ERCIM Alain Bensoussan Fellowship Programme

ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

Scientific Report

First name / Family name	Brenton Walker
Nationality	USA
Name of the *Host Organisation*	SICS Swedish ICT
First Name / family name of the *Scientific Coordinator*	Anders Lindgren
Period of the fellowship	22/09/2014 to 22/09/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Brenton Walker contributed to several projects in several ways, such as software and system development for the MOSES and Efficient IoT projects, cellular network trace analysis, experimentation on distributed network-coded content caching in Information-Centric Networks (ICNs), and handover-based cellular network reconstruction inspired by topological coverage testing problems.

MOSES and Efficient IoT

Brenton set up a development environment for Intel Edison devices, and built a base system image for the Edison containing the software needed for running and monitoring DTN stacks on these devices. He also cross-compiled the IBR-DTN stack for the Edison, and developed a node.js (javascript webserver) application that displays and manipulates the contents of the DTN stack's data store. He also participated in the project meeting discussions and filled in as a presenter at an EIT project review.

Analysis of Cellular Network Traces

Analysis of cellular traces, and simulations driven by these traces were the bulk of Brenton's work this year. The traces cover a country-scale cellular network over several months and are many terabytes in size, and contain billions of events from millions of users, which makes any analysis challenging. Some of the general statistics we investigated include:

- The relationship between user-cell co-location and inter-cell distance
- The distribution of upload/download speeds throughout the network
- The correlation between total throughput and peak upload/download speeds
- The distribution of total up/down data throughput across all cells
- User mobility profiling and profile-matching
- Common sequences of cell associations across all users

One in-depth project was to reconstruct the network proximity structure based on inferring handovers between cells. This required identifying and eliminating various statistical artifacts in the trace data, and employing dimension-reduction and graph embedding techniques. This work is partially described in our HotPlanet paper.

Another major undertaking was indexing the traces in a database and developing visualization tools. We experimented with different free database tools that could work within the limits of our data access agreement, and set up a database system that allows us O(1)-time access to activity data based on user ID and cell ID. This allowed us to build visualization tools for aggregate traffic load throughout the network and other network experiment results.

A final analysis challenge was computing inter-contact time distributions. This is the probability distribution of how long any pair of users goes between subsequent meetings. In our case we consider users to have “met” when they associate to the same cell at the same time. Computing it only requires keeping a runnling list of the last meeting time of each pair of users, but it is a challenge because even with 1,000,000 users, there are 499,999,500,000 possible pairs of users, and the processing cannot be fully parallelized. We implemented a hybrid RAM/databse-based system that has computed the user inter-encounter time distribution for 1,000,000 users over several months, a scale we believe has not been achieved before.

Finally Brenton published some of his thesis work, and extended on its theme by performing large-scale trace-driven opportunistic network coded data dissemination experiments. These are simulations using real user mobility and association data derived from the cellular traces. We experimented with different network coded dissemination strategies, and different cache-seeding strategies in a geographically distributed opportunistic ICN. This led to the paper that is currently in submission to the CCDWN workshop at CoNEXT.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

(ACCEPTED) *Deriving Cellular Network Structure From Inferred Handovers in a Cellular Association Trace*; Brenton Walker; Anders Lindgren; HotPlanet 2015 (MobiCom workshop)

> A cellular association trace consists of timestamped events recording user activity in labeled cells in a cellular network. From such data one can infer that if a user appears in two different cells within a short span of time, that a handover took place, and that the coverage areas of the two cells overlap. That is, one can infer geographic information from handover behavior. One would like to expand this kind of inference to a larger scale, perhaps reconstructing a proximity graph of the cellular sites, or creating an approximate 2-dimensional embedding of the cells. We have analyzed

a large-scale cellular association trace of several months of activity for several million users on a 3G network, and have found that handover behavior is actually incredibly diverse and complicated, making it very difficult to make any sort of global inferences, even in small sections of a network. In this paper we present some stable elements of handover behavior, and present several methods one can use to extract proximity information from such a trace.

(ACCEPTED) *Long-Term Country-Scale Opportunistic Network Coded Data Dissemination*; Brenton Walker, Anders Lindgren; (submitted to CCDWN workshop at CoNEXT 2015)

> We conduct large-scale cellular trace-driven experiments comparing different opportunistic network coded data dissemination strategies and different cache seeding strategies for distributing a large data object across a country-scale network of thousands of local repositories. We compare fragmentation, source-only erasure coding, cache coding, network coding, and propose two new dissemination strategies motivated by performance issues. We also experiment with several strategies for pre-seeding information to the local repositories, and examine the time/work trade-offs involved.

(IN PREPARATION) *Comparison of Mobile Traffic Features Across Cities Regions and Countries*; Yuan Quiao, Jane Yang, Brenton Walker, Anders Lindgren; (to be submitted to Elsevier Journal of Computer Communications Special Issue on Mobile Traffic Analytics).

Papers that were based mainly on thesis work I did while at the University of Maryland:
(ACCEPTED TALK ABSTRACT) *The Topological Structure of Geographically Distributed Network Coded Data*; Brenton Walker; TOPONETS 2015

> We generalize work using topology to study the coverage of wireless and sensor networks to that of covering a geographic area with network coded information in an opportunistic data distribution network. While the problem of sensor network coverage has been essentially a geometric and statistical problem, the network coded data coverage problem has a multi-dimensional algebraic aspect that leads to some surprising structure and coverage-testing counterexamples.

(ACCEPTED) *Computing Network Coded Data Coverage in an Opportunistic Information-Centric Network*; Brenton Walker; Swedish National Computer Networking Workshop (SNCNW 2015)

> We consider an opportunistic network in which mobile users and stationary data repositories distribute information directly between each other when they come in range. In this setting, using network/erasure coding on large data objects can greatly improve the performance and robustness of the network, but it becomes more difficult to plan, coordinate, and analyze the distribution of information. We introduce a simplicial data structure, the coverage complex, that captures enough of both the structure of the code and the geometry of the network that it can be used for drawing conclusions about network coded data coverage. The coverage complex can be built using only local proximity and data inventory information, and can be computed by a distributed algorithm.

(PENDING) *Computing Network Coded Data Coverage in an Opportunistic Data Dissemination Network*; Brenton Walker; (submitted to INFOCOM 2016)

> We consider an opportunistic wireless network where data repositories provide mobile users access to locally-cached data objects. In this setting using network/erasure coding to disseminate large data objects can greatly improve the performance and robustness of the network, but it becomes more difficult to plan, coordinate, and analyze the distribution of information. We introduce a simplicial data structure, the coverage complex, that captures enough of both the structure of the code and the geometry of the network that it can be used to draw conclusions about network coded data coverage. We give a distributed algorithm for computing the coverage complex based on local information, prove results on using it for coverage testing, and study more complicated cases where coverage testing can fail.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- Swedish National Computer Networking Workshop (SNCNW 2015); 28-29 May 2015; Karlstad Sweden
- TOPONETS 2015; 2 June 2015; Zaragoza Spain
- MobiCom 2015 and HotPlanet & CHANTS workshops; 7-11 September 2015; Paris France

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Host Institute: Consiglio Nazionale delle Richerche – Instituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR/IEIIT)
Turin Italy

Local Scientific Coordinator: Marco Fiore

Project Summary:
Cellular activity traces are now being used by many researches to study cell phone usage and human activity and mobility patterns. However the means by which such activity traces are collected, anonymized, filtered, and aggregated means that the traces provide widely varying levels of detail and contain statistical artifacts introduced by the recording process.

We have access to two traces that record cellular activity over the same 16-day time period in the same European city, that were recorded by probes with different levels of access to the network. One type, the GGSN probe, records data a high level in the network. It is stable and aggregates data over a wide branch of the network, but does not update users' cell associations unless they perform a location update or detach/attach to the network (rarely). The other, the RNC probe, provides accurate data at the base station level, but many more probes are required to cover the network, and they are not yet in production phase and require frequent reboots, losing data in the process.

During the REP we compared the statistics gathered by the two types of probes, to see just how inaccurate the data gathered by the high-level GGSN probes is. If the data is accurate, it means that good cellular activity data can be gathered much more broadly, reliably, and cheaply. The details of this study will be of interest to other researchers who often use GGSN-type probe data in their analysis.

We analyzed the distribution of absolute and load-scaled errors for the two main statistics, and looked for time-varying spatial patterns in the errors. Unfortunately we found that the user-association and data traffic statistics reported by the GGSN probe have significant and random errors relative to the RNC probes. We are preparing a short paper which will quantify the inaccuracy.