Reliable data collection in highly disconnected environments using mobile phones

Brian DeRenzi, Yaw Anokwa, Tapan Parikh, Gaetano Borriello
Computer Science and Engineering
University of Washington, Box 352350
Seattle, WA 98195
{bderenzi,yanokwa,tapan,gaetano}@cs.washington.edu

ABSTRACT

Over four and a half billion people live in the developing world and require access to services in the financial, agricultural, business, government and healthcare sectors. Due to constraints of the existing infrastructure (power, communications, etc), it is often difficult to deliver these services to remote areas in a timely and efficient manner.

The CAM framework has found success as a flexible platform for quickly developing and deploying high-impact applications for these environments. Many of the applications built with CAM have relied on a model where a field worker with a mobile phone regularly returns from a disconnected environment to one with connectivity. In this connected state, the phone and a centralized server can exchange information and get the collected data backed up on reliable media.

We propose extending CAM's networking model to enable continual operation in disconnected environments. Using a set of heterogeneous paths made available through social and geographic relationships naturally present among workers, we describe a system for asynchronously routing data in a best-effort manner.

1. INTRODUCTION

The infrastructure and economic constraints of the developing world have resulted in a field worker based model for delivering financial, agricultural, business, government and healthcare services. By using the existing and under-utilized workforce, companies, non-governmental organizations (NGOs) and governmental organizations use these field workers asynchronously as an efficient channel for services.

Beyond services, the reality of these environments is that field workers are a reliable way to transfer data. The relationships workers build with people they serve and work with are crucial to understanding the underlying problems which current services may not address [7]. These relationships and encounters can be used as a conduit for data transfer.

The first author of this paper was fortunate to witness this field worker model firsthand while collecting data on the living environment and health of the people of rural Tanzania. For two to three days, groups of volunteers and workers walked through expansive farming villages. Paper surveys about all occupants in a household, their living environment, and the various resources available were completed at each home.

After visiting many of the houses in the village, a health clinic was held. With the aid and approval of doctors, volunteers diagnosed patients and distributed medication to the sick. The data collection process was inefficient and error prone. Upon returning from the field, paper forms that had not been lost or damaged were entered into a computer system. Any damage or loss would require volunteers to resurvey - a waste of both time and resources. There was a clear need to move the data to a central and secured store as quickly as possible so useful information could be extracted.

The CAM framework [13] has found success in solving some of the problems identified above - namely safety and latency. With CAM, workers fill out paper forms which are digitized by the phone in a disconnected setting. Once a field worker returns to a connected environment, the digitized data is uploaded to a server. Since the data is stored in two formats (paper and phone), safety is not a pressing issue. A field worker can return to the village in a relatively short time period and reenter the data.

Additionally, latency is addressed by immediately digitizing the data and preparing it for the institutions who use it. CAM has been deployed with microfinance workers in India and has worked well due to good cellular coverage. Unfortunately, it cannot serve resource constrained environments like Tanzania where workers have to journey much further to reach connectivity. Furthermore, lost data is more expensive to recollect as it involves reaching remote locations and redoing lengthy interviews.

In this paper, we argue that there is value in enabling workers to stay in disconnected environments for extended periods of time. First, we discuss why previous work has fallen short, and then present a best-effort system which asynchronously transfers data using the heterogeneous social links naturally present among workers. We conclude by demonstrating a number of applications in healthcare which could be enabled by extending the capabilities of CAM.

2. RELATED WORK

Although there are different areas where this work is applicable (e.g., agriculture, microfinance and governmental services), we focus on healthcare for the scope of this paper due to our familiarity with the domain. We look at related work which has approached the problem with different assumptions about connectivity.

Epihandy[1] is a survey tool for mobile devices. It addresses errors related to manual data entry and lack of validation by putting the

[1]Epihandy http://www.epihandy.com/

NSDR'07, August 27, 2007, Kyoto, Japan.

data collection in an electronic form. TRACnet [3] is an HIV/AIDS project in Rwanda developed in conjunction with Voxiva[2]. Patient level data is entered into the system through the Internet, via voice, or using forms on mobile devices. Answers to questions in Epihandy and TRACnet are sent if there is connectivity or are cached until the device has a connection to the Internet and can synchronize with a server.

