Instant Message locality and topic extraction in rural networks

Mariya Zheleva, David Johnson
Department of Computer Science, UCSB
{mariya, davidj}@cs.ucsb.edu

1 Abstract

Internet is being introduced to developing regions in the world and a few key questions arise: How well suited is it to low-bandwidth conditions typical of these areas and what is it being used for? In our previous work we analyzed network traces collected from a rural village in Macha, Zambia and discovered that the majority of users make use of the Internet for social networking. We continue with this analysis using 2 months of captured traffic and discover that 54% of the instant messages being sent are between local users in the village. Converting user pairs and instant messages into a social graph representation revealed that only 35% of the users were local and of these 7% were travelers and acted as key contacts in the village. Topic analysis shows that there is a substantial overlap between topics typical to the local community and those discussed with external users. Nevertheless, there are topics such as church, short-term scheduling and health that are typical only to messaging between local users. Furthermore, there are not too many topics that are persistent over the entire trace, which means that likely topics vary significantly over time.

2 Introduction

The Internet has evolved both in terms of its size and its application since its birth in the early 1990's. Recent studies have shown that the average web page size in 2008 was twenty two times larger than its average size in 1995. There is also an increasing amount of off-pc storage such as photo storage using Picasa and file storage using Dropbox. Many applications which were usually 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 user's in developed countries closer to the vision of "anywhere anytime" computing but what are the implications for user's in developing regions still operating on average Internet speeds of 115 kbps typical of dial-up users of the 1990's?

One of the key consequences of web-based computing is an increase in traffic load. In the rural villages we have studied [6] there can be up to 60 users sharing a single slow satellite link. This means that these web-applications are going to become increasingly slow as more users come on-line. We have also seen that social networking is the most popular application for rural users and that user's are sharing messages and files with each other, often in the same village. When a user shares a file with another user in the same village using an off-pc storage server such as Dropbox, the same file will traverse the slow satellite gateway twice leading to congestion of an already strained satellite link.

Understanding the level of peer interaction between users in the village is best done by analyzing a social graph. This will give a much more detailed understanding of user interaction compared to only analyzing the binary content traversing the link as this does not have enough meta-data to associate the content with particular users. If there are interactions between users in the village and to users outside the village, it is useful to understand how often these occur or if there are key individuals in the village who are conduits to the outside world. Social graph analysis has typically been done by crawling Facebook to extract a graph of friend relationships. This does not, however, help you understand the amount of interaction between friends or friends that you may never interact with. In this work we are able to make use of network traces to extract Facebook instant messaging between users, we extract a social graph which reflects relationships which are active and how strong these relationships are in terms of number of conversations that occur between users. Although we did not analyze all the content sent through the satellite gateway, to check which content is destined for or consumed by local users, we postulate that a high degree of conversation between individuals will result in an increase in other types content shared between users, such as images.

A common question amongst the web community is how much value Facebook adds to a community. In the rural village of Macha, Zambia that has recently been exposed to the Internet, this question becomes even more important. We attempted to answer this question by carrying out topic analysis of conversations occurring between individuals inside the village and conversations to users outside the village. We found some common themes amongst all topics, such as scheduling appointments and staying in contact with friends, relatives and partners. We also found that local conversations tend to be about more short term events occurring within the span of a few days and there are also discussions about local issues such Church and medical facilities. External topics tend to focus on long term planning within the span of a year and issues about school and work are discussed.

Other than the obvious interesting anthropological inferences from this work, there are many interventions that could be implemented as a result of this analysis. A high degree of local conversations in this village provide strong motivation for an automated localization engine which is able to intercept content for local users and reroute it directly to the local user rather than traversing the satellite link.

3 Related Work

We have done extensive studies of the rural network in Macha, Zambia [7], [6] . Our previous studies [6] highlighted the fact that the majority of Internet traffic is web based and the most common application is social networking as shown in Fig. 1 . However we did not have any insight into the nature of the social interactions in the network, which this work has now highlighted.

