Social Behavior Prediction Through Reality Mining

Charlie Dagli, William Campbell, Clifford Weinstein

Human Language Technology Group
MIT Lincoln Laboratory

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work was sponsored by the DDR&E / RRTO under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- Introduction and Motivation
- Corpora
- Identification Based on Reality Mining
- Tensor Analysis of Reality Mining Data
- Other Efforts
- Conclusions & Future Potential

Acknowledgement:
Sandy Pentland, MIT Media Lab

Overview & Motivation

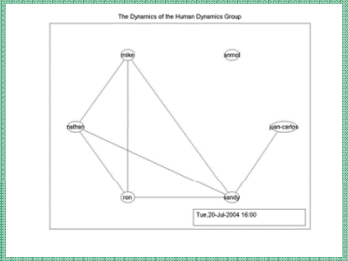
Sociometric Badges (microphone, IR, accelerometer, …)

Bluetooth

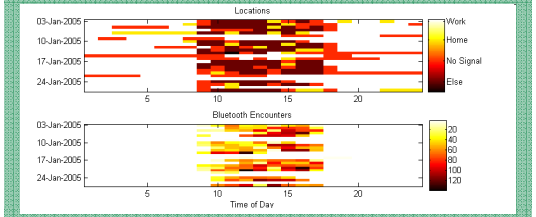
Smart Phones

- *Surveys*
- *Patient Records*
- *Call Info*

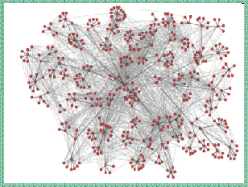
Sensors & Data



Dynamic Network Analysis



Anomalous Behavior



Group Behavior Prediction

Analytics

Focus:
- **Dynamic, real-time analysis of networks and behavior prediction**
- **Sensor rich multi-INT data**

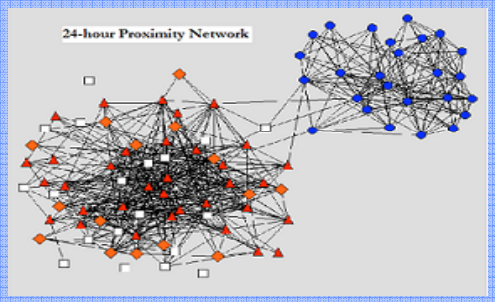
MIT LL Applications:
- **Correlation of Multiple INTs (cyber, cell, etc.)**
- **Persistent Surveillance**
- **Urban Warfare**
- **Blue force tracking**
- **Cognitive Radio**
- **CT**

Potential Future Sponsors:
- **DoD/IC**
- **DHS**

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

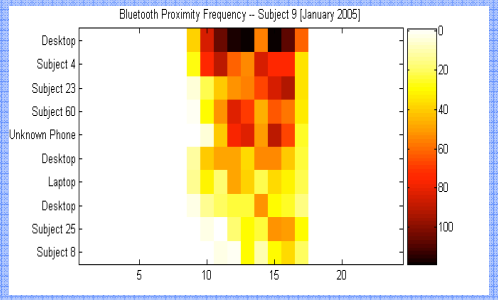
Recognition and Prediction

Multi-INT Fusion Produces Noisy Group Data

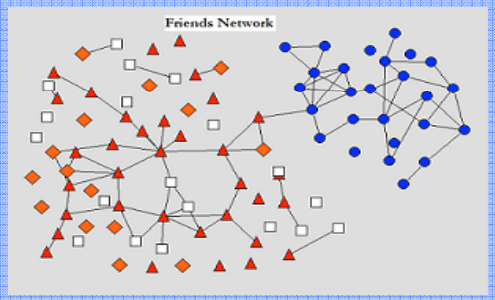


Single Devices Produce Fine-Grained Temporal Dynamics

Social Network Analysis Extracts High-Level Features



Proposed Effort: Real-Time Analysis of Network Features



Corpora

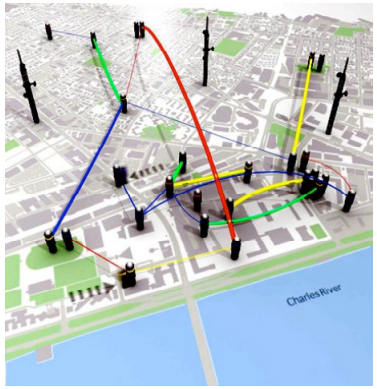
- 94 students were given Nokia 6600 smart cell phones [1]
- Sensor data collected 09/04 – 06/05
  - Cell phone location
  - Phone activity
  - Bluetooth scans
- Data anonymized for release
- Additional data:
  - Student surveys on health
  - Surveys on social network

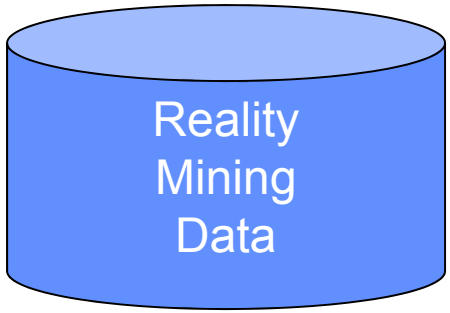
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Reality Mining Data