Systems like Epihandy and TRACnet would fail in the mountainous farming villages in Bumbuli, Tanzania where cellular or Internet service is rare. The high mountains combined with the sprawling farming villages make it difficult to provide consistent connectivity.

Our approach to this problem utilizes different paths to connectivity in the event where the phone does not return to a connected environment. Work in delay tolerant networks (DTNs) [4] and data muling [15], demonstrate the viability of such an approach in a coarse-grained manner.

Other approaches have centered around the kiosk model, where data is routed between an Internet gateway in town and a kiosk in the disconnected village. DakNet makes use of mobile access points on buses, motorcycles, bikes and even ox-carts. Tetherless computing [14] builds on DakNet by defining a solution which includes naming, addressing, forwarding, routing, identification, application support and security. These techniques do not offer fine-grained (pedestrian vs. vehicular) data transfer. Additionally, because field workers are mobile, relying on fixed locations like kiosks is inadequate.

ZebraNet [5] outfits zebras in Kenya with GPS and other sensors. As zebras come in contact with each other, sensors exchange all data. Thus, each zebra eventually carries the entire herd's information. Any zebra near a base station automatically uploads data to a server. The flooding protocol ZebraNet uses would fail with the amount of data workers often generate. Additionally, determining the zebra who is the best neighbor to exchange data with occurs only when a research vehicle is present. Without the mobile base station (a vehicle), the data is flooded amongst all zebras with no upload path.

Again, our goals are to enable a system which uses a set of heterogeneous paths made available through social relationships naturally present. The system should intelligently and asynchronously route data in a best-effort and fine-grained manner over mobile phones.

3. SYSTEM COMPONENTS

Our proposed system considers a number of questions that must be answered before deploying our solution. How is data routed properly? How do we determine likely candidates for transfer of data? How are acknowledgments transmitted to the original sender so that data which has reached a safe store can be removed from the original device? Finally, what are the user considerations regarding privacy and incentives?

3.1 Data Routing

We leverage work done in geographic routing [2] and location-aided routing on mobile devices [8]. We propose routing data using a best-effort "sneaker net" which sends data through the person most likely to be in a connected location at some point in the future.

Forwarding messages over social links is not novel [11], but our notion of social routing facilitates timely transfer of information to a safe store. This in turn allows field workers to work in disconnected environments for extended periods of time while still maintaining the ability to transmit information. Also, since the cost of collecting data is high, our routing also ensures data is retained on more than one device as serendipitous backups.

[2]Voxiva, Inc. http://www.voxiva.net/

3.2 Location Profile

As a field worker's phone moves through the world, it creates a *location profile* describing where it has been. Using GPS traces, Wi-Fi access points and cell towers [9, 10], useful places that the mobile device travels to [6] and is likely to travel next [1] are extracted.

When a field worker meets another worker, their phones automatically exchange location profiles and analyze the probability of their returning to connected environments. Bluetooth, Wi-Fi and (soon) UWB offer different bandwidth, power, and range tradeoffs for phones to exchange data. Data that needs to return to a connected environment is transferred to the phone most likely to return soonest.

In Figure 1, we illustrate how data moves from a disconnected field worker A, to a server I connected to the Internet. A comes in contact with B and C. C comes in contact with F and G. G finds H, who in addition to finding J, also finds the server I. I now has all the information from A. After this occurs B eventually finds D and then E, but the data is ignored as it is already in the system.