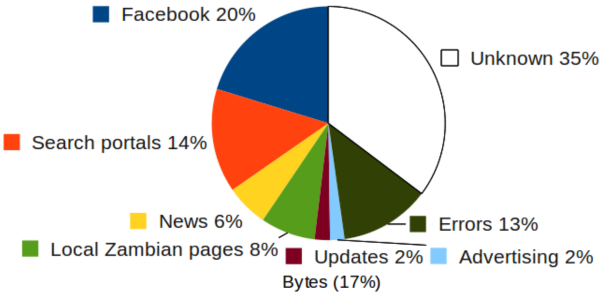


Fig. 1: **Analysis of web site usage** Facebook received most of the hits followed by search portals at like Google and Yahoo. Unknown traffic is traffic that could not be categorized by port analysis such as Skype traffic

There have been previous attempts to analyze the characteristic of text content in instant messages to find or track hot social topics. Want *et al* [10] synthesize related instant messages as a conversation and then perform enhancement of the conversation with closely related words. They then apply a k-means algorithm to cluster conversations into topics. They show that their algorithm outperforms traditional TF-IDF algorithms when creating topic clusters. Unlike our work, however, this is not used to discover topics but rather to categorize conversations into predefined topics. They also don't combine social graph analysis with topic analysis which we did to understand if there is a difference between topics in intra and extra-village communication.

In our analysis we focus on extraction of common topics that appear in the conversations over the entire trace duration. Previous work related to topic extraction from instant messaging (IM) [4] looked at topic extraction based on an enhanced TF-IDF algorithm. One of the results in this paper indicates that topics often change over time. According to our study on the Macha trace, there are not too many topics that remain constant over time; a result that is well aligned with the discoveries of [4].

Previous work on social network extraction based on IM [8] [9] relies on users' status change over time (i.e. on-line, off-line, idle, busy). Given such records for each user, along with user's friend list, the proposed algorithm indirectly extracts links between users based on the similarity of status change patterns over time. In our project we look at the source and destination user of each conversation to directly reconstruct the links between users. We are not familiar with previous work that had looked at conversations to extract the underlying social graph.

4 Dataset

4.1 Data Collection

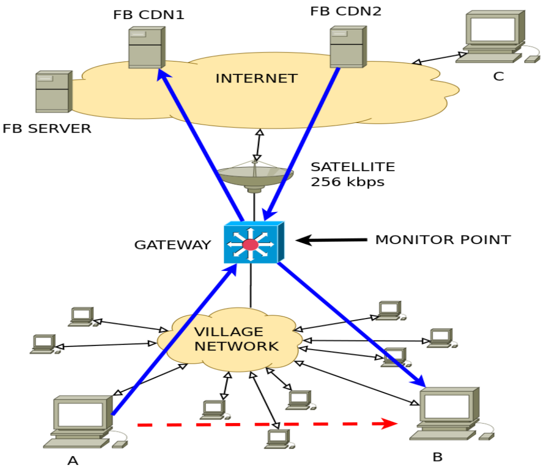


Fig. 2: **Traffic capture scenario**

The local network in the Macha community in Zambia is depicted on Fig. 2. Users' computers are connected through a wireless network to the gateway, which is connected to the Internet through a slow satellite link. All the traffic that goes out and in the network, passes through this gateway. As showed on Fig. 2 if a user A wants to send something to another user B in the community, the traffic needs to go through the wireless network to the gateway, then through the slow satellite link to the server and then back through the satellite link and the wireless network to user B. This scenario helped us capture all the traffic going out and coming in the network by monitoring the local network interface of the gateway. Over the course of two months (February and March, 2011) we captured full packet size traffic using tcpdump. The total traffic collected amounts to 200 Gb and is stored as multiple pcap files.

4.2 Data Characteristics

As described earlier, one of our main goals is determining how much of the traffic from the Macha trace is both generated and consumed locally. In order to extract such information we needed to compare up-link and down-link flows looking for similar patterns going out and coming in the network.

This is a very difficult task, as we are facing various types of traffic each of which follows different format and requires different techniques for preprocessing and comparison. Fig. 3 presents the types of traffic we observed in the Macha trace by looking at packet's MIME type. Binaries (e.g. javascript) are often downloaded and executed locally while users are browsing, and can be compared by checksum matching. Other very common type of traffic consists of images and video. However web sites usually compress and change the format of the images/video, thus looking at payload size/format is no longer a base for comparison. One solution is to compare image histograms over their color distributions. The latter remain the same regardless of the image format and compression. Next, there is a lot of encrypted content generated and exchanged in the network. We are not aware of a method for similarity search between encrypted data-sets. Lastly, vast majority of the traffic is text, which is associated with web browsing and instant messaging.

