sent directly between users without the use of the Internet, or the fraction of traffic being used to upload content to Internet file-sharing services in order to determine how effectively or ineffectively users are utilizing the local network.

To this end we collected traffic traces over a period of two months in Macha, Zambia that includes all local traffic within the network and all traffic to and from the Internet. Facebook instant messaging was common in our trace; however there was no direct local file sharing and a very small fraction of traffic using file synchronizations (0.94%) or file sharing services (0.65%). We believe this is due to the poor nature of the network and the relatively high cost of bandwidth ($30/GB) rather than a desire not to share these objects; on-line interviews substantiate this hypothesis.

Our interaction graphs of Facebook instant message chats reveal that 54% of chats are between local users in the village, even though only 35% of the observed users were local to the village. We also find that people who travel to areas outside the village are strong sources and sinks for local information exchange. We extract other statistics, such as social degree and clustering coefficients, for the social graph. Packet flows are analysed to understand the percentage of traffic sent directly between users in the village and via central web servers. Finally, we correlate some of our findings with on-line interviews that were collected from 77 users in the village from a wide range of age groups.

Other than interesting anthropological inferences from this analysis, there are many technical implications. The high degree of local conversation provides strong motivation for an automated localization engine that is able to intercept content for local users and re-route it directly to a local user rather than traversing the satellite link. Relocating services that enable user-to-user interaction from the Internet to servers in the local village is also promising for improving local performance.

2. RELATED WORK

There is a small set of studies that examine traffic usage in rural networks. One of the first studies on Internet cafès and kiosks in rural regions in Cambodia and Ghana [4] revealed that many web applications — often monolithic in nature — are poorly suited for low bandwidth environments and suggest smart proxies on either end of the bandwidth-limited link to reduce bandwidth usage. Our previous work includes extensive analysis of performance and traffic characterization in Macha, Zambia in 2010 [6, 7]. This work highlighted some negative effects of the satellite link, such as high round trip times that lead to web request time-outs. We analyzed usage patterns and found that the most commonly used service is social networking; we have since confirmed that this usage pattern continued into 2011 (see Figure 3). In this paper we analyze these social interactions in the network to study the level of local interactivity between users in the village.

More recent work to improve Internet usability in rural regions has been done by Chen *et al* [1, 2]. In [1], an asynchronous queuing model was used where users can queue web requests as well as a cache search feature with predictive text. Users responded positively even though a custom web frame was used to search a local cache or queue content. In [2], a browser plugin was used to pre-fetch pages and serve stale cached pages. This achieved an average acceleration of 2.8x for users browsing non-video web pages. These results bode well for interventions that may need the web access paradigm to be modified somewhat. Our study looks at traffic locality in a rural network and its implications on improving network performance and, as such, represents another tool in the toolbox of techniques to improve network performance in rural regions.

A recent media and society study of the localization in the Internet highlights the fact that as the Internet continues to grow, it is becoming “more local” [12]. This phenomena is beginning to blur the boundaries between online and offline social domains, and it is this trend that demands a localization approach to network design, especially in isolated rural communities.

There are many studies on social network interactions, both at a structural level using friend lists and at an interaction level using wall posts [8, 11]. Wilson *et al* argue that social links created by friend lists are not valid indicators of user interactions [15]. This is shown by the fact that the number of “friend adds” account for 45% of the activity per day whereas comments only account for 10% of the activity. Interestingly the common notion of small-world clustering, which is present in a social graph derived from “friend adds”, is absent from the interaction graph. Our work captures the physical locality of the users, which adds a new dimension to interaction graph analysis.

Locality of interest has been primarily studied in the domain of peer-to-peer networks. For example, a semantic clustering technique is employed by Handurukande *et al* in [5]. The semantic relationship is either implicit, using information such as peer-history, or explicit, using meta-information about the file, such as whether it is music or video. In a rural village, a P2P mechanism that first looks for a peer in a local subnet may be a possible solution to making file sharing more efficient.

3. MACHA NETWORK

Macha, Zambia, highlighted in Figure 1, is a resource-limited 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. Macha Works, through the LinkNet project, has deployed a wireless network that pro-

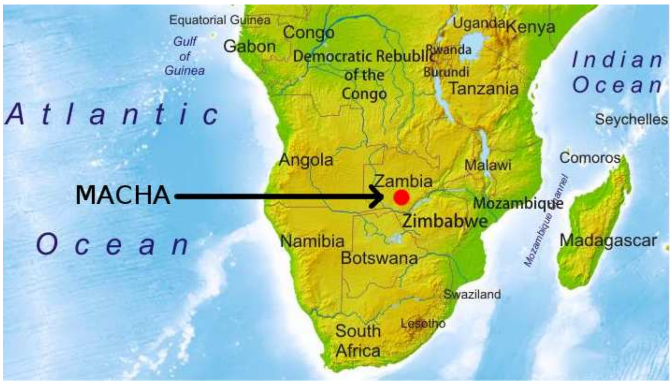


Figure 1: *Location of Macha in the southern province of Zambia.*

vides connectivity to approximately 300 community workers and visitors living around a mission hospital and medical research institute using a satellite-based Internet connection [9]. The majority of users access the Internet at work but half the users access the Internet in their homes.

