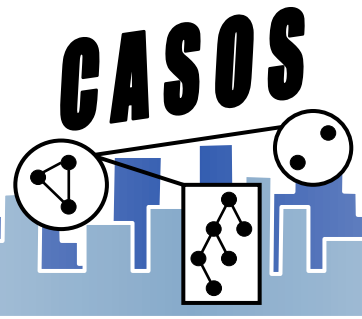


Predicting the Group: Data Science for Human Socio-Cultural Understanding and Prediction

Prof. Kathleen M. Carley

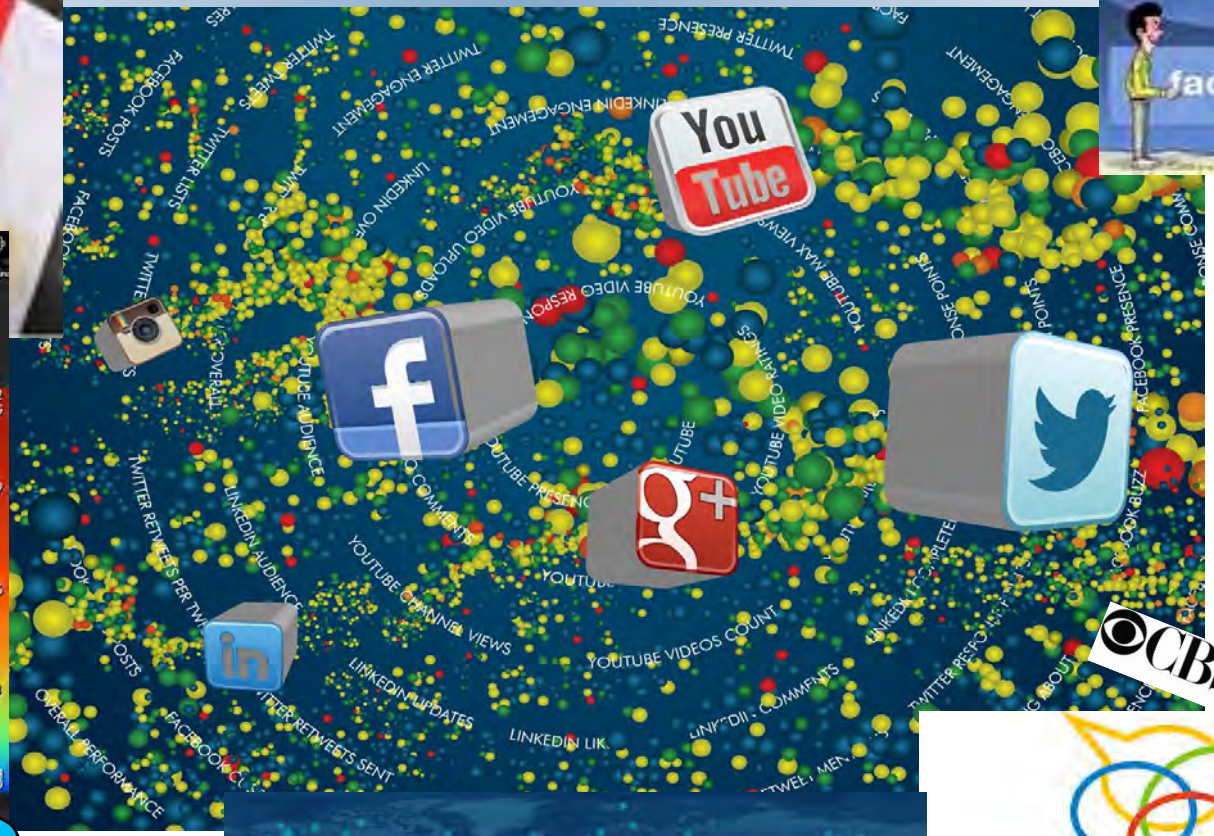
kathleen.carley@cs.cmu.edu

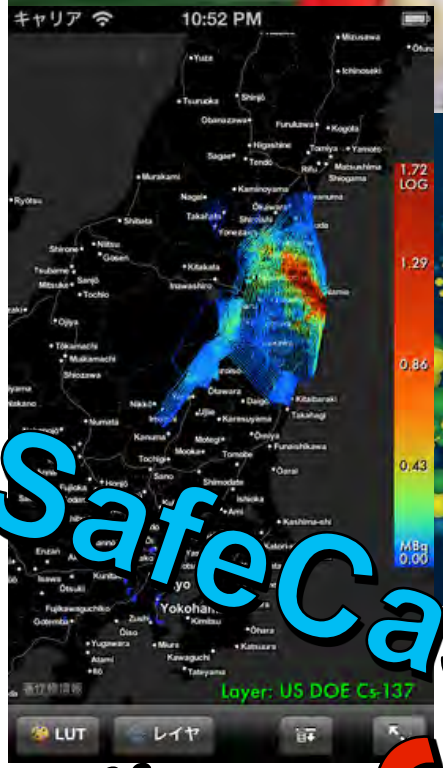
institute for SOFTWARE RESEARCH

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Big Data is Every Where





SafeCast

GDELT

CBS NEWS

OpenGovHub

Transparency | Community | Technology

CASOS

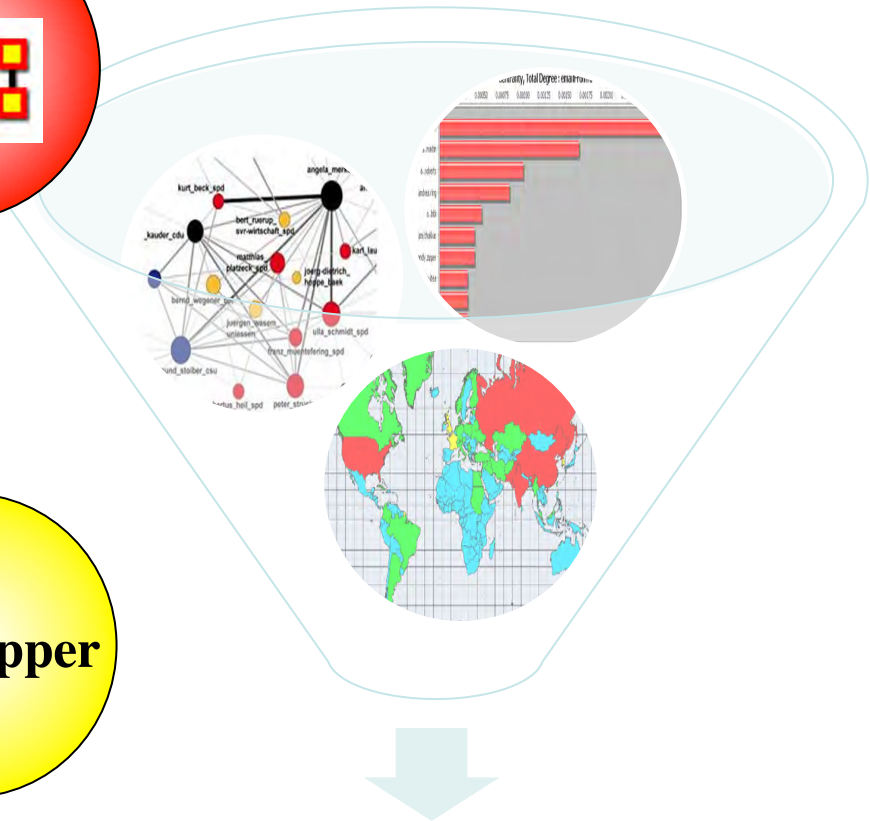
Methodological Challenges

- 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

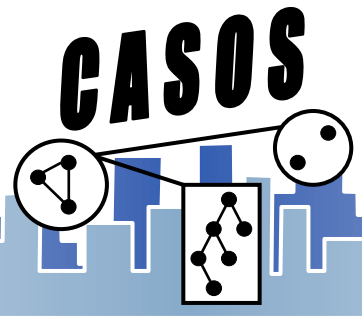
- Social and Dynamic Network Analytics
  - Graph metrics & algorithms
  - Statistical metrics & algorithms
  - Simulation
- Visual Analytics
- Text Analytics
- Machine Learning

ORA



NetMapper

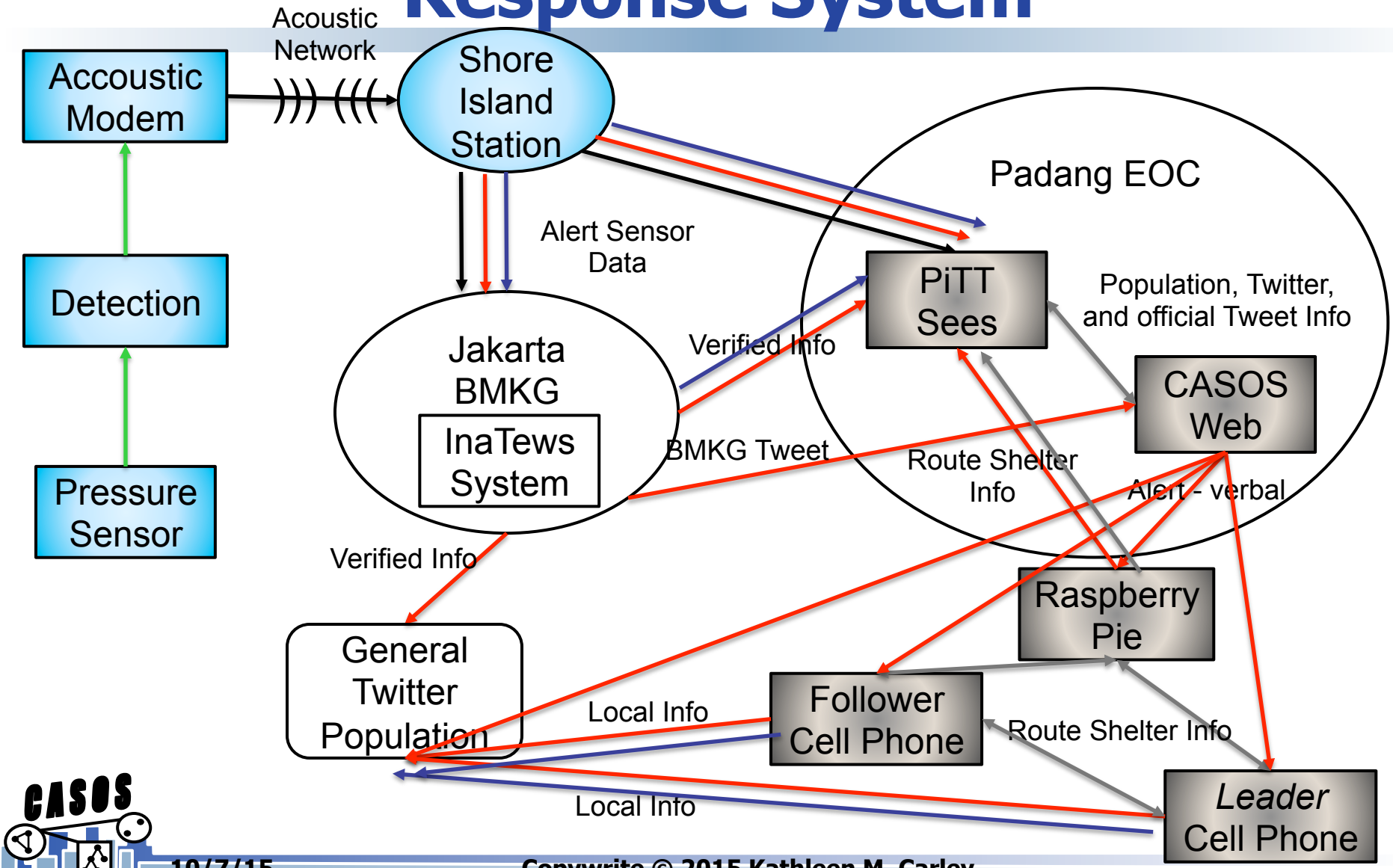
Analysis of who communicates, influences or did / will do what to whom - when, how, and why



Sampled Data

Tsunami Warning and Response

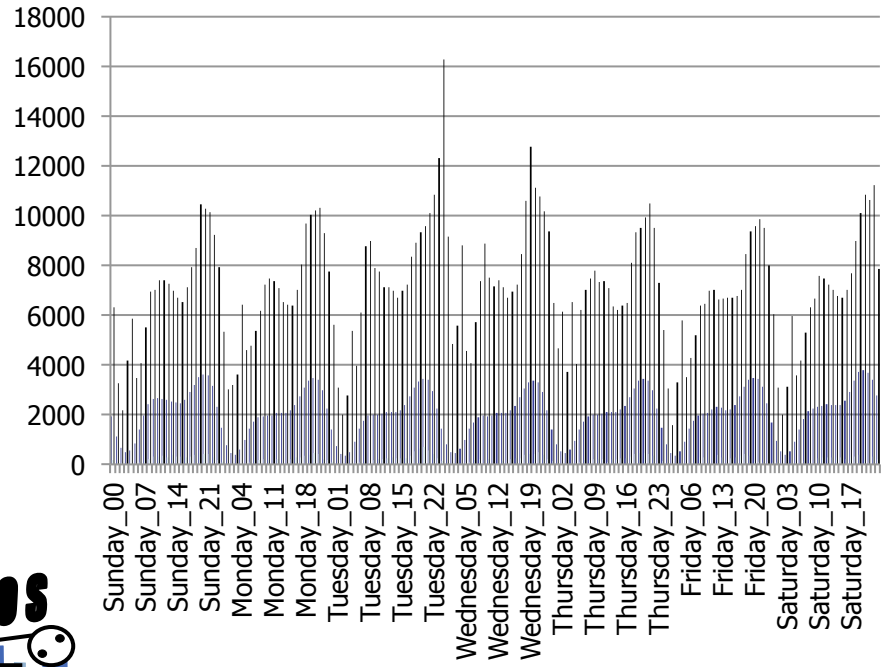
Proposed Tsunami Warning and Response System

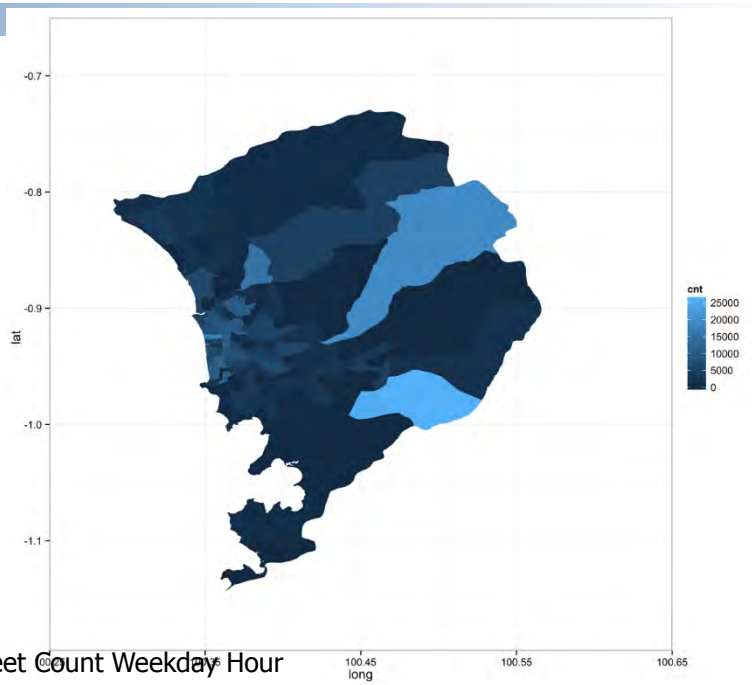


Tsunami Warning and Response Social Media System

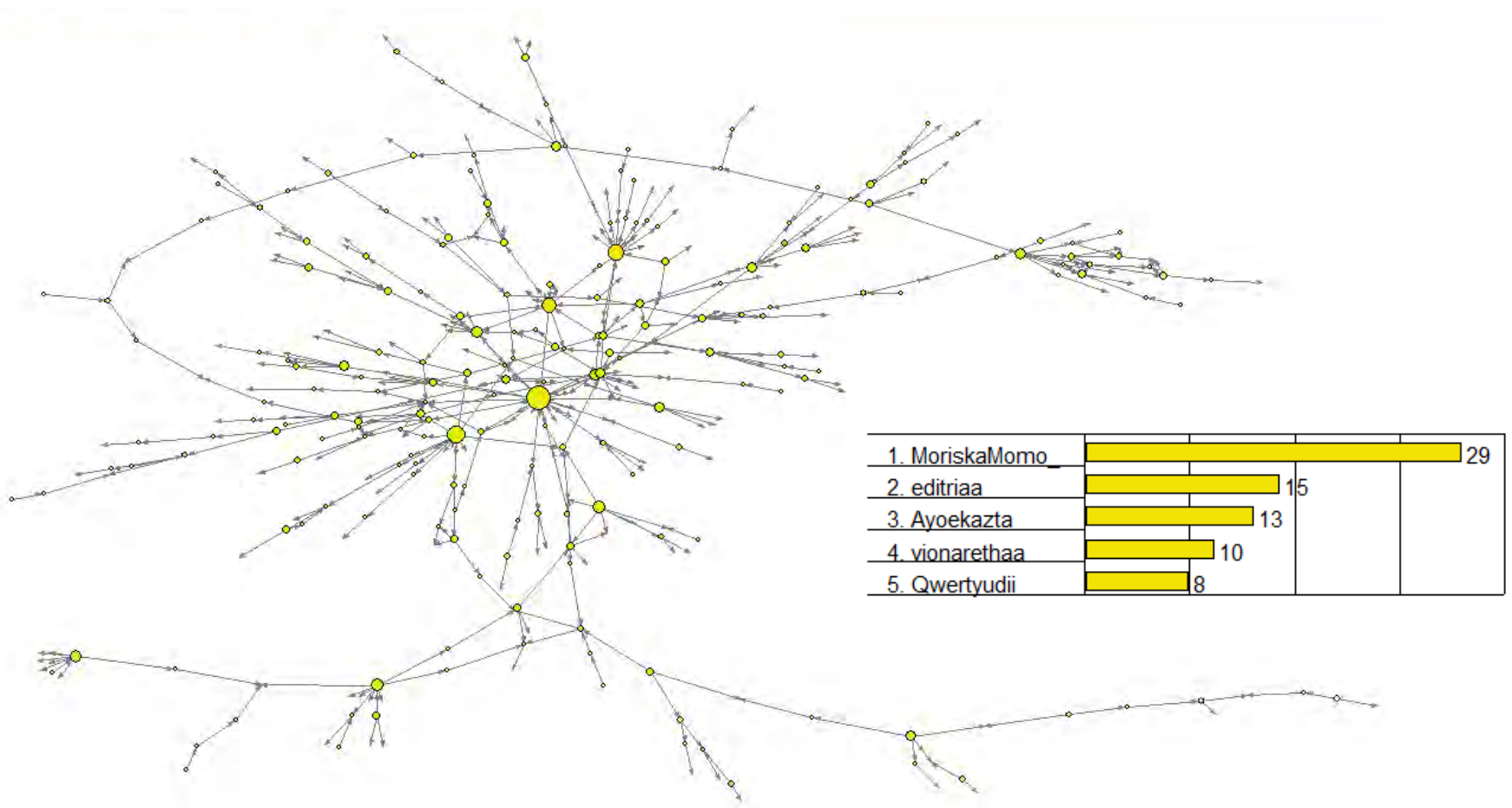
Illustration of Where Tweets are Coming From

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





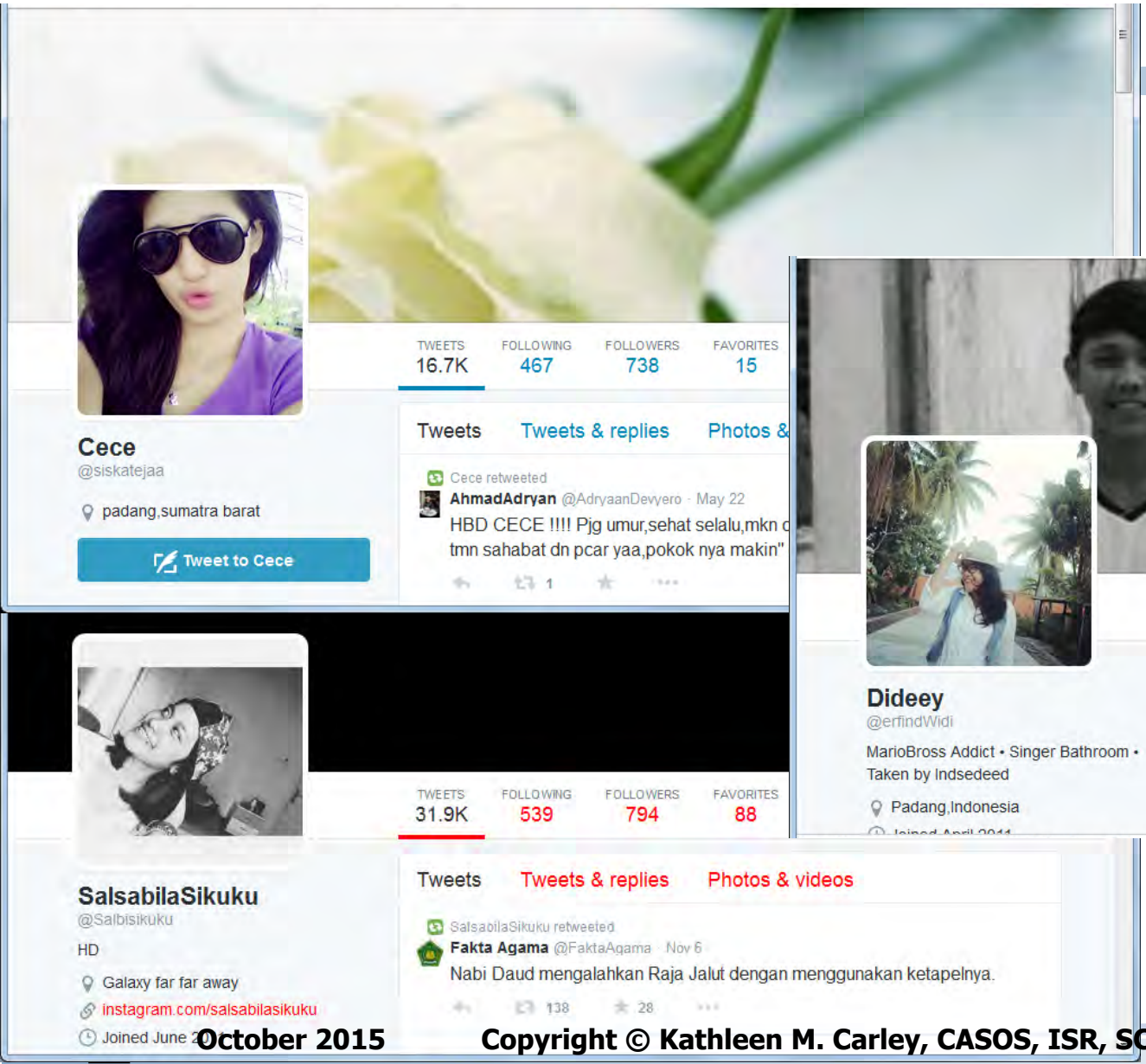
Who is Most Likely to Know Others/Influence and Respect

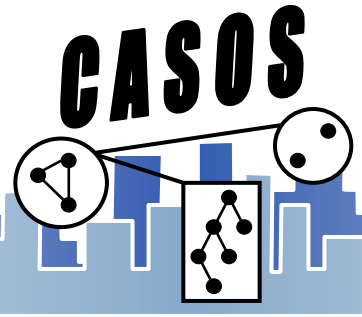


Size = Out-Degree

Carnegie Mellon
institute for SOFTWARE RESEARCH

Key Users





Geo-Temporal Data

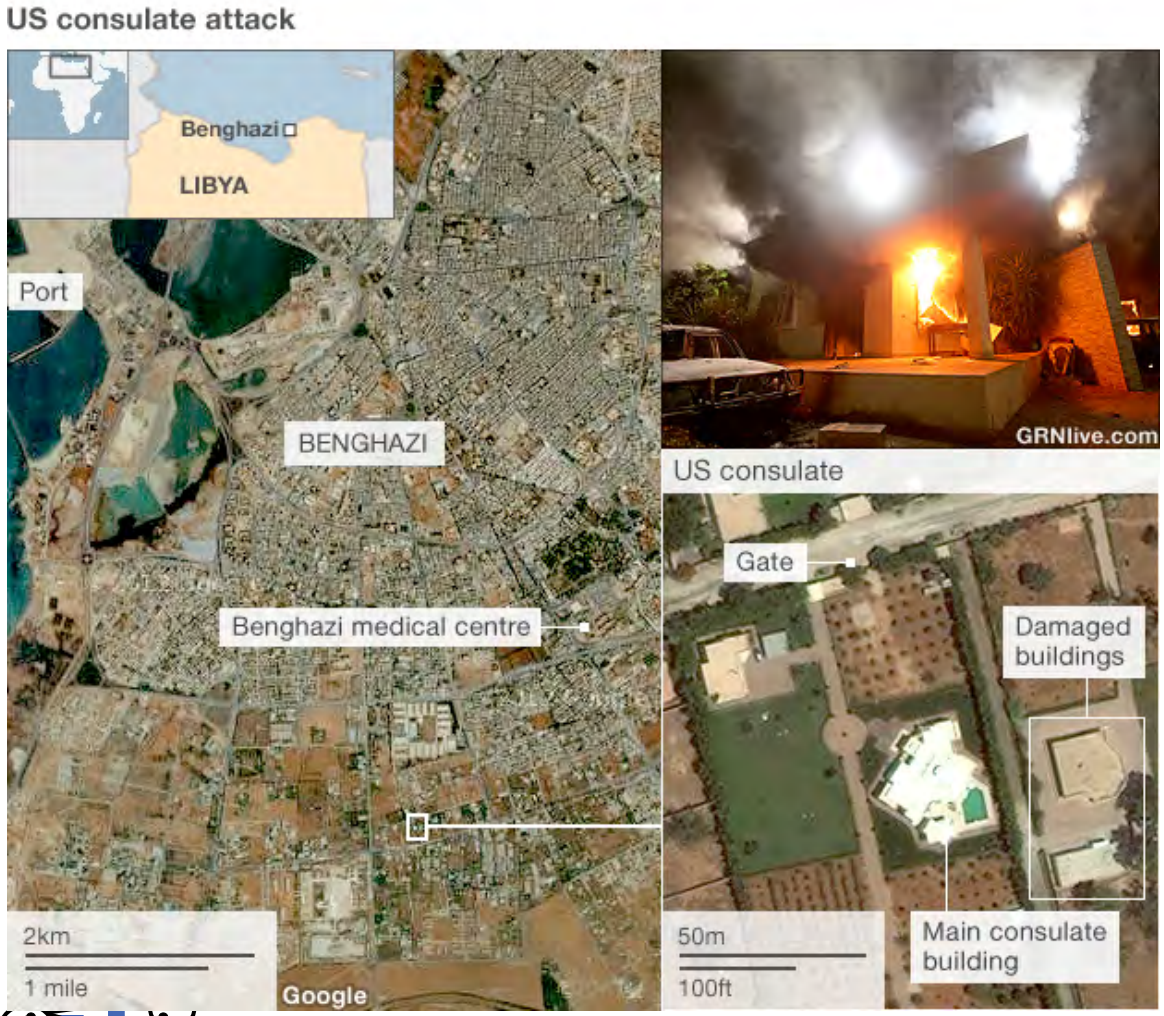
Understanding State Stability

institute for SOFTWARE RESEARCH

Carnegie Mellon

Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Benghazi



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

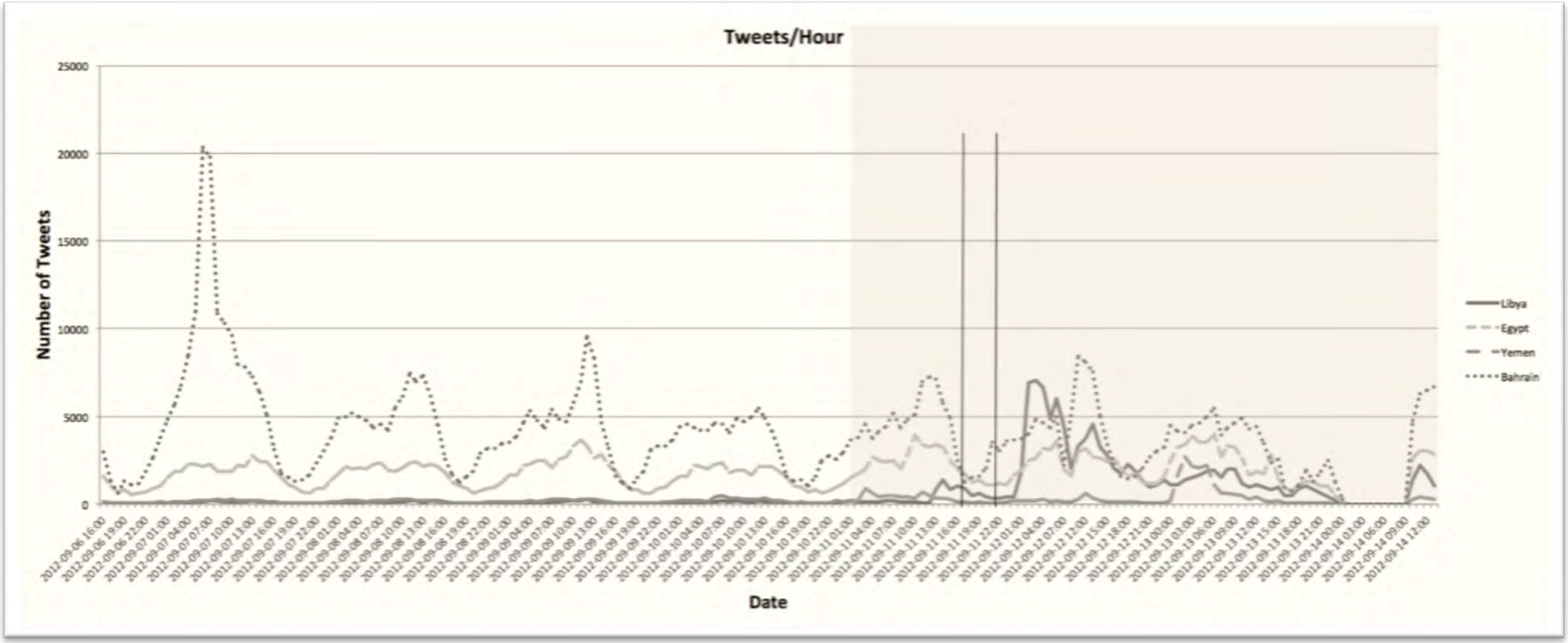
01:15: Mr Stevens arrives at hospital

Dies from smoke inhalation.

02:30: Security forces regain control of annex.

Tweets Per Hour

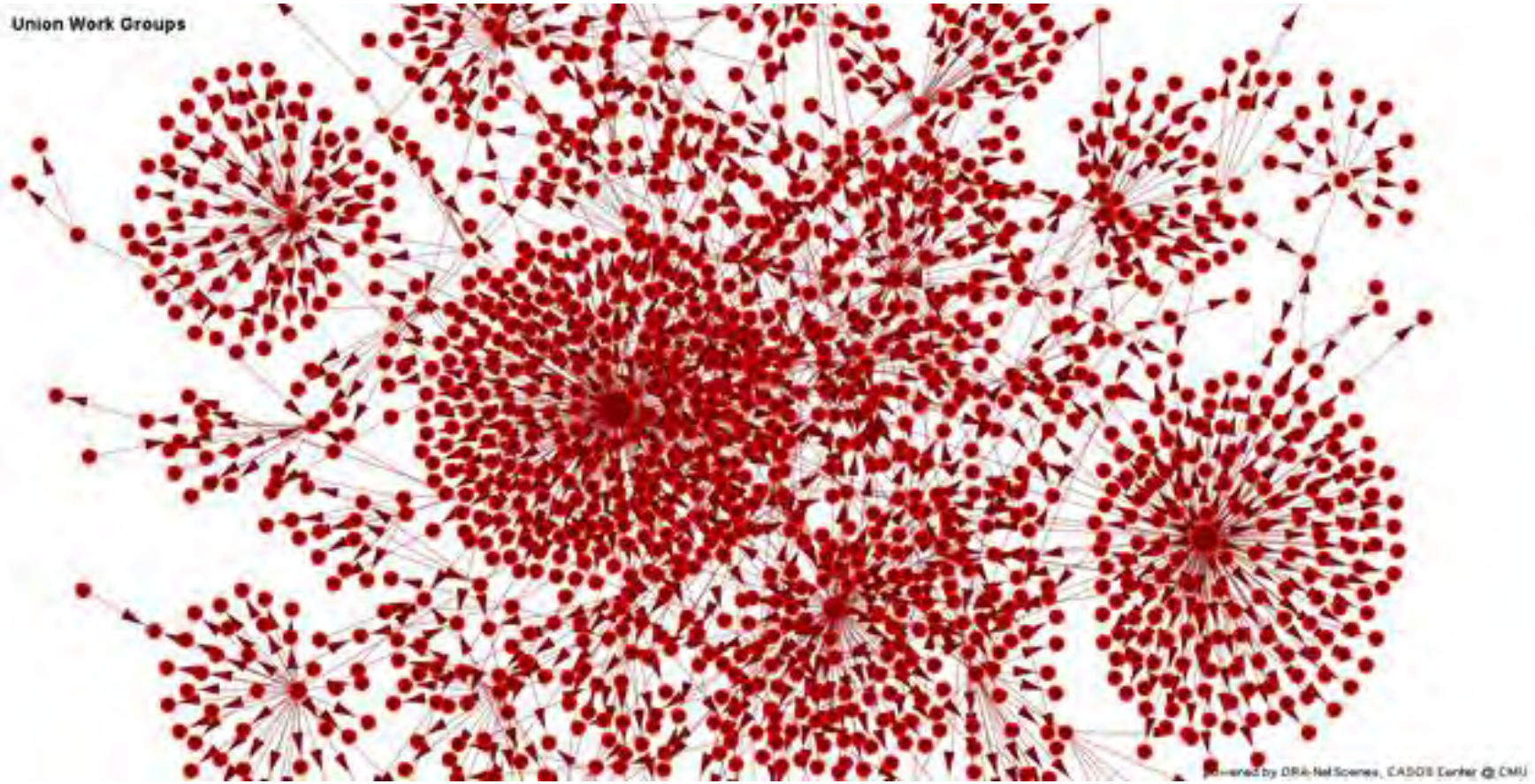
note Egyptian attack is against a background of violent events, Libya and Yemen are anomalous



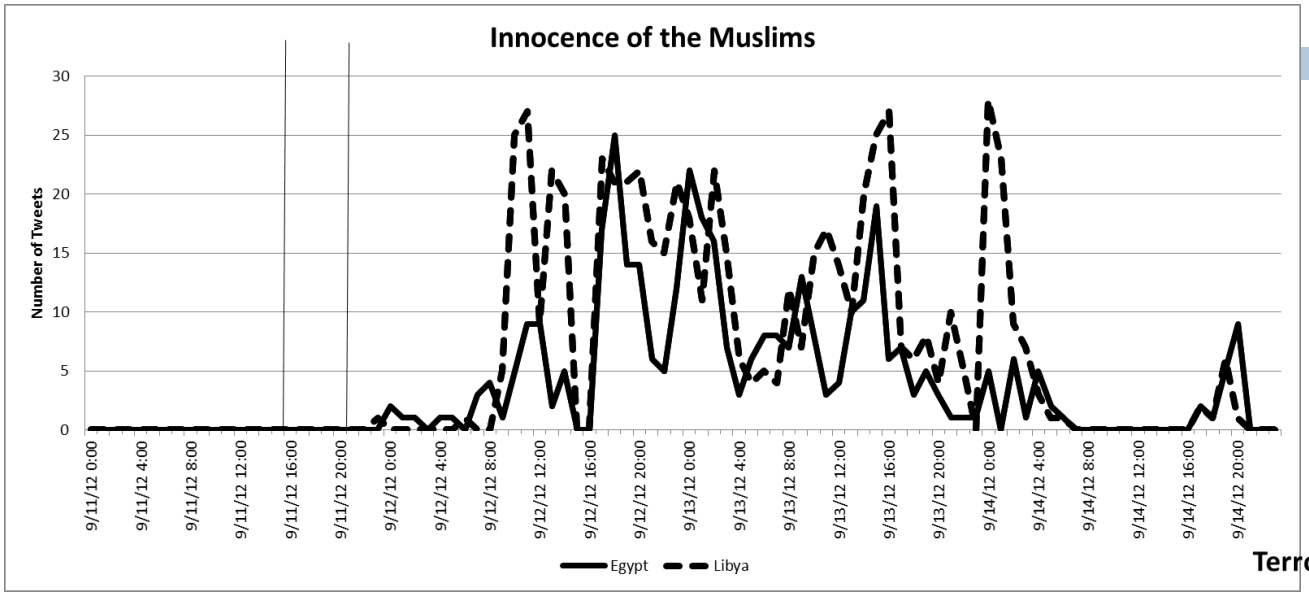
- News and Twitter peak near the same time
- Strong periodicity to the data
- Dramatic variation by country
- Most retweeted are news but they are small fraction of actors

Overall Tweet Network

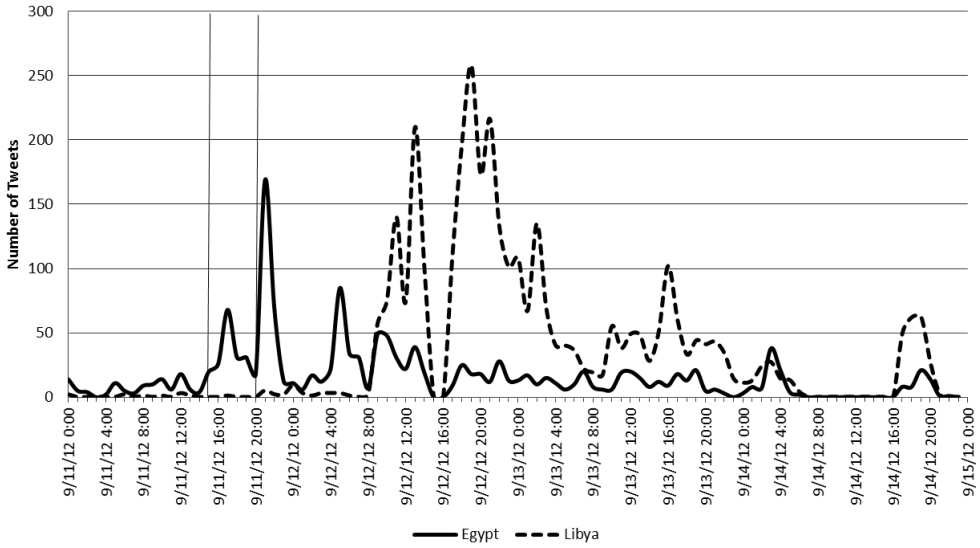
Note there are a few sources that are picked up



Benghazi Consulate



Terrorism



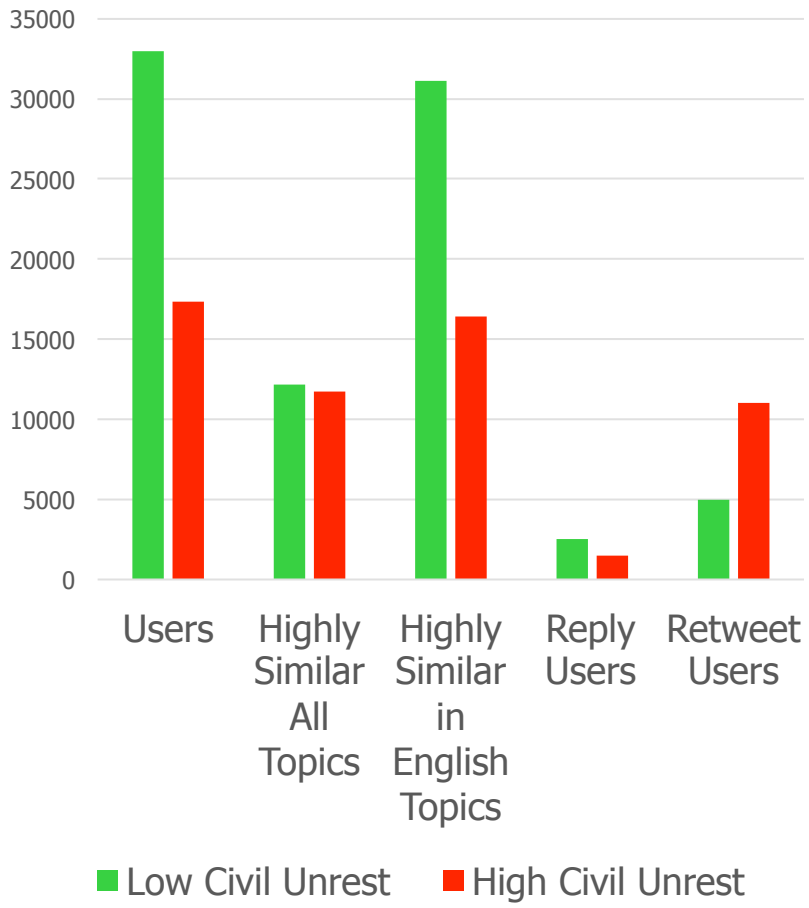
Trust & Source

	Adaptation	Nationalities	Protest	Terrorist Orgs	Ethnic Groups	Violence	War	Youth
Jordan	0	-1	0	2	1	0	0	2
Kuwait	2	2	0	0	-2	0	0	0
Saudi Arabia	2	0	0	2	-1	0	0	2
UAE	2	0	0	0	-2	0	0	2
Yemen	-2	2	1	1	2	2	-1	1
Bahrain	0	0	1	0	-1	2	0	2
Egypt	1	0	1	1	-2	2	0	1
Iran	0	1	0	-2	1	0	-2	0
Iraq	0	0	0	-1	-2	0	-1	0
Lebanon	0	1	0	-1	1	2	2	0
Libya	1	1	0	0	0	0	-1	0
Syria	0	1	1	2	0	1	-2	0
Tunisia	1	1	2	0	-2	2	0	1

-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

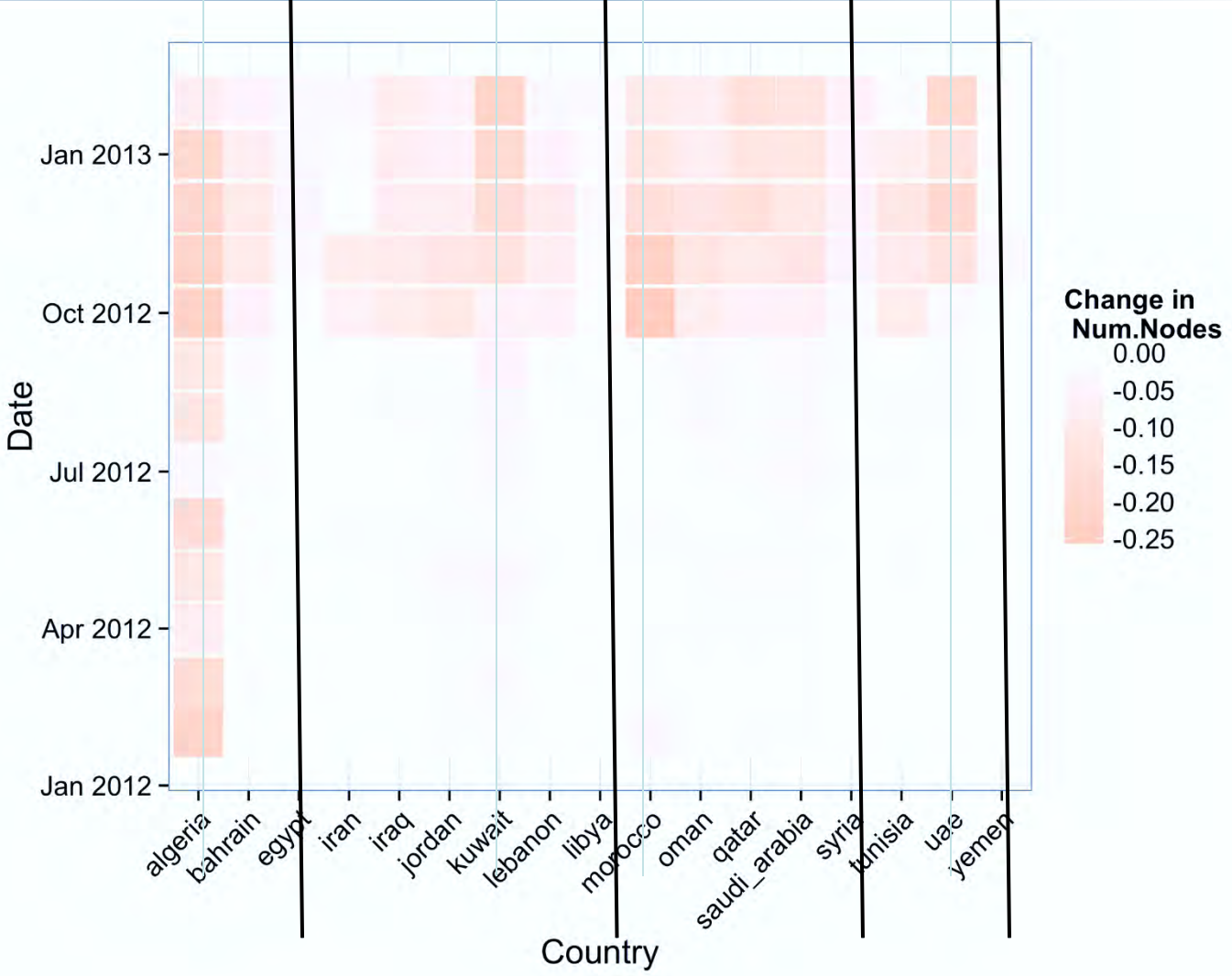
Relation of Media Users to Revolutionary Activity



- 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

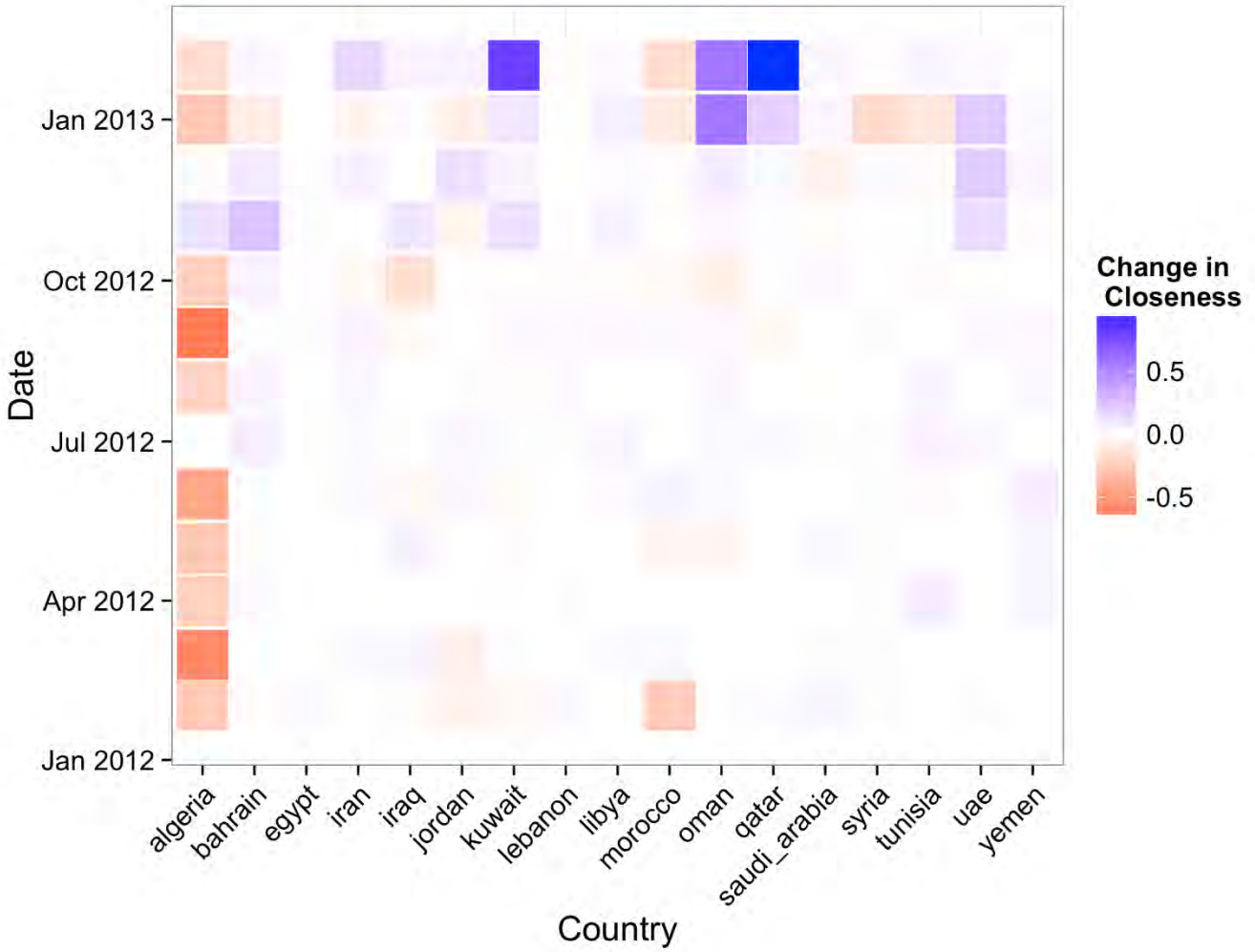
Change of Number of Tweeters after Suspended Users are Removed

- 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



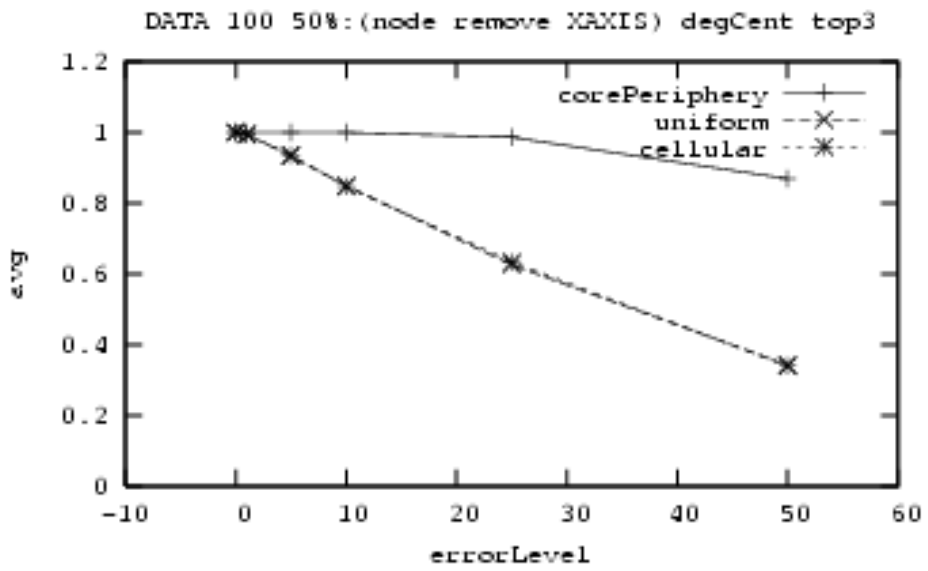
Change of Closeness after Suspended Users are Removed

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

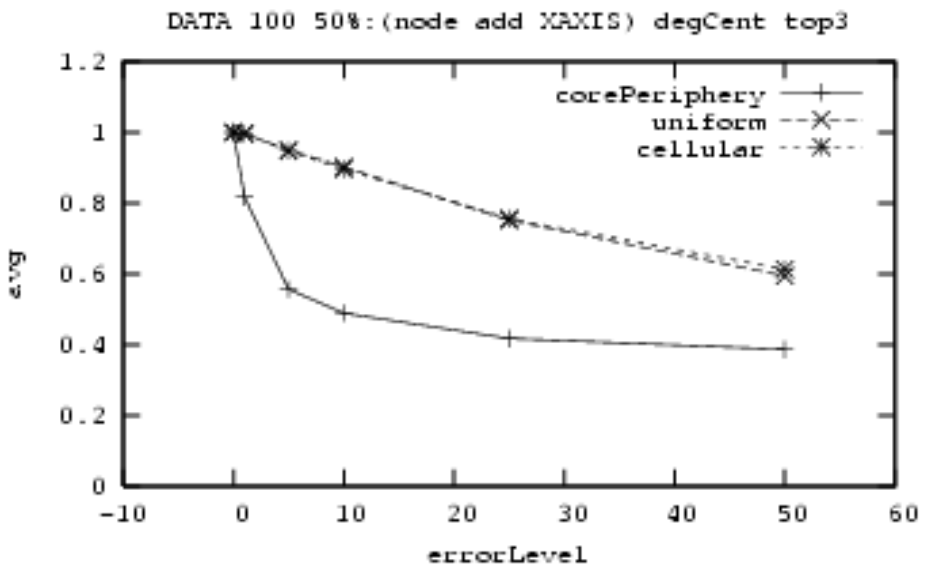


Impact on Network Metrics

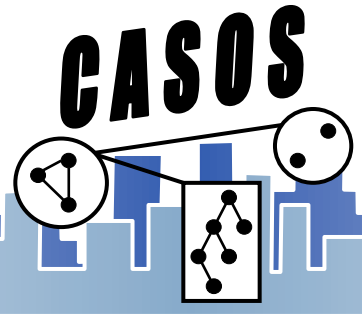
NODE REMOVE



NODE ADD



- 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

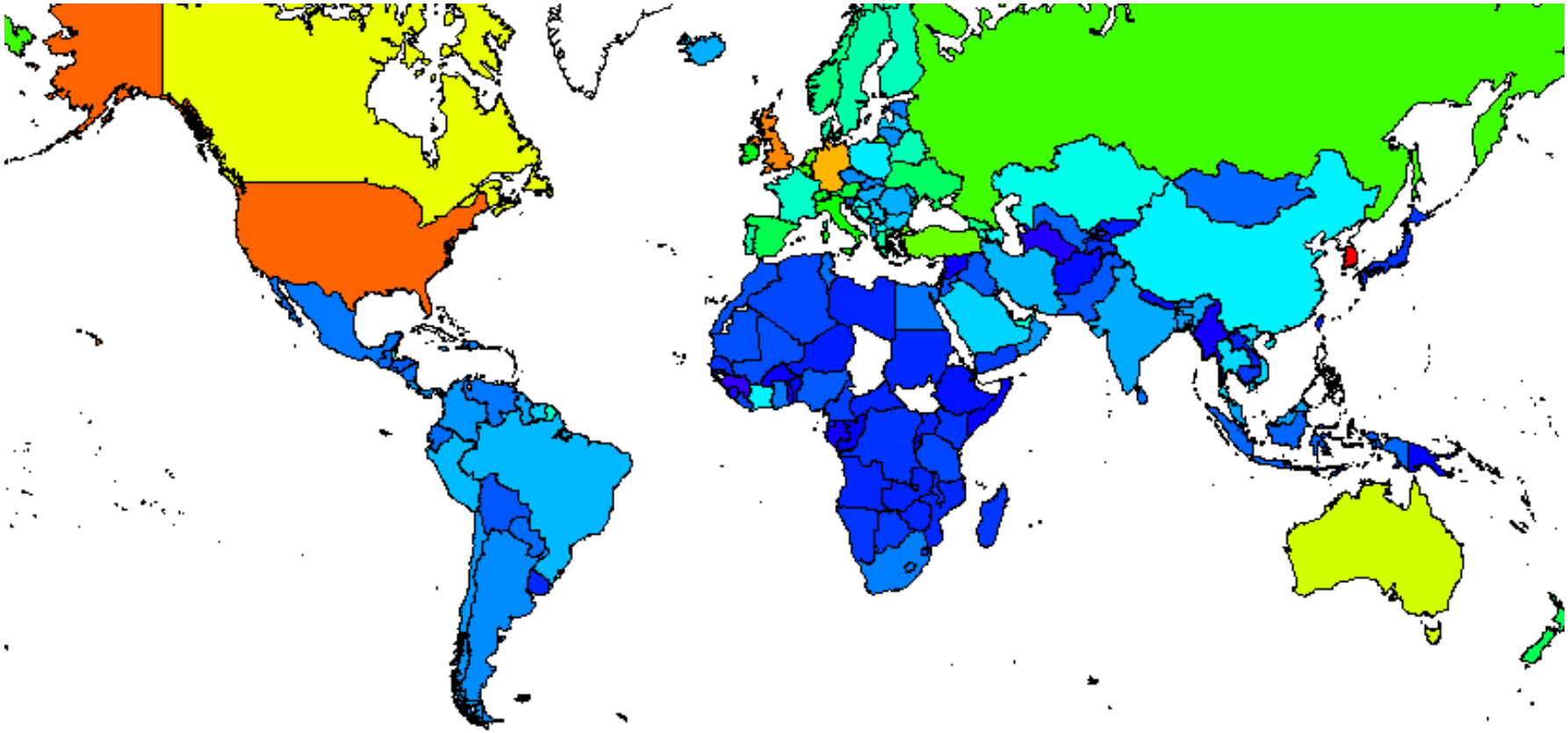
Cyber Attacks, ELFS –v- Trolls and ISIS

Problem and Approaches

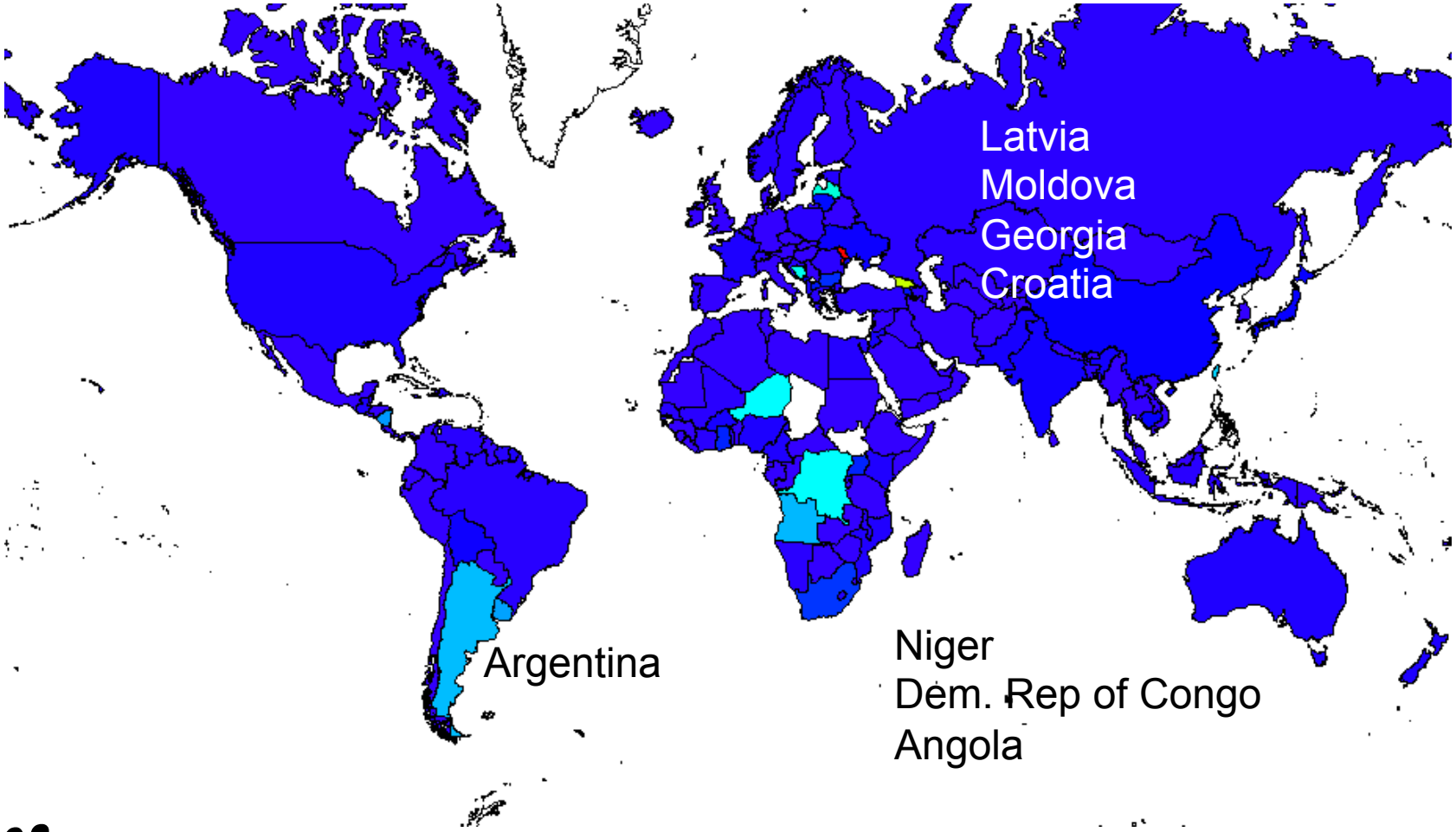
- Small Signal, often spatially and temporally distributed
- Ocean of data
- Interesting "stuff" is in the middle

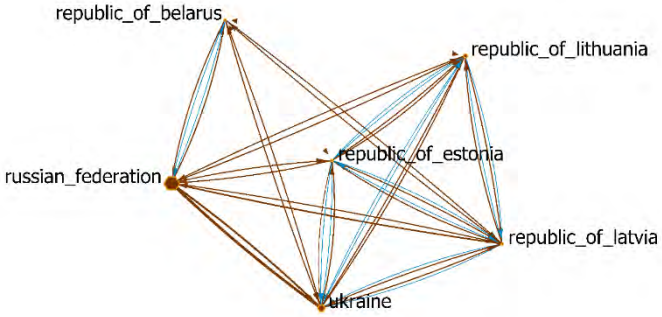
- Spectral Clustering
  - Combine different kinds of things such as network position and use of certain words
- Lasso
  - Often used in machine learning – to identify the network of connections among variables
- Geo-temporal visualization
  - Movies!

Web Site Threats Encountered

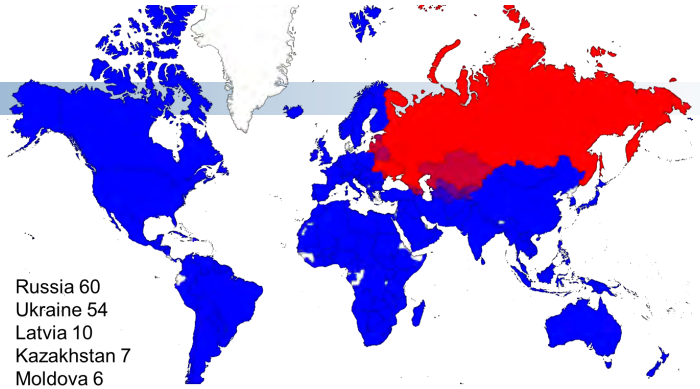


Exploits: Countries that act as wayports

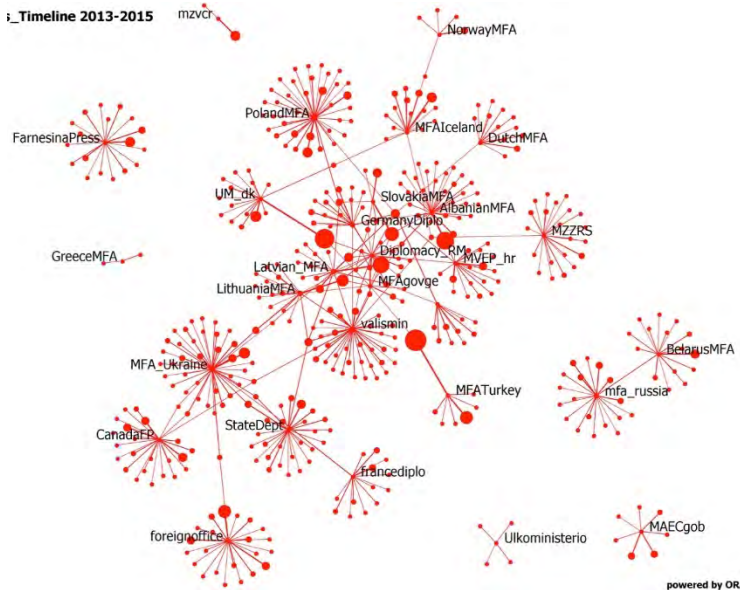




Cyber Connection



- 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



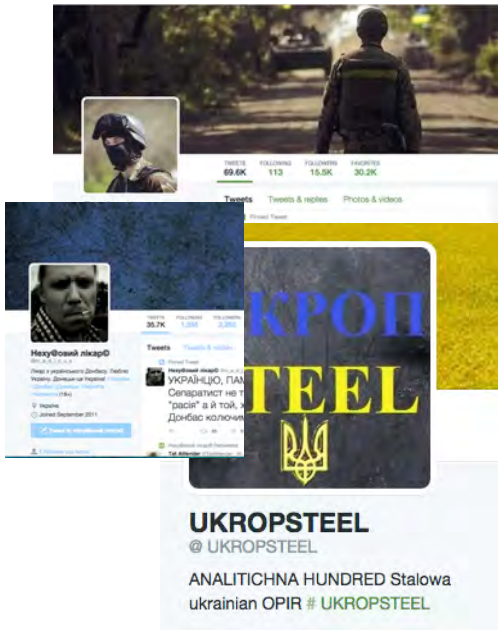
Carnegie Mellon
institute for SOFTWARE RESEARCH

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.

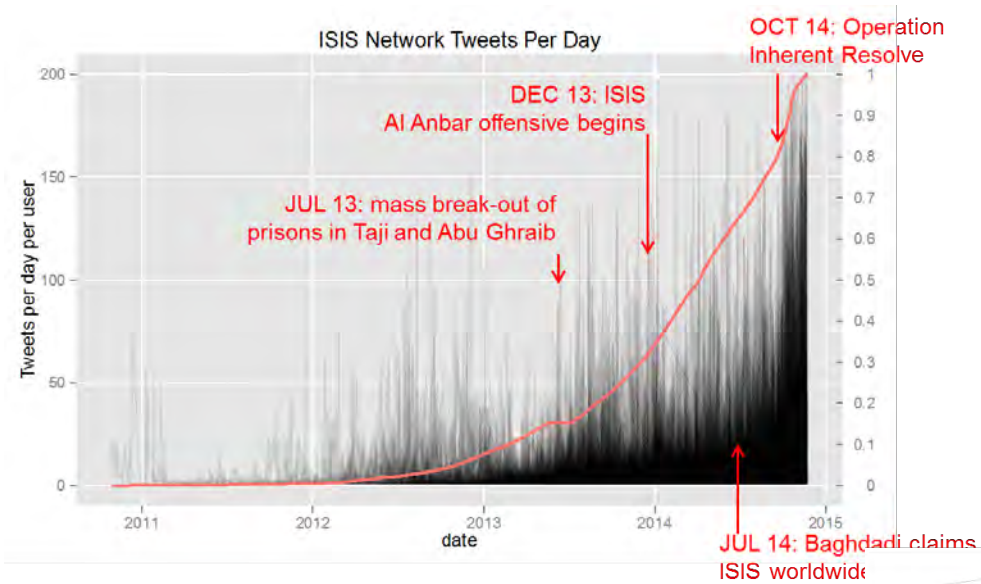
Excerpt from report by Ainis Gurevicius on Lithuanian news website Delfi on 21 September

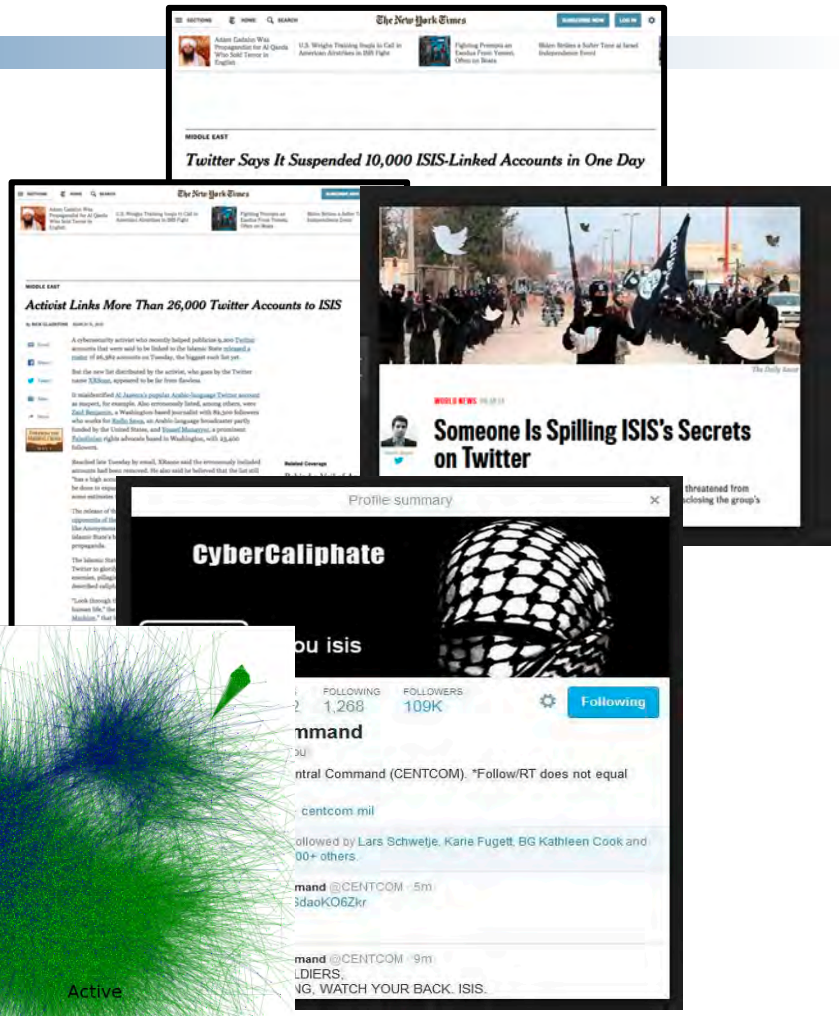


- The CASOS Lab at Carnegie Mellon University has identified ~4000 Twitter accounts engaged in anti-Russian dialogue on Twitter
- These accounts tweet predominantly in Russian, Ukrainian and English and focus on Russian intervention in the Ukraine and Syria

CASOS

ISIS Social Media Network





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