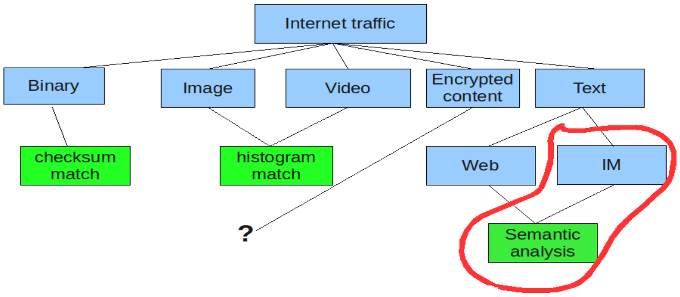


Fig. 3: **Captured traffic types**

Of the outlined types of traffic, we focus our attention on instant messages (IM). First of all, instant messages travel in the network in plain text, thus obtaining them from the traffic trace is relatively easy. Second, they are timestamped and can easily be associated with specific users which makes it possible to track the separate conversations. Having the text messages in plain format not only allows similarity search by the means of semantic analysis techniques, but it also allows extensive study on some social aspects of the local community related to Internet usage. We use this data to extract common topics discussed on-line, and reconstruction of the underlying social graph within the community. We hypothesize that the trends discovered in instant messaging can be extrapolated to other types of traffic as well: people who chat a lot are likely to exchange other type of content (e.g. pictures and video).

4.3 Data Preparation

In order to prepare the data, we needed to extract all the IM conversations from the pcap traces. As mentioned, we have 200 Gb of data, consisting of various types of traffic, where both up and down-link messages are in the same set of traces. Furthermore, one conversation might be split in multiple flows; each flow could be transmitted through multiple packets and packets are likely to be captured out of order. The main challenge then is how to reassemble an IM conversation out of the pcap traces?

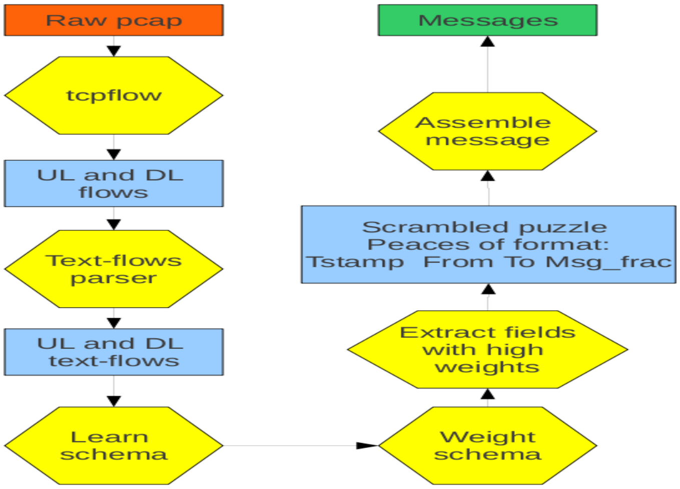
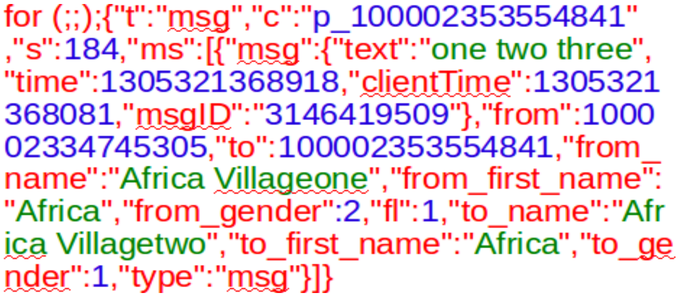


Fig. 4: **Data preparation pipeline**

Fig. 4 presents our pipeline for conversation extraction. It starts with raw pcap files, as collected by tcpdump [1]. We then use tcpflow [2] to convert the raw pcap files to flows and classify them as up-link or down-link. Next we use a text filter that based on MIME type, extracts only the text flows.