The satellite connection 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 satellite connection is $1200 (US dollars). Internet access is based on a pay-as-you-go model. Users buy voucher-based data bundles at a cost of $30 for 1GB or 31 days (whichever expires first). The community has a large enough user base to make the sale of Internet access vouchers sustainable in a 'not-for-profit' and 'not-for-loss' environment. There are, however, many frustrated users due to the poor network performance, particular during peak usage times.

4. DATA COLLECTION AND PROCESSING

4.1 Data collection

A simplified model of the network in the Macha community in Zambia is depicted in Figure 2. Computers are connected through a bridged wireless network to the gateway, which is connected to the Internet through a slow satellite link. All traffic in the network passes through the bridge that is monitored passively by a data collector. This allows us to capture all local traffic as well as traffic to the Internet.

Local or non-local direct links between machines using services such as secure copy are very rare in networks without advanced computer users and we found no evidence of this traffic. P2P traffic also creates direct links between machines and is common amongst many users in the developed world (between 42.51% and 69.95% in 2008/2009 [13]), but because of the high cost of data and low bandwidth capacity, it makes up a small portion of traffic in Macha (1% of total aggregate traffic). The majority of traffic is either to and from an Internet web server or between two local clients through an Internet web service. When using a web service like Facebook, a content delivery network (CDN) is typically employed to receive and deliver content. As shown in Figure 2, it is fairly common for a user in the village (Client 1) to send a packet to a service like Facebook through a CDN (CDN 1) and for a different CDN (CDN 2) to deliver a packet to another user in the village (Client 2).

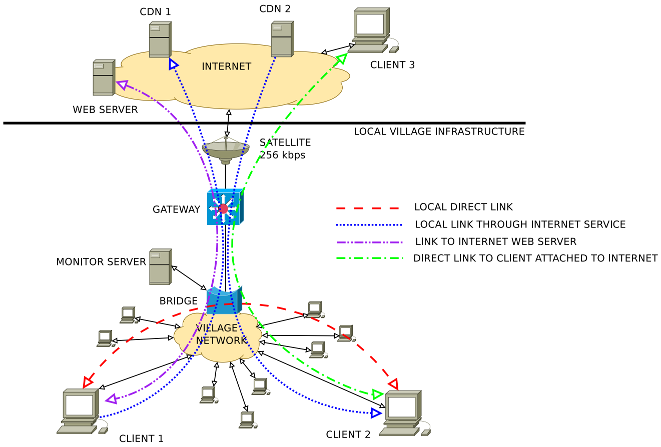

CDN 1
CDN 2
INTERNET
CLIENT 3
WEB SERVER
LOCAL VILLAGE INFRASTRUCTURE
SATELLITE
256 kbps
GATEWAY
LOCAL DIRECT LINK
LOCAL LINK THROUGH INTERNET SERVICE
LINK TO INTERNET WEB SERVER
DIRECT LINK TO CLIENT ATTACHED TO INTERNET
MONITOR SERVER
BRIDGE
VILLAGE NETWORK
CLIENT 1
CLIENT 2

Figure 2: *A simplified model of the network architecture in the Macha network. All traffic in the village passes through a bridge. Four types of possible traffic are highlighted. All traffic types are captured by our monitor server.*

Over the course of two months (February and March 2011), we captured traces of all traffic passing through the bridge, which connects all machines in the network, using tcpdump. The total traffic collected is 250 GB and is stored as multiple pcap formatted files.

4.2 Data characteristics

As described earlier, one of our main goals is to determine the quantity of traffic, from the Macha trace, between two local machines in the village. This is trivial to determine for traffic sent directly between two machines by checking whether the source and destination IP address are both in the local village subnet. However, determining this for traffic that passes through an Internet server is far more difficult. Hence, we need to compare uplink and downlink flows to look for similar patterns leaving and entering the network. These patterns are characterized by features such as IP addresses, port numbers flow sizes, content headers and, when these are not sufficient, features unique to specific objects, such as colour distribution in images. This traffic may be real-time, as in the case of VOIP calls or instant messaging, or it may be non-real time when using a file sharing service, for example.

Real-time traffic between two local machines routed through an Internet server is relatively simple to detect using IP address pairs for flow classification. Returning to Figure 2, if Client 1 routes a real-time connection to Client 2 through CDN 1, there will be a steady stream of packets from Client 1 to CDN 1 through the gateway. This same stream of packets will appear from CDN 1 to Client 2 within a time-window typical of the round trip time of the satellite link. If two different CDN servers are used, as is also shown in Figure 2, the CDN domain, rather than IP addresses can be used as a flow feature. For example, two Facebook CDN's for IM traffic could be "im-a.ak.fbcdn.net" and "im-b.ak.fbcdn.net"; we would use the domain "fbcdn.net" rather than the IP address to define the Internet server being used in the flow.