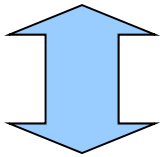
- **Raw data is Bluetooth Addresses, Cell Phone calling, and Cell Phone location**
- **No content—voice or message**
- **Typical Query of Location Data**

```
select * from locs where (myhn=44) order by d, t limit 4;
-----------------------------------------------
| myhn | date       | time     | areaid | cellid|
+------+------------+----------+--------+-------+
| 44   | 2004-11-01 | 18:25:30 | 5119   | 40813 |
| 44   | 2004-11-01 | 18:26:01 | 5119   | 40332 |
| 44   | 2004-11-01 | 18:26:18 | 5119   | 40811 |
| 44   | 2004-11-01 | 18:27:13 | 5119   | 40332 |
```





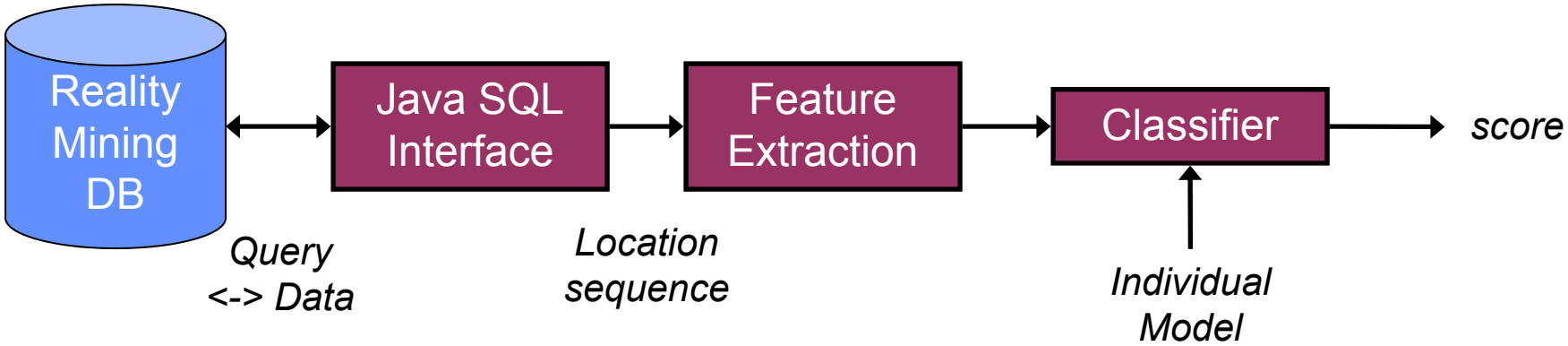
MySQL DB



Algorithms
& Experiments

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Identification via Reality Mining Data



- **Typical Testing Scenario**
- **Location sequence is over a period of time—typically 1-5 days**
- **Feature extraction summarizes location information as a vector**
- **Classifier used for detection of the individual**

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Features for Recognition

Input Stream

..... 5119_40811 5119_40811 5119_40811 5188_60241 5188_40811 5188_40813 5188_40811 5188_42171 5188_40811 5188_40332 5188_40811 5188_40332 5188_40811 5188_40332 5188_40811 5188_40811 5188_40811 5188_40811 5188_42171 5188_40811 5182_43861 5182_43861 5188_40811 5188_40332 5188_40811 5188_40332 5188_40811 5188_40332 5188_43032 5188_40333 5188_43032 5188_43033 5188_40333 5188_43033 5188_43031 5188_43033 5188_40333 5188_43031 .....

Window over k days

5119_40811 5188_60241 5188_40811 5188_40813 5188_40811 5188_42171 5188_40811 5188_40332

N-grams

5119_40811| 5188_60241 5188_60241| 5188_40811 5188_40811|5188_40813 5188_40813| 5188_40811
5188_40811|5188_42171 5188_42171| 5188_40811 5188_40811|5188_40332

Probabilities

P(5119_40811| 5188_60241) = 1/7
P(5188_60241| 5188_40811) = 1/7
…

Vector

V = [0 … 0 1/7 0 … 0 1/7 0 … 0 …]