To be able to extract the message fraction as well as its sender and receiver, we need to learn the schema of the flow. Fig. 5 shows a typical schema of a Facebook IM flow in down-link direction. The fields marked with red maintain constant format over time, those marked with blue remain the same length but change in terms of content, and those marked with green change both in terms of length and content. The schema learning algorithm uses a couple of messages as a training set to learn which fields are red, blue and green. It then weights the fields depending on their format persistence



```
for (;;);{"t":"msg","c":"p_100002353554841"
,"s":184,"ms":[{"msg":{"text":"one two three",
"time":1305321368918,"clientTime":1305321
368081,"msgID":"3146419509"},"from":1000
02334745305,"to":100002353554841,"from_
name":"Africa Villageone","from_first_name":
"Africa","from_gender":2,"fl":1,"to_name":"Afr
ica Villagetwo","to_first_name":"Africa","to_ge
nder":1,"type":"msg"}]}
```

Fig. 5: **Learning the schema**. Facebook message format in down-link direction.

and finally extracts the fields that vary a lot over time (i.e. have higher weight).

The result from the latter step is a file with message fractions of the format: *timestamp : fromUser : toUser : messageFraction*. The last step of the pipeline is conversations reassembling. We do this by looking at messages exchanged between unique pairs. If the timestamp difference between two messages is smaller than a threshold, then we consider them as a part of the same conversation. We chose as a threshold 1 hour. A quick test that we did confirmed this threshold as feasible - we counted the number of conversations generated by 1 hour threshold and 24 hour threshold and the results were comparable.

5 Data mining methods. Results

5.1 Identifying local traffic

As discussed in Section 4.2, we focused our attention on identifying local traffic for instant messages in Facebook. 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 analyzing the incoming packets due to an inherent behavior in instant message web clients. From Fig. 6, you can see that for every message being sent, the same message will be sent back to the user and displayed on the IM web client.

As shown by Fig. 6, checking if a conversation is local is done by simply checking if we receive two packets with the same message and user pair to two different machines (machines with different IP addresses). Our analysis revealed that 54% of the instant messages are between local users in the village even though only 35% of the user's are in the village. This shows that instant messaging is used extensively for intra-village communication. If the satellite connection is off-line due to poor weather or equipment damage, for example, all these intra-village messages would not be able to reach user's in the village due to client-server model of Facebook. This provides strong motivation for a localized messaging and file-sharing service in the village to save bandwidth as well as allow the local network to still be useful when the satellite connection is off-line.

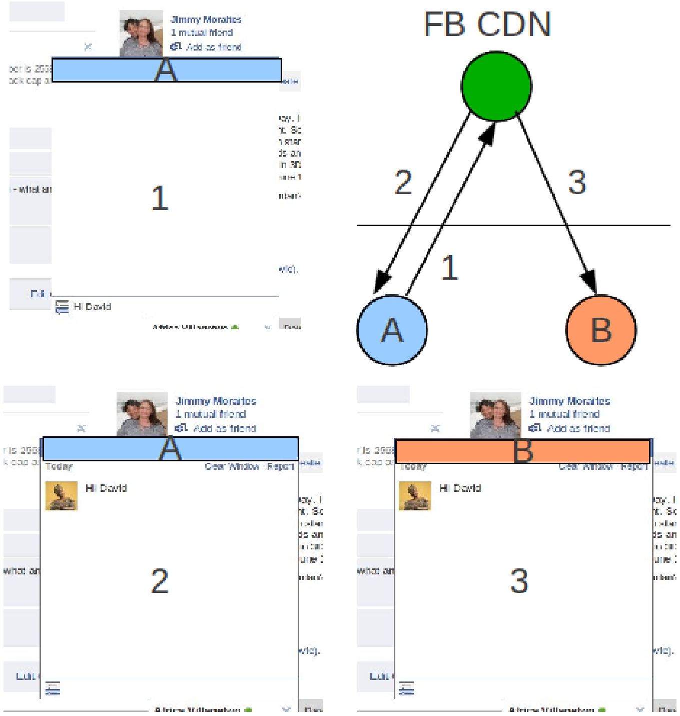


Fig. 6: **IM messages in Facebook**.Outgoing and incoming packets when sending and IM message in Facebook .