The task becomes more challenging with non real-time traffic where features such as IP addresses and flow lengths

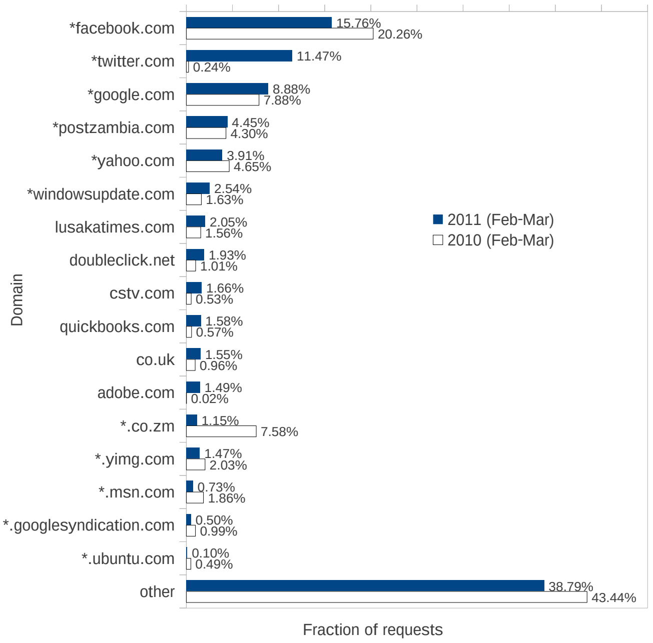


Figure 3: *Internet usage analysis in Macha, Zambia for February and March 2010 and 2011.*

may not be enough to correlate an upload with a download. For example, when using services that host pictures or videos such as Picasa and Youtube, the hosting server will compress or manipulate the object in some way, such that the uploaded object does not match the downloaded object. Fingerprinting can help solve this problem. For example, an image can be fingerprinted using a histogram of colours and will be immune to compression. However, a further complication is introduced when encryption is employed, rendering even fingerprinting techniques useless. In this study we were only able to analyse flows between local machines and real-time flows employing an Internet service. Extracting locality in flows that have been manipulated or which employ encryption is left as an exercise for future work. However to establish the potential quantity of local traffic using services on the Internet, we analyse the domains visited by large outgoing flows, to establish whether web sites are being visited that enable sharing between users.

We extract outgoing and incoming Facebook instant messages from the trace as the headers of instant messages travel in the network in plain text. They are also timestamped and can easily be associated with specific users, which makes it possible to build an interaction graph with temporal information embedded. We ensure privacy by anonymising the user names before performing our analysis.

4.3 Identifying local Facebook traffic

As mentioned in Section 2, we analysed Internet usage statistics in 2010 and found that Facebook traffic accounted for the majority of traffic in the network. We carried out the same analysis using our new 2011 network trace and found that this trend continued; however, some new interesting usage trends appeared. We compare the traffic analysis from 2010 and 2011 traces in Figure 3. Although Facebook declined slightly from 20.26% to 15.76%, Twitter, another social networking site, gained a strong following. While constituting only .24% of the network traffic in 2010, it has become the second most popular web service in 2011. Considering that both Facebook and Twitter are social networking services, social networking now accounts for almost three times the number of page visits compared with Google. The continued dominance of Facebook gives credence to our notion that Facebook traffic is a good representation of social connections in Macha, Zambia.

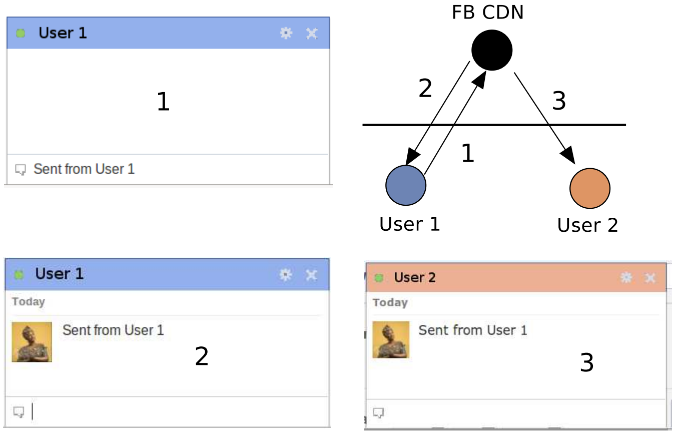


Figure 4: *Outgoing and incoming packets when sending an IM message in Facebook. (a) A message is typed on user A's Facebook IM console. (b) The message is displayed on user A's Facebook IM dialogue area. (c) The message is displayed on user B's Facebook IM dialogue area. (d) A flow diagram of packets sent between client A and B and the Facebook CDN for each of the 3 events in (a),(b) and (c).*

Facebook instant message traffic was extracted by searching for a header, unique to Facebook, in the HTML body of the packets in our network trace. Although Facebook uses a different schema for sending and receiving a packet during an IM conversation, we were able to simplify the analysis by only analysing the incoming packets due to the inherent behaviour shown in Figure 4. For every message transmitted, the same message is sent back to both the sending user and receiving user as it is displayed on the IM web client. To determine whether a conversation is local, we check whether we receive two packets with the same user pair on two different local machines (machines with different IP addresses).