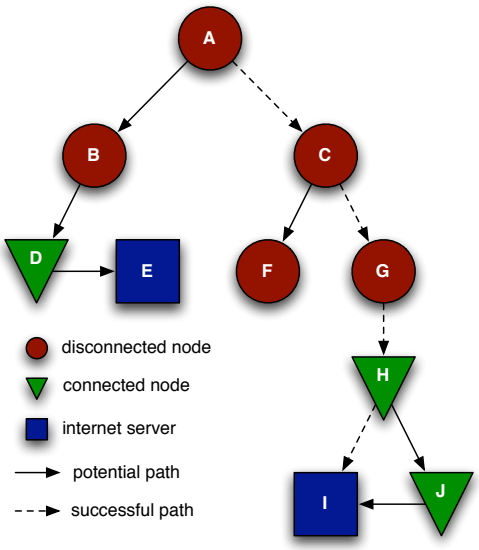

Figure 1: Data moving from a disconnected field worker to a server.

3.3 Data Prioritization

When the worker returns to a connected environment, the data is synchronized with a centralized server using USSD, SMS, GPRS, or Wi-Fi. Depending on the bandwidth available, we may choose to upload only first tier data (limited bandwidth, send only surveys and core data), or both first and second tier data (high bandwidth, send surveys, images and all other meta-data).

This prioritization can be application specific or determined heuristically. For example, we may want to send all pictures that are taken during the medical examination portion of a form, but not worry about pictures taken during an inspection of the home. Perhaps it would be best to send smallest data files first to increase the chances of completing file transfers. Determining what to send and when is

an open question.

3.4 Receipt Acknowledgments

Recollecting data in this domain is an expensive procedure as it involves returning to remote locations and conducting lengthy interviews. The cost of this process means workers must be made aware of all successful data transfers.

When data has been successfully received at a central location, an acknowledgment is generated and sent to the source field worker. Any phone in a connected state can receive acknowledgments from the central server. These acknowledgments are delivered to the source field worker along any path in the system. If along the chosen path the acknowledgment encounters a local copy of the data, that copy will be deleted.

Once the source node receives the acknowledgment, the related data is deleted. The source node must then pass the acknowledgment to all nodes the data was shared with. This ensures that first-level data mules which are one hop away from the source are able to reclaim storage space. The following section discusses the reclamation strategy in greater detail.

Figure 2 shows the acknowledgments (or other information) returning to A. Because E and I are on the web, they can communicate. In this example, E returns information to A through an alternative path. Additionally, although A and B have changed locations, the acknowledgments are still returned to A as routing is done in a location agnostic manner.

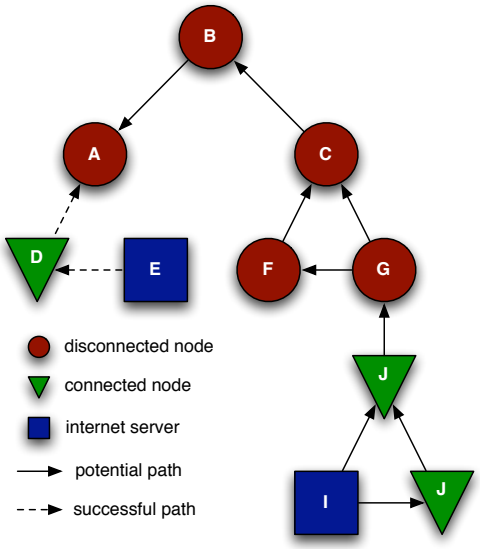

Figure 2: Data moving from an Internet server to a disconnected field worker.

3.5 Storage Reclamation

As data is accumulated, it will be necessary to employ a reclamation strategy so workers do not run out of space. A simple strategy would be to remove data which is farthest from the source. Each time data is transmitted from one device to the next, a counter is incremented. Assuming that data spreads in a tree structure, as the counter increases, the number of nodes with that data increases exponentially. As a result, the probability of removing that data goes up, allowing highly replicated data to be removed from devices first.

All devices one hop away from the source are to keep the data until the source explicitly requests a deletion after an acknowledgment is received as described in the previous section. In the event that data on a device one hop away needs to be deleted, the device will explicitly notify the source that the data will be deleted. This allows the source device to ensure future propagation and increase the chances of the data making it to the final destination.

Devices that are more than one hop away will combine the deletion probability (as defined by their distance from the source) with a most recently transfered or total number of times transfered heuristic to ensure that 'healthy' data which is being propagated through the network is deleted first.