5.2 Social Graph Analysis

In order to understand the relationships between users in and out the village, we created a social graph which captures users as nodes and conversations as edges. The weight of the edge depicts the number of conversations in the 2 month measurement period between two specific users. Fig. 7 shows the central section of the Facebook instant message social graph for the village. The social graph has been arranged so that users well connected with other users 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 traveler. A user was detected as a traveler when they originated a message from within the village or from the outside the village at first and changed status at a later stage. Fig. 8 shows the edge of the social graph with isolated communities of users often having one to six local village users connecting to up to 20 outside users.

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

- There are key contact people in the village which act as linkages to the outside world. Often these key people are travelers. 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

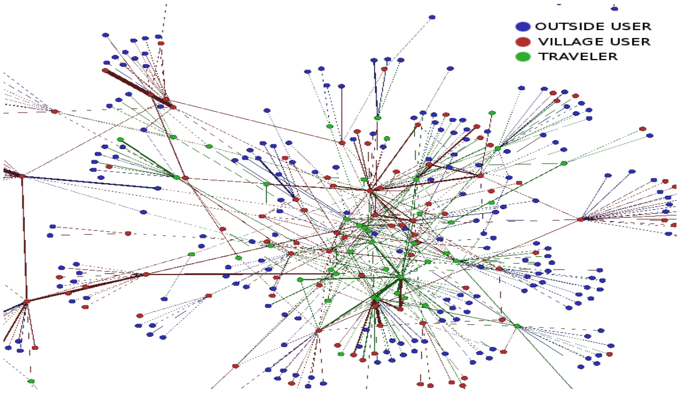


Fig. 7: **IM social graph**. Social graph showing users as nodes who either remain in the village, travel or who are always outside the village. The edges denote conversations and thicker edges mean more instant messages were sent between users.

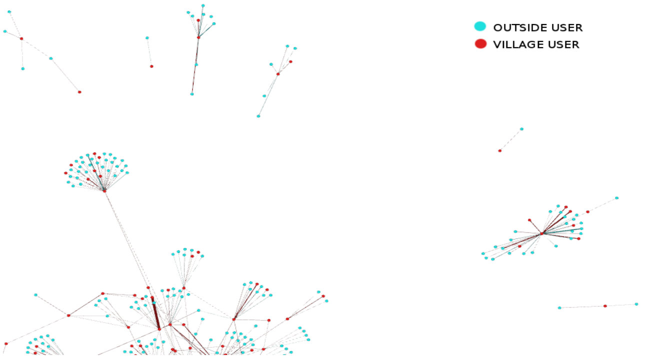


Fig. 8: **IM social graph**. A section of the social graph showing isolated communities of users.

little interaction with local people.

- The strongest bonds (highest number of conversations) are between local users, especially those that travel and key contact people.
- 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 have 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 travelers from Macha tend to use only one person as their contact in the village.
- There are a number of isolated communities of users who either have a small set of local users with no contact through IM to other local users or have a very weakly coupled link through one local user to the set of well connected local users.

5.3 Topic extraction

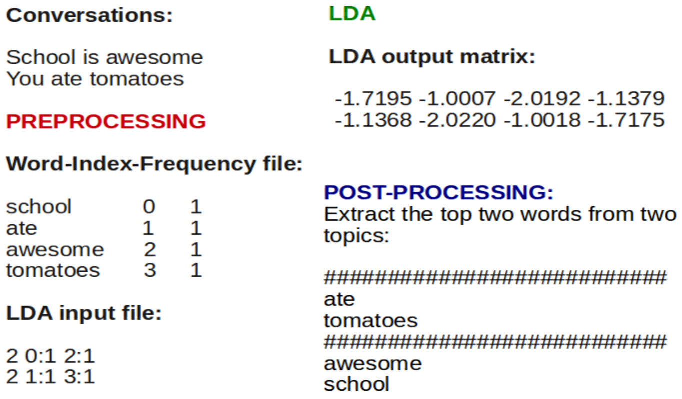
Part of our IM analysis aimed at extracting the common topics that are being discussed through Facebook IM. One question we wanted to answer in particular was whether there is a difference between most popular topics in the community and outside of it. For this purpose we were looking at the overlap between topics discussed by unique user pairs from the community and topics discussed with external people. Our analysis shows that there are not too many distinct topics discussed both internally and externally. There are topics that are common for both external and internal conversations, nevertheless there are specific themes discussed in the community that are not typical to external conversations.