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

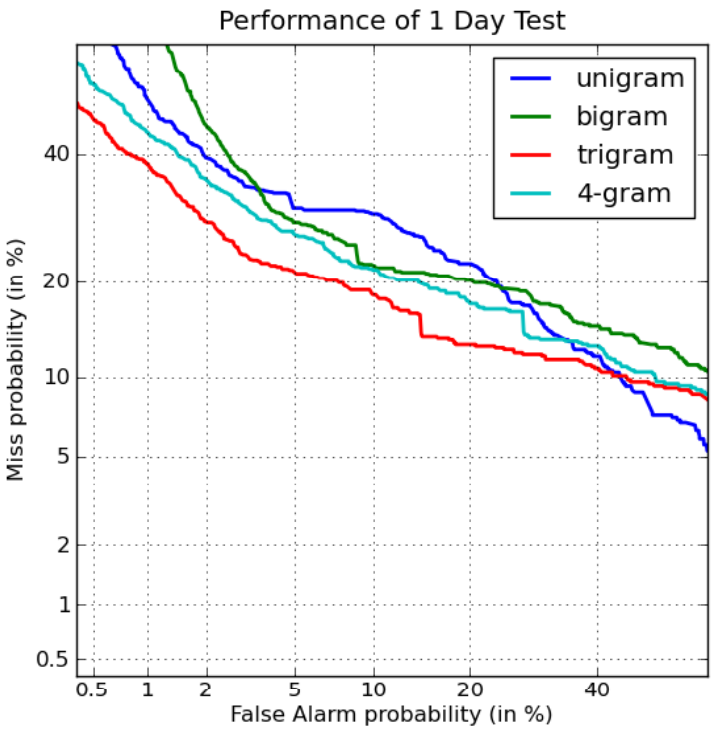
Experimental Setup

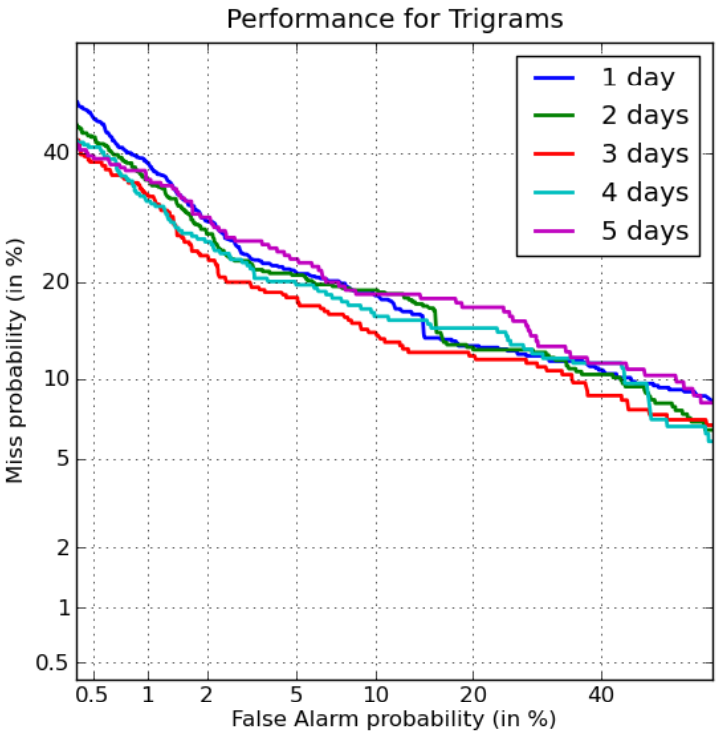
- N-gram Features weighted with log of inverse frequency in "background"
- SVM Classifier, Linear Kernel
- Training Set 1 (TR1)
  - Data collected between September 2004 and December 1, 2004
- Training Set 2 (TR2)
  - Data collected between March 1, 2005 and March 31, 2005
- Test sets at about 1 month intervals:
  - TST1, Jan 1–Feb 1, 2005
  - TST2, Feb1–March 1, 2005
  - TST3, March 1–April 1, 2005
  - TST4, April 1–June 1, 2005

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Identification Experimental Results #1

- Initial experiments looked TRN1 and TST1
- Varied *n*-gram order
- Varied window for accumulating statistics from 1-5 days

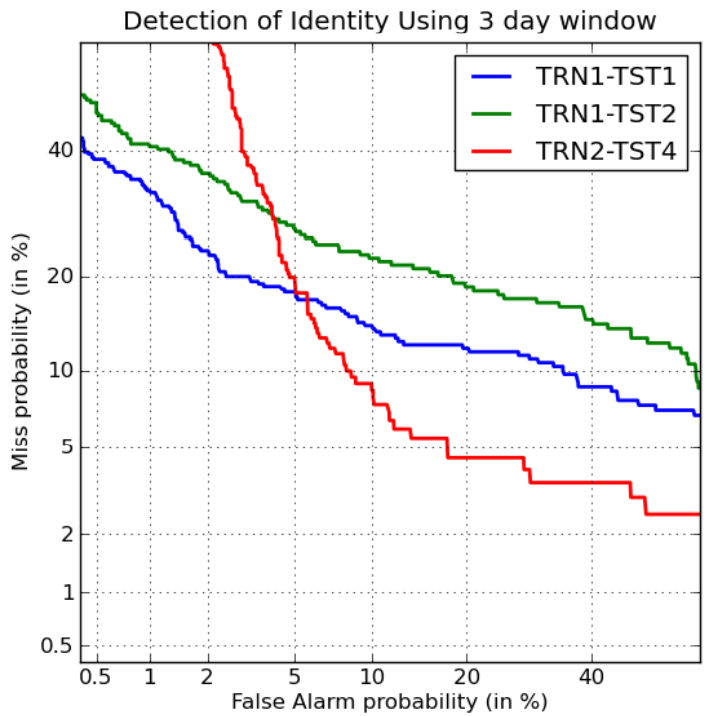




1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Identification Experimental Results #2