3.6 Usage Incentives

As with any networking technology, it is important to provide incentives and answer the question of why users should adopt our new approach. For field workers, there is a mutual advantage for sharing data. Workers do not have to risk losing data and are able to ensure employers get data in a timely manner.

Other users of the system (e.g., truck drivers delivering supplies to the remote locations) would potentially enter into a contract with the organizations whose data they are muling. The contract would provide compensation based on the amount of data muled by that individual. It is not difficult to imagine schemes which would benefit all parties involved. If users do not have to actively manage the data transfer, minimal compensation (provided by the organizations) may be acceptable.

Finally, for companies, NGOs and governmental organizations, the solution we are proposing is cost effective as there is no expensive infrastructure. Only field workers and drivers for a given application will need to be outfitted with low-cost mobile phones while the benefits could potentially be significant. Today these devices can easily be purchased for $200. While we do expect this price to drop as time goes on, it is not an unreasonable cost for an organization. With a handful of field workers serving hundreds of patients, the one time purchase cost of a mobile phone would be amortized over the amount of data collected.

3.7 User Privacy

Privacy and security are major issues that must be addressed. In health applications, it is important that patient data remain secure when transferring between devices. Using a standard public key infrastructure would be a good solution. Each device can encrypt data with a public key of the centralized server. When the data is received, it can be decrypted using the private key.

The privacy concern of sharing location profiles raises questions about tracking people. To introduce a sufficient amount of plausible deniability, we imagine introducing variable granularity of location. For example, locations could be reported on village level instead of precise locations of connectivity.

Depending on the application, it may be that field workers wish to be tracked so they can be managed more efficiently by their organizations. Being able to tell where field workers are can reduce overlap and duplication of data collection. This especially important for transport workers muling data.

4. ENABLING NEW APPLICATIONS

Healthcare efforts rely on information about diseases patients face and the resources available. Moving this data around quickly and efficiently is important in maintaining a high level of care in resource scarce environments.

At the same time, it is extremely important that data is not lost on the way. Health data is difficult to collect because it involves

surveying people directly. If collected on a mobile device, there may only be one record so it becomes imperative that the data be protected.

The system components discussed previously enable two kinds of health care applications discussed below.

4.1 Satellite Hospitals

Colleagues working with Partners In Health[3] in Rwanda have identified consistent communication as a recurring challenge. With one hospital serving several remote satellite clinics (a model used frequently [12]), reliable communication over heterogeneous paths is a priority.

It is often the case that drug stock needs to be checked between clinics or that hospitals need to be alerted of patients in grave danger who are being transferred to another hospital for advanced care. Unreliable internet and phone connections mean that communication can take any number of routes. A system to deliver messages quickly and efficiently would increase the standard of care and could be enabled by our system.

4.2 Surveying Remote Locations

Much of healthcare and public health is informed by data collection. It is necessary to understand where diseases are and how they spread in order to respond to them with preventative measures. Logistics for resource management are informed by patient-level statistics of common disease and current equipment stock at the respective clinic. The more information that an organization has, the better.

In the previous example of Tanzania, each survey result produced a few KBs of data. To augment the answers with meta-data (e.g., a few pictures taken with the mobile phone), would mean that number could easily jump to 1 MB per survey. With several field workers visiting homes, this could easily be 100 MB per day. In just over a week and a half, 1 GB of data may be easily collected. This meta-data is useful because questions, such as those about the cleanliness of the area, building materials of the house, or state of a pit latrine are more easily and fully described with an image. Being able to capture and transmit this amount of data in an extremely disconnected environment would be possible with our proposed system.

5. NEXT STEPS

The next steps of this research will occur in the summer when we return to Tanzania to collect more data to inform our design. There are many open questions that still remain. We want to observe examples of actual data collection in order to determine the right parameters for our design.

Having a better sense of the connectivity which is available in the region will help inform our decisions about routing. How often do we run into other field workers? For a given location, how many hops would it practically take to reach a connected environment? How well do we manage the tradeoff between keeping copies of the data safe with the storage constraints? What issues are introduced by using explicit acknowledgments?