The process of topic extraction consisted of three steps - preprocessing, topic extraction and post-processing. Given the conversations as extracted according to Section 3.3, we need to pre-process them to a format that is suitable for our topic extraction algorithm. For the actual topic extraction we use the Latent Dirichlet Allocation (LDA) algorithm. Finally we post-process the output from LDA to display the top words among the extracted topics. Following are details about each step.

The preprocessing engine is written in Python and consists of two steps. In the first step, we create a word index based on the set of extracted conversations. The program takes as an input a file with all the conversations - each one on a separate line, and a list of stop-words and returns as an output a word index of the format **[WORD WORD-INDEX WORD-FREQUENCY]**, which does not contain any stop-words. The second step of the preprocessing takes as an input the set of conversations, a stop-words list and the word index created in step 1 and returns an LDA input file in which each line corresponds to a separate conversation. The format of each line is **[COUNT-OF-UNIQUE-WORDS-IN-CONV [WORD-INDEX : WORD-FREQUENCY]]**.

For the actual topic extraction we used algorithm called Latent Dirichlet Allocation (LDA) [5]. LDA is a two step algorithm: it first takes a set of documents and a number of topics to be extracted and based on these learns a word/topic distribution model; then based on this model it can associate any given document to a topic by looking at the likelihood of a word to be generated by a given topic. In our project we utilize the first step of the LDA algorithm to learn the most famous topics discussed in our set of conversations. In particular, we use an implementation of LDA called GlibbsLDA++ [3]. It takes as an input an lda-input-file as generated in the preprocessing step, and a number of topics to be extracted; LDA returns as an output a model that indicates the likelihood of a word to be generated by a given topic. The output is formatted as a Markov matrix, where the number of rows is equal to the number of topics specified by the user and the number of columns is equal to the size of the dictionary generated by the input set of conversations. Each element of the output matrix is a log of probability indicating if a given word belongs to a topic.

In the post-processing step we extract a human-readable version of the top-k words from the extracted topics. The post-processing engine is also written in



Conversations:

School is awesome
You ate tomatoes

PREPROCESSING

Word-Index-Frequency file:

```
school     0   1
ate        1   1
awesome    2   1
tomatoes   3   1
```

LDA input file:

```
2 0:1 2:1
2 1:1 3:1
```

LDA

LDA output matrix:

```
-1.7195 -1.0007 -2.0192 -1.1379
-1.1368 -2.0220 -1.0018 -1.7175
```

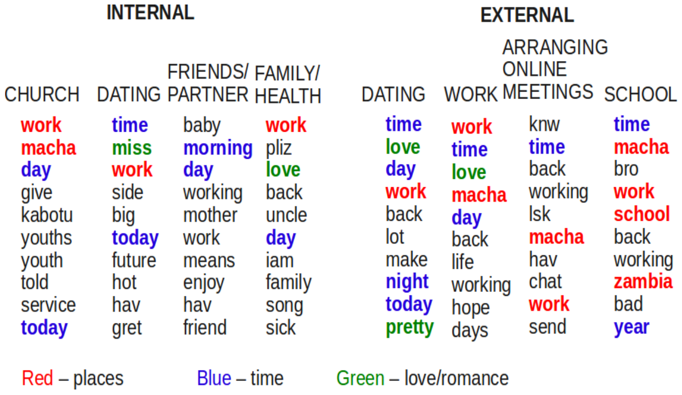
POST-PROCESSING:
Extract the top two words from two topics:

```
############################
ate
tomatoes
############################
awesome
school
```

Fig. 9: **Topics extraction example.**

Python; it takes as an input the matrix generated by LDA and the word index extracted in the preprocessing step. It then returns a list of the top-k words from each extracted topic. An example of a topic extraction from a set of conversations is given on Fig. 9.

Our results show that there are not that many disctinct topic discussed through IM. We varied the number of topics extracted by LDA from 3 to 20 and around 4 topics the output made the most sense. Topics like *scheduling* and *keep in touch* are common for both internal and external conversations. Nevertheless, there are some specific themes like *church*, *health* and *short-term scheduling* that are unique for the internal community. At the same time the external conversations often contain place-related words. Fig. 10 shows the top-10 words in the four extracted topics, as well as our interpretation of each of the topics.