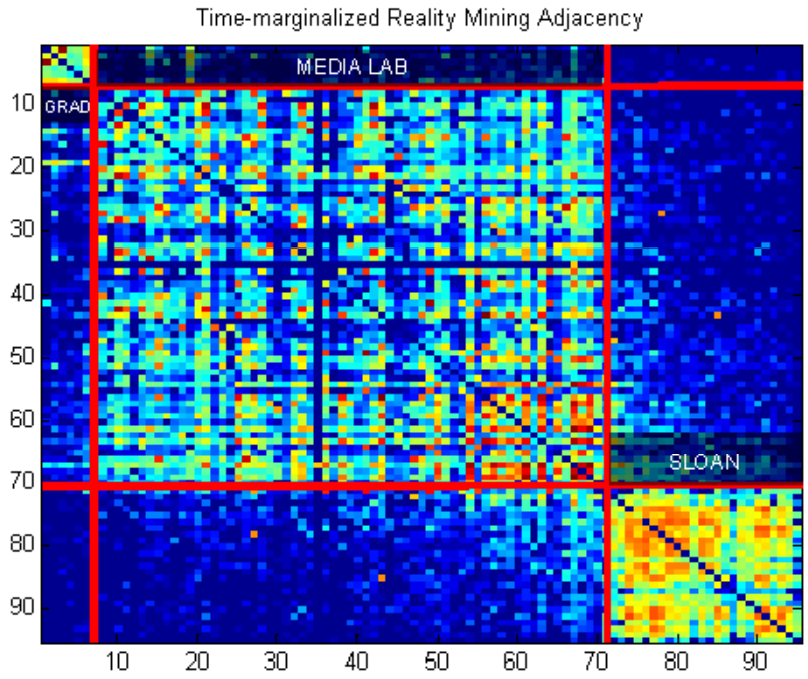
- **Considered training duration:**
  - **1 month versus 3 months**
- **More months of training data changes curve shape—better results at low Pfa**
- **Equal Error rates < 10% !**



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

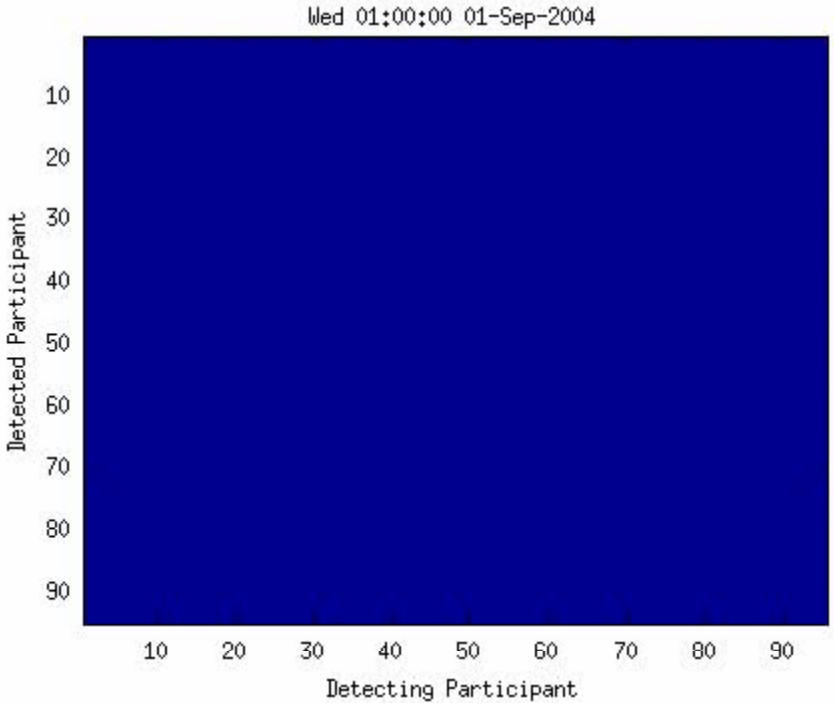
Tensor Representation and Analysis

- Dynamic social network (DSN) [2] as a 3-dimensional data cube (node x node x time)
- Corresponds to 3-mode *tensor* (multi-way array) [3]
- Can use *tensor analysis* tools to gain meaningful insights into DSNs [4]
- Why these tools are useful, what can they do?
  - Give high-order, meaningful summarizations of input data
  - Computationally fast
  - Find co-clusters in multiple dimensions when they exist



