12-2005

CRAWDAD: a Community Resource for Archiving Wireless Data at Dartmouth

David Kotz
Dartmouth College, David.F.Kotz@Dartmouth.EDU

Tristan Henderson
Dartmouth College

Editor: Mary Baker ■ HP Labs ■ mary.baker@hp.com

CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth

David Kotz and Tristan Henderson

At MobiCom 2005, in Cologne, Germany, we held a workshop to launch the new *Community Resource for Archiving Wireless Data at Dartmouth*. In the evening of the last day of the main conference, about 30 people gathered to learn more about CRAWDAD and share their thoughts on its direction. But before we describe the workshop, let's look at the CRAWDAD project's genesis.

THE NEED FOR DATA

Researchers who work with wireless networks or mobile computing are seriously starved for data. Data captured from live wireless networks would help us all understand how real users, applications, and devices use real networks under real conditions, and how mobile users actually move about. This data helps us to identify and understand the real problems, to evaluate possible solutions, and to evaluate new applications and services.

On the other hand, most research today is based on analytical or simulation models. These models are severely limited by the complexity of real-world radio propagation and the lack of understanding about behavior of wireless applications and users. Experimental studies, however, are extremely difficult to set up. To collect data about real users on real networks requires a considerable amount of equipment, specialized software for collecting and

anonymizing data, organizational permission and assistance to collect data, and human-subjects research clearance from the appropriate institutional review board (IRB).

At Dartmouth College we are fortunate. We have a campuswide wireless infrastructure, with comprehensive data-collection mechanisms to gather traces of wireless users and their behavior. We have developed an extensive toolset for collecting, anonymizing, and analyzing the trace data. We have a cooperative network-management organization, and experience with the IRB process. We have a history of sharing our (anonymized) data with the research community. Several other research groups from around the world, in both academia and industry, have used our data. In our experience, the need for this sort of data is great.

To meet this need, the US National Science Foundation is funding an effort to turn this Dartmouth resource into a true community resource: an archive with the capacity to store wireless trace data from many contributing locations, with the staff to develop better tools to handle the data. The resulting CRAWDAD project will work with

- community leaders to ensure that the archive meets the research community's needs,
- the other leading centers that develop network tracing tools and metadata, and
- research organizations and corporations to ensure continuing support for the archive after NSF funding ends.

THE WORKSHOP

The CRAWDAD workshop consisted of an invited talk by Ravi Jain, of DoCoMo Labs USA, and then group discussion.

The importance of measurement

Jain gave an inspiring and educational talk about the importance of measurement in our field. He sees the mobility and networking research communities beginning to mature, as evidenced by the increased interplay between theoretical and experimental

Published by the IEEE CS and IEEE ComSoc ■ 1536-1268/05/$20.00 © 2005 IEEE

research. In his view, the two "dance" in a supportive cycle: experimental data allows analysis and modeling, which enhances and enables new theoretical research, which in turn generates new requirements for data, which inspires new research. He noted that many early wireless-network data-collection studies, beginning with Diane Tang and Mary Baker in 2000,[1] brought realism to the study of wireless-network traffic and user mobility. On the theoretical side, the MANET (mobile ad hoc network) community has long used arbitrary mobility models, such as random waypoint, that have no basis in reality but are theoretically tractable. Now, these communities are beginning to meet in the middle, attempting to build realistic—but usable—mobility models based on real mobility traces.

Jain made an interesting analogy to the human-genome project. The genomics research world is exploding because the availability of a detailed, common data set keeps the entry barrier low for a wide variety of research. In contrast, many aspects of wireless-network research have been very difficult because it is hard or impossible for most researchers to obtain realistic wireless data. Wireless-network providers generally do not want to release their data, and measuring a network yourself takes a tremendous amount of time and resources. So, he encouraged the community to support and engage with the CRAWDAD project.

He noted many challenges for the community. User privacy, he emphasized, is crucial. Everyday network users are unwittingly caught in traces of live networks, typically without consent or even being informed. Most careful research groups do obtain the necessary human-subjects research approval and take great care to anonymize the data and respect users' privacy. However, Jain predicted that the data-collection community will need to find more effective ways to obtain informed consent. He also indicated that data providers, particularly in the commercial world, will insist on protection. Could a provider's competitors gain an advantage by mining a data set released for research purposes? Could hackers learn enough to improve their capability to attack the provider's network? Protecting a data set against these forms of misuse might be technically difficult.

