Social Behavior Prediction Through Reality Mining

Charlie Dagli, William Campbell, Clifford Weinstein

Human Language Technology Group
MIT Lincoln Laboratory

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

**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

**Sociometric Badges**
(microphone, IR, accelerometer, ...)

**Smart Phones**
- Surveys
- Patient Records
- Call Info

**Analytics**

**Dynamic Network Analysis**

**Anomalous Behavior**

**Group Behavior Prediction**
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
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
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;
```

<table>
<thead>
<tr>
<th>myhn</th>
<th>date</th>
<th>time</th>
<th>areaid</th>
<th>cellid</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>2004-11-01</td>
<td>18:25:30</td>
<td>5119</td>
<td>40813</td>
</tr>
<tr>
<td>44</td>
<td>2004-11-01</td>
<td>18:26:01</td>
<td>5119</td>
<td>40332</td>
</tr>
<tr>
<td>44</td>
<td>2004-11-01</td>
<td>18:26:18</td>
<td>5119</td>
<td>40811</td>
</tr>
<tr>
<td>44</td>
<td>2004-11-01</td>
<td>18:27:13</td>
<td>5119</td>
<td>40332</td>
</tr>
</tbody>
</table>
Identification via Reality Mining Data

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

Typical Testing Scenario

Classifier

Individual Model

Score

Query

Data

Location sequence

DB

Java SQL Interface

Reality Mining
Features for Recognition

Input Stream

Window over k days

N-grams

Probabilities

Vector

V = [0 ... 0 1/7 0 ... 0 1/7 0 ... 0 ...]
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, Feb 1–March 1, 2005
  – TST3, March 1–April 1, 2005
  – TST4, April 1–June 1, 2005
Identification Experimental Results #1

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

Performance of 1 Day Test

Performance for Trigrams
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% !
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
Multi-linear (Tensor) Algebra

$$N_1 \times N_2 \times N_3 = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} y_{i,j,k}$$

$$X \approx \sum_{k=1}^{K_3} v_j w_k$$

$${\chi} \approx X_k$$
Toy Problem

\[ \mathcal{X} \]

\[ SN \text{ Nodes} \quad \text{SN Nodes} \quad \text{Time} \]

\[
\begin{align*}
\sum_{j=1}^{K_1} \sum_{k=1}^{K_2} \sum_{l=1}^{K_3} y_{i,j,k} & = u_i^T \mathbf{Y}^k \mathbf{v}_j \\
\end{align*}
\]

\[ \text{Rank-}(K_1,K_2,K_3) \text{ Approximation} \]
Reality Mining – Bluetooth Proximity

Time-marginalized Reality Mining Adjacency
Reality Mining

\[ \sum_{k=1}^{\frac{K}{2}} X^k \]

\[ N_1 \leftarrow N_3 \rightarrow W_k \]

Positively Correlated Social Interactions for Time-Profile 1

Positively Correlated Social Interactions for Time-Profile 2

Time-Profile 1

Time-Profile 2

4/27/2011
Course-Attending Behavior

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
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
References