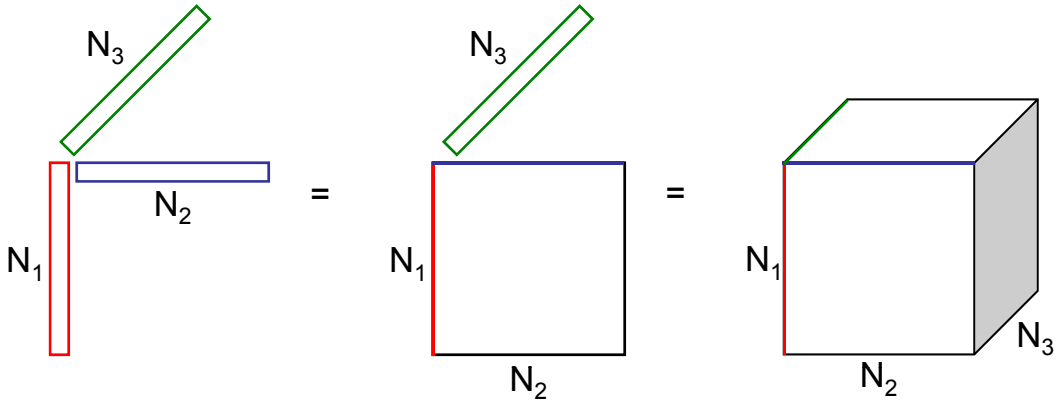
Tensor Representation and Analysis

- Dynamic social network (DSN) [2] as a 3-dimensional data cube (node x node x time)
- Corresponds to 3-mode *tensor* (multi-way array) [3]
- Can use *tensor analysis* tools to gain meaningful insights into DSNs [4]
- Why these tools are useful, what can they do?
  - Give high-order, meaningful summarizations of input data
  - Computationally fast
  - Find co-clusters in multiple dimensions when they exist

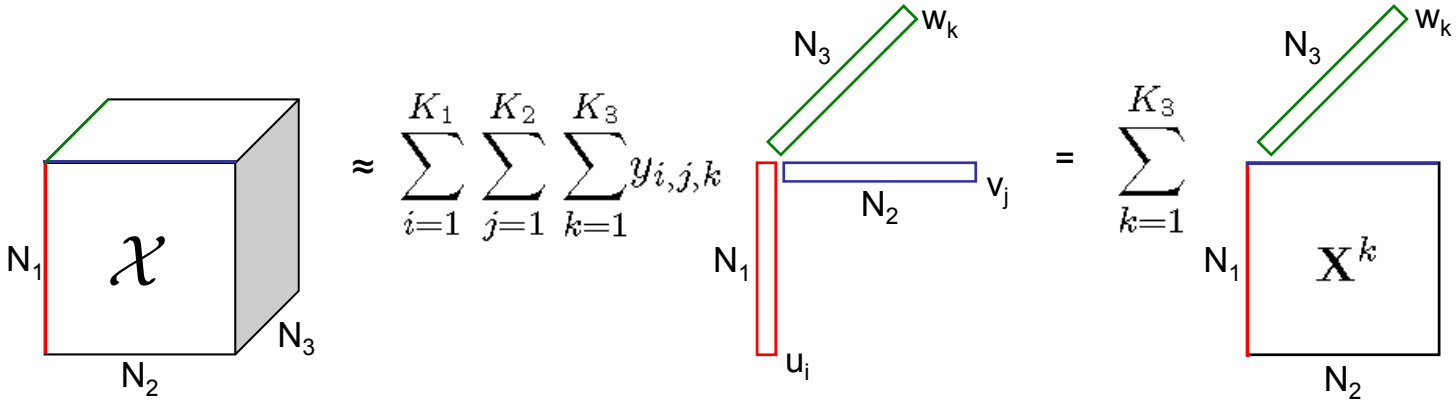


1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Multi-linear (Tensor) Algebra