Tackling the important questions

After a break for refreshments, the assembled group got down to business. We asked for discussion on these important topics:

- What sort of data does the community need?
- What metadata do we need?
- What tools do we need?

By measuring wireless-network users, we are potentially invading their privacy, and protecting their privacy is paramount, and indeed a legal requirement.

- How do we protect human subjects?
- What sort of educational uses does this data have?

Regarding data, many participants were interested in user mobility and thus requested that CRAWDAD contain user- or device-mobility traces as well as traffic traces. Some were interested in MANETs, including data from both controlled experiments and test beds. One participant was interested in active measurements, such as an interference map generated from a campus Wi-Fi network by using probes to learn about the interference between access points. Some people focused on security and wanted to find data that exhibited network attacks. Many were interested in data from large network providers, both cellular networks and wireless ISPs, although those in the room who were close to that industry were not optimistic that such traces were obtainable. A few were interested in data from mesh and community wireless networks. In fact, some suggested that mesh network operators would be very interested in collecting data. Many of these networks are just getting off the ground, and the operators might find performance data useful in network design and deployment.

The question regarding metadata provoked a long discussion. What information must be supplied along with a network trace to understand that trace's context? It depends on the use of the data, to be sure, but might include

- the network topology and geography;
- the number and type of users and the character of the user population;
- configuration information about network components including brand, model, and firmware versions; and
- information about the data-collection methodology and glitches such as power failures.

The discussion on tools focused particularly on those for sanitizing data in ways that protect users' privacy and (where necessary) data providers' anonymity. Many participants requested tools and documentation about how to collect wireless-network and mobility data. Others wanted visualization tools to help examine the data, analysis tools to extract information from the data, or educational tools that could allow use of the data in a classroom.

The discussion on protecting human subjects emphasized the need to obtain their approval for conducting wireless-network-measurement studies. By measuring wireless-network users, we are potentially invading their privacy, and protecting their privacy is paramount, and indeed a legal requirement. Because the workshop took place in Europe, some discussion concentrated on European privacy law such as the European

CONFERENCES

Union Data Protection Directive. European privacy laws are much stricter than their American counterparts, and data collectors should be aware of the relevant regulations.

Unfortunately we reached the final discussion topic—educational uses of data—late in the evening, so the discussion was short. We would be interested in hearing from any readers who are or will be using wireless data in their classes.

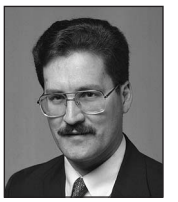
The workshop was a great success and resulted in many action items for us. We have begun building a Web site and encouraging wireless-network operators and researchers to contribute their data to our archive. Next, we will start compiling how-to documents on various topics such as collecting data, sanitizing data, and writing an IRB proposal, so that other researchers can easily conduct measurement studies. We also plan to set up working groups to deal with specific issues—for instance, particular types of data such as MANET and VANET (vehicular ad hoc network) data—or topics such as education. Please contact us if you are interested in helping.

If you are interested in learning more about CRAWDAD, please visit the Web site. You can access our data collection, view relevant published papers, and subscribe to a mailing list. We also welcome suggestions and volunteers to help collect and organize data.

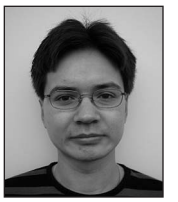
FOR MORE INFORMATION

See the CRAWDAD Web site, http://crawdad.cs.dartmouth.edu.

David Kotz is a professor of computer science at Dartmouth College. Contact him at dfk@cs.dartmouth.edu.



Tristan Henderson is a research assistant professor of computer science at Dartmouth College. Contact him at tristan@cs.dartmouth.edu.



REFERENCE

1. D. Tang and M. Baker, "Analysis of a Local-Area Wireless Network," *Proc. 6th Int'l Conf. Mobile Computing and Networking* (Mobicom 00), ACM Press, 2000, pp. 1–10.