5. SOCIAL GRAPH ANALYSIS

In this section, we present an analysis of a social graph built from instant messages sent in Facebook. We call this social graph an "interaction graph" as it captures the actual interactivity between users rather than passive links. Instant messages represent the strongest possible relationship linkage between users in Facebook as they indicate a one-to-one mapping between users, and they capture true user interaction rather than passive relationships represented by "friend adds" and wall posts. Previous studies support the notion of different levels of relationships in Facebook [15]. They showed that wall posts and photo comments represent a much stronger indication of relationship bonds between Facebook users than "friend adds".

5.1 Social graph

In order to understand the relationships between users within and outside the village, we create a social graph that captures users as nodes and conversations as edges. The weight of the edge depicts the number of conversations in the two month measurement period between two specific users. Figure 5 shows the central section of the Facebook interaction graph for the village. The graph has been arranged so that users who are well connected with each other are closer to the centre. The color, as shown by the key, depicts whether the user was always in the village, always outside the village, or a traveller. Users were detected as travellers when they originated a message both from within and from outside the village during the trace. This was possible to detect as users were tracked using Facebook IDs rather than associated IP addresses, which are not persistent. Figure 6 shows the edge of the interaction graph with isolated communities of users, often having one to six local village users, connecting up to 20 outside users.

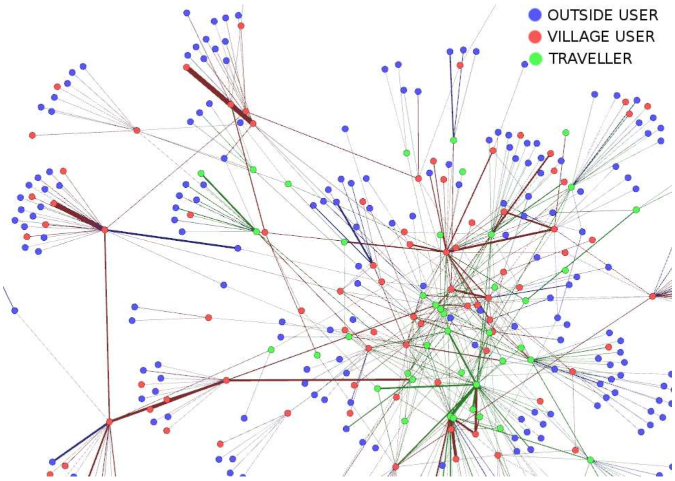


Figure 5: *Interaction graph for Facebook IM traffic showing users as nodes who either remain in the village, travel but return to the village or who are always outside the village. The edges denote conversations. Thicker edges indicate that more instant messages were sent between users.*

There are a few key conclusions we can draw from these interaction graphs:

- There are key users in the village that act as strong links to the outside world. Often these key people are travellers. These users are easily identified by the fan shape motifs in the graph.

- There are a number of isolated communities where the majority of nodes are external, often linking to only one or two local users. From personal observations made in the village, these are likely researchers who visit the village to carry out their research and collaborate with overseas colleagues and have very little online interaction with local people.

- The strongest bonds (highest number of conversations) are between local users, especially those that travel and key contact people. These users most likely form information brokers enabling the flow of information between users outside and inside the village.

- There are surprisingly few cases of one external user connecting to many users in the village. One interpretation is that users who arrive in the village come from a diverse set of communities, where people in each community only know one person in the village. Another is that outside users who meet travellers from Macha tend to use only one person as their key contact or gatekeeper in the village.

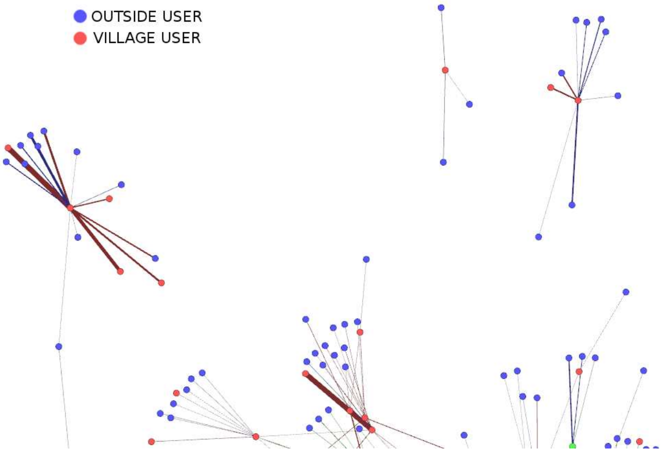


Figure 6: *A section on the periphery of the interaction graph showing isolated communities of users.*

It is clear from this interaction graph that users in the village form a very close knit community with stronger links between other local users than to users outside the village. However, the graph also reveals that Macha is not a homogeneous community. There is a small set of isolated individuals and communities; on-the-ground observations reveal that these are often non-locals living in the village. The presence of non-locals in the community is common due to international visitors and relocations and rotation of Zambian health and education personnel nationwide.