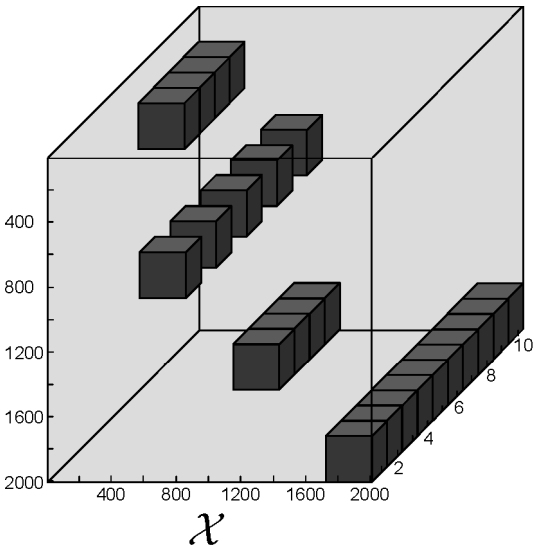
Rank-1 Tensor

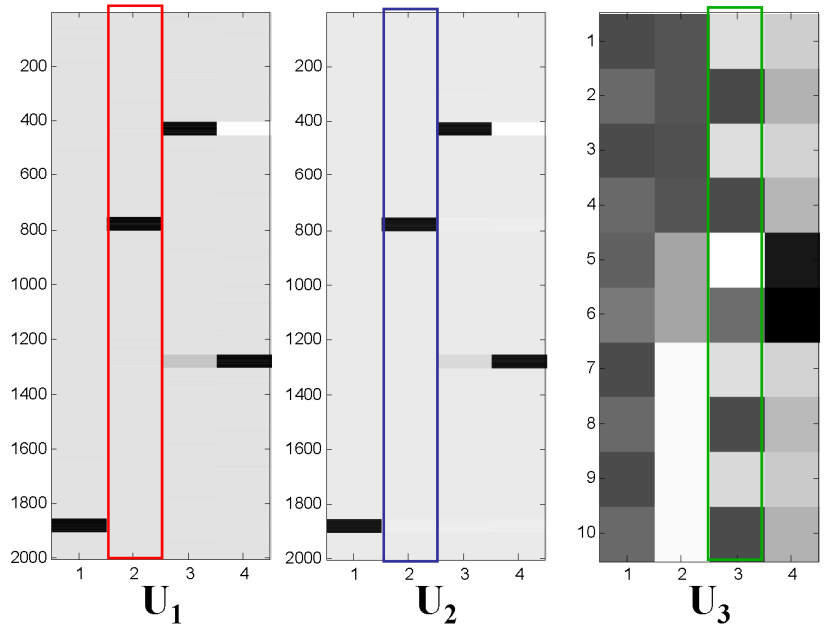


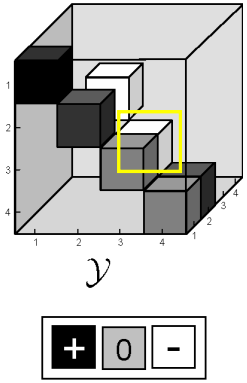
Rank-(K_1,K_2,K_3) Approximation

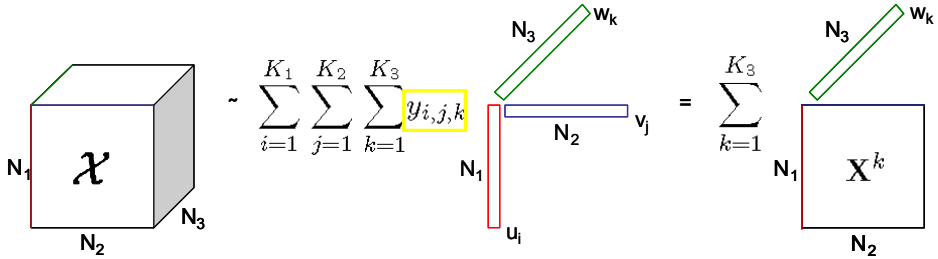
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Toy Problem