INTERNAL				EXTERNAL			
CHURCH	DATING	FRIENDS/ PARTNER	FAMILY/ HEALTH	DATING	WORK	ARRANGING ONLINE MEETINGS	SCHOOL
work	**time**	baby	**work**	**time**	**work**	knw	**time**
macha	**miss**	**morning**	pliz	**love**	**time**	**time**	**macha**
day	**work**	**day**	**love**	**day**	**love**	back	bro
give	side	working	back	**work**	**macha**	working	**work**
kabotu	big	mother	uncle	back	**day**	lsk	**school**
youths	**today**	work	**day**	lot	back	**macha**	back
youth	future	means	iam	make	life	hav	working
told	hot	enjoy	family	**night**	working	chat	**zambia**
service	hav	hav	song	**today**	hope	**work**	bad
today	gret	friend	sick	**pretty**	days	send	**year**

Red – places Blue – time Green – love/romance

Fig. 10: **Topics extraction example.**

6 Lessons learned and Future work

One should be very careful when choosing the number of topics to be extracted by LDA. If too many topics are specified at LDA's input, this results in high redundancy in the extracted topics. At the same time if too few topics are specified, there are words from possibly different topics mixed together, which makes it impossible to extract the meaning of a topic.

Language issues are more complex in the Macha village environment due to the prevalance of local African languages as well as some European languages due to volunteers and visitors from Europe. Although 95% of the conversations we analysed were english, preprocessing in order to carry out language translation of non-english words would improve the accuracy of the topic extraction as well as help create a more generic solution for villages which may use more non-english words. There are many word contractions which appear to be local to Macha or Zambia and are not necessarily universally used IM contractions. If ongoing traffic analysis is done, a local contraction dictionary could be built up which displays unrecognized contractions to researchers who can use local expertise to interpret these contractions.

Combining the social graph structure which has underlying temporal information due to the time stamped data set and topic analysis can produce new insights such as monitoring topic dispersion through a community over time. This could be very useful in understanding how rumours spread or how political ideas spread through a community.

Protecting privacy is a key issue with this research and although anonymizing user names is an obvious first step, replacing all names of people in the text may be more complex than it appears; there are many user's in African villages who use abstract nouns as names such as prudence or fortune.

7 Conclusions

Facebook is extensively used for intra-village conversation. Although only 35% of the users were local to the village with 7% of these being travelers, 54% of the instant messages sent were between local users. Building a social graph to represent the instant message interactions revealed that there are key people in the community who act as linkages to the outside world and these people are often travelers. We also noticed that there are a number of isolated communities in the village where one or two village users communicate with a five to twenty outside users and have no connection through instant messages to any other local users. Topic analysis revealed a limited set of topcis which mostly centered around scheduling and relationships between friends, family and partners. There were, however, a few subtle differences between topics in local conversaions and conversation to outside users. Conversations to outside users tended to use more long-term time related words such as year and local coonversations had more emphasis on local community issues such as church and health.

References

[1] http://www.tcpdump.org/.
[2] http://www.circlemud.org/ jelson/software/tcpflow/.
[3] http://gibbslda.sourceforge.net/.

[4] P. Adam and C. Martell. Topic detection and extraction in chat. In *The IEEE International Conference of Semantic Computing*, September 2008.

[5] M. Jordan D. Blei, A. Ng. Latent Dirichlet Allocation. In *Journal of Machine Learning Research 3*, 2003.

[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 *Proceedings of the 20th international conference companion on World wide web*, Hyperabad, India, March 2011.

[8] J. Resig, S. Dawara, C. Homan, and A. Teredesai. Extracting social networks from instant messaging populations. In *LinkKDD*, Seattle, Washington, USA, August 2004.

[9] J. Resig and A. Teredesai. A framework for mining instant messaging services. In *SIAM*, Lake Buena Vista, Florida, USA, September 2004.

[10] L. Wang, Y. Jia, and W. Han. Instant message clustering based on extended vector space model. In *Proceedings of the 2nd international conference on Advances in computation and intelligence*, Wuhan, China, September 2007.