In order to understand the characteristics of this interaction graph, such as the fraction of local messaging and clustering, we now perform more detailed statistical analysis.

5.2 Statistical analysis

Over the two month measurement period 573 unique Facebook users were identified, of which 140 were local users who never left the village and 43 were users who travelled. There were 14,217 unique instant messages sent between 726 unique user pairs. Our analysis reveals that 54% of the instant messages were between local users in the village even though only 35% of the users were in the village, with 7.5% of these local users occasionally travelling. This shows that instant messaging is used extensively for intra-village communication.

To understand the statistical distribution of these relationships and messages among the users in the village, we plot the cumulative distribution function (CDF) of the social degree between local users, and from local to external users in Figure 7(a). Social degree is a measure of the number of edges connected to a node; this correlates to the number of users with which a specific user has communicated. Only a local user's perspective is shown as the social degree of the external users cannot be fully known. We also plot the CDF of messages sent between local users, and from local to external users, in Figure 7(b).

Node degree between local users reveals a large number of small cliques and a small set of well connected community members who act as messenger hubs. Local users who have connections with external users generally tend to have a large number of connections. 70% of local users have a degree of 3 or less, compared to 7 or less for local to external users. The average degree of a local user is 3.6 and the average degree from a local user to an external user is 5.3. However, Figure 7(b) shows that local users send more messages to each other than to external users even though they have a lower node degree. Local to local user communication also has a heavier tail revealing a small subset of very active local users who message each other. Looking at the social graph, we see that well connected users are often also responsible for high message counts, confirming their role as information brokers. These results confirm that it is not enough to look at the connections between users in a social graph. Interaction between users reveals the true behavioural characteristics of users, and in our specific domain of interest, potential of users to interact locally.

In order to understand the level of cohesiveness of users in the social graph, a number of social graph related metrics are employed:

- **Clustering coefficient** measures the tendency of nodes in a graph to cluster together. For a node with N neighbours and E edges between these neighbours, the clustering coefficient is (2E)/(N(N-1)). A higher clustering coefficient can be interpreted as nodes forming tightly connected localized cliques with their direct neighbours.

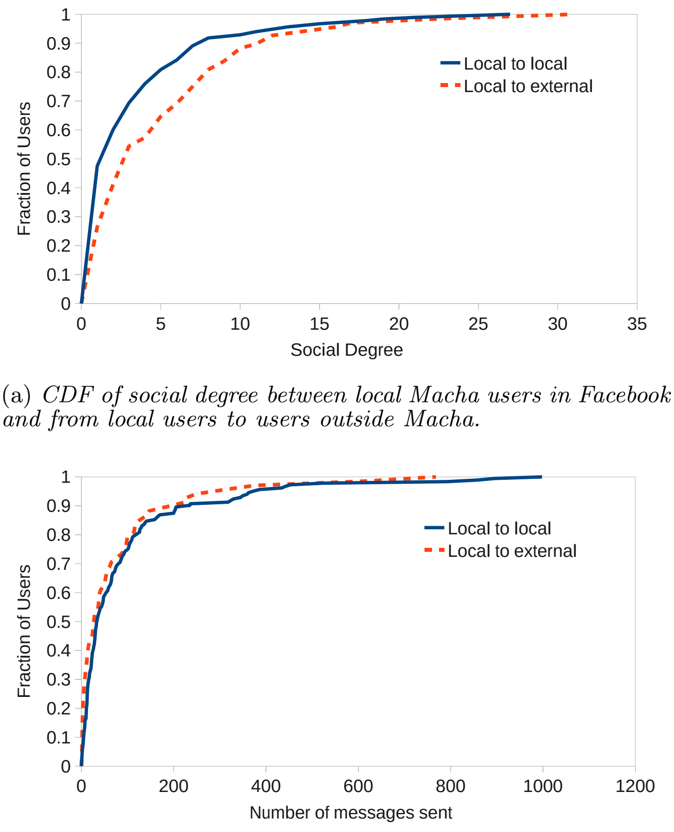
- **Average path length** defines the number of edges along the shortest path between all possible pairs of nodes in the network. This measures the degree of separation between users. A lower average path length indicates a community that is closely connected and through which information can spread quickly.

- **Eccentricity** of a graph is the maximum distance between any two nodes.

- **Diameter** is the maximum of all eccentricities.

The clustering coefficient is only measured for local users in the network as we do not have a complete picture of Facebook interaction for external users. This value effectively measures the cohesiveness of the local community rather than checking for a "small-world effect" in the larger Facebook community. The average clustering coefficient is 0.1 (average on facebook is 0.164 [15]) for all the local nodes. Taking into account that, on average, regular interactivity only occurs with one fourth of a user's Facebook friend list [15], this value still represents a local community that is strongly connected.

The average path length between all nodes in the graph is 3.798 (average on facebook is 4.8 [15]). This is far less than the six-degrees of separation hypothesis for social graphs [10] as we have an incomplete graph, with edges only extending to the first tier of external users. This, together with the clustering coefficient, provides further proof for a strongly connected local community.



(a) *CDF of social degree between local Macha users in Facebook and from local users to users outside Macha.*

(b) *CDF of instant messages between local Macha users in Facebook and from local users to users outside Macha.*

Figure 7: *Statistical distributions for all Facebook instant messages sent in February and March 2011.*

The diameter of the local graph is 8 (average on facebook is 9.8 [15] and average on Orkit is 9 [15]). Considering that this is a localized graph, the value is surprisingly high. This is due to a set of outlier users who are weakly connected to other users in the network (low social degree). They form a a category of users who do not use Facebook IM as a regular means of communication and most likely use other channels such as different IM clients, SMS or email.

From this social graph analysis, it is clear that the potential for local interaction, other than instant message interaction captured in the social graph, is very high. If the infrastructure was supportive of local connectivity, users would share music, pictures, videos and software as a natural consequence of these relationships as is seen in well provisioned networks in developed countries. In the next section, we determine if indeed there are any attempts to send data between local users.

6. LOCAL TRAFFIC PATTERNS

Information can be exchanged between users using a deliberate or passive action. A deliberate action includes activities such as making a Skype call, transferring a file directly between two computers, or sending an email. A passive action involves a user uploading a file to a server, such as an image upload on Facebook, and another user downloading this file at a later point in time. In this section we analyse our traffic traces to detect deliberate actions to share content with other users in the village or use of services which may lead to passive content sharing.

We begin with an analysis of all traffic transmitted directly between local machines in the network.

6.1 Traffic sent directly between machines

The local wireless network makes use of approximately 100 802.11b/g wireless routers to connect close to 300 users in the village. These routers are typically able to operate between 1 Mbps and 5 Mbps depending on the distance between them and level of interference [6]. Hence the local network has substantially higher capacity than the satellite gateway (128kbps) and, as such, represents a large amount of unused spare capacity. In order to understand whether this local capacity is utilized we now analyse the amount of traffic transmitted between users in the local network.

Table 1: *Statistics for traffic between local machines.*

Description	Quantity
Total flows	50,355,759
Total Data (GB)	249,338
Including gateway servers	
Total local flows	13,546,582 (26%)
Total local data (GB)	12,548 (5%)
Excluding gateway servers	
Total local flows	6,726 (0.013%)
Total local data (GB)	7 (0.0029%)

We summarize the high level findings of local traffic in Macha in Table 1. Traffic, including gateway servers (DNS server, capture portal server), consists primarily of HTTP traffic as users sign onto the Internet using the local web "capture portal". There were also many DNS requests to the local DNS server to resolve host names. This traffic constitutes 26% of the total aggregate traffic passing through the gateway. Due to the local network having between five and twenty times more bandwidth than the satellite link and the opportunity for local connections to create isolated flows between local users in different parts of the network, we conclude that the local network will have a large amount of residual capacity.

In order to understand whether users connect directly to each other in the network, we exclude traffic to the gateway servers, which perform common tasks of Internet login and DNS resolution, and calculate the remaining residual traffic transmitted directly between client machines. This makes up a very small portion of the traffic, constituting only 0.0029% of the total traffic seen during the measurement period. Further investigation of this traffic revealed that 44% is Netbios traffic, a Windows networking protocol, and 2.4% is samba traffic, a Windows network and printer sharing service. These are services which are automatically activated by Windows and do not represent an explicit desire by a user to connect to another user. No evidence was found of files being transferred using Samba or any other protocol such as the "File Transfer Protocol" or "Secure Copy". There was also no evidence of any VoIP calls being made directly between two local machines without the use of the Internet.

Clearly the best solution to conserve the bandwidth-limited and costly Internet gateway is to make use of services that can enable direct user-to-user communication or host content on local servers in Macha. For example, a simple FTP server could host music files that users wish to share or a Asterisk VoIP server could enable local calls between users in Macha using soft-phones. However, this requires advanced computer skills and breaks the current paradigm of web-based computer usage [7]. As a result, we primarily observe flow patterns where users make use of web services to share content or make phone calls. We discuss this scenario in the following section.

6.2 Traffic sent locally through Internet servers

As highlighted in Section 4.2, the process of extracting local traffic routed via servers on the Internet is a complex process. In order to determine the quantity of potential local traffic, we first analyse the destination domains of large outgoing flows (flows greater than 100KB). These flows have the potential to contain objects that may be downloaded later or streamed in real-time to local users in the village. We found that 15% of the outgoing traffic consisted of large outgoing flows. We divide this portion of traffic into domain categories in Figure 8. We are also able to check whether any of the Skype flows embedded in the "Skype and Bittorrent" category are routed between local users through an Internet super-peer. We leave the task of matching up and down flows in manipulated objects, such as photos and images, and encrypted flows for future work.