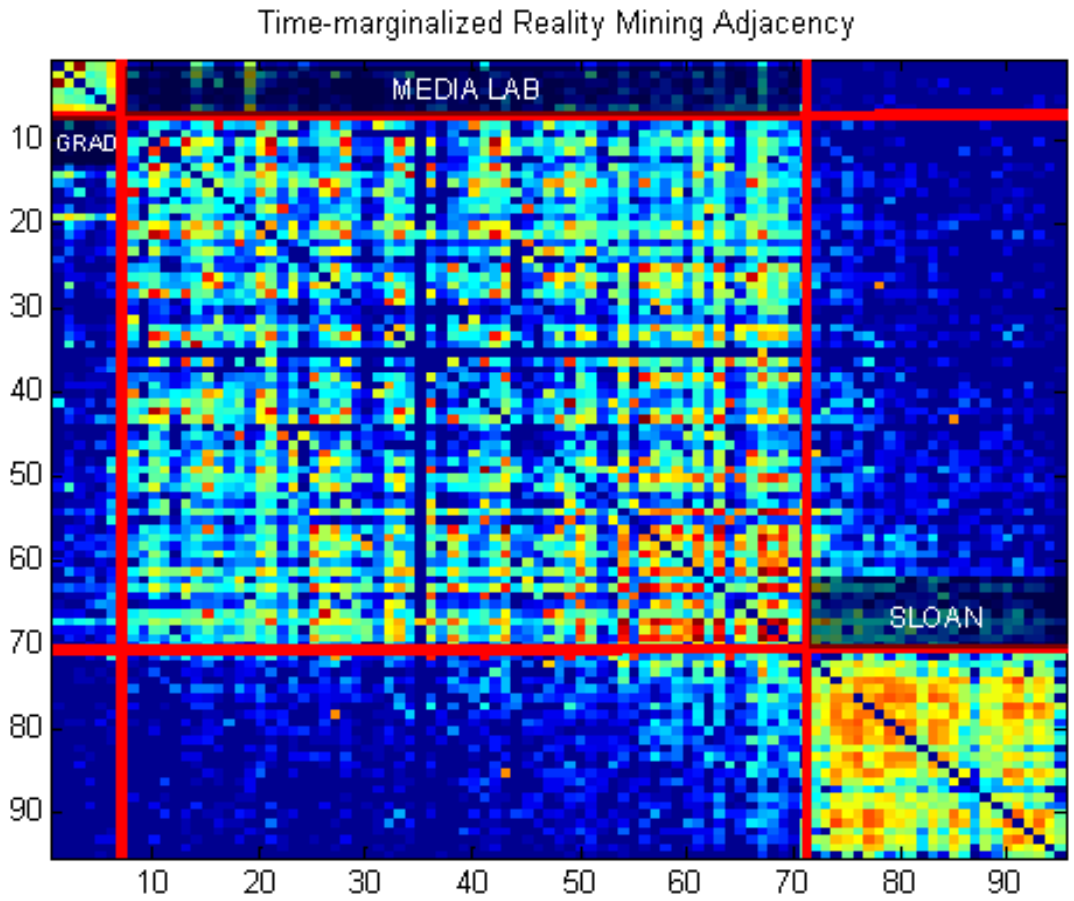




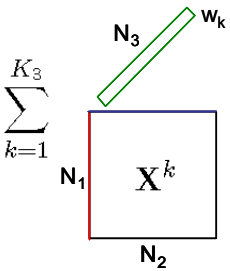
Rank-(K_1,K_2,K_3) Approximation

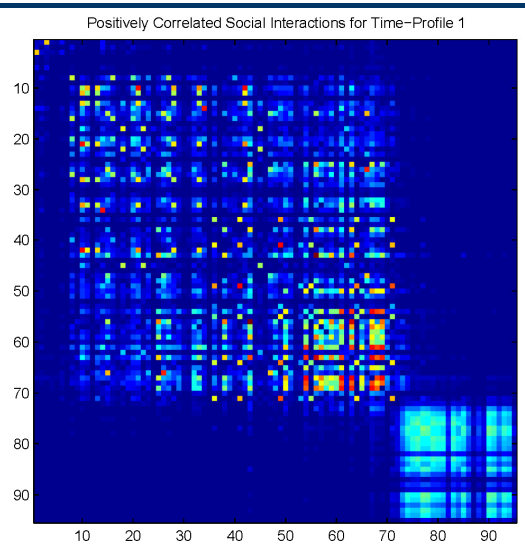
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

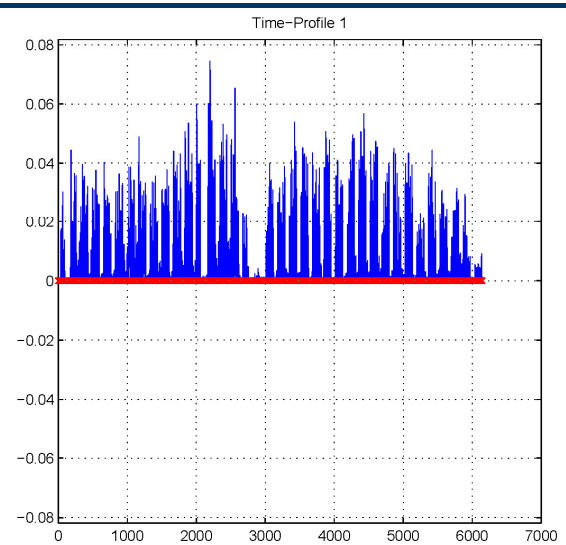
Reality Mining – Bluetooth Proximity

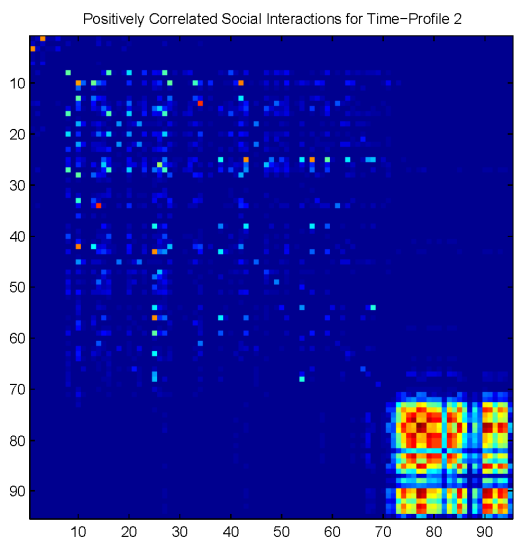


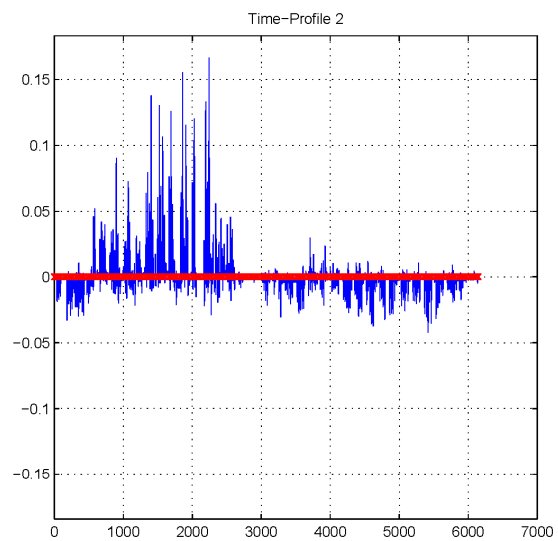
Reality Mining











1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Course-Attending Behavior

```
Thu 11:00:00 07-Oct-2004
Thu 16:00:00 14-Oct-2004
Tue 10:00:00 26-Oct-2004
Thu 11:00:00 28-Oct-2004
Tue 11:00:00 02-Nov-2004
Thu 10:00:00 04-Nov-2004
Tue 11:00:00 09-Nov-2004
Tue 11:00:00 16-Nov-2004
Thu 11:00:00 18-Nov-2004
Tue 11:00:00 23-Nov-2004
Tue 11:00:00 30-Nov-2004
Thu 11:00:00 02-Dec-2004
Mon 10:00:00 06-Dec-2004
```

