Data Warehousing for Social Network and Human Mobility Research and Detecting Real-World Events in Space and Time Using Feature Clustering

Alec Pawling

Department of Computer Science and Engineering
University of Notre Dame

December 7, 2009

Overview

Overview: Dynamic Data Driven Application Systems[1]

Utilize simulations and models that can

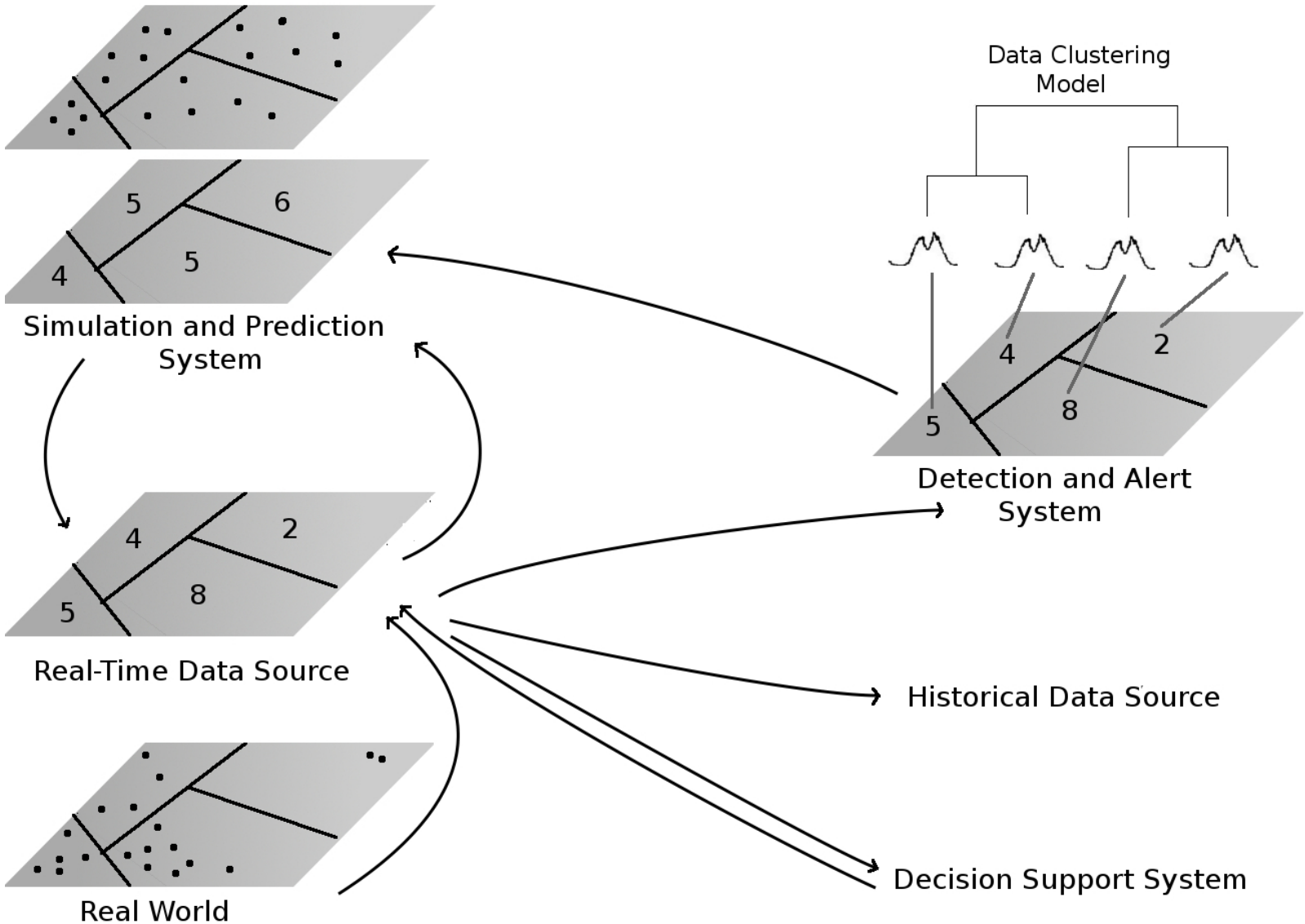
- Incorporate newly available data online, and
- Steer the data collection mechanism to obtain the most useful data

Areas of application:

- Population movement forecasting (Schoenharl and Madey 2008)
- Forecasting the spread of fire: wildfire (Mendel *et al*, 2007), in buildings (Chaturvedi, *et al.*, 2005)
- Weather (Ramakrishnan *et al.*, 2007), hurricane (Allen, 2007) forecasting
- and many others.

[1]Portions of this work were funded by the National Science Foundation: DDDAS Program, Grant CNS-0540348.

Overview: Wireless Phone Based Emergency Response System



Overview: Road Map

In this presentation, we will cover two aspects of the WIPER system:

- The Historical Data Source
  - Data partitioning aspect of the warehouse design.
  - Issues arising from the data merge.
- Detection and Alert System
  - Feature clustering approach for event detection.

A Data Warehouse for Social Network and Human Mobility Research

Overview: Data Warehousing

Data warehouses store long term historical data organized in such a way that it can be effectively used for analysis (traditionally decision support in a business environment).

- Data residing in a warehouse are often from different sources (departments).
- The warehouse is non-volatile: once loaded, the data are not updated.

Overview: The Phone Data

Two major components:

- Usage data
  - Call data: billing records of service usage initiated by customers of the company.
  - Interconnect data: billing records of service usage initiated by users that are not customers of the company and received by customers of the company.
  - Call Record Data (CDR): usage records with tower information. Four types of records:
    - MOC: record for a voice call from the originating tower
    - MTC: record for a voice call from the terminating tower
    - SOM: record for an SMS from the originating tower
    - STM: record for an SMS from the terminating tower
- Customer data
  - Defines the set of "in-company" users.
  - Users that are not in the customer data are "out-of-company".

Motivation

Data used for:

- Development of the Simulation and Prediction and Detection and Alert Systems of WIPER.
- Social network and human mobility research.

Problems:

- Large (5 TB) set of flat files.
- Inefficient de-identification mechanism (phone numbers replaced by strings)
- Data preparation is time consuming, error prone, and often redundant.

Goal: Develop a repository that allows efficient extraction of relevant subsets of the data.

Related Work:

Gray *et al.*, 2005. Advocate for the use of database management systems instead of file formats traditionally used in scientific research.