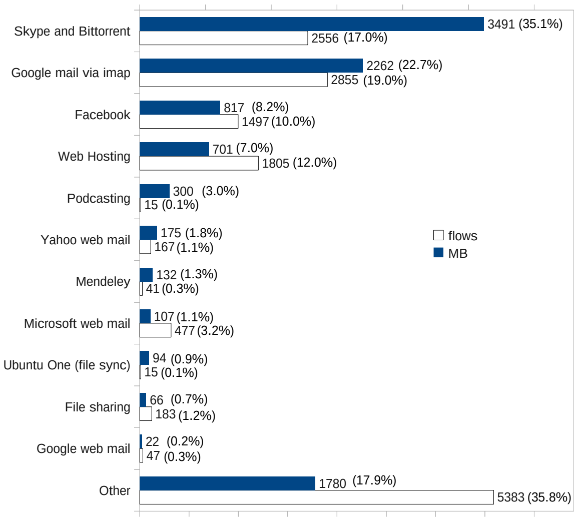


Figure 8: *Domains visited by outgoing flows greater than 100KB.*

Skype and Bittorrent applications create the largest amount of upload activity. It was not possible to separate these two traffic types, as they both behave in a similar manner. They use a mix of UDP and TCP and random source and destination ports, and they exhibit balanced throughput for incoming and outgoing streams. Bittorrent over a satellite gateway with cap-based-Internet access is a costly option as users pay for somebody else downloading fragments of the file. This is due to the 'tit-for-tat' nature of Bittorrent. Hence we expect that most of this traffic is Skype rather than Bittorrent (strong evidence for Skype dominance is also provided in our interviews in Section 7), but further advanced data mining techniques will be required to extract the exact breakdown of these two traffic types. Skype and Bittorrent have the potential to be used to make calls between users in the village or exchange files. However this is far more likely to occur with Skype as Bittorrent is mostly used to download software, music or video that was not sourced by local friends.

The "Google mail via imap" category is for users who connect through an IMAP email client and send file attachments. This traffic type is encrypted using SSL. Analysis of the Facebook traffic category shows primarily image uploads which are usually compressed by the Facebook server. Matching up and down flows for both these forms of traffic, containing encryption and image manipulation, is left as a future exercise. There were a number of web-hosting sites used in Macha, such as softlayer and freewebs; these are grouped together to form one category called "Web hosting". Although no Youtube uploads were found, users in Macha regularly upload podcasts to Podomatic, which accepts short audio or video clips that users have generated using mobile phones or web cams. Only a small portion of file sharing was found as well as some file synchronization services through Ubuntu One. Mendeley is a program for managing and sharing research papers, and some academic paper uploads occurred. There was a surprisingly small fraction of uploads using web-based email clients; most of the mail attachments occur through non web-based email clients.

Using a set of basic heuristics, it was possible to identify two local Skype calls in the measurement period that were routed between two local users via an international super-peer. Each flow is time-stamped and a simple check was made to see whether a local user connects to an external super-peer, and this same super-peer connects back to a different local user. The second heuristic is to check that the time-stamp between the outgoing and incoming flows is not more than a few seconds. These two skype calls lasted 10 minutes, and 3 minutes respectively, and make up an insignificant proportion of the total number of Skype and Bittorrent traffic seen.

Although it is difficult to conclusively determine the amount of local content sharing in Macha due to much of the traffic being encrypted or manipulated, we conclude that there is most likely only a small portion of overall uplink network traffic that is being used to share large objects between local users. One of the key reasons for this is the high cost of bandwidth, together with the Internet-cap-model being used. This is in stark contrast to the large amount of instant messaging between users. For example, sharing a 500M video costs a user $15. It is far more cost effective to put the video on a flash drive and walk to another local user. However, if the architecture of the network made it possible to exchange these files for free using a local web-based application or through intelligent routing, local file sharing would be sure to increase as it has in developed regions.

7. ON-LINE INTERVIEWS

To understand whether the demographics and qualitative usage behaviour of the users in the village have any bearing on the behaviour we observed in our trace, we conducted an on-line survey in Macha. This survey was conducted during June and July 2011 and collected data broadly focused on access and usage of Web 2.0 applications and services. The survey was implemented on the SurveyMonkey tool, and was based upon an open source example. It was the first time that an extended questionnaire was offered in Macha with only an online version available. Users were invited to participate via email and Facebook links. Local support to help respondents in interacting with the online survey was provided upon request. 77 users living in Macha participated in the survey consisting of 89 questions.

Some demographic findings from the survey:

- 69% of respondents were between 20 and 30 years old.
- 34% of the respondents were female, and 66% male.
- 88% of respondents were able to use computers and the Internet to achieve most of their objectives. 12% regarded themselves as novice to computers.
- 67% of respondents use the Internet more than 3 hours a day.
- 49% of respondents have Internet connectivity at home.
- 87% of respondents use the Internet at work.
- 71% of respondents use the Internet for learning.
- 51% of respondents use the Internet for entertainment.
- 91% of respondents wish to access the Internet more frequently. 34% are prevented from doing so because of Internet connectivity costs, 33% because of bandwidth limitations, and 25% because of the (institutional) regulations for use.