In order to explore the balance between geographic and social routing, it will be necessary to determine how often workers travel in the field. Do they generally work in one region? Do they maintain the same social network when traveling? To some degree this will be determined by the organization and application. By collecting this data, we will be able to expand the networking model in the CAM system to work in these highly disconnected environments and enable a new set of high impact applications.

[3]Partners In Health http://www.pih.org

6. CONCLUSION

We have proposed an extension to CAM's networking model that will use a set of heterogeneous paths available through the natural social network among field workers to perform best-effort routing in the disconnected environments like those encountered in rural Africa.

Previous research in data collection has omitted transmitting from completely disconnected environments in a timely manner. Networking projects in this area have focused on the kiosk model where data in the disconnected environment is transmitted and received from a fixed location.

To further inform our design choices, we have listed a concrete set of steps to take during the summer of 2007 when the first author will return to Tanzania to begin assessing the importance of various design factors.

Expanding CAM will enable a new set of applications that require safe storage and timely data collection from disconnected environments, furthering our goal of creating a flexible data collection device for use in resource constrained environments.

7. REFERENCES

[1] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with gps. In *ISWC '02: Proceedings of the 6th IEEE International Symposium on Wearable Computers*, page 101, Washington, DC, USA, 2002. IEEE Computer Society.

[2] Tomasz Imieliński and Julio C. Navas. Gps-based geographic addressing, routing, and resource discovery. *Commun. ACM*, 42(4):86–92, 1999.

[3] Donner J. Innovative approaches to public health information systems in developing countries: An example from rwanda. *Presented at "Mobile Technology and Health: Benefits and Risks"*, 2004.

[4] Sushant Jain, Kevin Fall, and Rabin Patra. Routing in a delay tolerant network. *SIGCOMM Comput. Commun. Rev.*, 34(4):145–158, 2004.

[5] Philo Juang, Hidekazu Oki, Yong Wang, Margaret Martonosi, Li Shiuan Peh, and Daniel Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet. *SIGPLAN Not.*, 37(10):96–107, 2002.

[6] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting places from traces of locations. *SIGMOBILE Mob. Comput. Commun. Rev.*, 9(3):58–68, 2005.

[7] Tracy Kidder. *Mountains Beyond Mountains: The Quest of Dr. Paul Farmer, a Man Who Would Cure the World*. Random House, 2003.

[8] Young-Bae Ko and Nitin H. Vaidya. Location-aided routing (lar) in mobile ad hoc networks. *Wirel. Netw.*, 6(4):307–321, 2000.

[9] K. Laasonen, M. Raento, and H. Toivonen. Adaptive on-device location recognition. In *Proceedings of PERVASIVE 2004, Second International Conference on Pervasive Computing*, Vienna, Austria, 2004.

[10] Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, Fred Potter, Jason Tabert, Pauline Powledge, Gaetano Borriello, and Bill Schilit. Place lab: Device positioning using radio beacons in the wild. In *Proceedings of the Third International Conference on Pervasive Computing*, May 2005.

[11] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. From the Cover: Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

[12] A. Martinez, V. Villarroel, J. Seonane, and F. del Pozo. Analysis of information and communication needs in rural primary health care in developing countries. *IEEE Transactions on Information Technology in Biomedicine*, 9(1), 2005.

[13] Tapan S. Parikh and Edward D. Lazowska. Designing an architecture for delivering mobile information services to the rural developing world. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 791–800, New York, NY, USA, 2006. ACM Press.

[14] A. Seth, D. Kroeker, M. Zaharia, S. Guo, and S. Keshav. Low-cost communication for rural internet kiosks using mechanical backhaul. In *MobiCom '06: Proceedings of the 12th annual international conference on Mobile computing and networking*, pages 334–345, New York, NY, USA, 2006. ACM Press.

[15] Rahul C. Shah, Sumit Roy, Sushant Jain, and Waylon Brunette. Data mules: modeling and analysis of a three-tier architecture for sparse sensor networks. *Ad Hoc Networks*, 1(2-3):215–233, 2003.