Examples of scientific databases:

- Bioinformatics: INTERACT (Eilbeck, *et al*, 1999)
- Astronomy: Sloan Digital Sky Survey (O'Mullane, *et al.*, 2005)
- High Energy Physics: BaBar (CERN) (Becla and Wang, 2005)

Warehouse Design

Design considerations: (Inmon, 2005)

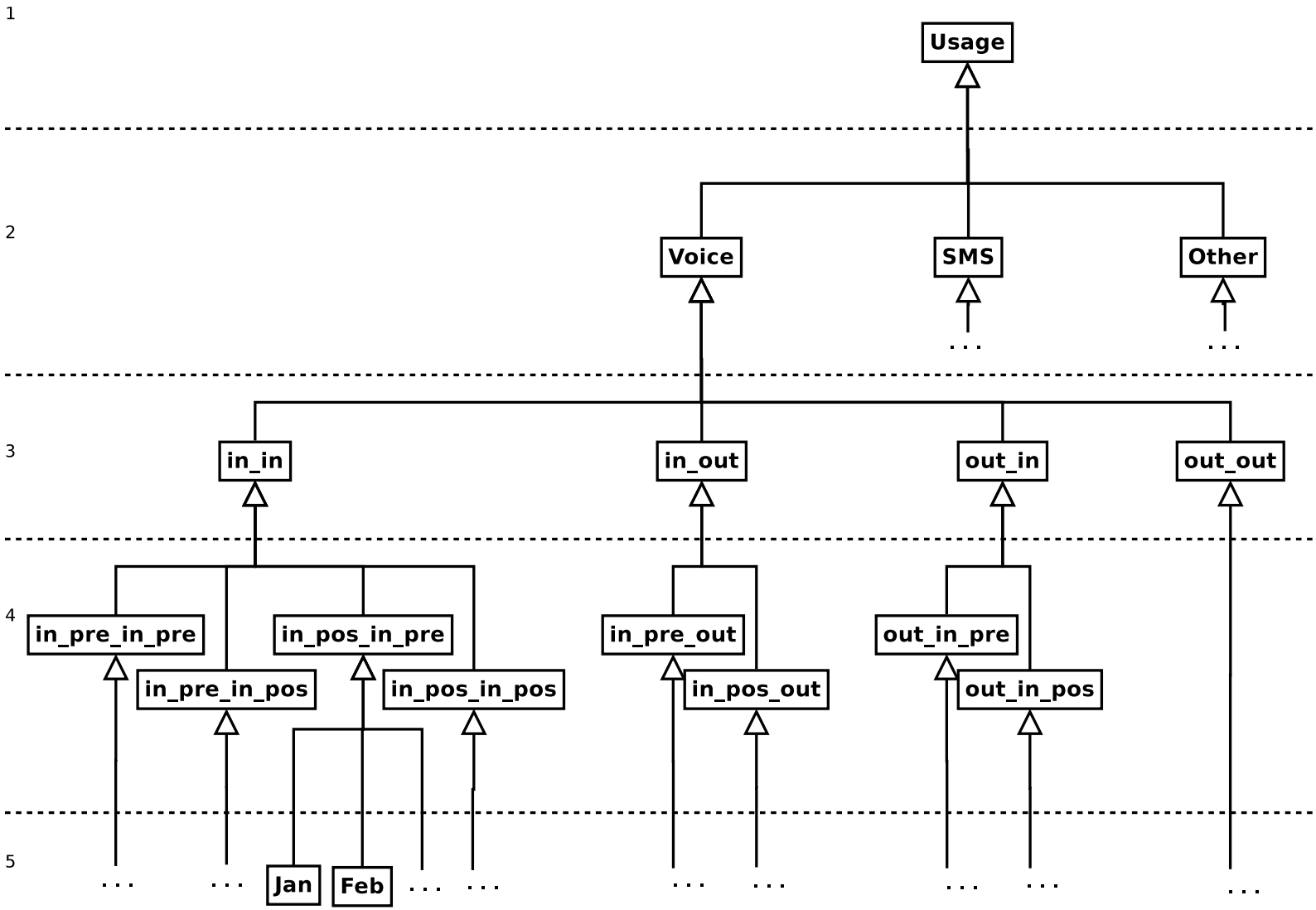
- Granularity: Level of detail.
- Partitioning: Divide the data into manageable chunks.

Data Partitioning

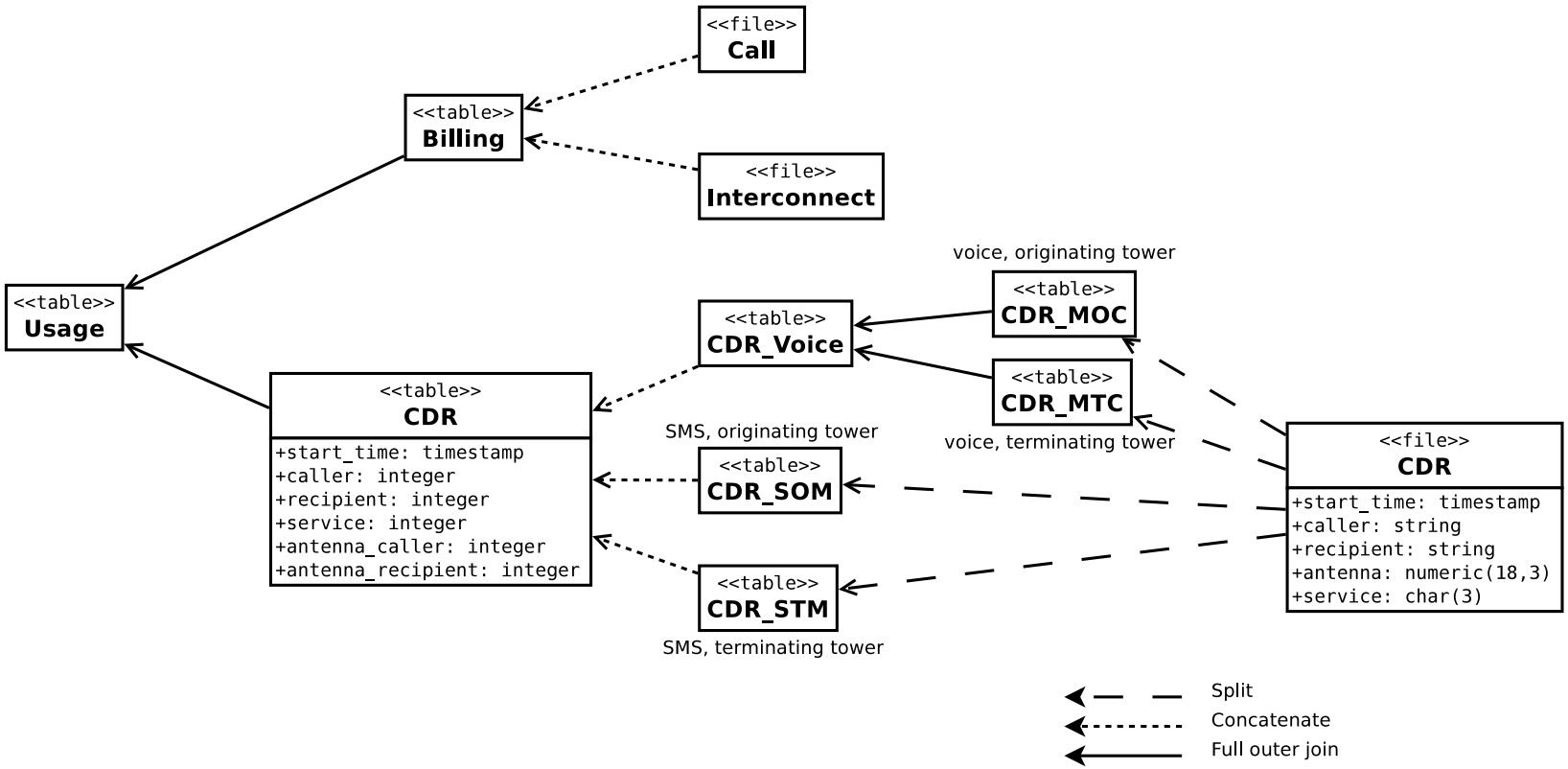
Guided by the way in which the data has been historically used: groups of greatest interest are voice calls and "in-company" users.

- Onnela *et al.* 2007. Structure and Tie Strengths in Mobile Communication Networks. *Proceedings of the National Academy of Sciences*.
- González *et al.* 2008. Understanding Human Mobility Patterns. *Nature*.
- Wang *et al.* 2009. Understanding the Spreading Patterns of Mobile Phone Viruses. *Science*.

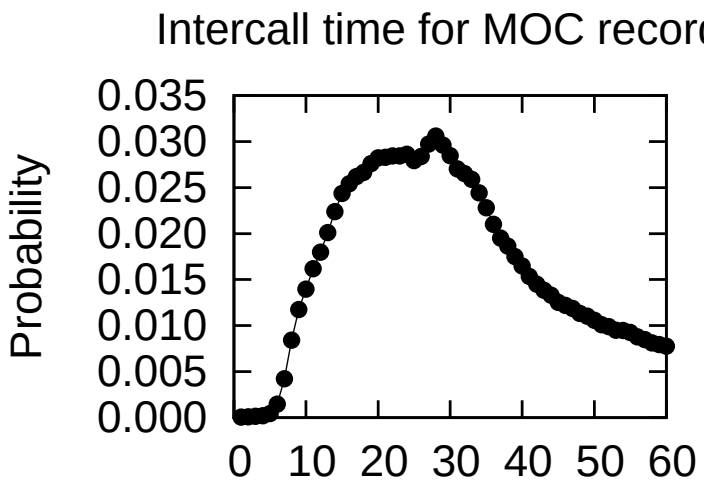
Data Partitioning

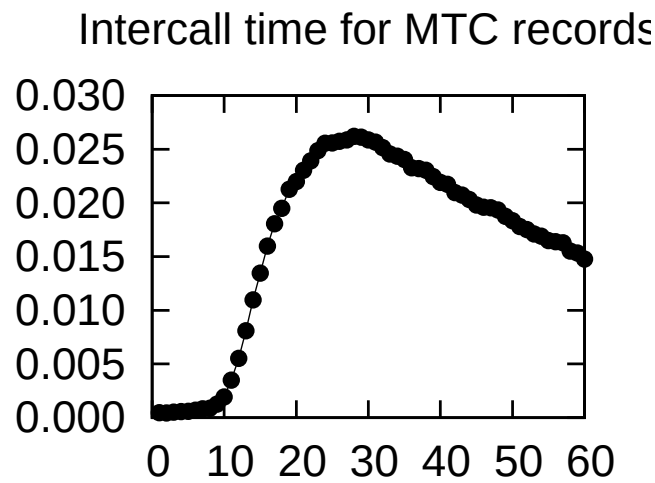


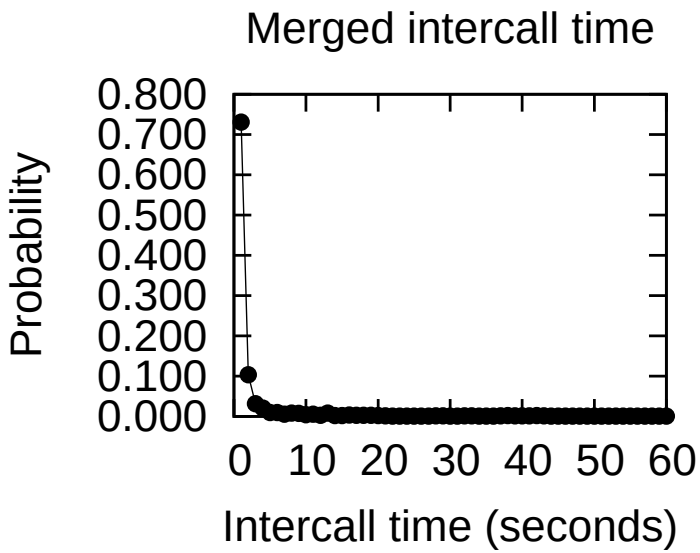
Usage Data Integration: Partition and Merge

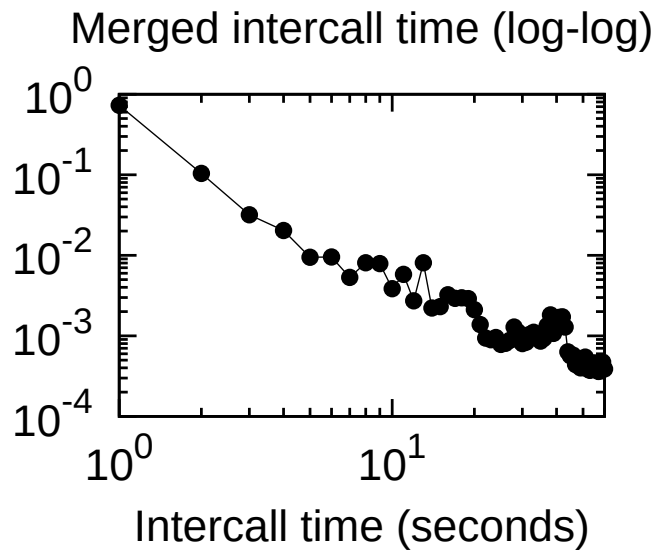


Introduction of Duplicates by Merge





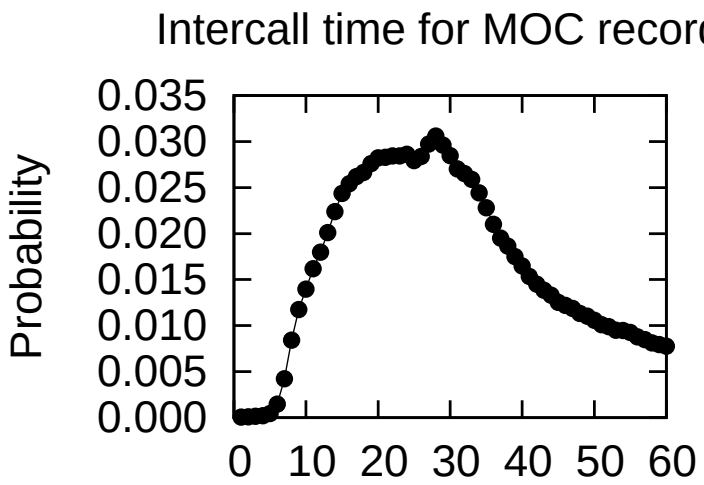


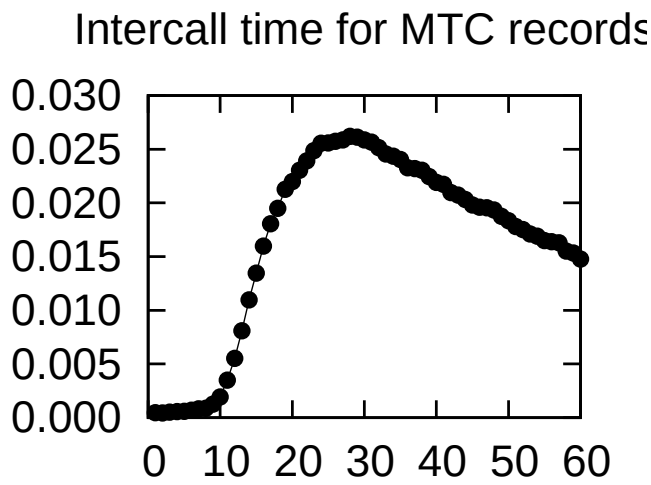


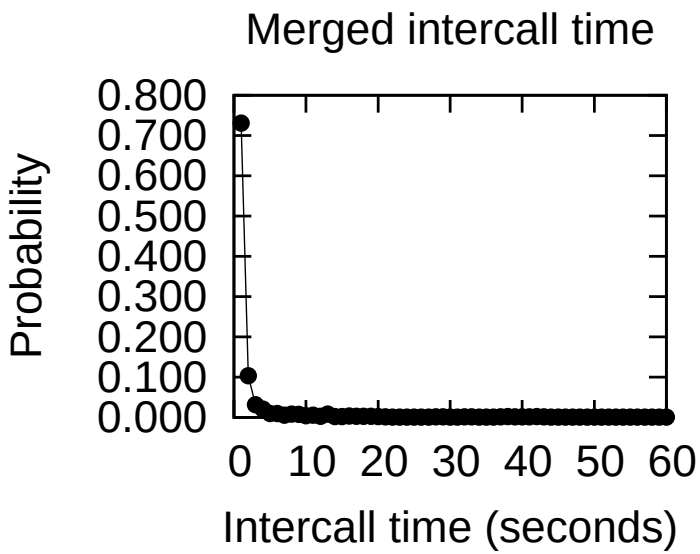
Cleaning

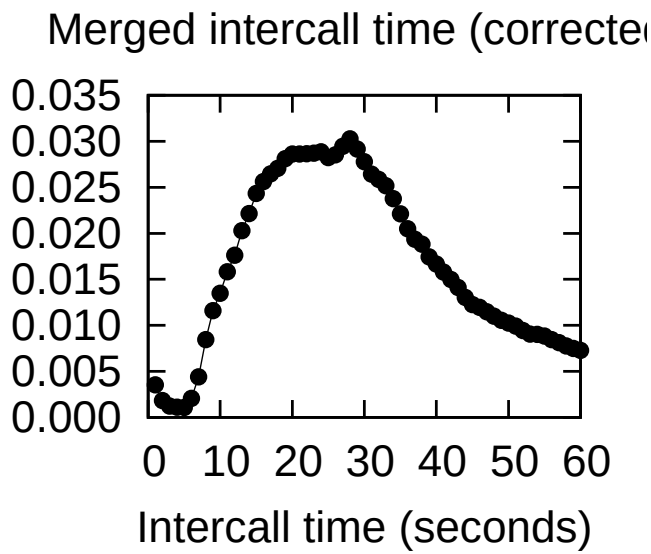
- Consider only records that haven't been matched (initially records with a value in exactly one of the two antenna fields).
- Add perturbation to the start time of records having a value in only the terminating antenna field and attempt to match with records having a value in only the originating antenna field.
  - Sequence of perturbation: $+1, -1, +2, -2, \cdots + 60, -60$.
- Fields matched at each step are removed from consideration:
  - Match records with smallest possible perturbation.
  - Match each record only once.

Results

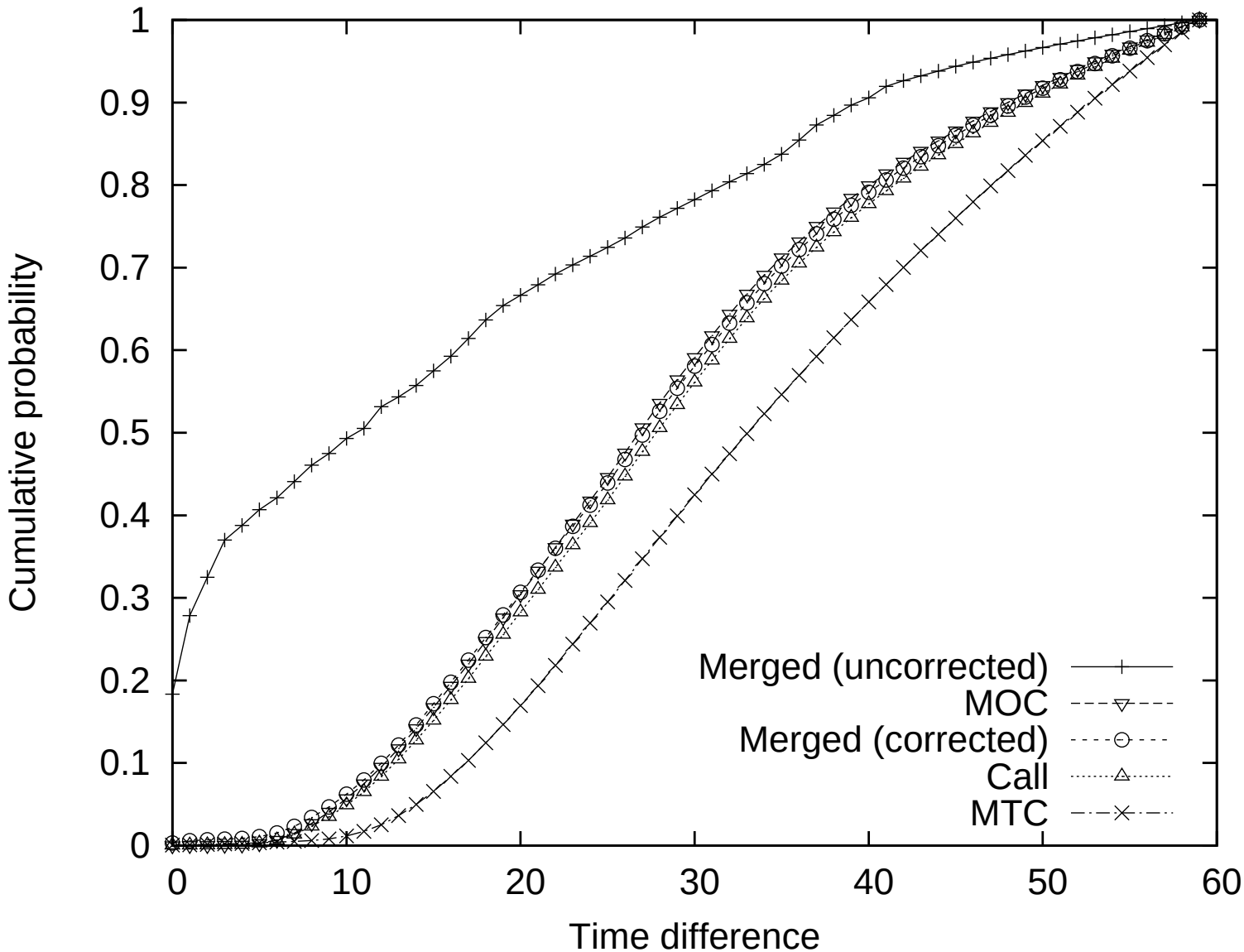

Intercall time for MOC records
Probability
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60


Intercall time for MTC records
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60


Merged intercall time
Probability
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
0 10 20 30 40 50 60
Intercall time (seconds)


Merged intercall time (correcte
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60
Intercall time (seconds)

Results


Cumulative probability
Time difference
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
10
20
30
40
50
60
Merged (uncorrected)
MOC
Merged (corrected)
Call
MTC

Consequences

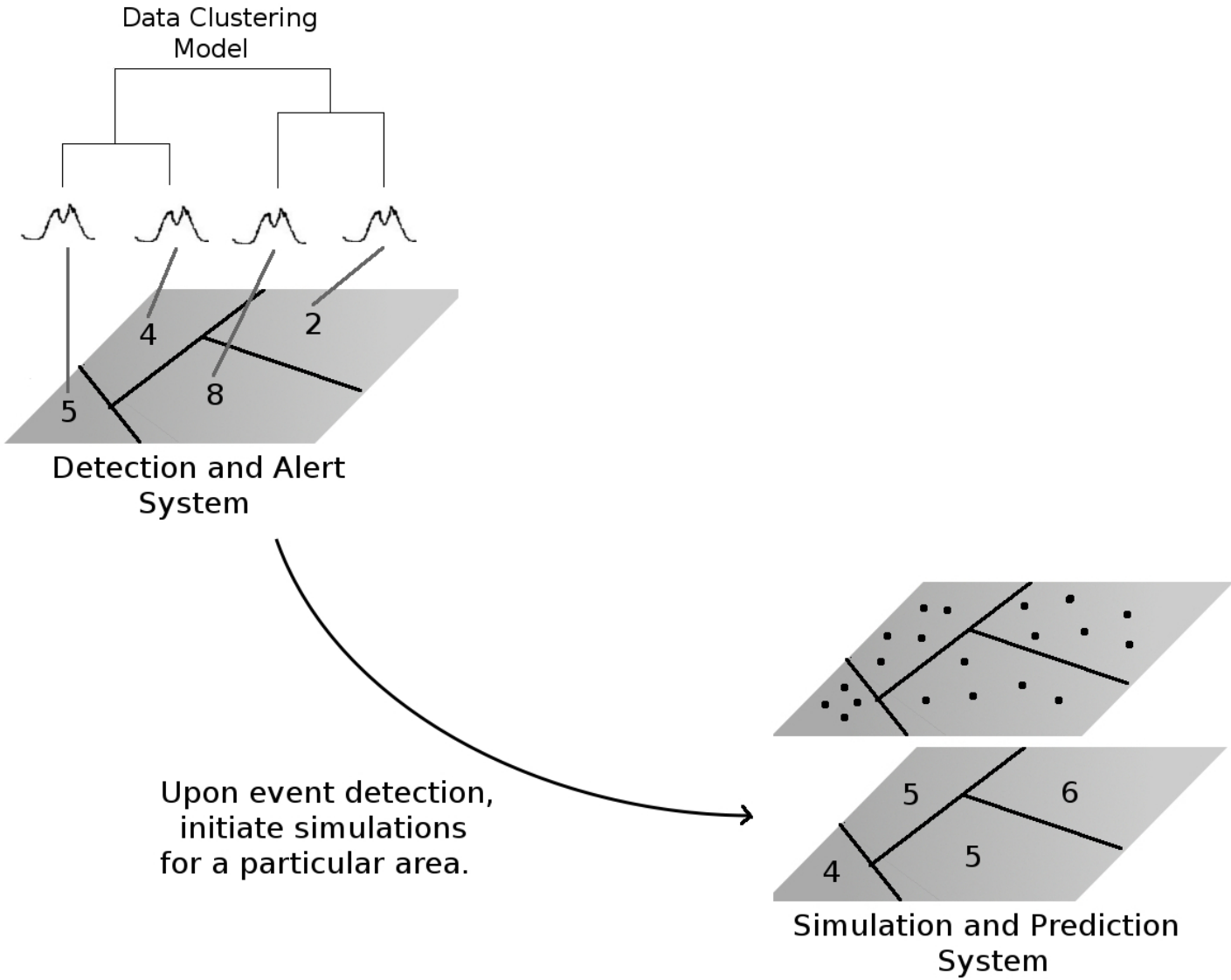
- In many cases, these duplicates can be ignored, *e.g.*
  - Building graphs with edges weighted by total duration or cost (Onnela *et al*, 2007).
  - Generating user trajectories (González *et al*, 2008).
- Can be problematic when generating time series of call activity.

Summary

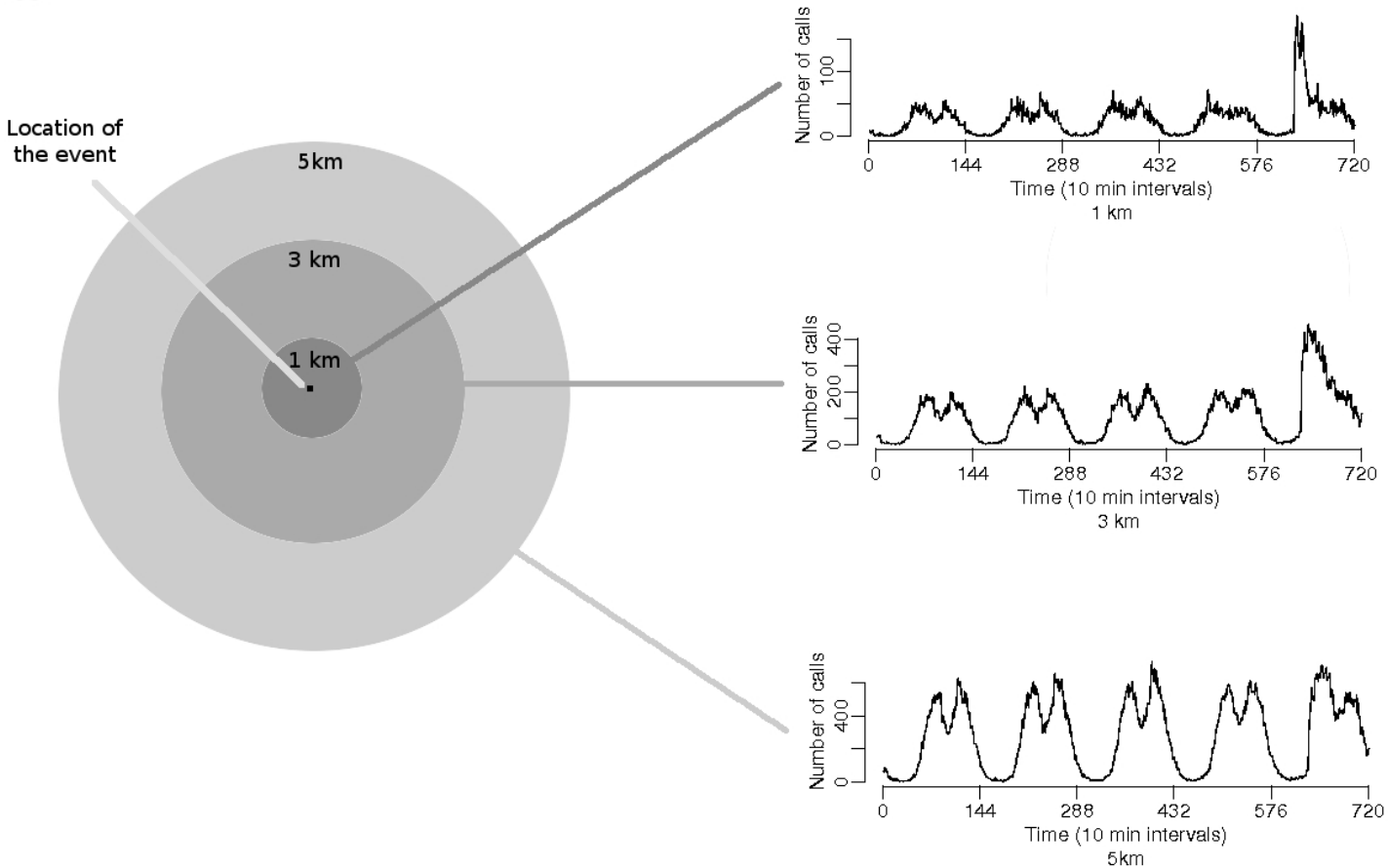
- Data warehouse contains 14 months of data.
- Data loading process in highly automated:
  - Two scripts: customer and usage data
  - Each script:
    - validates fields
    - loads the initial tables
    - replaces user strings with integers
    - integrates the various record types
    - usage script partitions the records
- Data extract from the warehouse is in use for a social network study
  - Comparable data extraction from text files takes weeks: code is written to identify relevant users, replace user ID strings, filter data.
  - One year of data can be extracted in one day using a single SQL query.

Detecting Real-World Events in Space and Time Using Feature Clustering

Overview


Data Clustering
Model
4
2
8
5
Detection and Alert
System
Upon event detection,
initiate simulations
for a particular area.
5
6
5
4
Simulation and Prediction
System

Key Observation: Certain Events Cause Localized Changes in Call Activity



Clustering

Let $\mathbf{D}$ be an $n \times m$ matrix:

- n is the number of observations
- m is the number of features per observation
- In this case, features are the call activity at each spatial area (postal code, tower). Data item is recorded every 10 minutes.

Goal: partition the n rows $\mathbf{D}$ into a natural grouping.

- Minimize intra-cluster distance.
- Maximize inter-cluster distance.

We want to cluster the time series, so we transpose $\mathbf{D}$ before clustering; this is called feature clustering.

Related Work

- Rodrigues *et al.* (2004). Stream feature clustering algorithm.
  - Correlation dissimilarity: sufficient statistics require quadratic space with respect to the number of features.
- Aggarwal *et al.* (2003).
  - If entire history is used, stale data dominates results.

Approach

Dataset:

- Aggregate records based on spatial area: postal code or tower.
- Time series: number of calls in 10 minute intervals for each spatial area.

Clustering:

- Single link clustering over a 1 day sliding window.

Event Detection:

- Outliers are detected by examining maximum distance between clusters.
- Baseline is computed using the maximum distance between clusters over the 2 months of data.

Two events:

- An explosion.
- Celebration following a sports victory.

Results: Bombing

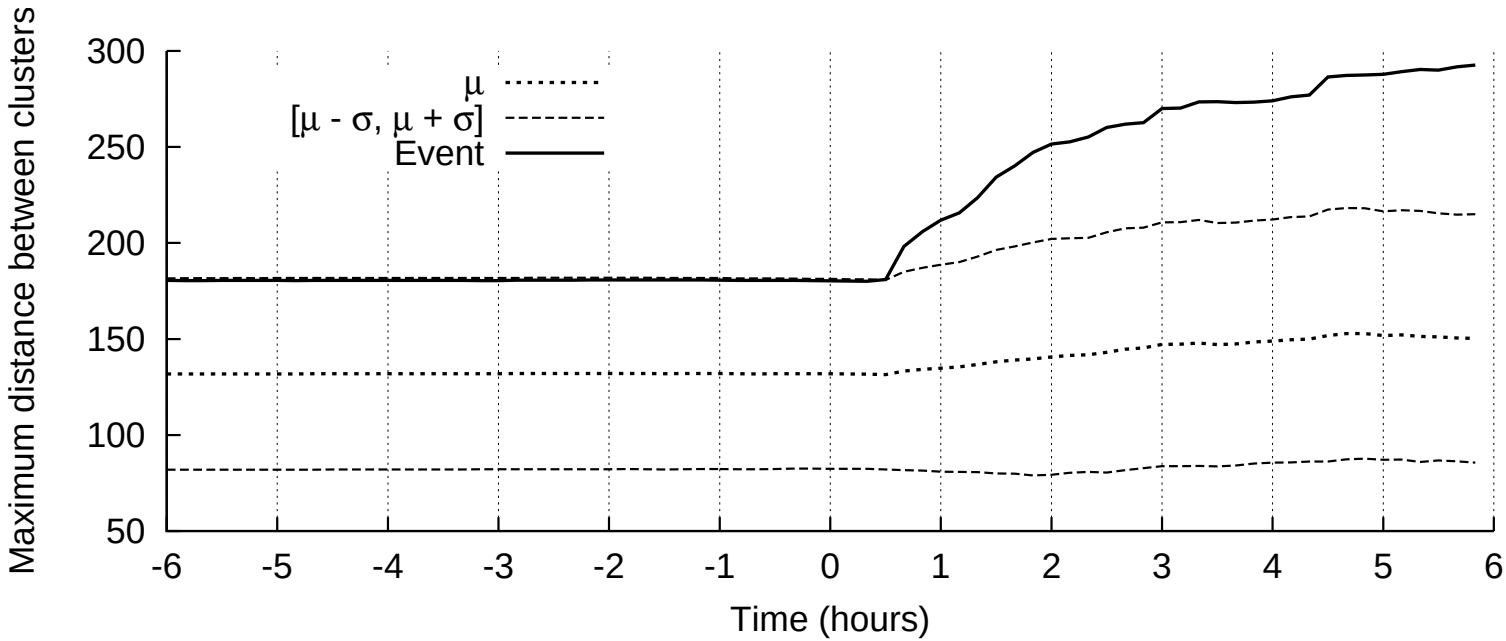
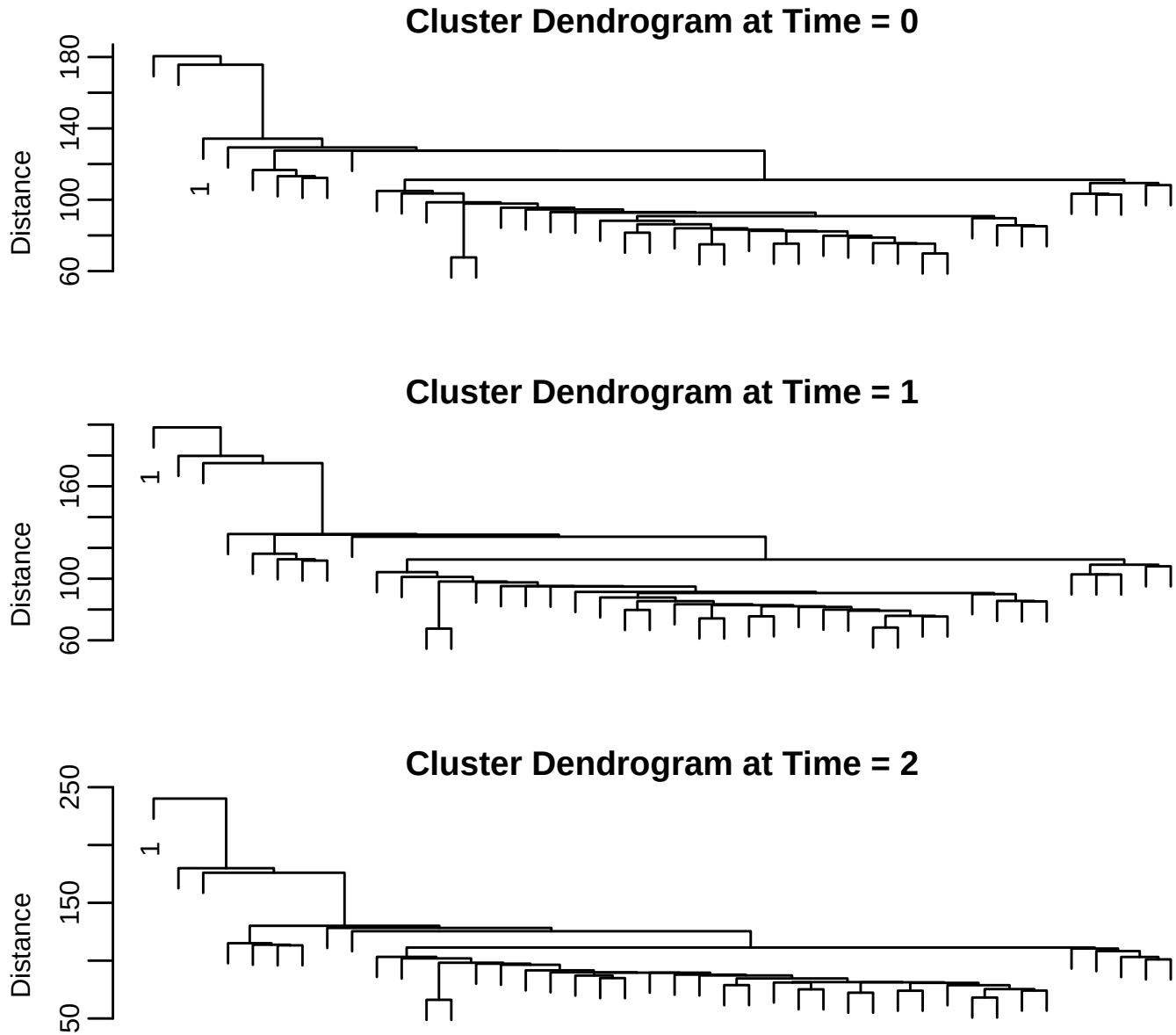


Figure: The maximum distance between clusters for a 12 hour window around the bombing with the mean and $\pm$ 1 standard deviation over 2 months. The bombing occurs between time $= 0$ and time $= 0.5$ (conflicting news reports)

Results: Bombing



Results: Celebration

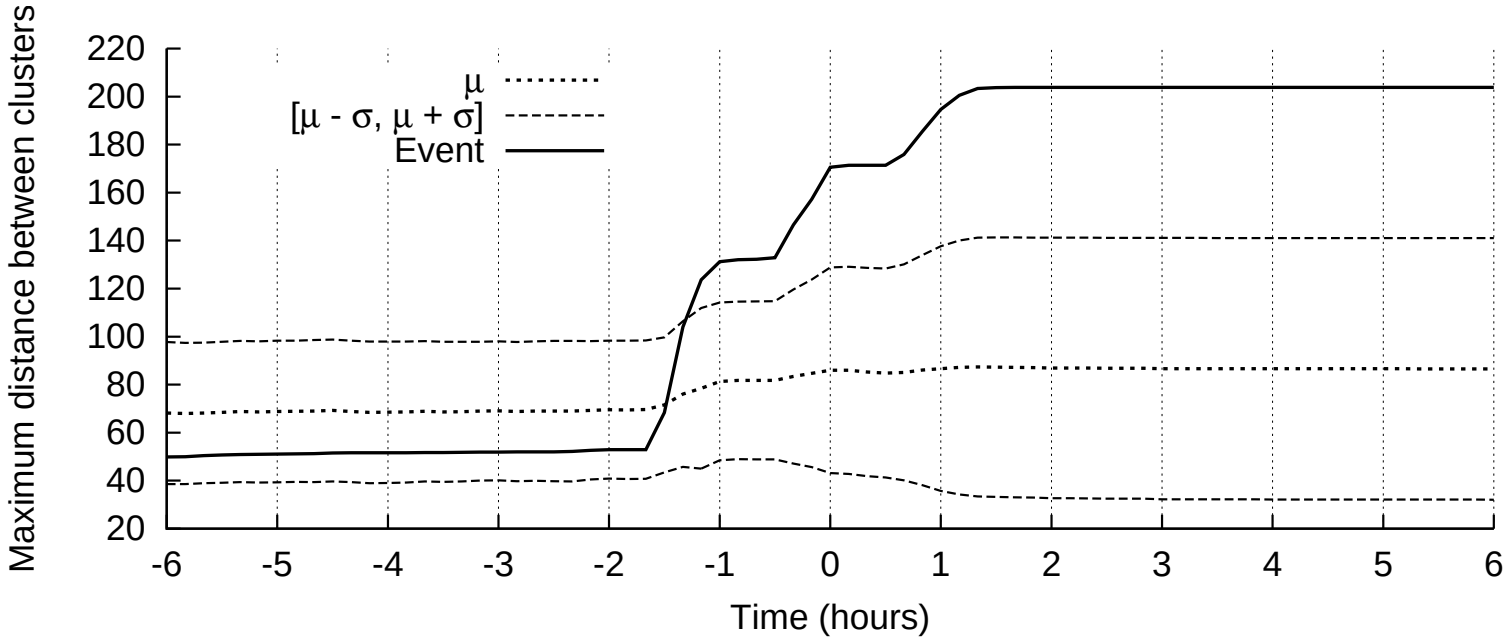
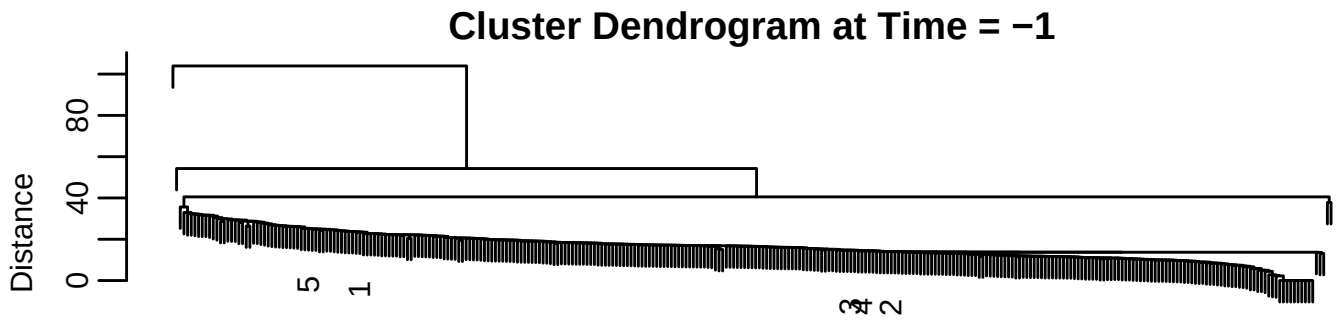
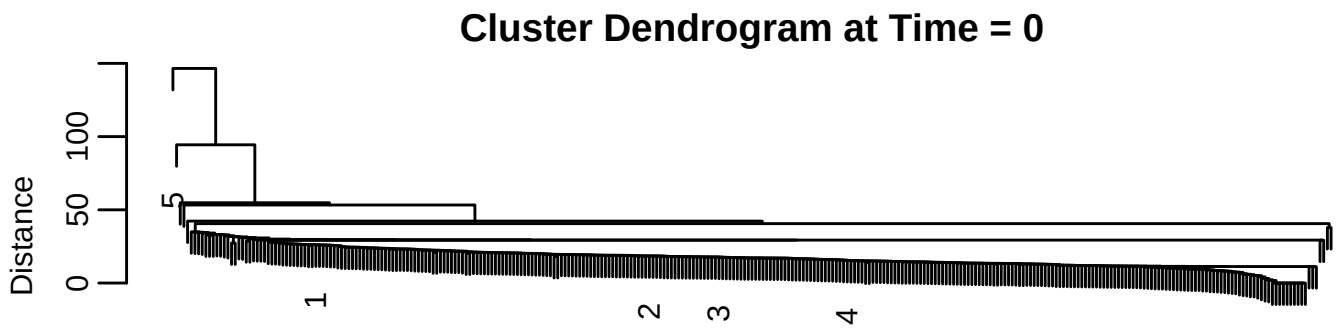
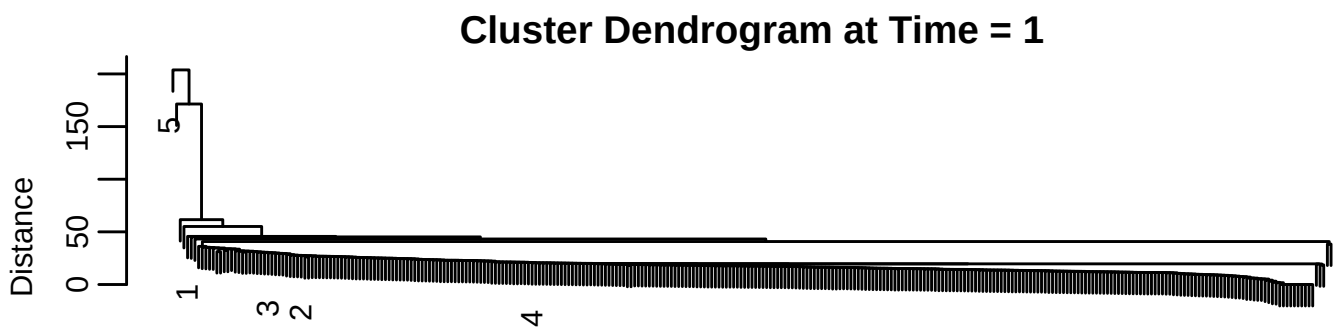


Figure: The maximum distance between clusters for a 12 hour window around the celebration with the mean and $\pm$ 1 standard deviation over 2 months. Important events are: at approximately time -1 the sporting event ends, at time $= 0$, the crowd has gathered to meet the team, and at approximately time $= 1$ the team leaves the celebration.

Results: Celebration







Summary

- In cases were events cause localized changes in call activity, we can use feature clustering to detect the events and their approximate location.
- This achieves our goals of automating event detection and constraining the area that must be simulated by the Simulation and Prediction System.

Summary

Summary of Research

Two components:

- Historical Data Source. Data repository for:
  - development of WIPER system components
  - social network and human mobility research
- Feature clustering algorithm for event detection
  - In cases where event causes a localized change in call activity, the event and it's approximate location can be identified.

First Author Publications

- Alec Pawling, Ping Yan, Julián Candia, Tim Schoenharl, and Greg Madey. "Anomaly Detection in Streaming Sensor Data," In: *Intelligent Techniques for Warehousing and Mining Sensor Network Data.* Alfredo Cuzzocrea, Ed., 2009.
- Alec Pawling and Greg Madey. "Feature Clustering for Data Steering in Dynamic Data Driven Application Systems." Proceedings of the 9th International Conference on Computational Science, 2009.
- Alec Pawling, Tim Schoenharl, Ping Yan, and Greg Madey. "WIPER: An Emergency Response System." In: Proceedings of the 5th International ISCRAM Conference, 2008.
- Alec Pawling, Nitesh Chawla, and Greg Madey. "Anomaly Detection in a Mobile Communication Network." Computational & Mathematical Organization Theory. 13:4, December, 2007.

First Author Publications

- Alec Pawling, Nitesh Chawla, and Greg Madey. "Anomaly Detection in a Mobile Communication Network." In: Proceedings of the Annual Conference of the North American Association for Computational Social and Organization Sciences. 2006. (Received Best Student Paper Award).
- Alec Pawling, Nitesh Chawla, and Amitabh Chaudary. "Evaluation of Summarization Schemes for Learning in Streams." In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases." 2006.
- Alec Pawling and Nitesh V. Chawla, and Amitabh Chaudhary. "Computing Information Gain in Data Streams." In: Proceedings of the ICDM Workshop on Temporal Data Mining: Algorithms, Theory, and Applications. 2005.

Other Publications

- Gregory R. Madey, Albert-László Barabási, Nitesh V. Chawla, Marta González, David Hachen, Brett Lantz, Alec Pawling, Timothy Schoenharl, Gábor Szabó, Pu Wang and Ping Yan, "Enhanced Situational Awareness: Application of DDDAS Concepts to Emergency and Disaster Management ", in International Conference on Computational Science, serial Lecture Notes in Computer Science (LNCS 4487), Y. Shi, G. D. van Albada, J. Dongarra, and P. M. A. Sloot, Eds., May 2007, pp. 1090-1097.

Publication Plan

- Data warehouse implementation (Chapters 3, 4): Planned submission to the International Journal of Data Warehousing and Mining, pending revision.
- Feature clustering approach for event detection (Chapter 7): Planned submission to IEEE Transactions on Patern Analysis and Machine Intelligence, pending revision.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. CNS-0540348 and BCS-0826958, Northeastern University, and the Defense Threat Reduction Agency under Grant BRBAA07-J-2-0035.

I would also like to thank:

- My advisor: Dr. Madey.
- My committee: Dr. Chawla, Dr. Chaudhary, and Dr. Poellabauer.
- The outside chair: Dr. Hachen.
- Colleagues in Dr. Madey's Group: Tim Schoenharl, Ping Yan, Ryan Bravo, and Ryan McCune.
- Dr. Barabási and the members of the Center for Complex Network Research at Northheastern University. In particular, I am grateful to Dr. Marta González, Dr. Julián Candia, Dr. Sune Lehmann, Dr. Jim Bagrow, Nick Blumm and Dashun Wang.