The survey examined current online activity in 16 categories. Key findings of the survey relevant to social networking are:

- In online engagement, the use of photo sharing stands out, with 60% of the respondents sharing pictures on Facebook and 52% commenting on them, at least several times per month or more;
- 34% of respondents do watch videos on a video sharing website several times per month or more.
- 53% of respondents indicate that they work collaboratively online using tools such as Google Docs.
- 54% of respondents interact on social networks like Facebook several times a week or more. A mere 9% of respondents never interact on social networks, while 24% never used Facebook. 72% use instant messaging.
- All respondents use e-mail. 59% of respondents make phone calls over the Internet several times per month or more, and 59% of respondents search for news online at least several times a week.
- 94% of respondents never used the Internet for an online business. 73% have never used the Internet to purchase goods.

In conclusion, the survey and measurements give indications of an Internet community of 300 users with 200 regular users in Macha. Most people desire more access and interaction, but costs and bandwidth limitations restrain them from doing so. Many of these findings correlate well with what we have observed in our analysis. The 140 local Facebook users active on instant messaging correlates with 72% of the 200 regular users who claim to actively use instant messaging. The well connected interaction graph in Facebook is supported by the fact that only 9% of the community does not use social networking. The dominance of email traffic in our trace is also well supported by the fact that all respondents claim to use email.

The lower proportion of users who have Internet access at home versus users who have access at work will most likely have a negative effective on local user connections as activities like local VoIP and file sharing often occur after work. Internet users in Macha are primarily under 30 and represent a new burgeoning group of trend setters for the rest of the surrounding rural community. As these users, and others that will join, begin to generate and share content and become active members of the global digital village, every effort should be made to mask the crippling effect of the slow Internet gateway. Building a novel new localized network software architecture, that takes full advantage the unique strong clustering revealed in our traffic analysis, is one such mechanism to do this.

8. CONCLUSION

Social networking has grown in Macha and the network in Macha is used extensively for intra-village conversation in the form of instant messaging. For Facebook instant messaging, our statistical analysis revealed that 35% of the users were local to the village, with 7% of these being travellers, and 54% of the instant messages sent were between local users. Interviews also revealed a large proportion of users who make extensive use of social networking with instant messaging being especially popular. Although there is strong social cohesion in this community, no direct local networking traffic related to sharing files or making voice or video calls was found. There was also very little use of Skype between users in the village via Internet super-peers (2 calls within the two month measurement window) or file sharing (0.65% of large outgoing flows). File synchronization services on the Internet were also minimal (0.94% of large outgoing flows) due to the high cost of bandwidth ($32/GB) and bandwidth limitations of the satellite gateway.

It is clear that the current hub and spoke architecture of the web, with most services running as monolithic servers, is not well suited to rural networks with low-bandwidth gateways. However, the strong social cohesion and interdependence often found in rural communities is a key differentiator which, if taken advantage of, can make substantial improvements to network performance when bandwidth-limited Internet gateways are used. A new localized network software architecture is needed to take full advantage of this strong clustering found in rural communities. This software should place services that enable user-to-user interaction and file sharing in the village. Once in place, rural communities will disseminate information amongst each other more efficiently and Internet responsiveness will improve due to offloading "local" traffic from the gateway.

9. ACKNOWLEDGEMENTS

The work described in this paper was funded in part by NSF NetSE award CNS-1064821. The authors would also like to thank Linknet in Macha, Zambia for their cooperation on this project.

10. REFERENCES

[1] 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*, London, December 2010.

[2] J. Chen, D. Hutchful, W. Thies, and L. Subramanian. Analyzing and accelerating web access in a school in peri-urban india. In *WWW*, Hyperabad, India, March 2011.

[3] J. Domčnech, A. Pont, J. Sahuquillo, and J. A. Gil. A user-focused evaluation of web prefetching algorithms. *Computer communications*, 30(10):2213–2224, 2007.

[4] Bowei Du, Michael Demmer, and Eric Brewer. Analysis of WWW traffic in Cambodia and Ghana. In *WWW*, Edinburgh, UK, May 2006.

[5] S. B. Handurukande, A. M. Kermarrec, F. L. Fessant, and L. Massoulié. Exploiting semantic clustering in the edonkey p2p network. In *Proceedings of the 11th workshop on ACM SIGOPS European workshop*, Leuven, Belgium, September 2004.

[6] 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.

[7] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. Traffic characterization and internet usage in rural africa. In *WWW*, Hyperabad, India, March 2011.

[8] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. *Link Mining: Models, Algorithms, and Applications*, pages 337–357, 2010.

[9] 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.

[10] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, San Diego, CA, 2007.

[12] J. Postill. Localizing the internet beyond communities and networks. *New Media & Society*, 10(3):413, 2008.

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

[14] S. Sounders. The HTTP archive. http://httparchive.org.

[15] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B.Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer systems*, pages 205–218, 2009.