Predicting the Group: Data Science for Human Socio-Cultural Understanding and Prediction

Prof. Kathleen M. Carley

kathleen.carley@cs.cmu.edu
Big Data is Everywhere
Methodological Challenges

• sampled data
  – Available data may have 30% or more missing, and biases are not known
  – Metrics assume < 10% missing data, known biases

• geo-temporal data
  – Data is “big data” and it varies with time and location
  – Most tools metrics assume
    • Little data but many time periods
    • Lots of data but one time period
    • Controls for either location or time, not both
    • Spatial and temporal analytics are very computationally costly

• “wide” data
  – Few cases - lots of variables
  – Tools assume – lots of cases few variables
Supporting Technologies

- Social and Dynamic Network Analytics
  - Graph metrics & algorithms
  - Statistical metrics & algorithms
  - Simulation
- Visual Analytics
- Text Analytics
- Machine Learning

Analysis of who communicates, influences or did / will do what to whom - when, how, and why
Sampled Data

Tsunami Warning and Response
Proposed Tsunami Warning and Response System

Accoustic Modem

Detection

Pressure Sensor

Shore Island Station

Acoustic Network

Alert Sensor Data

Alert - verbal

Raspberry Pie

General Twitter Population

InaTews System

Populated, Twitter, and official Tweet Info

BMKG Tweet

Route Shelter Info

Route Shelter Info

Leader Cell Phone

Follower Cell Phone

Verified Info

Verified Info

Casos Web

Leader Cell Phone

Cell Phone

Population

Jakarta BMKG

PiTT Sees

Casos Web

Alert - verbal

Local Info

Local Info

Copywrite © 2015 Kathleen M. Carley

10/7/15
Tsunami Warning and Response Social Media System

PADANG INDONESIA

- Development: CASOS at Carnegie Mellon University
- Sponsor: the National Science Foundation
- Design: HTML5 UP
- Powered by ©ORA
Illustration of Where Tweets are Coming From

- 7 days
- Heatmap showing the number of tweets by area
  - Overall
  - By day or hour - different

Tweet Count Weekday Hour Report
Tweet Count Weekday Hour Report
Tweet Count Weekday Hour Report
Who is Most Likely to Know Others/Influence and Respect

Size = Out-Degree

1. McriskaMomo 29
2. editriaa 15
3. Aypekazta 13
4. viotaretha 10
5. Qwertyuiddi 8
Key Users
Geo-Temporal Data

Understanding State Stability
Benghazi

US consulate attack

BBC
2200: Attackers open fire.
22:15: Assailants gain entry to complex
main building engulfed in flames.
22:45: Security staff try to retake come under heavy fire retreat.
23.20: Attempt 2 to retake Success
Fighting moves to the annex.
Midnight: Fighting continues at annex – 2 die
01:15: Mr Stevens arrives at hospital
Dies from smoke inhalation.
02:30: Security forces regain control of annex.
Tweets Per Hour

note Egyptian attack is against a background of violent events, Libya and Yemen are anomalous

- News and Twitter peak near the same time
- Strong periodicity to the data
- Dramatic variation by country
- Most retweeted are news but they are small fraction of actors
Overall Tweet Network

Note there are a few sources that are picked up
Benghazi Consulate

Innocence of the Muslims

Terrorism

Number of Tweets

Egypt

Libya

9/11/12 0:00
9/11/12 4:00
9/11/12 8:00
9/11/12 12:00
9/11/12 20:00
9/12/12 0:00
9/12/12 4:00
9/12/12 8:00
9/12/12 12:00
9/12/12 20:00
9/13/12 0:00
9/13/12 4:00
9/13/12 8:00
9/13/12 12:00
9/13/12 20:00
9/14/12 0:00
9/14/12 4:00
9/14/12 8:00
9/14/12 12:00
9/14/12 20:00
9/15/12 0:00
## Trust & Source

<table>
<thead>
<tr>
<th>Country</th>
<th>Adaptation</th>
<th>Nationalities</th>
<th>Protest</th>
<th>Terrorist Orgs</th>
<th>Ethnic Groups</th>
<th>Violence</th>
<th>War</th>
<th>Youth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jordan</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Kuwait</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>UAE</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Yemen</td>
<td>-2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Bahrain</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Egypt</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Iran</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>Iraq</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>Lebanon</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Libya</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>Syria</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>Tunisia</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

-2 twitter and not news

-1 more twitter than news

0 neither

1 more news than twitter

2 news and not twitter

- **Low civil unrest**
  - News & Twitter distinct
  - Ethnic groups more discussed in twitter

- **High civil unrest**
  - More twitter on war, ethnics, terrorists
  - Less reliance on news
Relation of Media Users to Revolutionary Activity

- Similar number high similarity users

- Low Civil unrest – personalized discussion
  - More users, more users with high similarity in non-Arabic tweet, more “social interaction” through replies in small groups

- High Civil unrest – get the message out!
  - Lower similarity among non-Arabic tweeters, less “social” interaction and more inciting through retweets
Change of Number of Tweeters after Suspended Users are Removed

- After Oct2012 or 10% or more tweeters are suspended
- Algeria often has 20-25% tweeters suspended
- Change in political reality
  - more violent states have fewer suspended
  - Less violent states have more suspended
Change of Closeness after Suspended Users are Removed

- Removing suspended tweeters has mixed impact
- In Qatar and Kuwait there is a .5 increase in closeness
- In Algeria closeness drops sometimes by .5
Impact on Network Metrics

- 25% nodes removed – chance you predicted top 3 correct drops to 60%
- 25% extra nodes (undetected bots) – chance you predicted top 3 correct drops to 80%
Wide Data

Cyber Attacks, ELFS –v- Trolls and ISIS
Problem and Approaches

- Small Signal, often spatially and temporally distributed
- Ocean of data
- Interesting “stuff” is in the middle

- Spectral Clustering
  - Combine different kinds of things such as network position and use of certain words

- Lasso
  - Often used in machine learning – to identify the network of connections among variables

- Geo-temporal visualization
  - Movies!
Web Site Threats Encountered
Exploits: Countries that act as wayports

- Latvia
- Moldova
- Georgia
- Croatia
- Argentina
- Niger
- Dem. Rep of Congo
- Angola
• Former USSR mostly attack each other in Cyber space
• Strongest attacks are on Russia, and from Ukraine
• In social media Russia mainly talks to itself and Belarus
• Ukraine, Estonia, Lithuania, Latvia use social media to build NATO sympathy and share approach
• Dirty bombs may be associated with cyber attacks
Anti-Russian “Elf Army” on Twitter

“Secret Forces Rising for Fight Against Kremlin Trolls”

As a response to the so called Kremlin trolls, who are very active on the internet, in Lithuania the elves have appeared - active, educated people who sacrifice their free time for the fight against Vladimir Putin regime's recruits from the Russian troll factories, who spread lies and slander in cyberspace.

Excerpt from report by Ainis Gurevicius on Lithuanian news website Delfi on 21 September

- The CASOS Lab at Carnegie Mellon University has identified ~4000 Twitter accounts engaged in anti-Russian dialogue on Twitter
- These accounts tweet predominantly in Russian, Ukrainian and English and focus on Russian intervention in the Ukraine and Syria
Group’s activity driven by events
Joining is through following Imams

Reciprocal Mentions: Trust
Arabic tweeters have since been suspended
Summary

- Data is everywhere
- New methodologies are changing the questions that can now be answered
- Understand what the data really represents – sampling
- Move from geo-correlation to geo-interpretation
- Move from temporal-correlation to temporal forecasting
- Move from standard statistics to wide data techniques
- Move from standard social network analysis to high dimensional networks