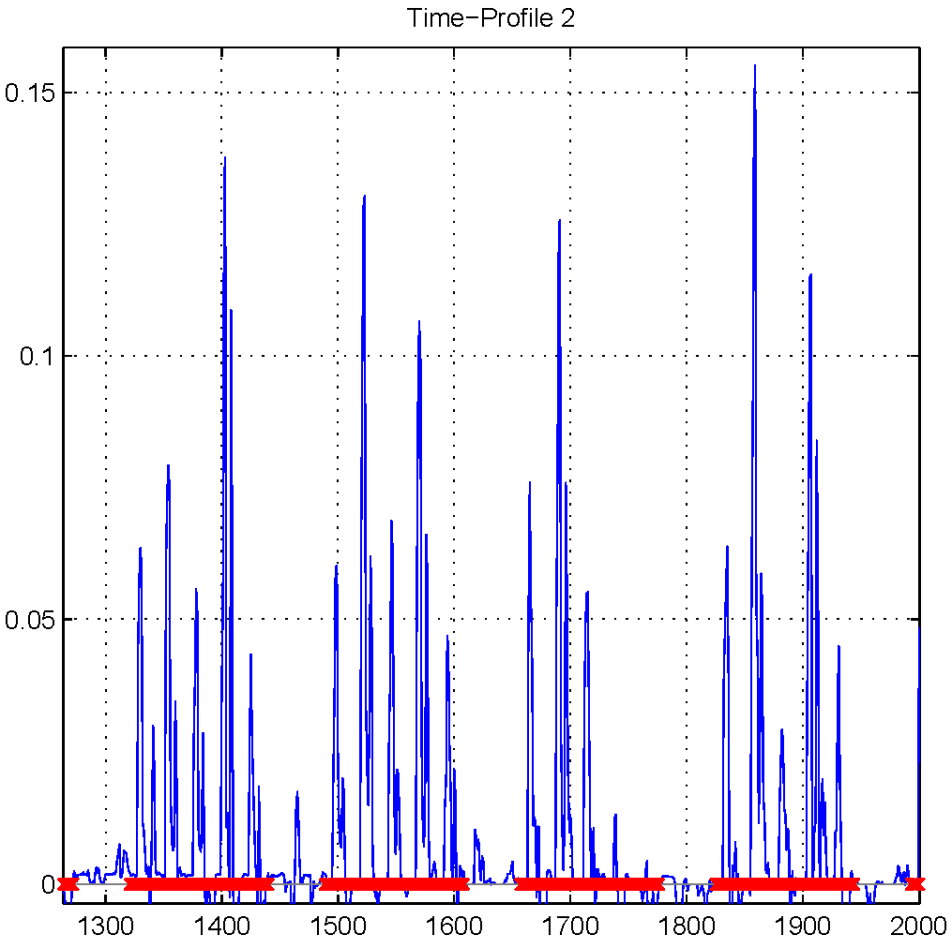
```
15.515 L01 E51-395 TR8.30-10
15.515 L02 E51-325 TR8.30-10
15.515 L04 E51-395 TR10.30-12
15.515 L05 E51-325 TR10.30-12
15.515 L06 E51-315 TR10.30-12
15.515 R01 E51-151 R3
15.515 R02 E51-085 R1
15.515 R03 E51-057 R1
15.515 R04 E51-085 R3
15.515 R05 E51-085 R4
15.515 R06 E51-315 R4
```

[5]

15.515 Financial Accounting
Prereq: —
G (Fall)
4-0-5

An intensive introduction to the preparation and interpretation of financial information. Adopts a decision-maker perspective on accounting by emphasizing the relation between accounting data and the underlying economic events generating them. Class sessions are a mixture of lecture and case discussion. Assignments include textbook problems, analysis of financial statements, and cases. Restricted to first-year Sloan Master's students. *R. Frankel, G. Plesko*

[6]



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Conclusion & Future Potential

- *Dynamic* social network analysis:
  - Interesting and "new" area at the tactical level
  - Amenable to many standard techniques in Machine Learning, Signal Processing, Social Network analysis

- Analysis Methods:
  - Good accuracy for detection of identity from network data
  - Tensor methods provide insight into patterns of life
  - Other analysis methods provide rules and generative models for dynamic data

- Future Potential:
  - Numerous applications for dynamic network data analysis
  - Fusion with other modalities
  - Large opportunity for doing novel work in this area

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

References

[1] N. Eagle and A. Pentland, “Reality mining: sensing complex social systems,” Personal and Ubiquitous Computing, vol. 10, no. 4, pp. 255–268, 2006.

[2] K. Carley, “Dynamic network analysis,” in Dynamic social network modeling and analysis: Workshop summary and papers, 2003, pp. 133–145.

[3] L. De Lathauwer, B. De Moor, J. Vandewalle et al., “A multilinear singular value decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 21, no. 4, pp. 1253–1278, 2000.

[4] J. Sun, D. Tao, and C. Faloutsos, “Beyond streams and graphs: dynamic tensor analysis,” in Proceedings
of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 374–383.

[5] MIT Fall 2004-2005 Class Schedule, Online, 2004. [Online]. Available: http://replay.waybackmachine.org/20041124034024/http:/web.mit.edu/registrar/www/schedules/csbindex.shtml#15

[6] MIT 2004-2005 Course Catalog, Online, 2004. [Online]. Available: http://web.mit.edu/catalog/archive/2004-2005/part 3.pdf

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY