Social Media Mining and Retrieval

Carlos Castillo

http://chato.cl/ http://bigcrisisdata.org/

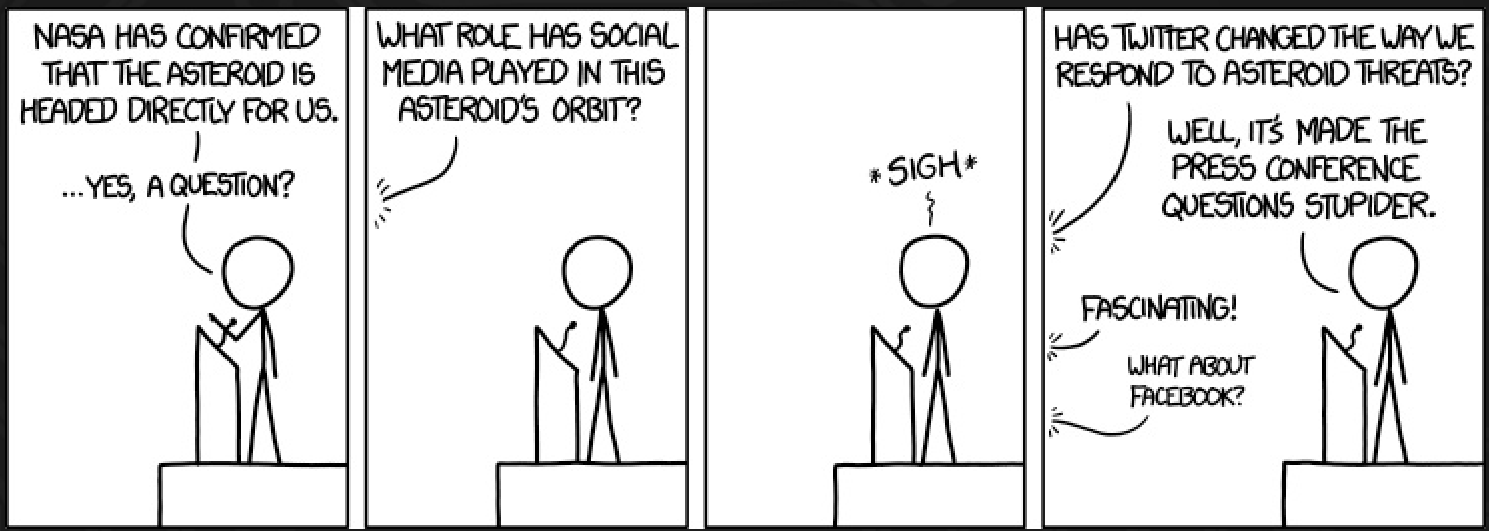
@chatox @bigcrisisdata

Outline

- Part 01: Preliminaries
- Part 02: Social media mining
- Part 03: Social networks
- Part 04: Information cascades
- Social media and natural disasters

Social Media Mining and Retrieval
Part 01: Preliminaries

Social media changes *everything*



https://xkcd.com/1239/

An attractive topic

- If you work in IR sooner or later you'll be dealing with documents from social media
- Many in science, technology and engineering have also interest in the humanities
  - Plus a bit of actual formal education on the subject
  - Plus a ton of intuitions, a few of them correct

Cut-off between 1 and 2 is 30 years old for social media.

An attractive topic (cont.)

- Social media is a “young” technology (~10 to 15 years old)
- Douglas Adams on new technologies:
  - Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works.
  - Anything that's invented between when you're 15 and 35 is new and exciting and revolutionary and you can probably get a career in it.
  - Anything invented after you're 35 is against the natural order of things.

This talk is about ...

- Social software
  - Software to facilitate or mediate social interactions
- Social networking sites
  - Web applications to maintain social connections
- Social media sites
  - Web applications to create, share, and exchange content
- Social media content
  - The content shared by users in social media platforms

Example

"Media must report about d alleged 20k RSS chaps off 2 #Nepal.here's a pic coz d 1 @ShainaNC shared isn't true.. ;)"

Example

"Media must report about d alleged 20k RSS chaps off 2 #Nepal.here's a pic coz d 1 @ShainaNC shared isn't true.. ;)"

Social media messages

- Social media is more like a transcript of a conversation than like text meant to stand on its own
  - Awkward entry methods:
    - Fragmented language and incomplete sentences
    - Many typographic and grammatical errors
  - Conversational:
    - Little or no context (hard to comprehend in isolation)
    - Code switching and borrowing
    - Internet slang

Slang

:)	Small smiley	ATB	All the best	EVRY1	Everyone	L8R	Later	SPK	Speak
:-)	Big smiley	ATM	At the moment	FTTB	For the time being	LMK	Let me know	SUM1	Someone
:-D	Laughter	B	Be	FYI	For your info	LOL	Laughing out loud	SUP?	What's up?
:-X	Kiss	B4	Before	GR8	Great	LUV	Love	THX	Thanks
;-)	Winking smiley	B4N	Bye for now	GTG	Got to go	LYL	Love you lots	U	You
:-(	Sad face	BCNU	Be seeing you	H&K	Hug and kiss	M8	Mate	UR	You are
:-O	Surprised	BCOZ	Because	VH8	Hate	MOB	Mobile	URAQT	You are a cutie!
4	For	BRB	Be right back	IAC	In any case	MSG	Message	WIV	With
+LY	Positively	BRT	Be right there	IDK	I don't know	NE1	Anyone	WKND	Weekend
2DAY	Today	BTW	By the way	IMO	In my opinion	NO1	No-one	WOT	What's up?
2MORO	Tomorrow	CIO	Check it out	IOH	I'm outta here	NRN	No reply necessary	XOXOX	Hugs and kisses
2NITE	Tonight	CSL	Can't stop laughing	IOW	In other words	OIC	Oh I see	YNK	You never know
AFAIK	As far as I know	CUL8R	See you later	IYD	In your dreams	PLS	Please		
AMBW	All my best wishes	DGT	Don't go there	KIT	Keep in touch	R	Are		
ASAP	As soon as possible	DKDK	Don't know, don't care	L8	Late	RU OK?	Are you okay?		

Alternatives to traditional text proc.

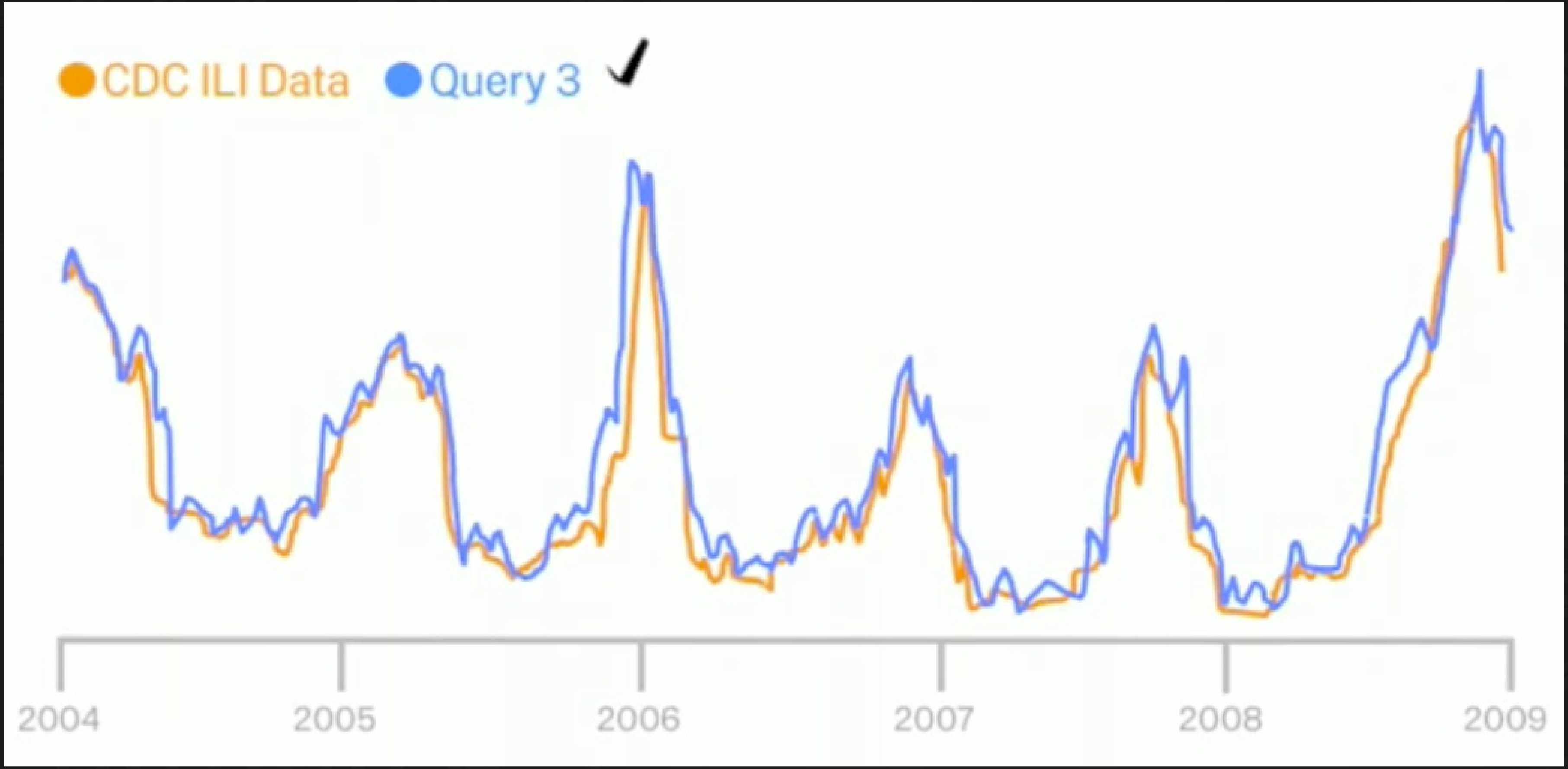
- Change the methods
  - Develop new methods that are aware of these particularities
- Change the queries and/or the documents
  - Pre-process: "r u ok m8" → "Are you OK, mate?"
- Change both

Social Media Mining and Retrieval
Part 02: Mining

Why mining social media?

- "What do people think / how do they feel about X?"
  - Sentiment analysis and opinion mining
- An alternative to traditional opinion polls?
- Attractive for many reasons including:
  - Lower latency (waiting time)
  - Lower cost
  - Larger population

Template: Google Flu Trends



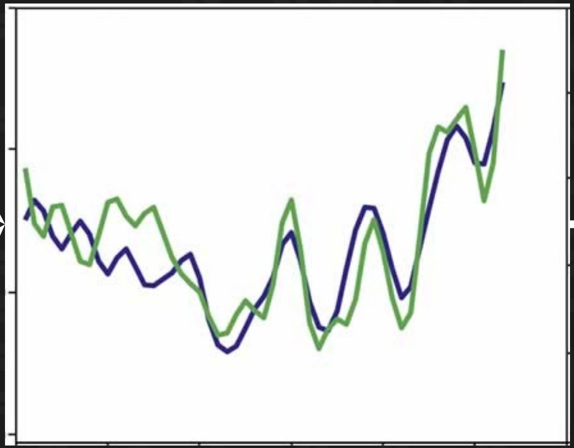
Low latency: people query symptomps before going to the doctor!

Many social media mining papers

Domain-specific data

Social media data

Correlation/Influence

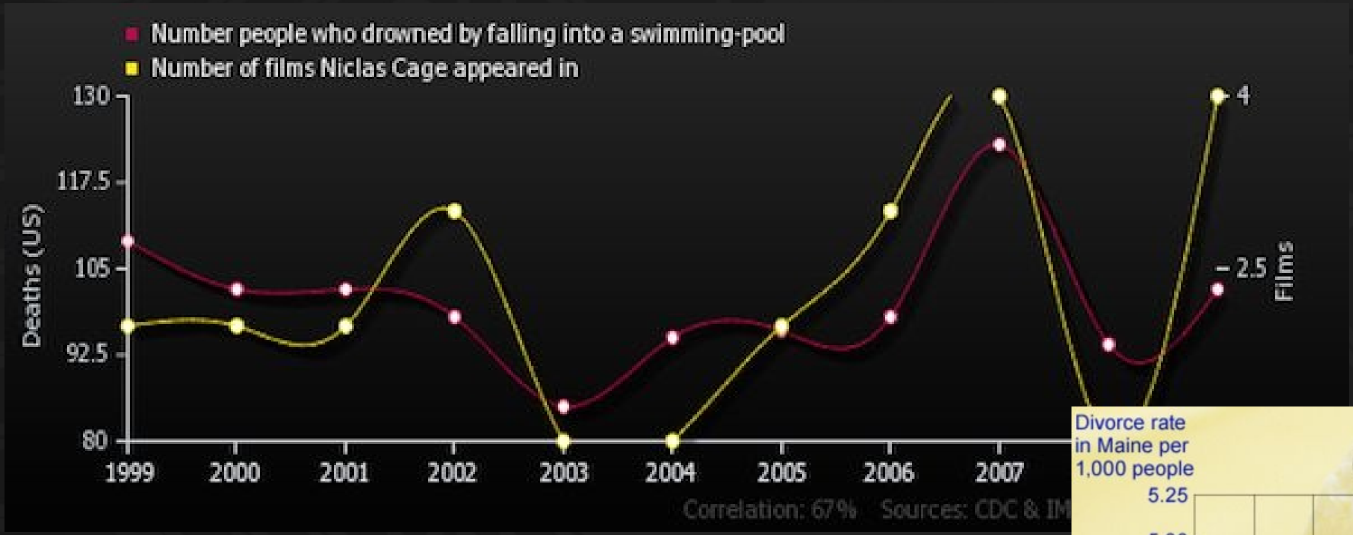


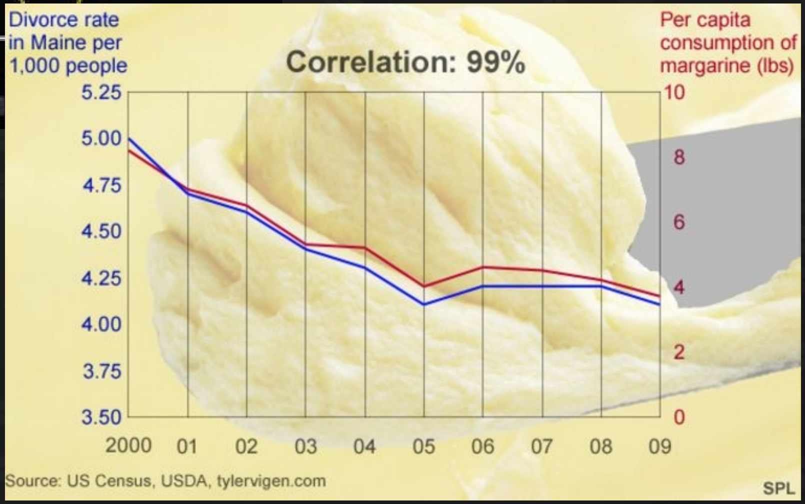
Profit?

The devil is in the details

- Which domain-specific data? This is not always readily available
- Mapping social media data to a time series?
  - Geolocation of messages
  - Mapping to topics/sentiments/intents or other characteristics
  - What is the variable: Volume? Sentiment? Other?
- Measuring correlation/influence
  - Correlation (lagged); Transfer entropy
- Finding a mechanism

Caveat 1: correlation might be spurious



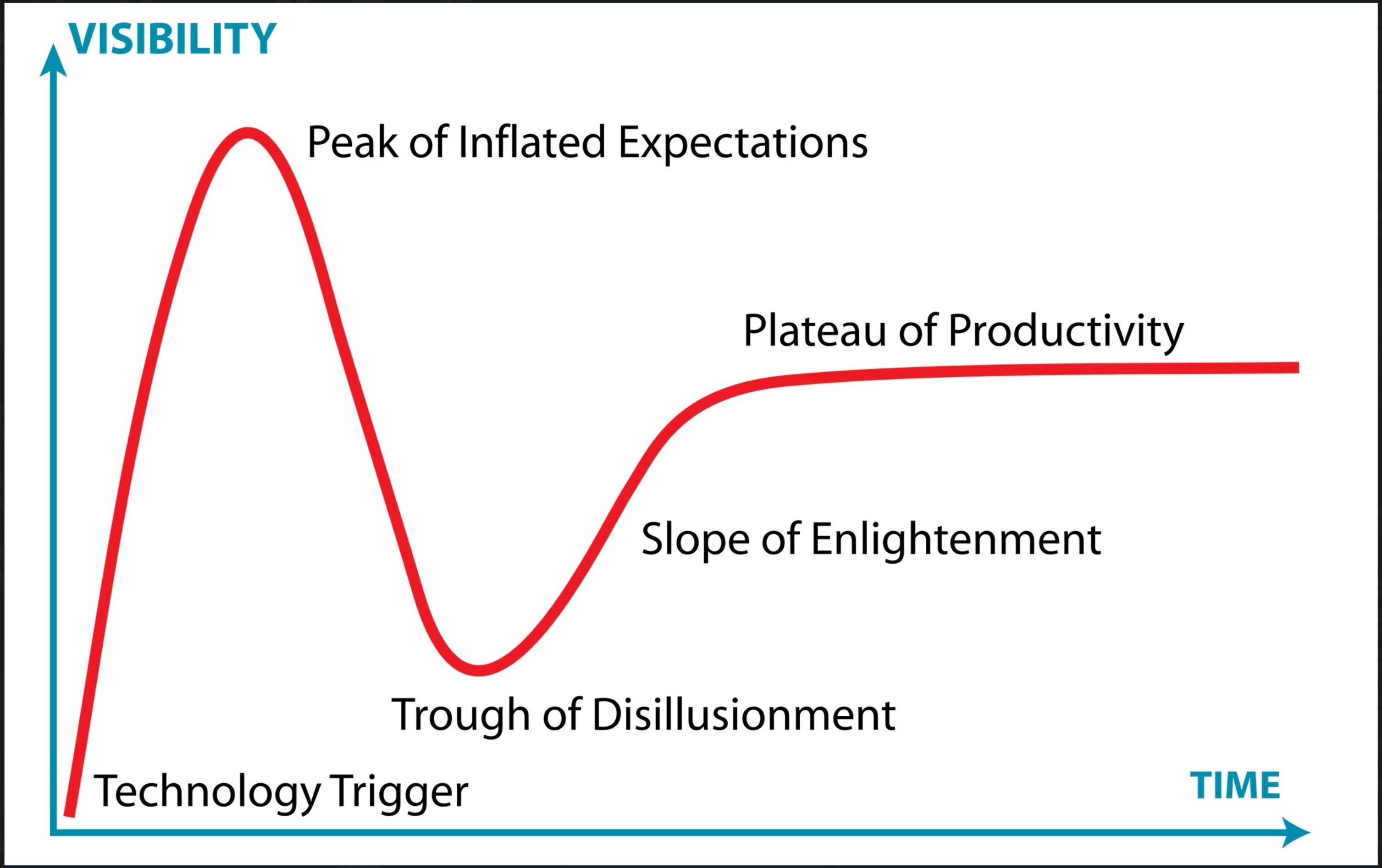


Caveat 2: correlation might be useless

- Sometimes there are much better predictors
- Social media can be used to predict box office revenue
  - But ticket sales on first weekend *almost* always determine total sales, with exceptions: Citizen Kane (1941), Blade Runner (1982), Fight Club (1999)
- Social media can be used to detect earthquakes
  - But seismographic sensors are quite dense in many areas of the world, the exception being underdeveloped areas

Caveat 3: the "war on terror"

The Hype Curve



Examples in the 1980s: (a) AI, (b) online learning.

Example social media mining topics

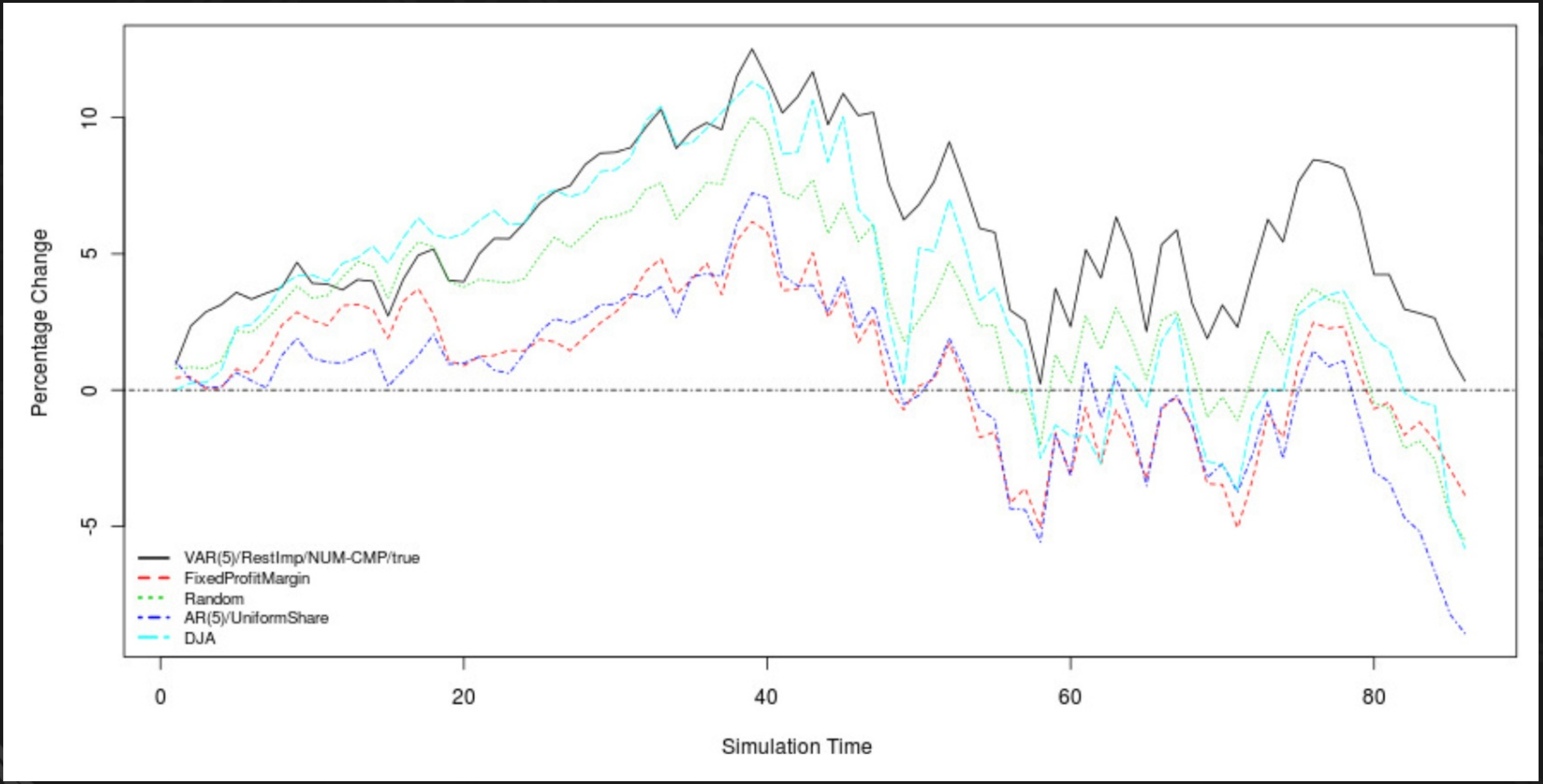
- Economics
- Politics
- Public health
- Smart cities
- Event detection

Most examples on this section come from https://sites.google.com/site/twitterandtherealworld/home

Examples in economics

- Financial success of movies
- Economic indices such as DJIA or NASDAQ
  - Words related to anxiety/worry/calmness/hope
- Stock option prices
  - Centrality in interaction graphs

Trading stock using social media

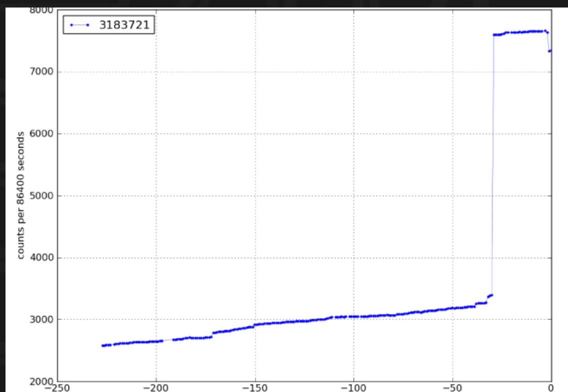


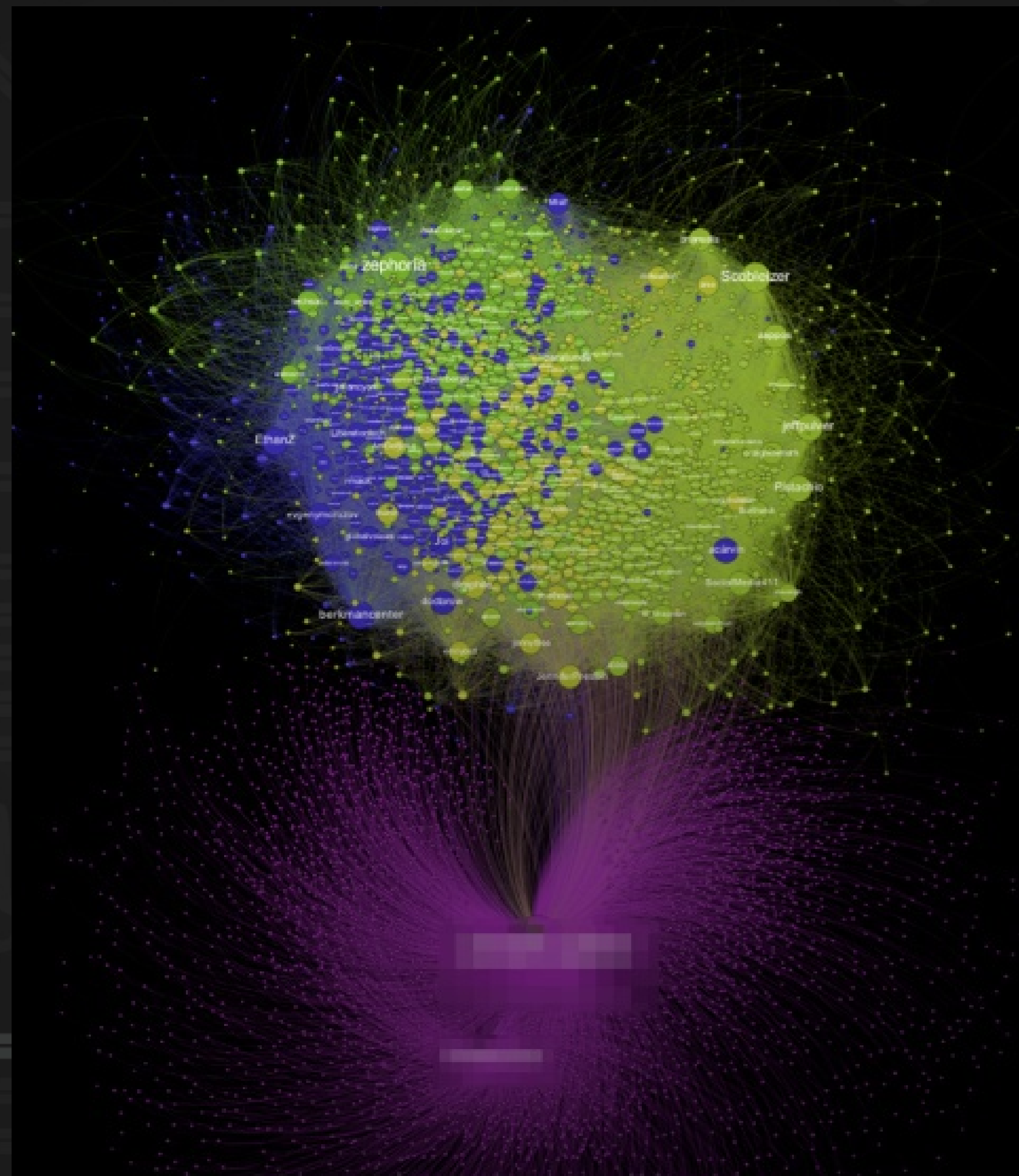
http://dx.doi.org/10.1145/2124295.2124358

Examples in politics

- Hashtags are a good indicator of political topics
- Signs of political leaning
  - Connections, profiles, conversations
- Political manipulation
  - Fake “grassroot” campaigns = “astroturfing”
- “No, you can't predict elections with Twitter”

Astroturfing (4K followers for USD 5)





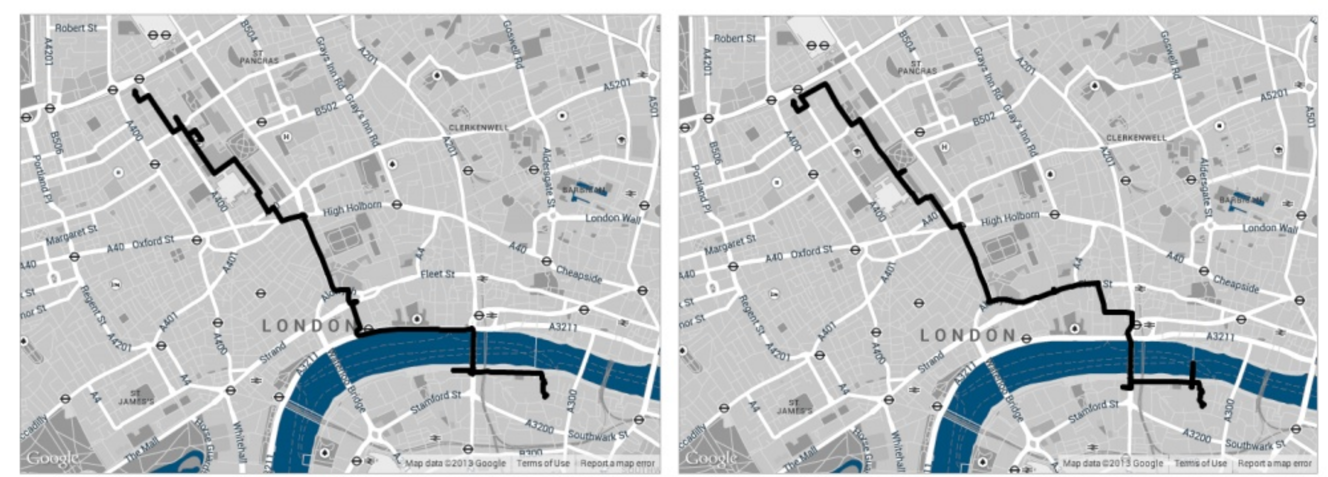
https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3

Examples in public health

- Many works derived from original Flu Trends
- Increasingly complex models of symptom-messages, treatment-messages
- Allergies, obesity, insomnia
- Mapping well-being in a city

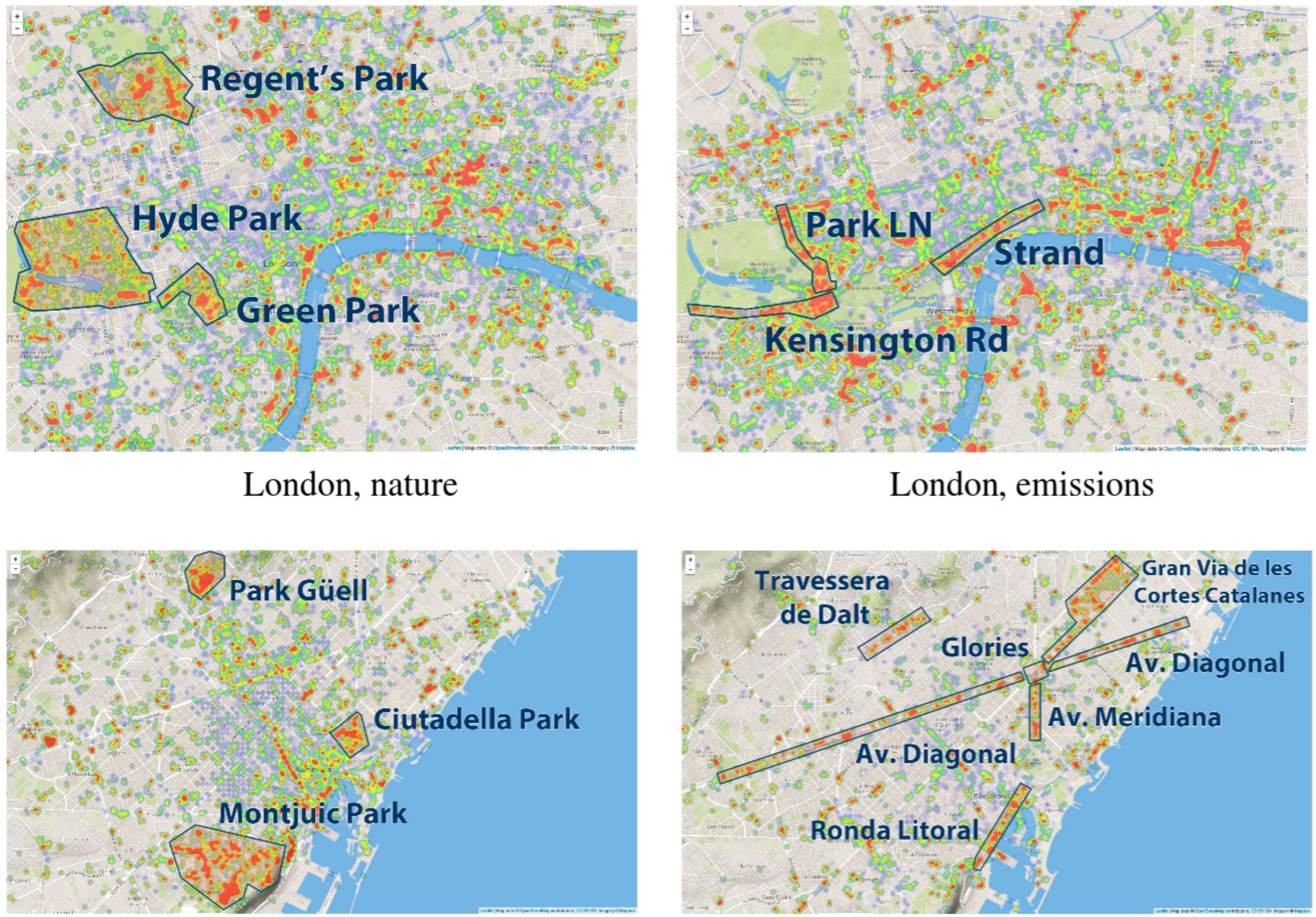
Examples in "smart cities"

- Data-driven neighborhood boundaries
- Data-driven residencial/commercial zones
- Tourism and beauty



(a) Shortest (b) Beauty

Smells



London, nature

London, emissions

Barcelona, nature

Barcelona, emissions

http://researchswinger.org/smellymaps/

Examples in event detection

- Mass convergence events, e.g. demonstrations
- Precursors of riots
- Traffic jams, accidents, or road blocks
- Man-made and natural disasters
  - And sub-events

Best practices in social media mining

- Interdisciplinary work
- Mixed methods: qualitative and quantitative
- Well-grounded in the domains' literature
- Recognize, measure, and possibly counter sample biases
- Robust to different settings, metrics, datasets
- Outcomes provide an advantage to practitioners
  - E.g. to make better decisions than without this data

Social Media Mining and Retrieval
Part 03: Social Network Analysis

Social media and retrieval

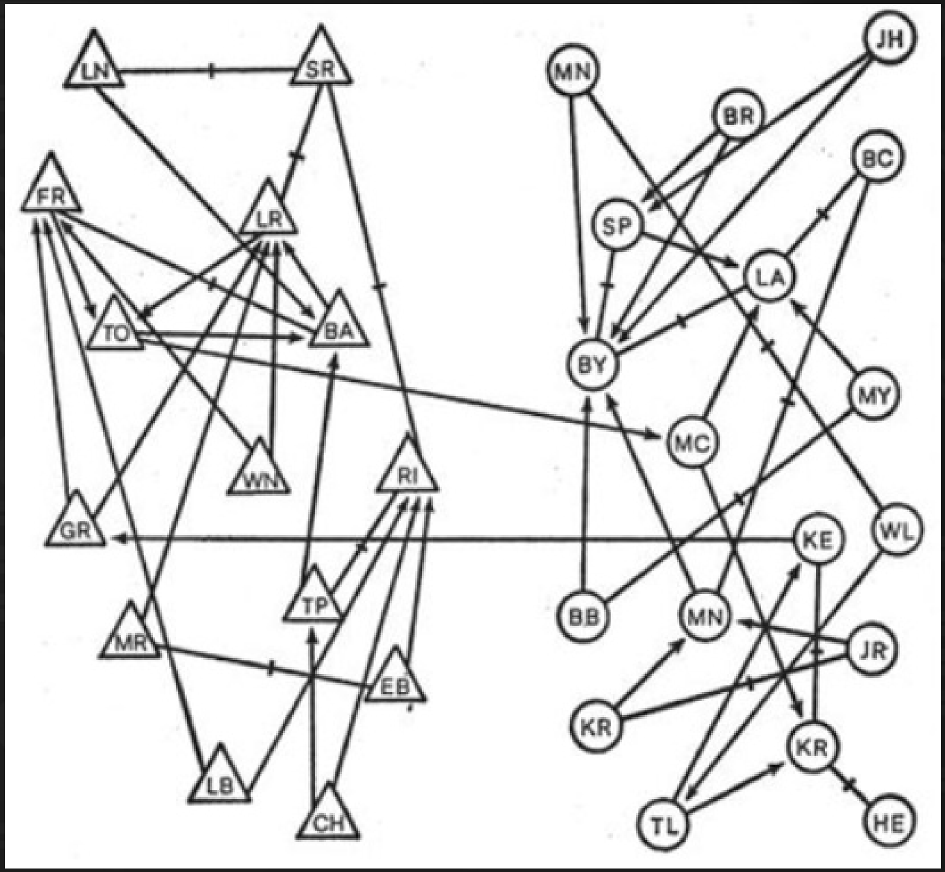
- *(In addition to the new characteristics of texts)*
- Social media is a prolific source of new relevance signals:
  - Social networks (structural signals)
  - Information cascades (propagation signals)

Graph essentials in next slides mostly from

http://dmml.asu.edu/smm/book/

Social networks

- Sociograms started to be collected systematically in the 1930s
  - E.g. Girls/Boys (triangles/circles)
- Built from interviews and direct observation
- Now we call them social networks

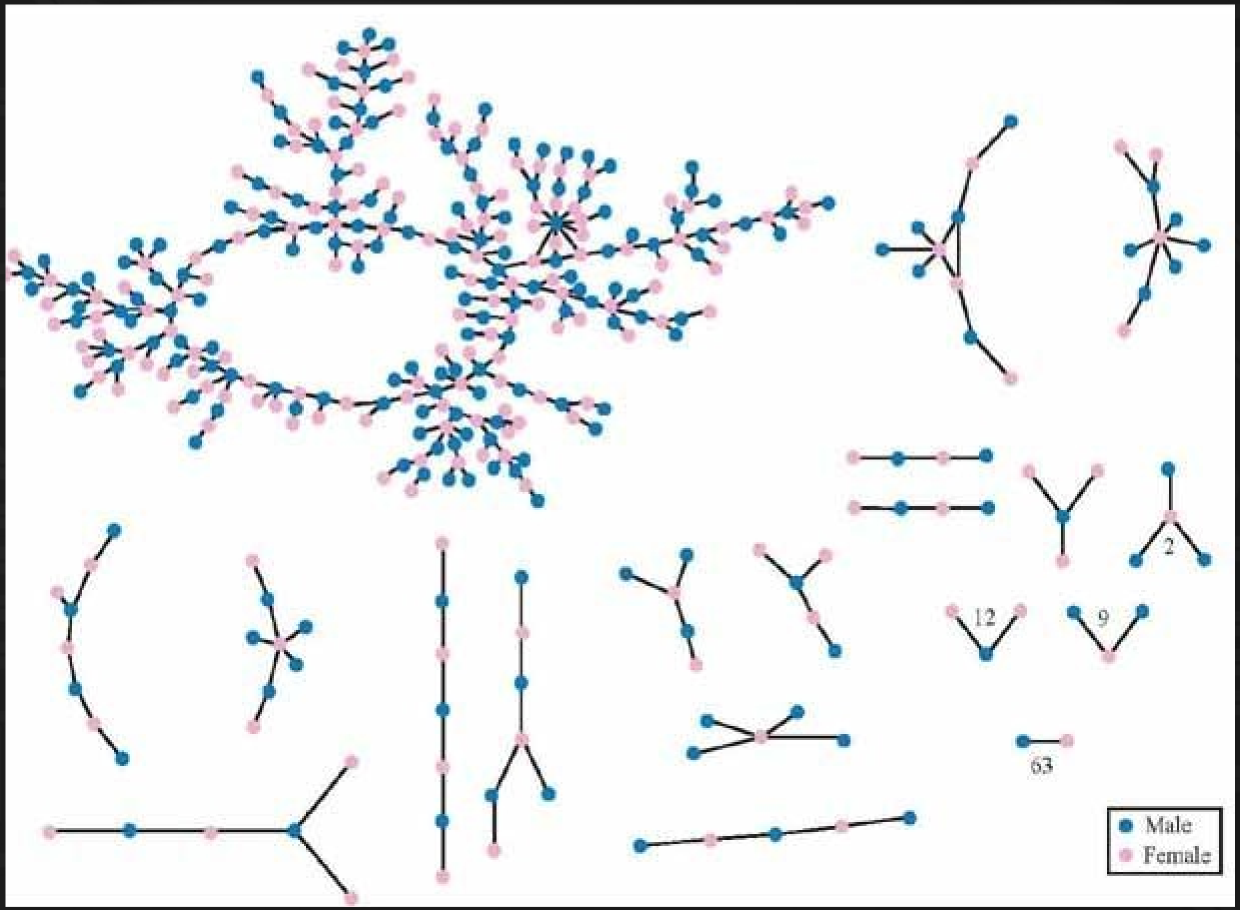


We still build some sociograms by hand!

Special romantic relationship

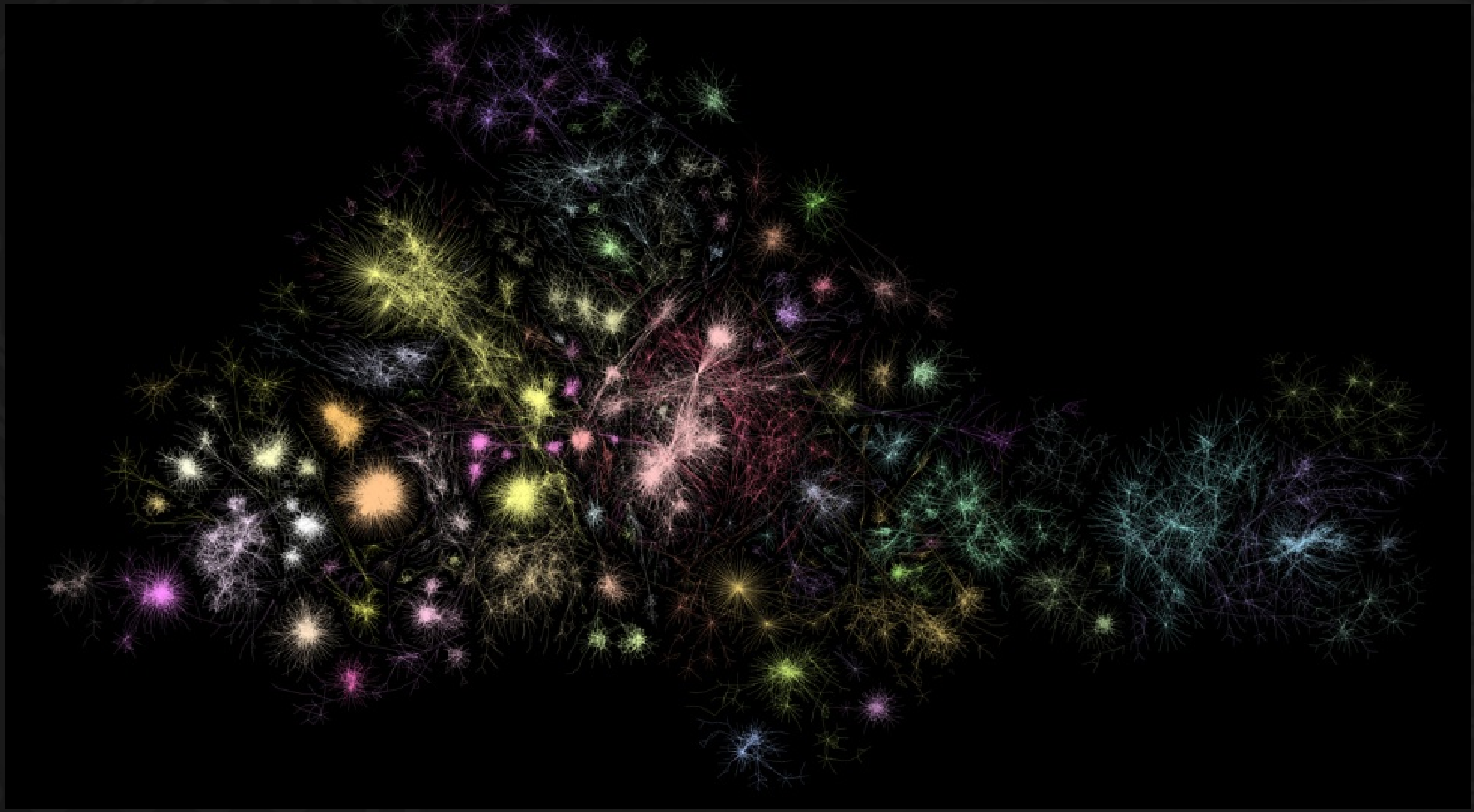
—OR—

Nonromantic sexual relationship



[Bearman et al. 2004]
http://www.jstor.org/stable/10.1086/386272

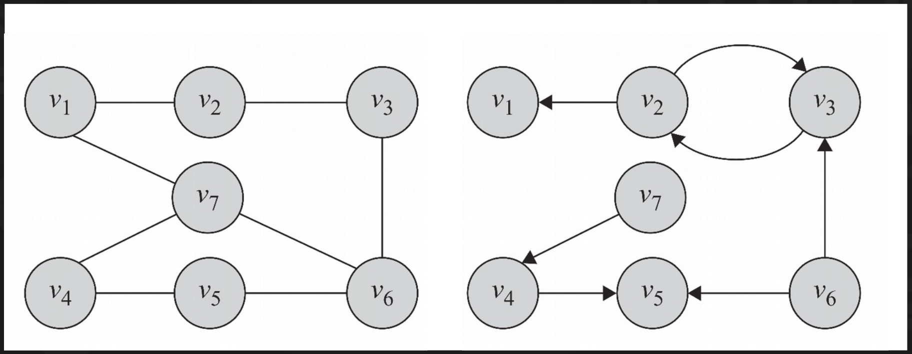
Lots of social network data to play with



https://snap.stanford.edu/data/

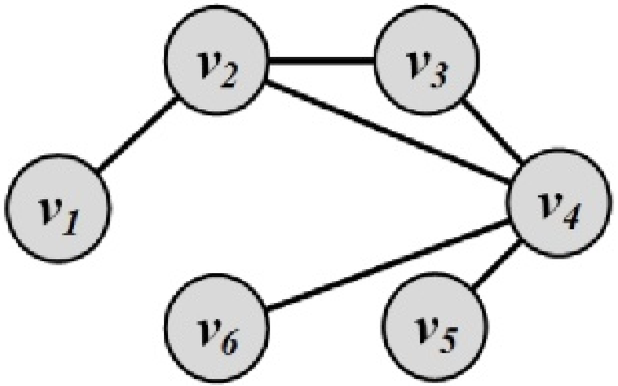
Graphs

- For a set A of objects, a graph provides a very convenient representation of relations in A, which are subsets of A x A
- Symmetry of relation determines type of graph
  - Symmetric relations: undirected graphs having vertices and edges
  - Non-symmetric relations: directed graphs having nodes and arcs



Matrix representation

$$A_{ij} = \begin{cases} 1, \text{ if there is an edge between nodes } vi \text{ and } vj \\ 0, \text{ otherwise} \end{cases}$$

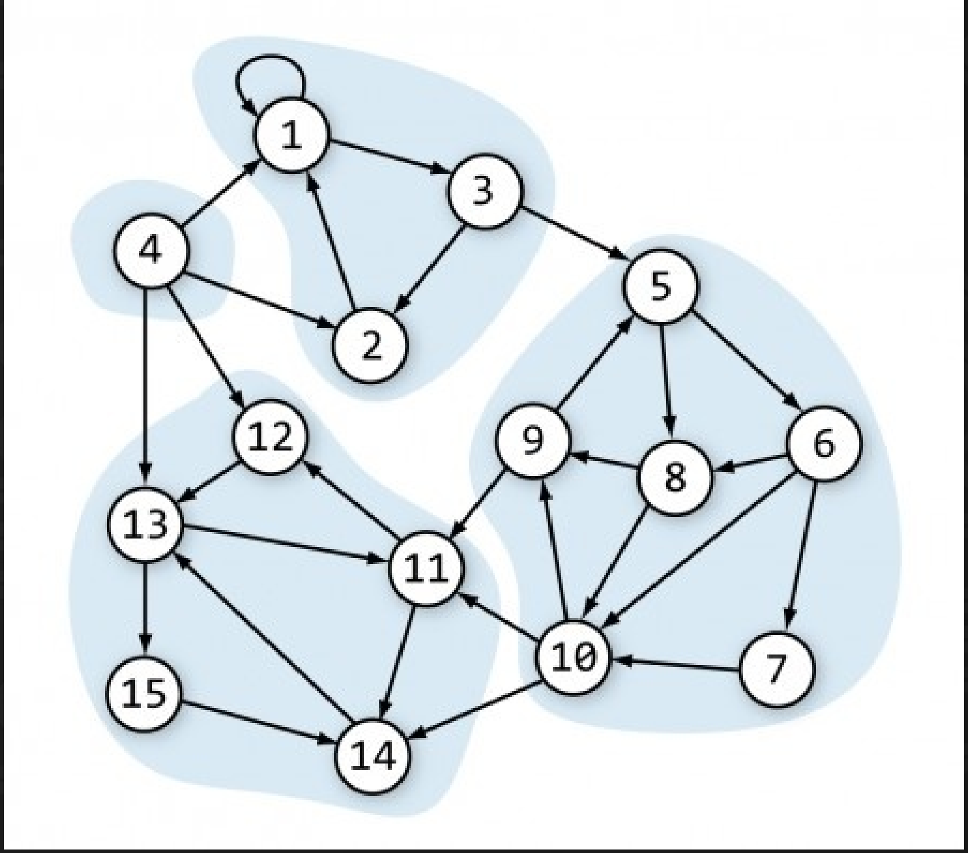


	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	0	0
v_2	1	0	1	1	0	0
v_3	0	1	0	1	0	0
v_4	0	1	1	0	1	1
v_5	0	0	0	1	0	0
v_6	0	0	0	1	0	0

- Can be extended to weighted graphs
- Social networks tend to be sparse matrices

Community structure

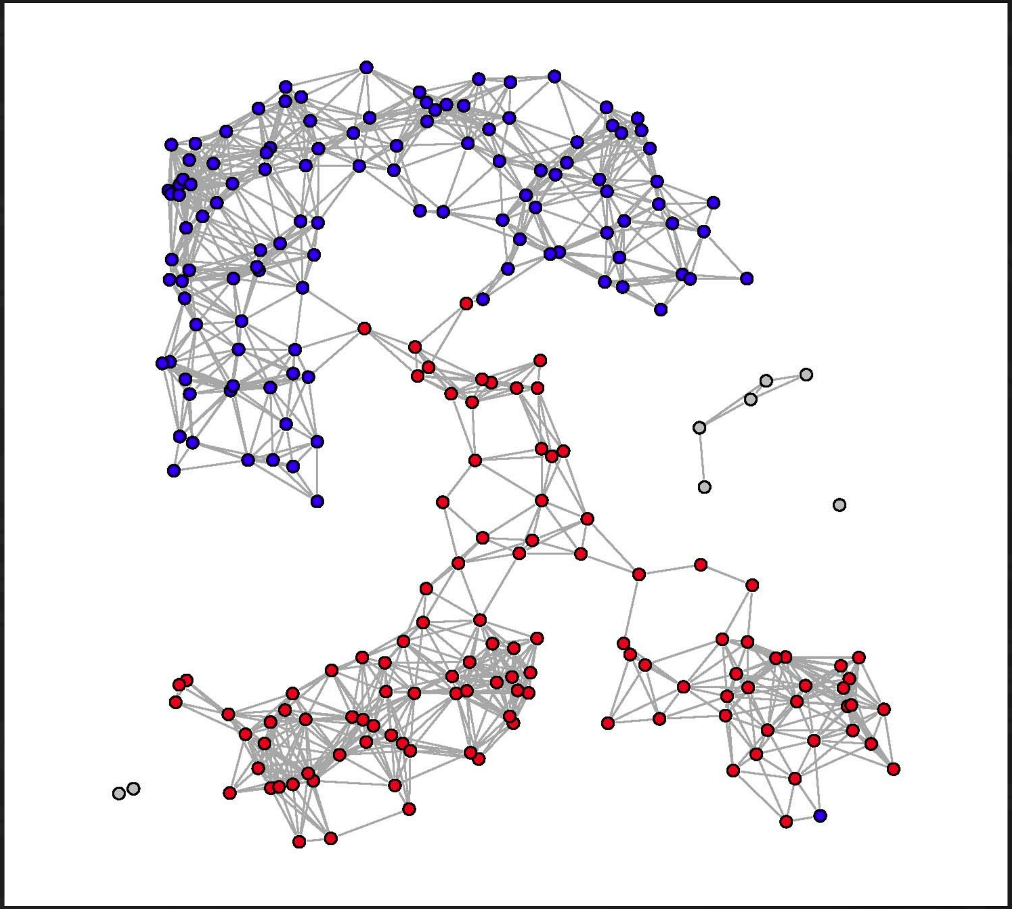
- Connected components (undirected graphs)
  - Nodes reachable
- Strongly connected components (directed graphs)
  - Nodes mutually reachable



http://www.greatandlittle.com/studios/index.php?post/2013/03/26/Strongly-Connected-Components-and-the-2-SAT-Problem-in-Dart

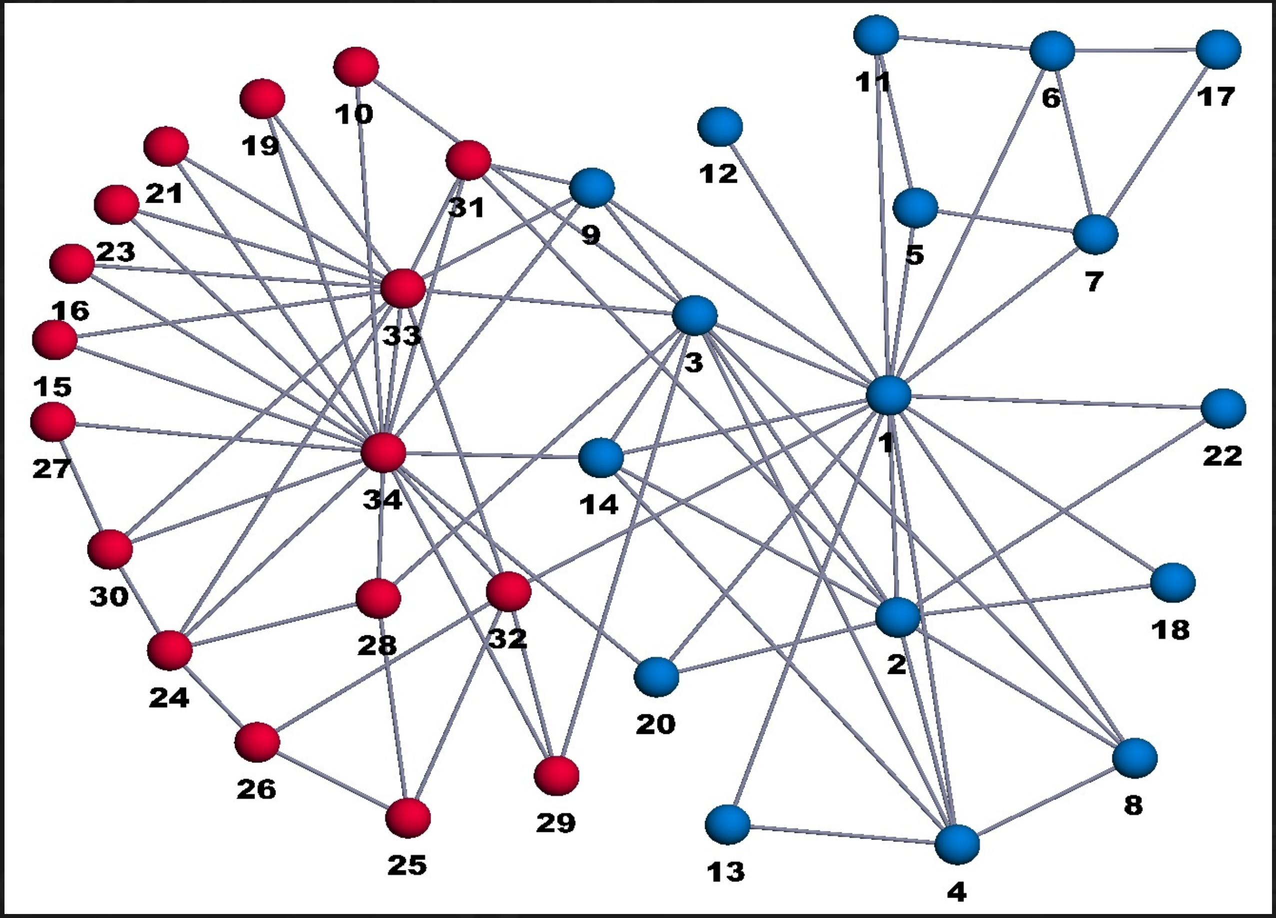
Community structure (cont.)

- A community, cluster, or partition is a group of nodes that is *more connected among them* than with the rest of the graph
- Many formal definitions and algorithms
- [Girvan & Newman 2002]: remove high-betweenness edges, keep track of connected components



http://math.aalto.fi/~lleskela/SDLRG/

Karate Club, US University in 1970



Nodes 1 and 34 were the karate instructor and an administrator from the university.

They had a big argument, and the club splitted in two.

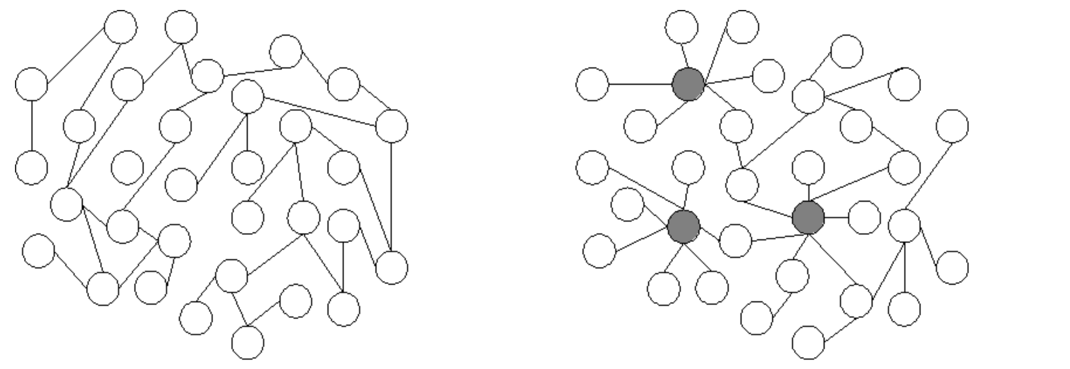
W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).

Degree

- Number of connections of a node
- In-degree, Out-degree in directed graphs
- Weighted (in-/out-)degree in weighted graphs

Degree distributions

- Social networks have skewed degree distributions
  - Scale-free networks, power laws
  - Many nodes with huge degree
- Plausible mechanism: preferential attachment



(a) Random network (b) Scale-free network

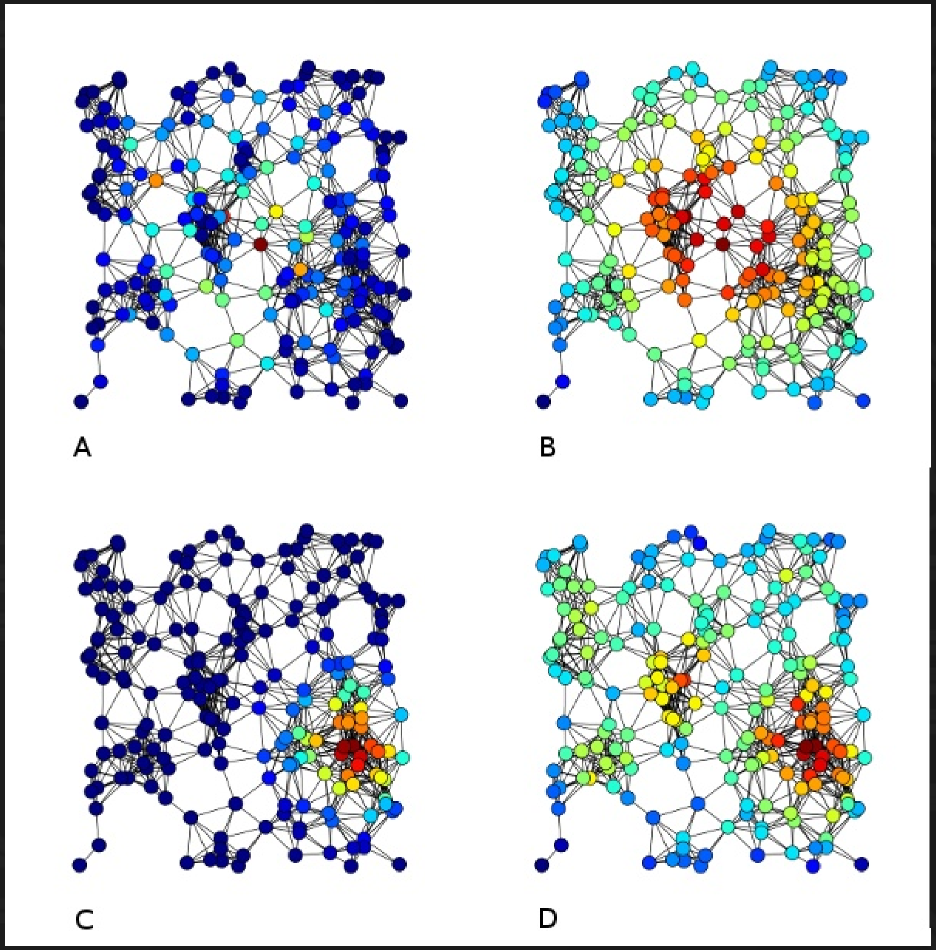
Centrality

A) Betweenness centrality
- Being in many shortest paths

B) Closeness
- Being close to many nodes

C) Eigenvector centrality
- End of many paths

D) Degree centrality
- High degree



Centrality = Quality?

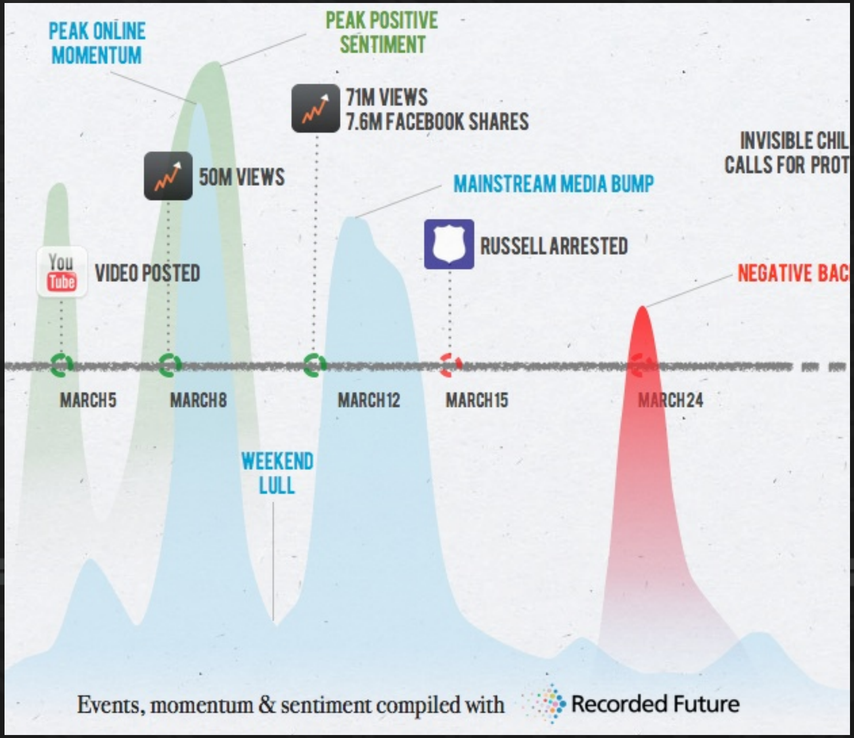
- Various hypothesis about high-centrality authors
  - They can produce “better” content
  - They are more “influential”
  - They are “experts” (within a community/cluster/partition)
- To some extent, yes, but …
  - How to combine this with other signals requires careful tunning
  - Ideally on a learning-to-rank framework

Social Media Mining and Retrieval
Part 04: Information Cascades

"Twitter Revolution"

- "Viral" calls to demonstrations against fraud in elections in Moldova and Iran in 2009
- Explosive "bursts" of messages that reach huge audiences
- Example:

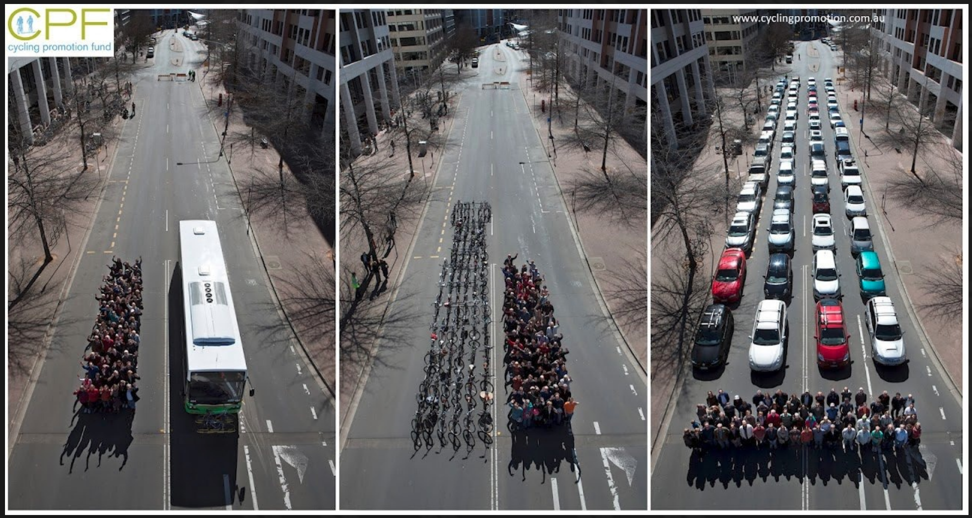
#Kony2012

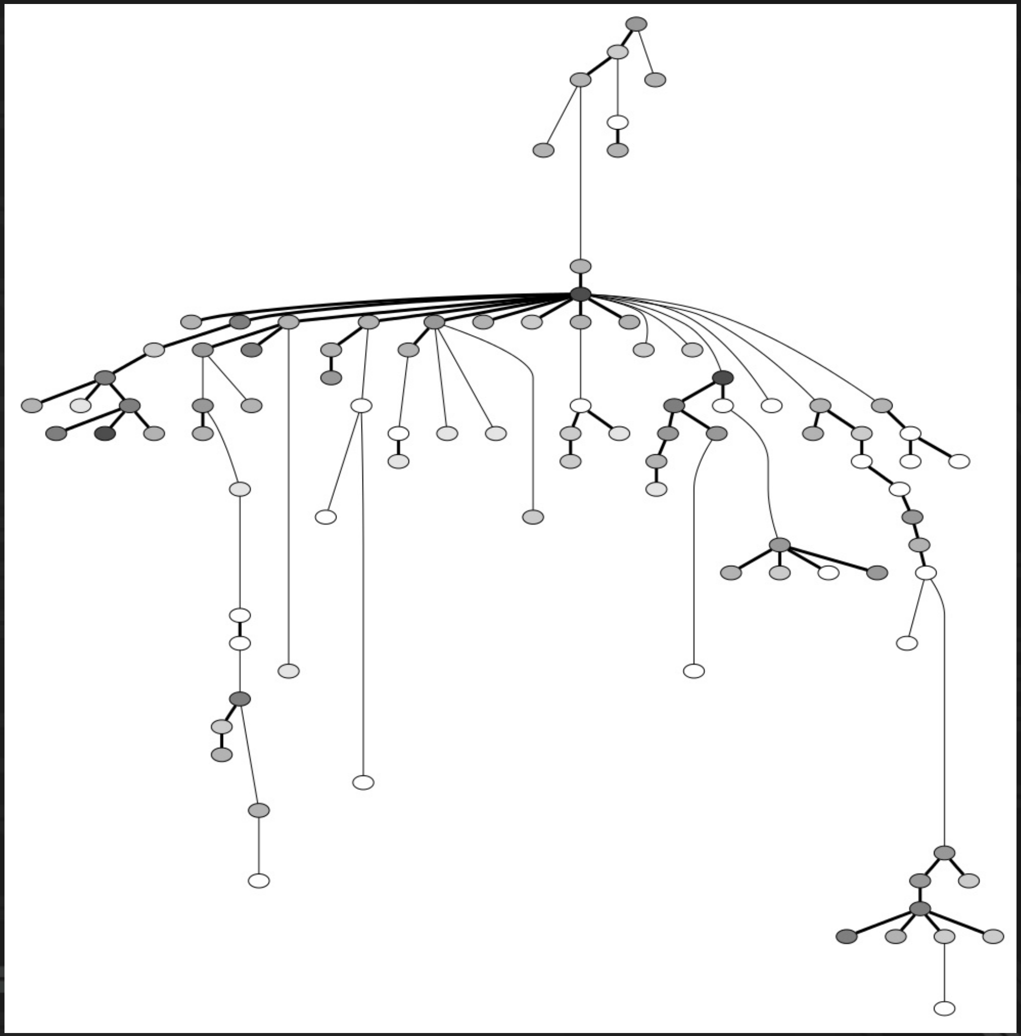


Viral content

Everybody wants their content to “go viral”

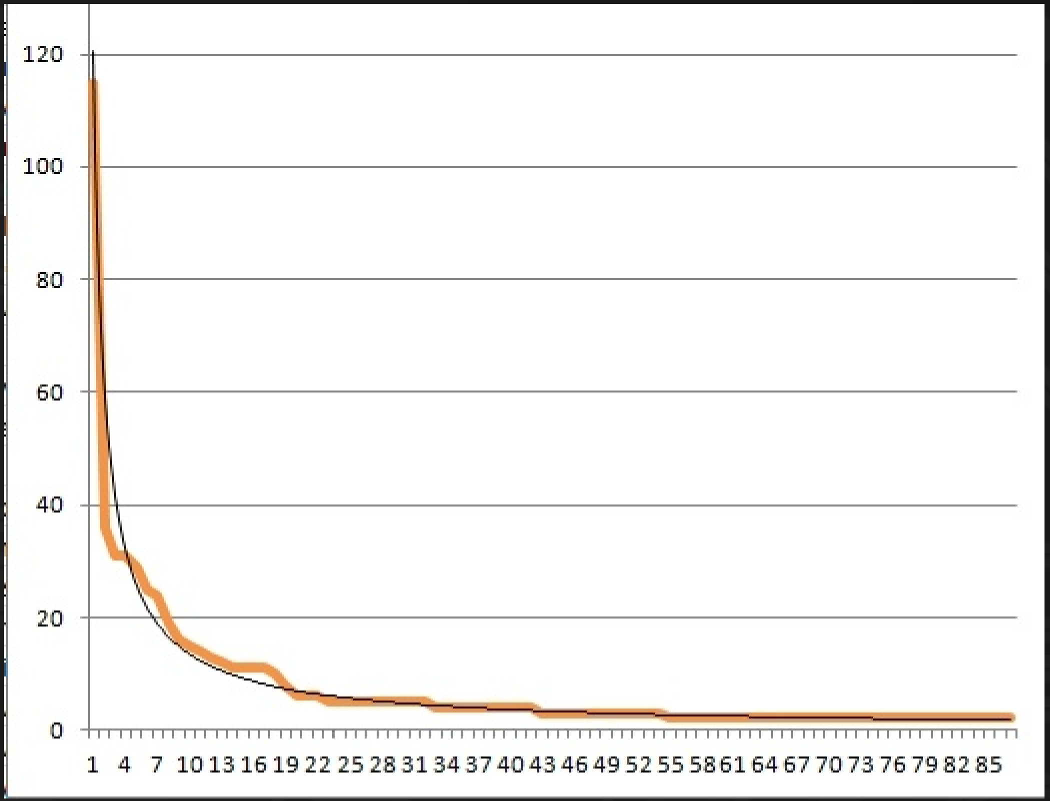
Example cascade [Ienco et al. 2010]





Large cascades are rare

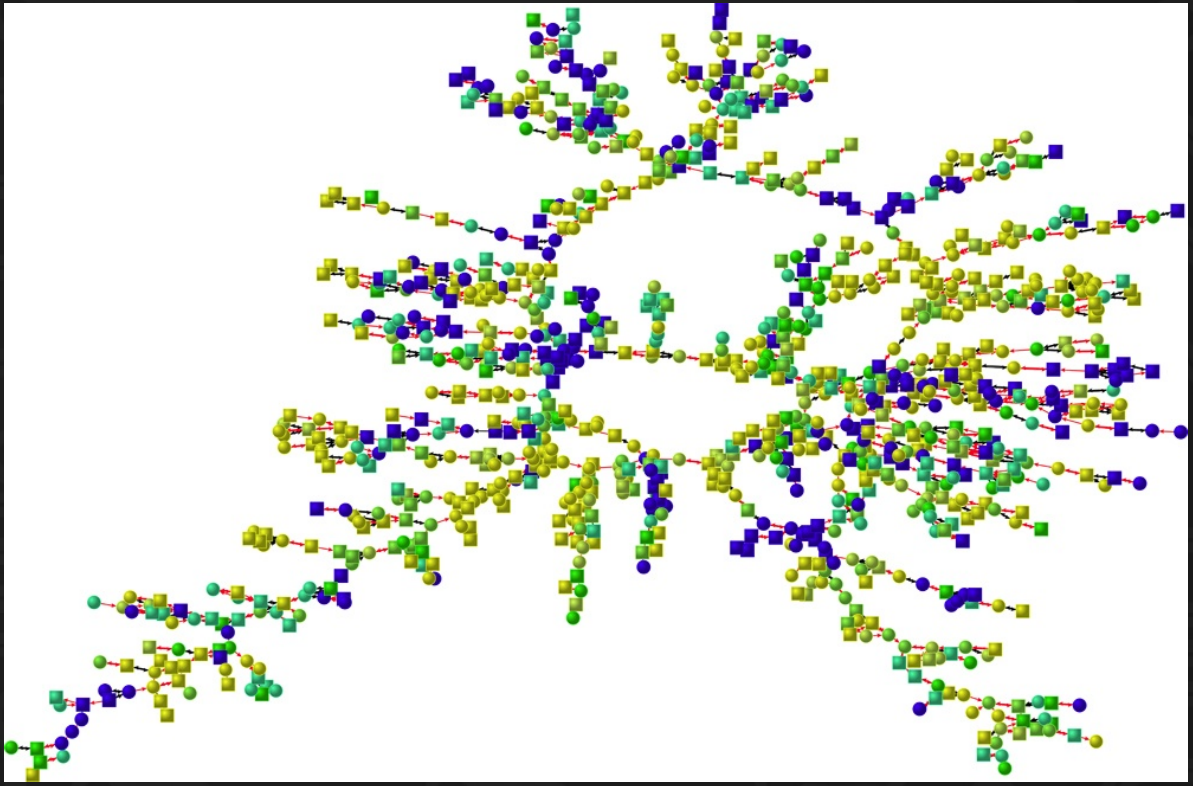
- Most content is never shared
- Content that is shared is mostly shared just once
  - Shares per item have a very skewed distribution



http://www.boyntons.us/website/new-media/analyses/concentration/focus-twitter-streams.html

Many Phenomena are Epidemic

- "Infected" can mean many things:
  - Buying a product, hiring an insurance, sharing a post, prefering a beverage, voting for a candidate, etc.
- Obesity is "contagious" [Christakis&Fowler 2007]
- Happiness, too! [Fowler&Christakis 2008]



Circles are female, squares are male; lines indicate relationships (black = siblings; red = friends, spouces). Color is happiness, with blue indicating "the blues," and yellow indicating sheer joy. Green is somewhere in between.

http://web.med.harvard.edu/sites/RELEASES/html/christakis_happiness.html

Epidemic models

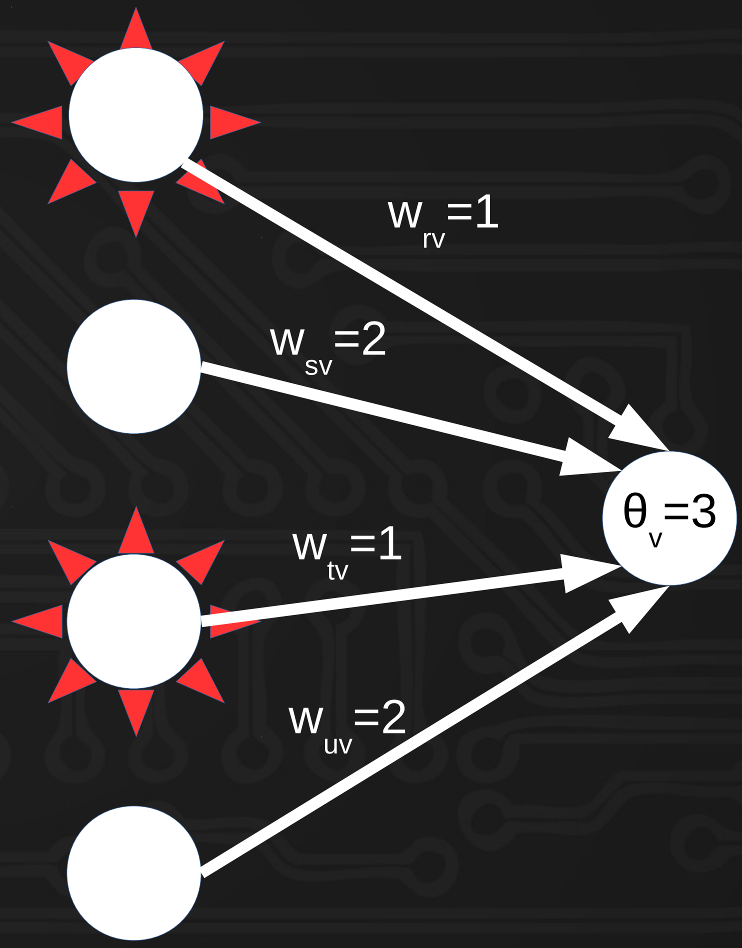
- Many possible stages: Susceptible-Infected simplest one
  - Susceptible-Infected-Susceptible / Susceptible-Infected-Inmune
- Populations described by simple differential equations
  - See e.g. https://en.wikipedia.org/wiki/Epidemic_model
- Current models are discrete, stochastic, and assume only certain propagations/contagions are possible

Discrete models of viral propagation

- Linear threshold
  - Activate if sum of weighted in-links exceeds a threshold
- Independent cascades
  - One attempt to activate through each probability-weighted out-link
- General activation functions

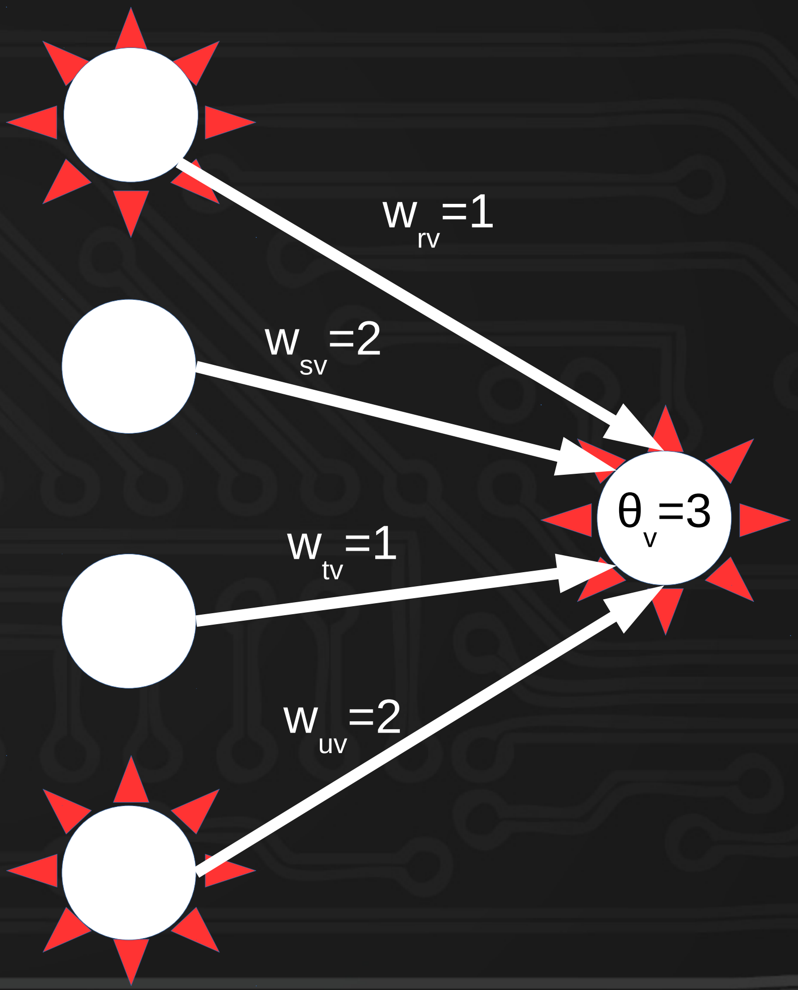
Linear threshold

- Each node has a threshold
- Activate if weighted sum of inputs reaches or exceeds threshold
- Arc weights represent influence



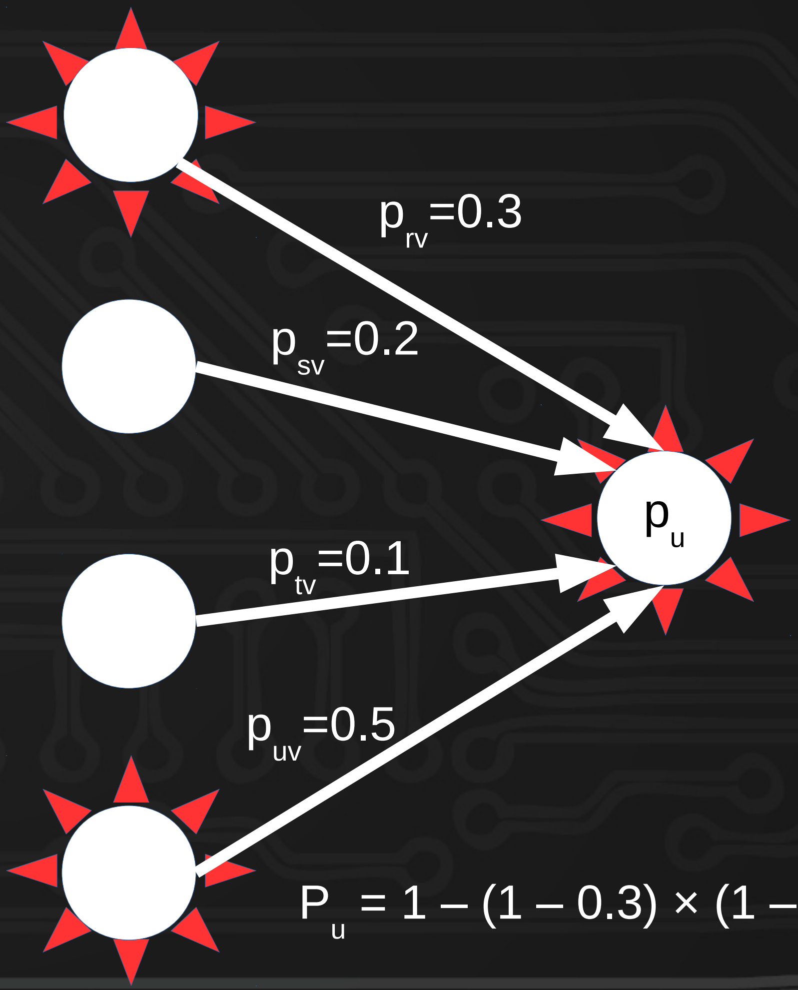
Linear threshold (cont.)

- Each node has a threshold
- Activate if weighted sum of inputs reaches or exceeds threshold
- Arc weights represent influence



Independent cascades

- Each active node has one chance of activating neighbors
- Arc probabilities are the chance of suceeding



$P_u = 1 - (1 - 0.3) \times (1 - 0.5)$

Some problems

- Determining expected size of cascades
  - Simulation is main approach
- Inferring influence probabilities
- Topic-specific cascades
- Time-critical cascades
- Competitive cascades

Engineering cascades?

- Strategy 1: invest money in convincing a few influentials
- Strategy 2: invest money in convincing random people
- Strategy 1 vs Strategy 2 still a controversy
  - Viral marketing business driven by outliers?
  - Models not faithful enough?
  - One model doesn't fit all cases?

Social Media in Natural Disasters

Carlos Castillo

Thanks to: Patrick Meier, Alexandra Olteanu, Muhammad Imran, Sarah Vieweg, Fernando Diaz, Aditi Gupta, Hemant Purohit

http://bigcrisisdata.org/

@bigcrisisdata

Humanitarian Computing

At least **775** publications:

- Crisis Analysis (55)
- Crisis Management (309)
- Situational Awareness (67)
- Social Media (231)
- Mobile Phones (74)
- Crowdsourcing (116)
- Software and Tools (97)
- Human-Computer Interaction (28)
- Natural Language Processing (33)
- Trust and Security (33)
- Geographical Analysis (53)

Source: http://humanitariancomp.referata.com/

ISCRAM 2014

Network Theory

Analytical Modelling and Simulation

Planning, Foresight and Risk Analysis

Ethical, Legal and Social Issues

Geospatial Data and Geographical Information Science

Command & Control Studies

Human Centred Design and Evaluation

Social Media Studies

Community Engagement

Decision Support Systems

Practitioner Cases And Practitioner-Centered Research

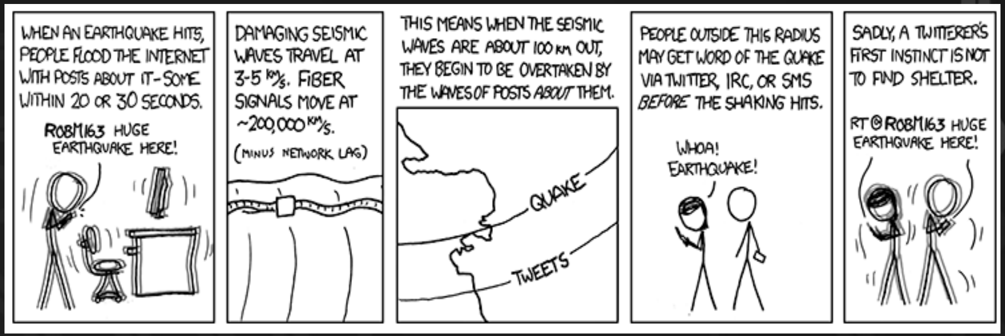
Researching Crisis: Methodologies

Serious Gaming

Understanding Collaborative Work Practices

An earthquake hits a Twitter user

- When an earthquake strikes, the first tweets are posted 20-30 seconds later
- Damaging seismic waves travel at 3-5 km/s, while network communications are light speed on fiber/copper + latency
- After ~100km seismic waves may be overtaken by tweets about them



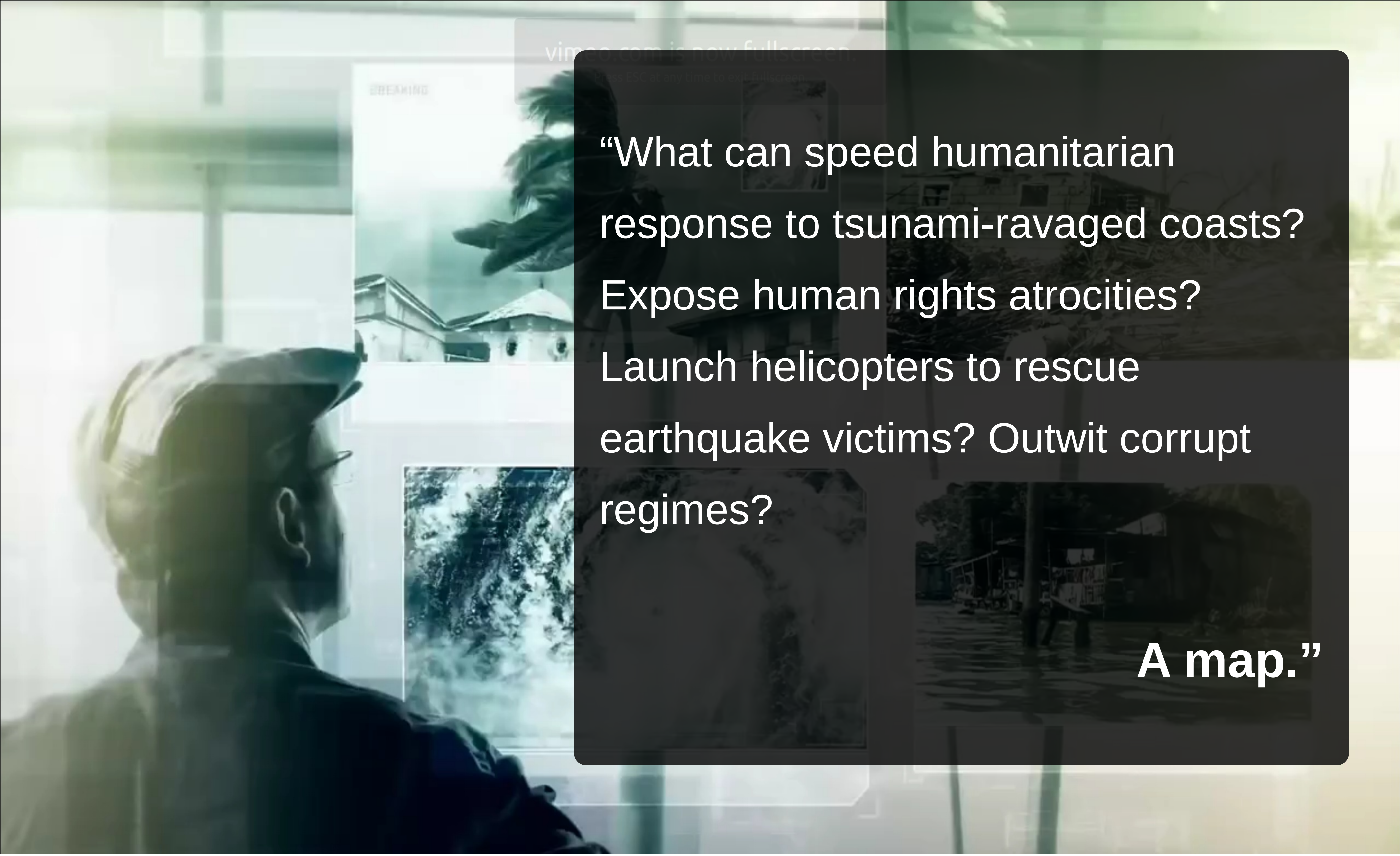
http://xkcd.com/723/

Actual messages during disasters

- **“OMG! The fire seems out of control: It’s running down the hills!”** (bush fire near Marseilles, France, in 2009)
- **“Red River at East Grand Forks is 48.70 feet, +20.7 feet of flood stage, -5.65 feet of 1997 crest. #flood09”** (Red River Valley floods in 2009).
- **“My moms backyard in Hatteras. That dock is usually about 3 feet above water [photo]”** (Hurricane Sandy 2013 - reddit)
- **“Sirens going off now!! Take cover...be safe!”** (Moore Tornado 2013)
- **“There is shooting at Utøya, my little sister is there and just called home!”** (2011 attacks in Norway)

Possible topics

- (Sub-)event detection
- Characterizing (sub-)events with structured data
- Summarizing (sub-)events
- Prioritizing/filtering messages
- Helping to evaluate severity of damage, urgency of needs
- Routing messages to responders
- Matching messages describing problems and solutions



“What can speed humanitarian response to tsunami-ravaged coasts? Expose human rights atrocities? Launch helicopters to rescue earthquake victims? Outwit corrupt regimes?

A map.”

NATIONAL GEOGRAPHIC

Crisis mapping goes mainstream (2011)

Syria Tracker

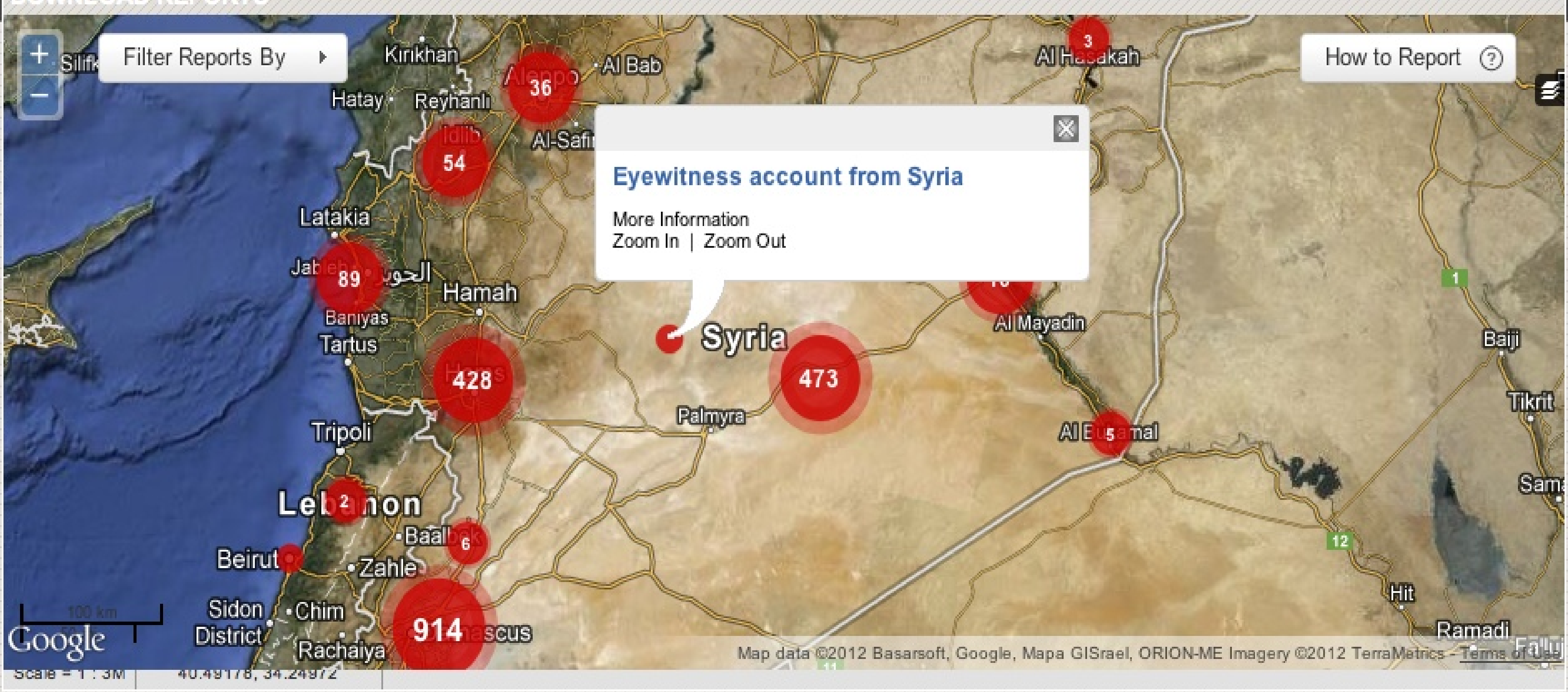
Missing, Killed, Arrested, Eyewitness, Report

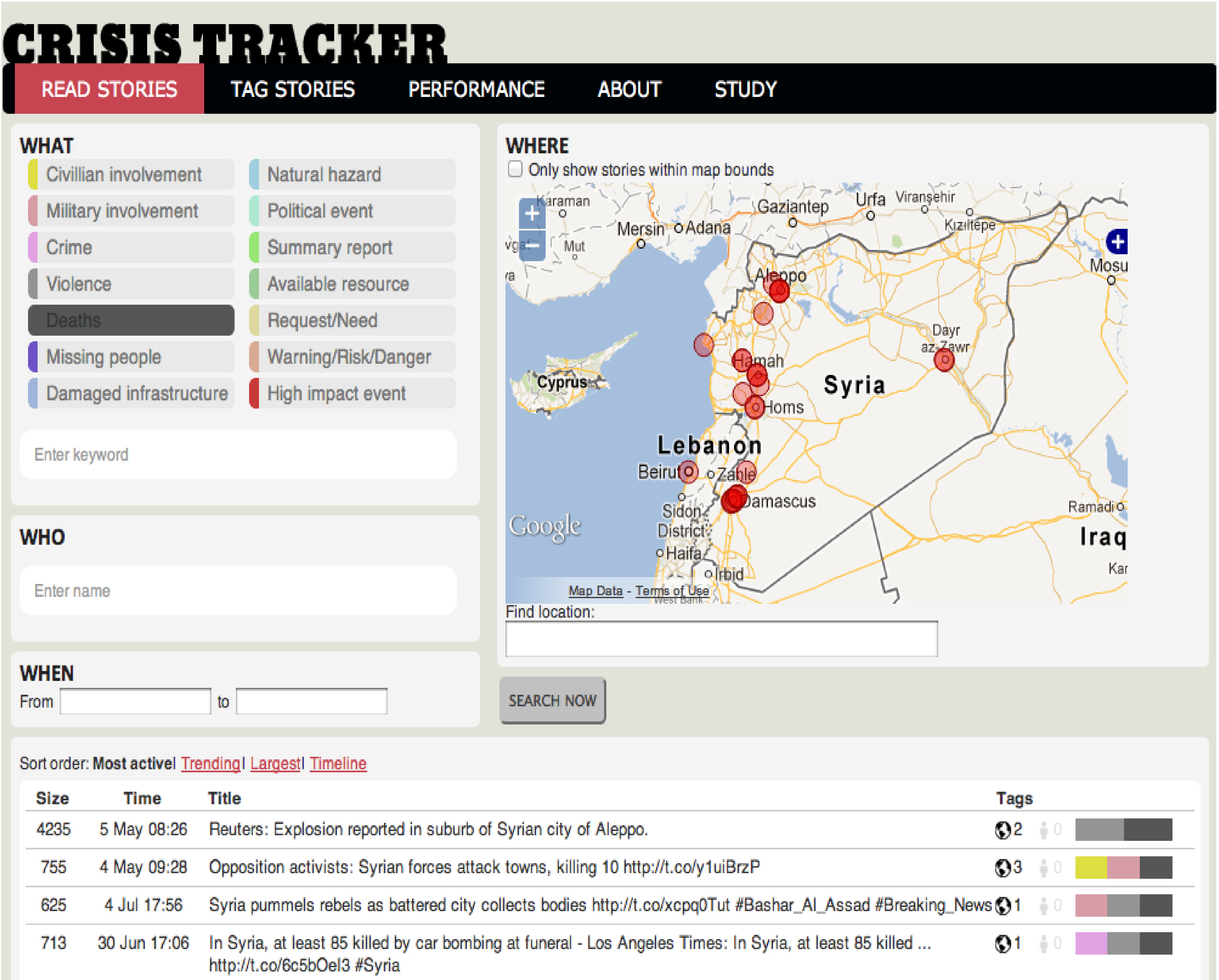
SUBMIT A REPORT

30

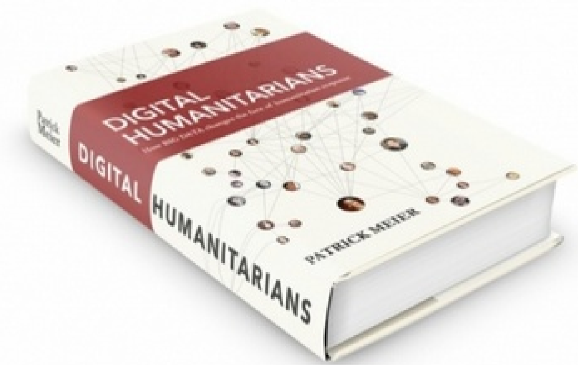
40,154 victims from March 18, 2011 to November 5, 2012. Maps for Syria Tracker's Crowdmapped Data from Mar 18, 2011 through Present. Please help us document the crimes in Syria. Here is a short tutorial on how to report. In addition, please see the Instructions page for security precautions to take while submitting reports from the field. Reports can be submitted anonymously or you have the option to provide your personal information [يمكنك إدخال التقرير بدون الكشف عن شخصيتك أو إذا أردت يمكنك إدخال معلوماتك الشخصية]. You can also submit reports via email to syriatracker@gmail.com or by adding the hashtag #basharcrimes to your tweet(s) (please make sure you include the location or geo-location of the report when submitting via email or twitter). Learn to protect your security online (Arabic version - English version). Reports can be downloaded here. You can subscribe to alerts here. To report cases of human rights violations of sexualized violence, please visit Women Under Siege.

HOME | REPORTS | SUBMIT A REPORT | GET ALERTS | CONTACT US | ABOUT US (من نحن) | INSTRUCTIONS (الأمن الرقمي) | DOWNLOAD REPORTS




CRISIS TRACKER
READ STORIES
TAG STORIES
PERFORMANCE
ABOUT
STUDY
WHAT
Civillian involvement
Natural hazard
Military involvement
Political event
Crime
Summary report
Violence
Available resource
Deaths
Request/Need
Missing people
Warning/Risk/Danger
Damaged infrastructure
High impact event
Enter keyword
WHO
Enter name
WHEN
From
to
WHERE
Only show stories within map bounds
Karaman
Gaziantep
Urfa
Viranşehir
Mersin
Adana
Kızıltepe
Mut
Aleppo
Mosu
Dayr az Zawr
Hamah
Cyprus
Syria
Homs
Lebanon
Beirut
Zahle
Damascus
Sidon District
Haifa
Irbid
Ramadi
Iraq
Kar
Google
Map Data - Terms of Use
Find location:
SEARCH NOW
Sort order: Most active| Trending| Largest| Timeline
Size
Time
Title
Tags
4235
5 May 08:26
Reuters: Explosion reported in suburb of Syrian city of Aleppo.
2
0
755
4 May 09:28
Opposition activists: Syrian forces attack towns, killing 10 http://t.co/y1uiBrzP
3
0
625
4 Jul 17:56
Syria pummels rebels as battered city collects bodies http://t.co/xcpq0Tut #Bashar_Al_Assad #Breaking_News
1
0
713
30 Jun 17:06
In Syria, at least 85 killed by car bombing at funeral - Los Angeles Times: In Syria, at least 85 killed ... http://t.co/6c5bOel3 #Syria
1
0

Leveraging digital networks for humanitarian response



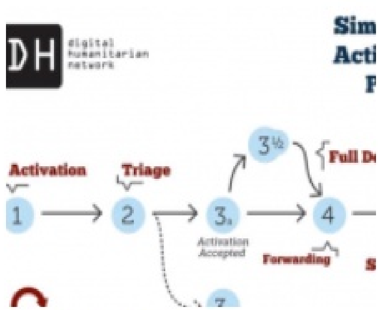
Digital Humanitarians: The Book

Patrick Meier is writing a book charting the sudden rise of Digital Humanitarians by sharing their remarkable, real-life stories, highlighting how their humanity coupled with innovative solutions is changing humanitarian response forever. Look for it spring 2015!

BECOME A MEMBER

Organizations working in this space can apply to become members. Find out how here.

Read More



ACTIVATE DHN

When disaster strikes, humanitarian organizations can apply here to activate a DHN team to support response.

Read More

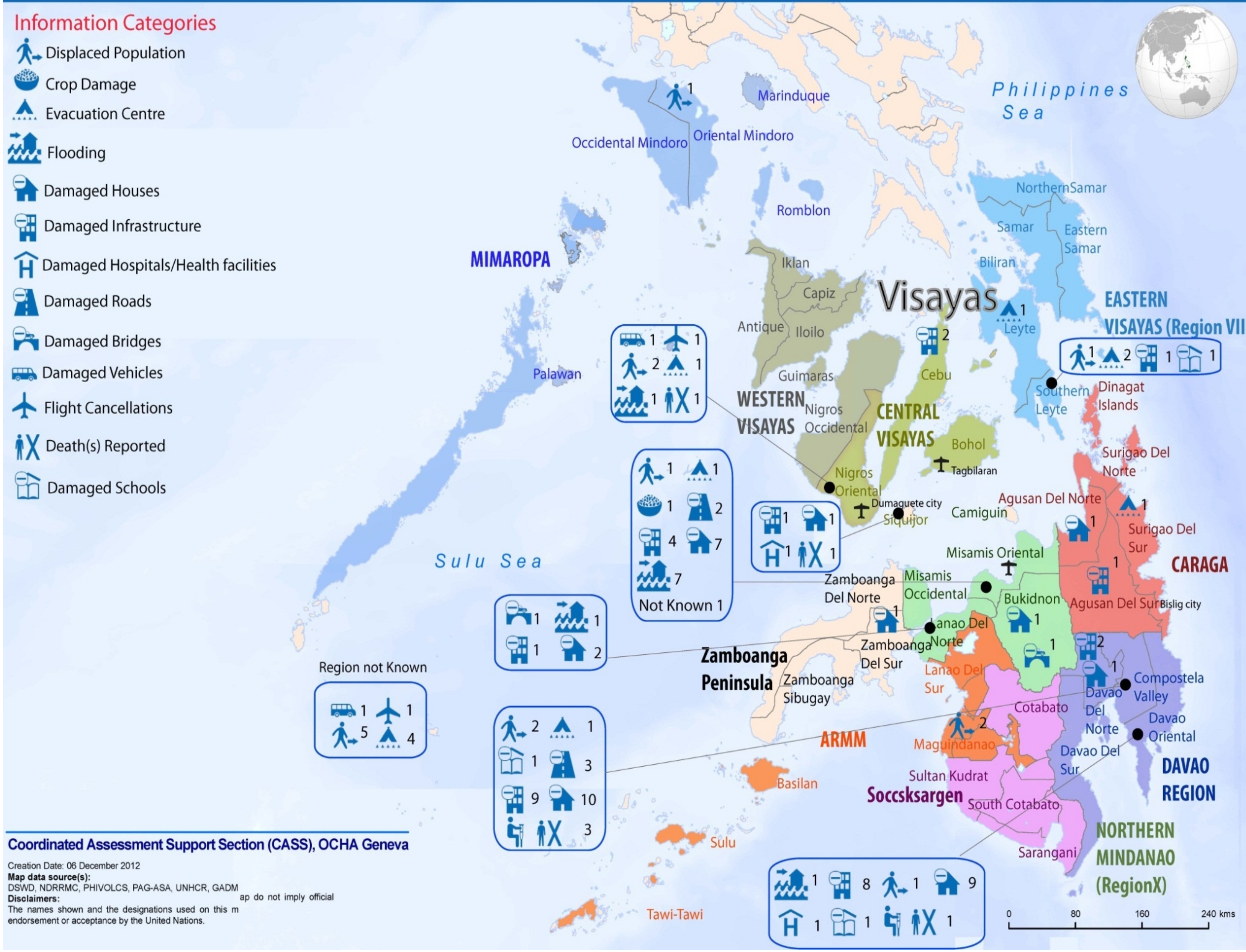
FIND USEFUL MATERIALS

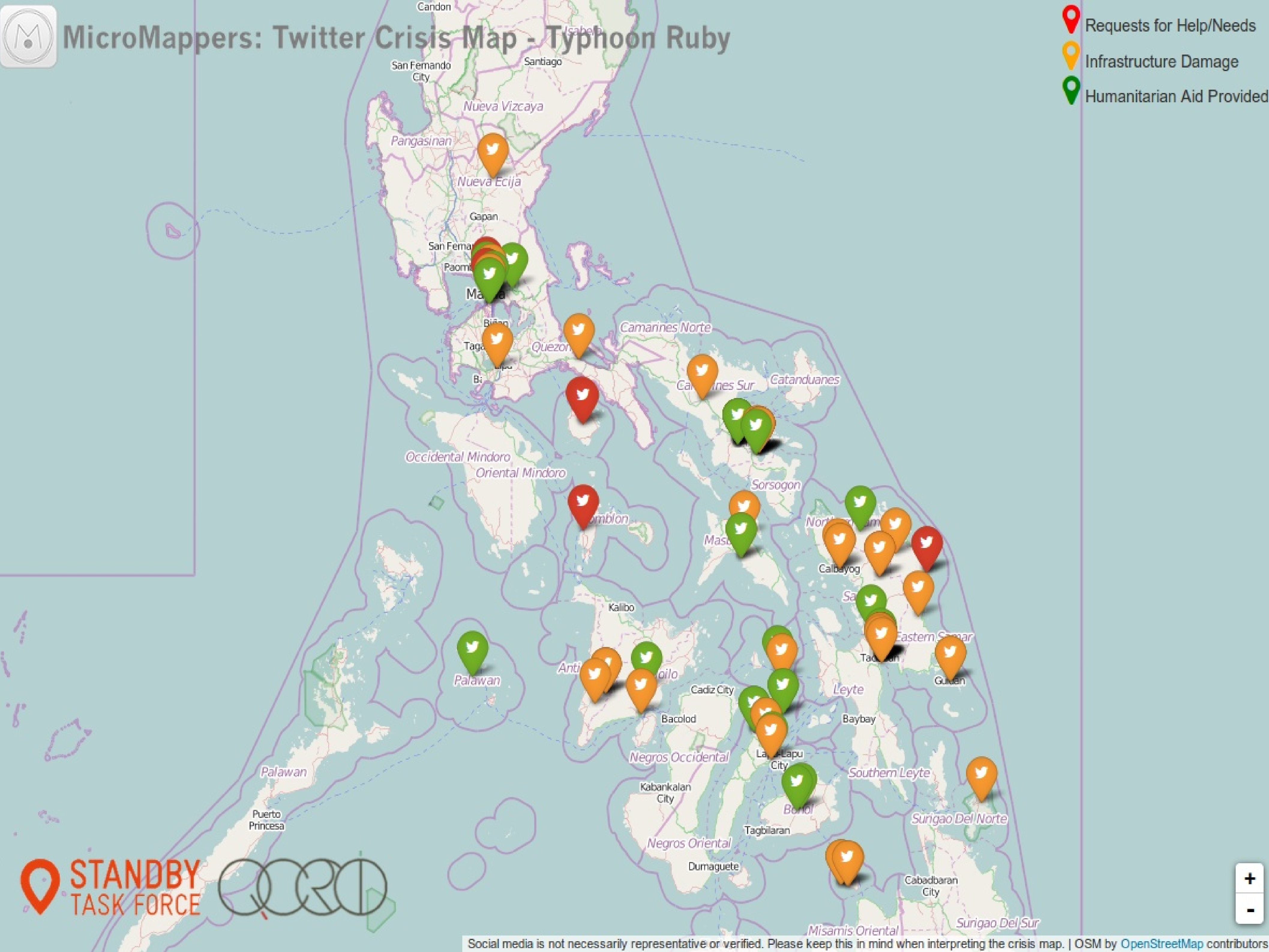
On volunteer & technical communities, past activations, effectively collaborating and more.

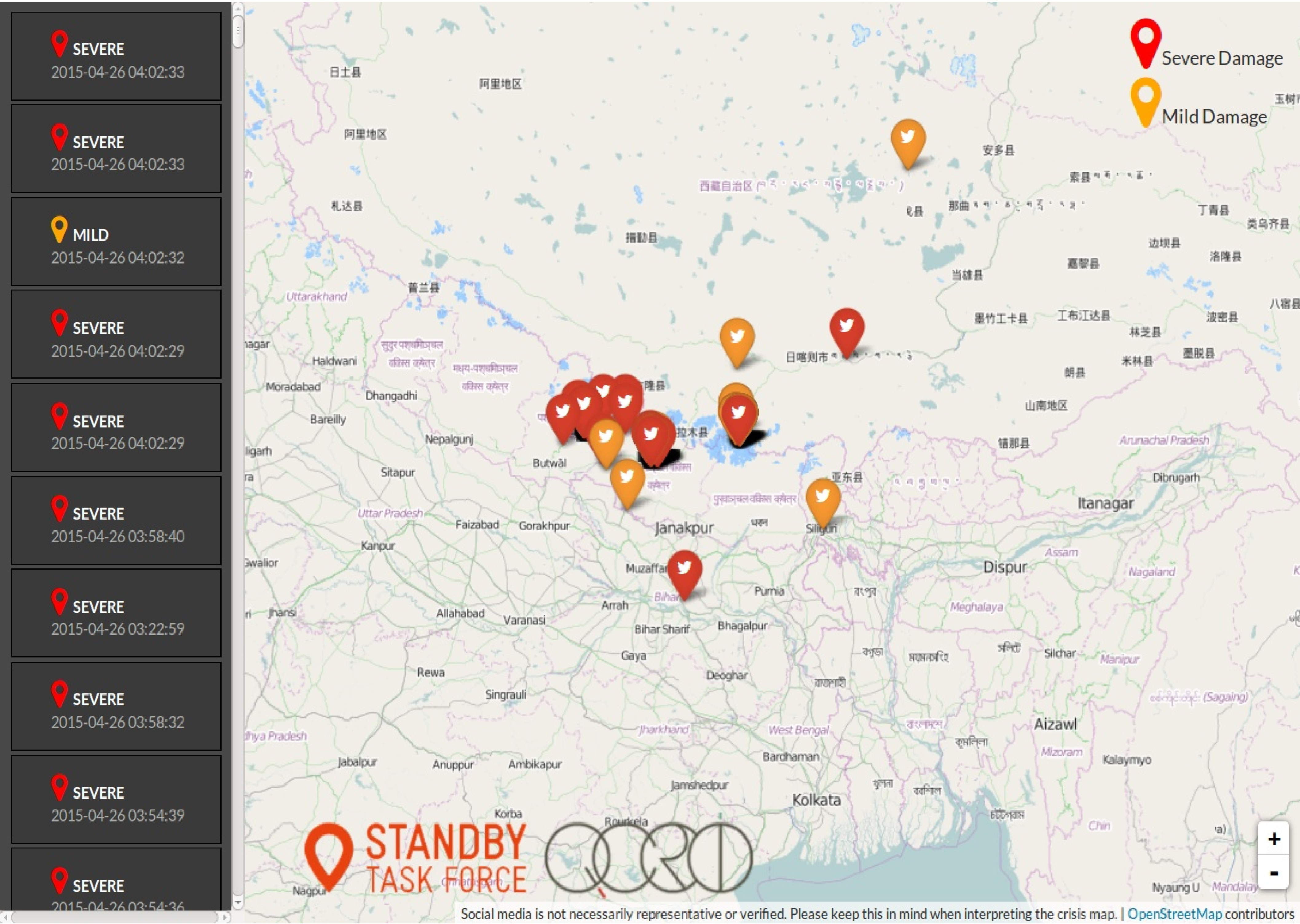
Read More

Philippines - Typhoon Pablo Social Media Mapping (6 December 2012)

OCHA




MicroMappers: Twitter Crisis Map - Typhoon Ruby
Requests for Help/Needs
Infrastructure Damage
Humanitarian Aid Provided
Candon
San Fernando City
Santiago
Nueva Vizcaya
Pangasinan
Nueva Ecija
Gapan
Camarines Norte
Catanduanes
Occidental Mindoro
Oriental Mindoro
Sorsogon
Calbayog
Eastern Samar
Kalibo
Palawan
Cadiz City
Bacolod
Leyte
Baybay
Guiuan
Negros Occidental
Southern Leyte
Kabankalan City
Puerto Princesa
Tagbilaran
Surigao Del Norte
Negros Oriental
Dumaguete
Cabadbaran City
Misamis Oriental
Surigao Del Sur
STANDBY TASK FORCE
Social media is not necessarily representative or verified. Please keep this in mind when interpreting the crisis map. | OSM by OpenStreetMap contributors



SEVERE
2015-04-26 04:02:33
SEVERE
2015-04-26 04:02:33
MILD
2015-04-26 04:02:32
SEVERE
2015-04-26 04:02:29
SEVERE
2015-04-26 04:02:29
SEVERE
2015-04-26 03:58:40
SEVERE
2015-04-26 03:22:59
SEVERE
2015-04-26 03:58:32
SEVERE
2015-04-26 03:54:39
SEVERE
2015-04-26 03:54:36
Severe Damage
Mild Damage
Uttarakhand
Haldwani
Moradabad
Dhangadhi
Bareilly
Nepalgunj
Butwāl
Sitapur
Uttar Pradesh
Faizabad
Gorakhpur
Janakpur
Kanpur
Muzaffar
Arrah
Allahabad
Varanasi
Jhansi
Bihar Sharif
Bhagalpur
Purnia
Gaya
Rewa
Deoghar
Singrauli
Jharkhand
West Bengal
Jabalpur
Anuppur
Ambikapur
Bardhaman
Jamshedpur
Kolkata
Korba
Rourkela
Siliguri
Itanagar
Dibrugarh
Arunachal Pradesh
Dispur
Assam
Nagaland
Meghalaya
Silchar
Manipur
Aizawl
Mizoram
Kalaymyo
Chin
Nyaung U
Mandalay
STANDBY TASK FORCE
Social media is not necessarily representative or verified. Please keep this in mind when interpreting the crisis map. | OpenStreetMap contributors

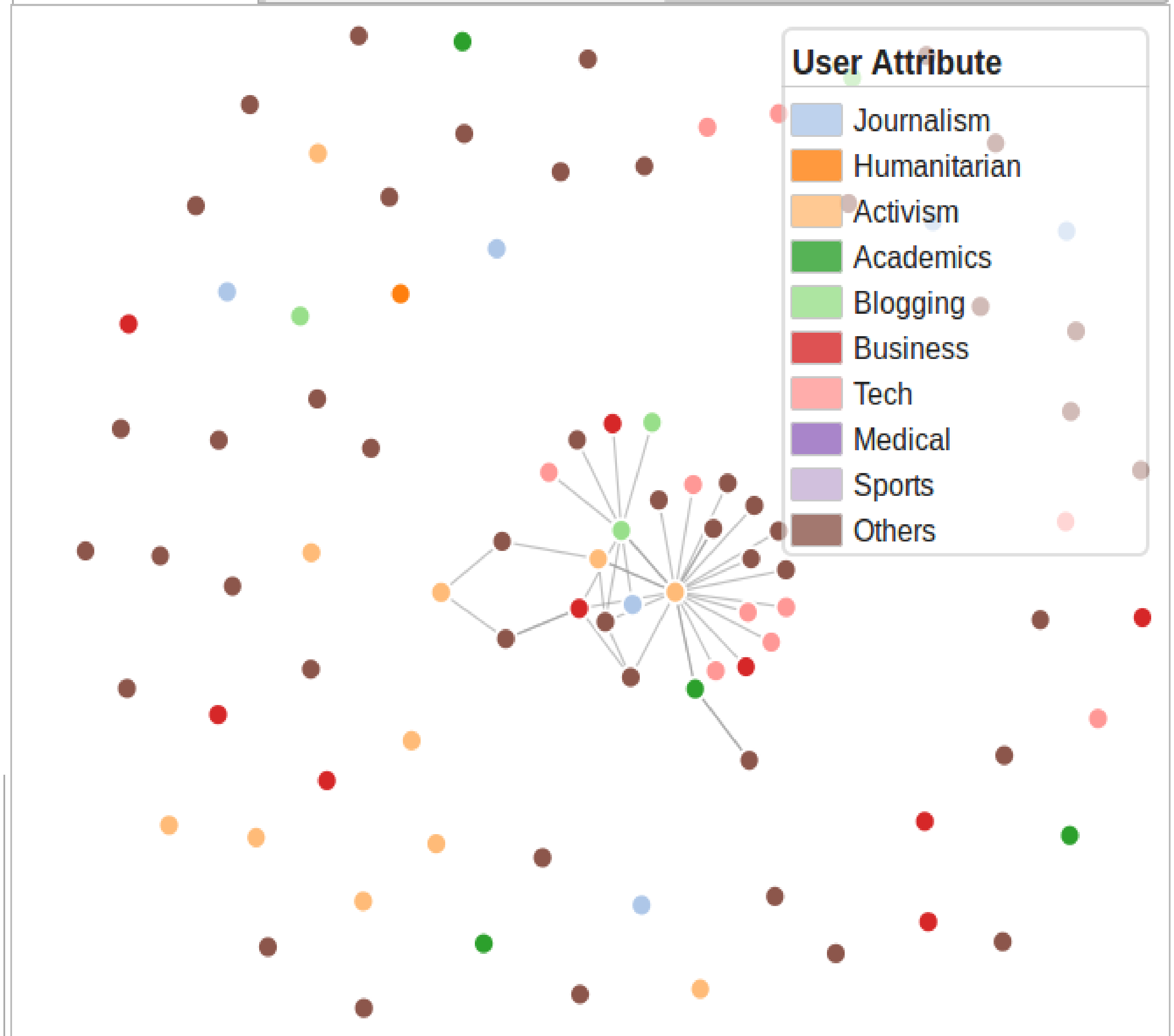
KNOESIS.ORG

Twitris Monitoring support for http://incrisisrelief.org for North East India Floods 2014

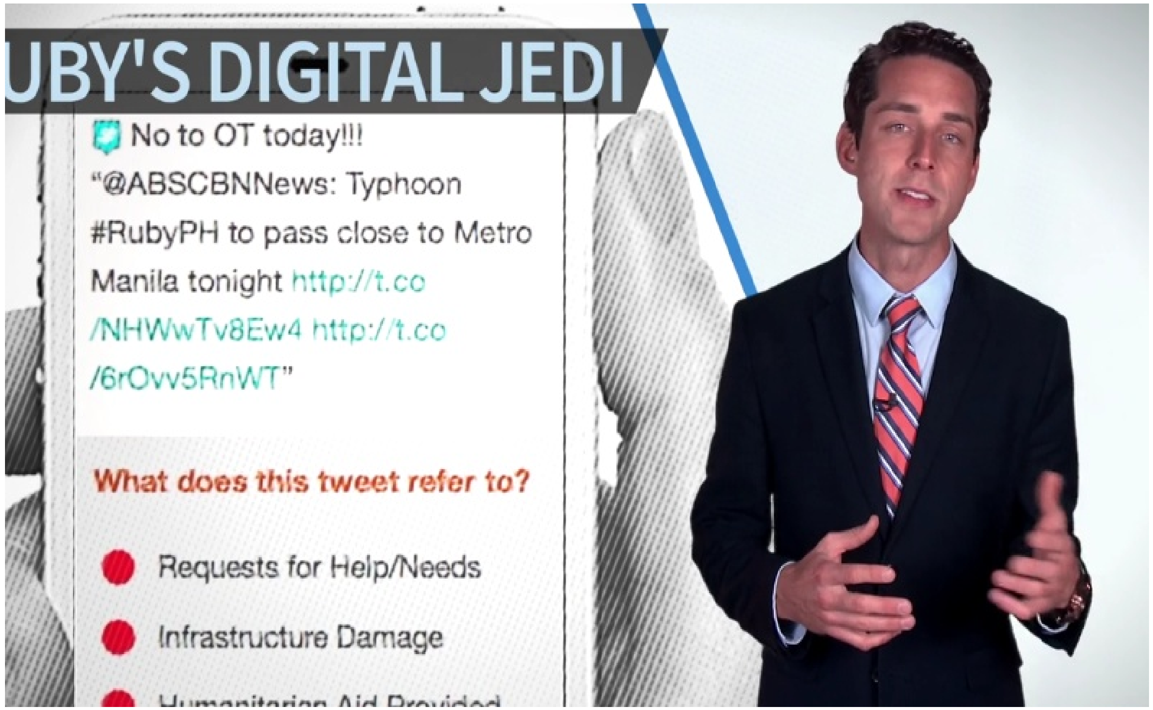
Emerging Community Leaders to engage with

- Money Donations
- Medical Needs
  - HUMANITARIAN
  - OTHERS
- General Needs
- Volunteer Help

Network Graph | User and Tweet Information

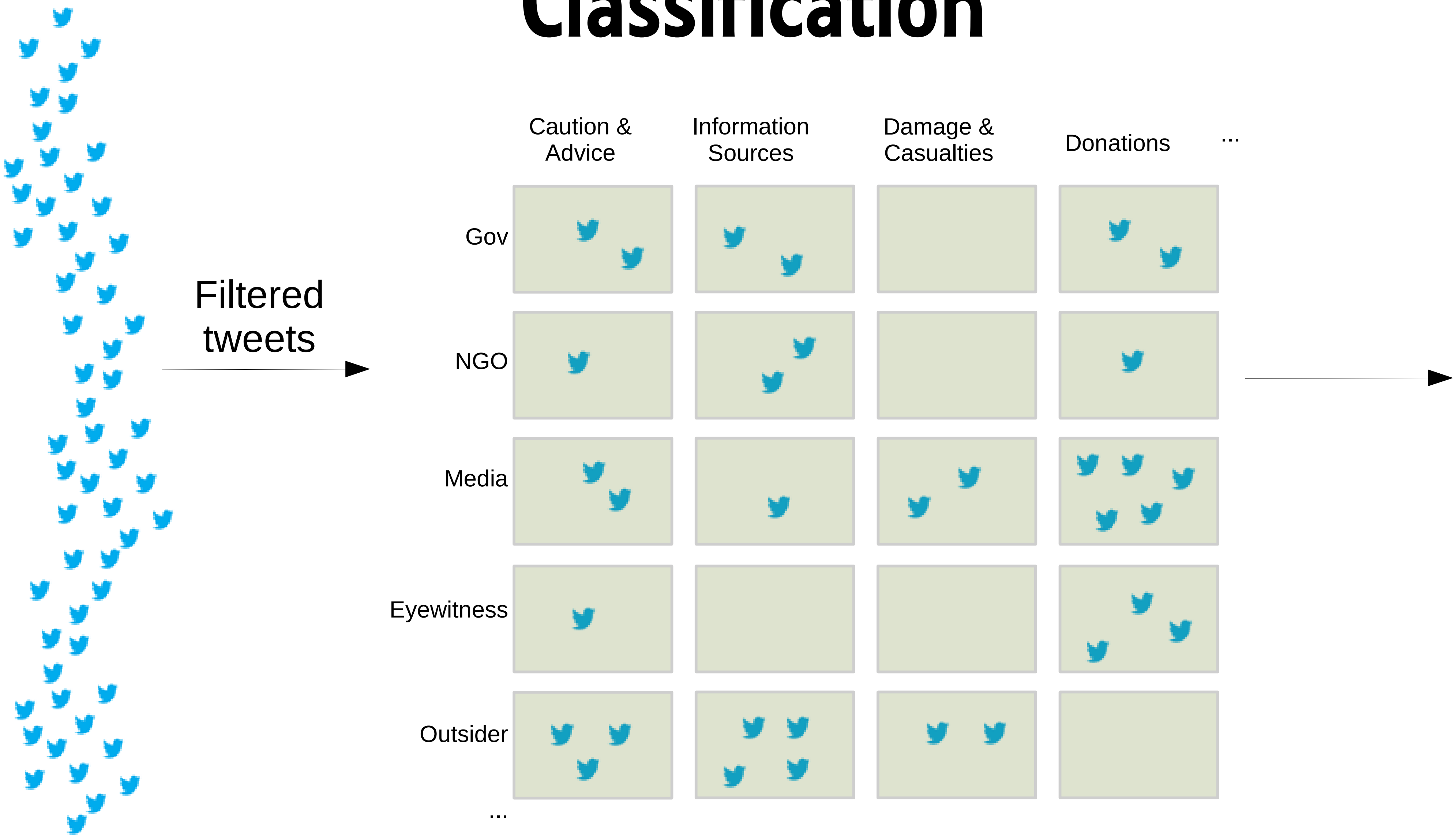


NewsBeat Social



http://newsbeatsocial.com/watch/0_s6xxcr3p

Classification



Classification Axes

- By usefulness (application-dependent!)
  - Not related, Related but useless, Useful
- By factual, subjective, or emotional content
- By information provided
- By information source
  - Government, NGOs, media, eyewitnesses, etc.
- By humanitarian clusters

Humanitarian Clusters

Of the 15% of tweets potentially relevant to the humanitarian community:

Overview

Number of tweets in the first 48 hours

442,560

4.1%
18,076 tweets

Education and Child Welfare

RT @AdamsonUni: Classes and work at all levels are suspended today Nov 8 in anticipation of Typhoon Yolanda. Stay safe Adamsonians. #wala

3.3%
14,602 tweets

Early Recovery

Doing relief efforts now for #YolandaPH.
Need free shipping line info.

1.8%
8,002 tweets

Telecommunication

MTSAT enhanced-IR satellite image of #YolandaPH
as of 2:30 am 09 November 2013: http://.../ RT @govph

Humanitarian Clusters (cont.)

1.8%
8,002
tweets

Telecommunication

MTSAT enhanced-IR satellite image of #YolandaPH
as of 2:30 am 09 November 2013: http://.../ RT @govph

1.8%
7,884
tweets

Safety and Security

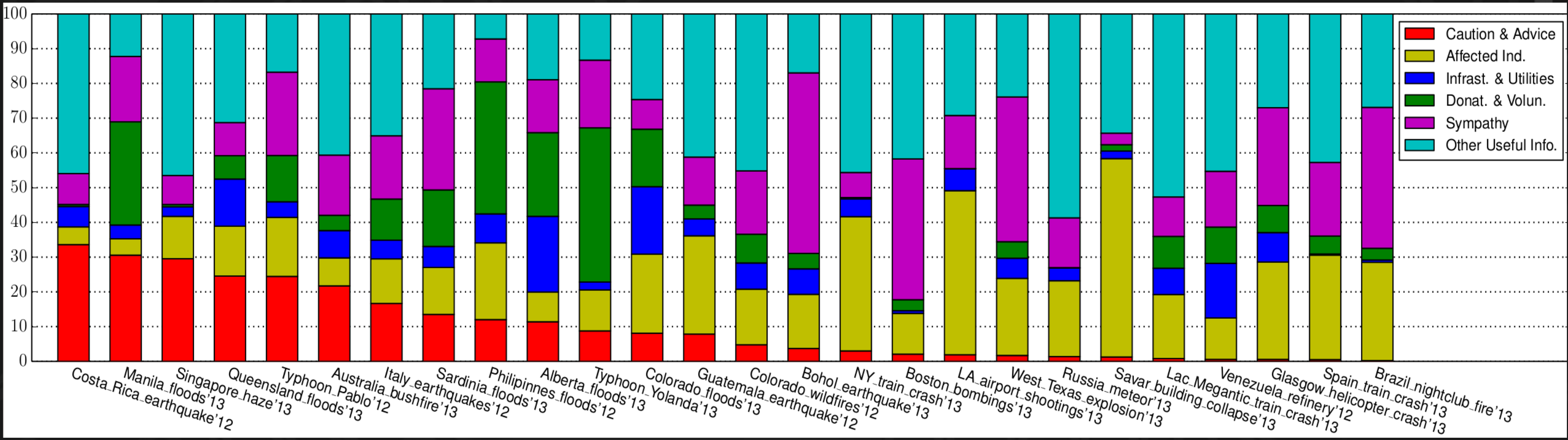
7000 kid's parents have been killed by the storm in the Philippines
and #StayStrongJustin is trending... Ridiculous http://.../

1.1%
4,712
tweets

Food and Nutrition

Red Cross asks for help from police / military. their trucks w/ food
and water for 25000 families are stopped in Tanauan

Information Provided in Crisis Tweets



N=26; Data available at http://crisislex.org/

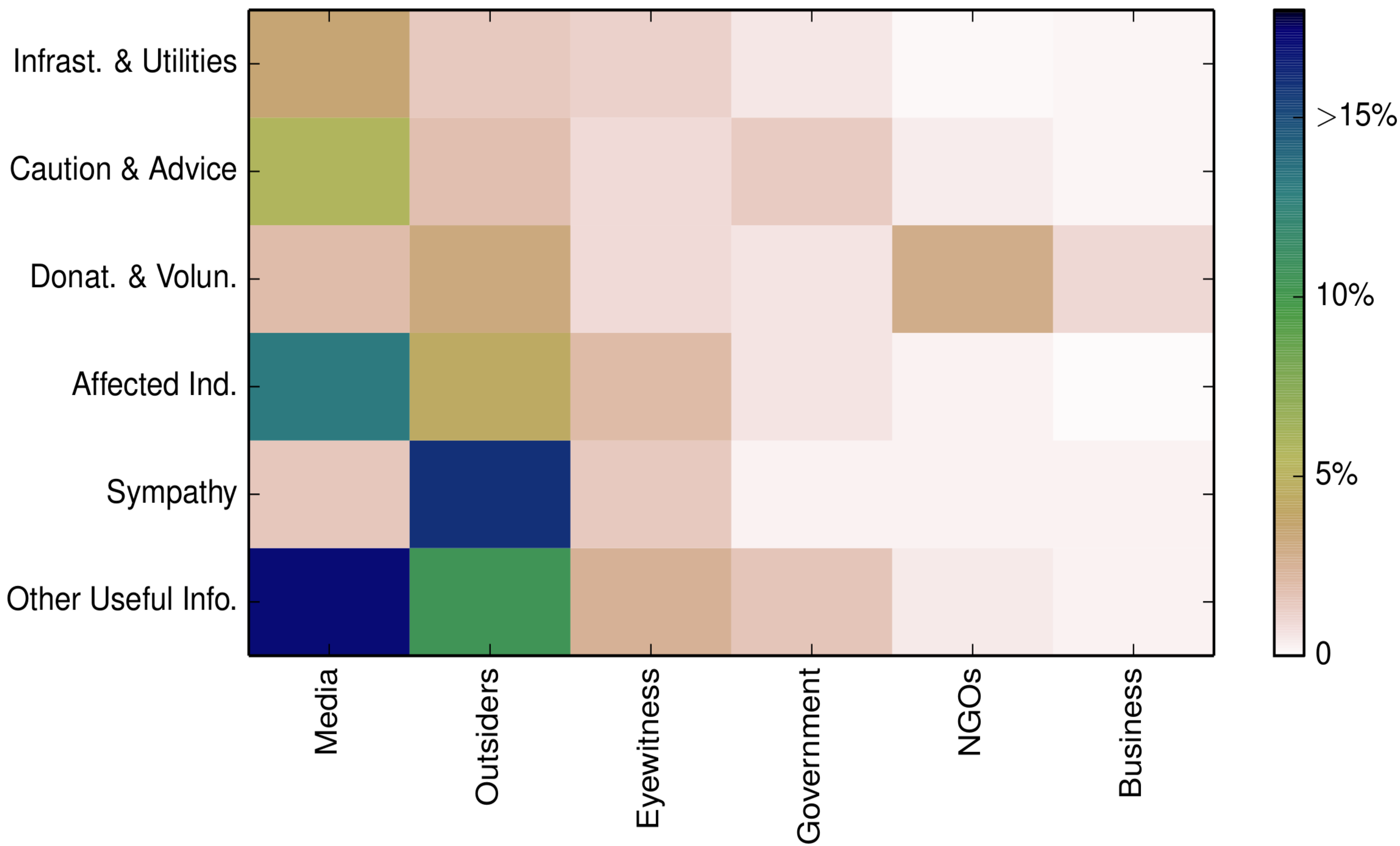
What do people tweet about?

- Affected individuals
  - 20% on average (min. 5%, max. 57%)
  - most prevalent in human-induced, focalized & instantaneous events
- Sympathy and emotional support
  - 20% on average (min. 3%, max. 52%)
  - most prevalent in instantaneous events
- Other useful information
  - 32% on average (min. 7%, max. 59%)
  - least prevalent in diffused events

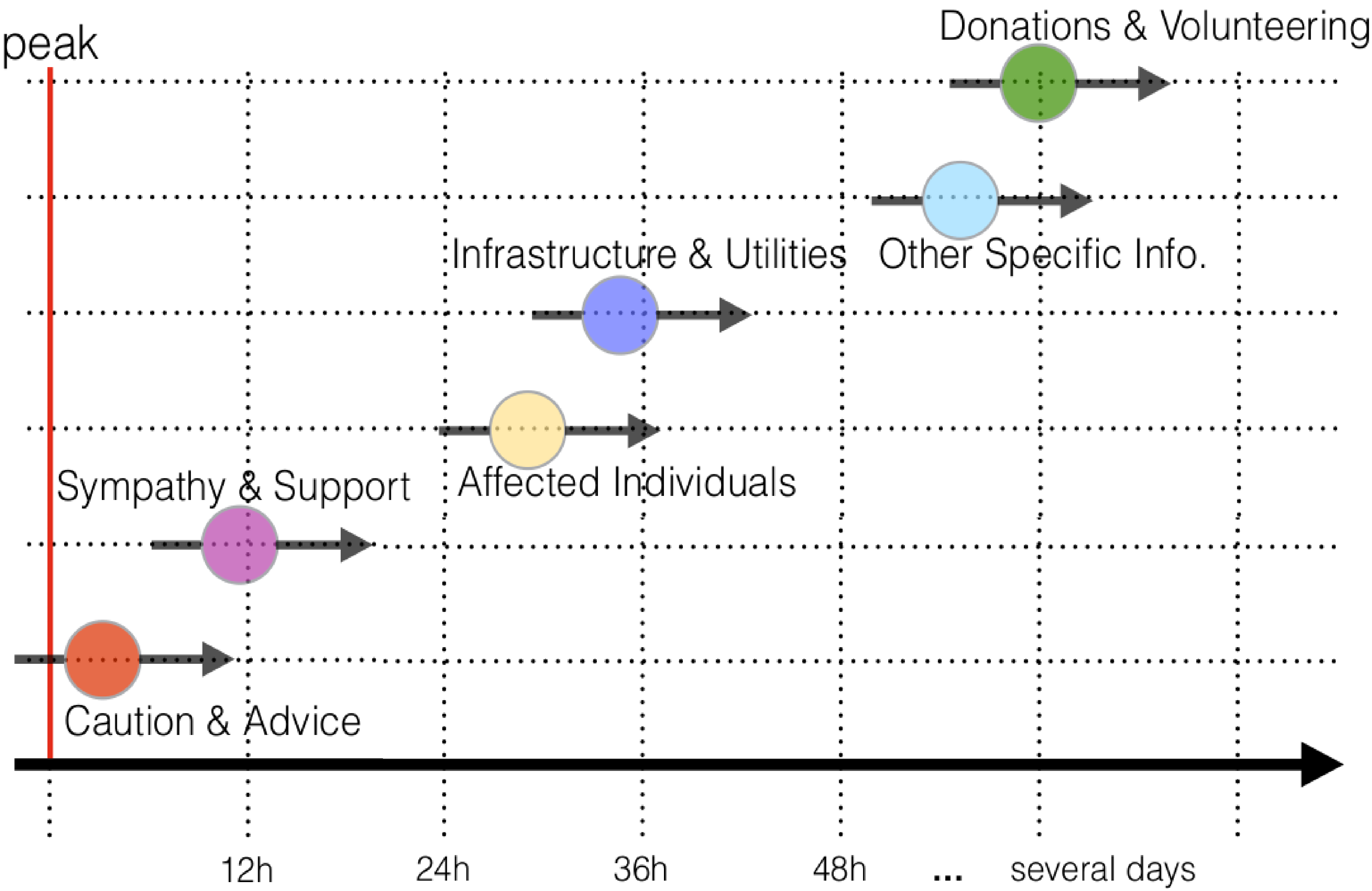
What do people tweet about? (cont.)

- Infrastructure and utilities
  - 7% on average (min. 0%, max. 22%)
  - most prevalent in diffused events, in particular floods
- Caution and advice
  - 10% on average (min. 0%, max. 34%)
  - least prevalent in instantaneous & human-induced events
- Donations and volunteering
  - 10% on average (min. 0%, max. 44%)
  - most prevalent in natural hazards

Distribution over information sources



Distribution over time

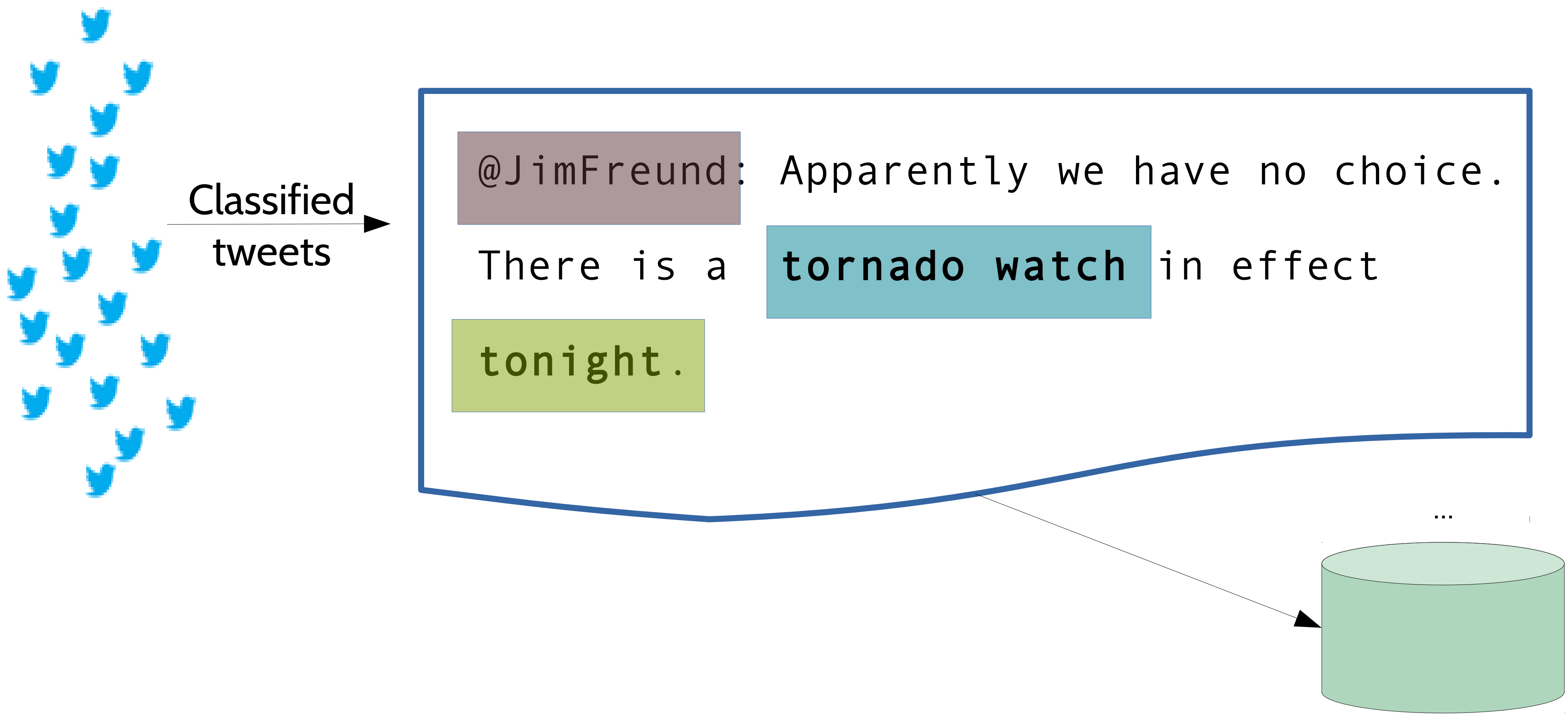


Dataset

CrisisLexT26

www.crisislex.org

Information Extraction



Output examples

✓ RT @weatherchannel: .@NYGovCuomo orders closing of NYC bridges. Only Staten Island bridges unaffected at this time. Bridges must close by 7pm. #Sandy #NYC

✗ Wow what a mess #Sandy has made. Be sure to check on the elderly and homeless please! Thoughts and prayers to all affected

✓ RT @twc_hurricane: Wind gusts over 60 mph are being reported at Central Park and JFK airport in #NYC this hour. #Sandy

✓ RT @mitchellreports: Red Cross tells us grateful for Romney donation but prefer people send money or donate blood dont collect goods NOT best way to help #Sandy

Outline of rest of this part

- Example 1: Readability
- Example 2: Credibility
- AIDR

Example 1/2: Readability

- The ease with which text can be understood
- History
  - Started in early 20th century
  - Purpose: grade school texts
  - Vocabulary, syntax, structure
  - Classical approach: readability formulae
- Modern approaches: machine learning

Readability (Plain English Campaign)

If there are any points on which you require explanation or further particulars we shall be glad to furnish such additional details as may be required by telephone. 28 words

If you have any questions, please phone. 7 words

Typical readability problems

- Misspellings
- Unknown or unfamiliar words
- Unknown abbreviations and acronyms
- Long sentences
- Too many hashtags
- Non-standard word ordering
- No connectives
- Ambiguous syntax
- Impersonal style and passive voice

Readability in Crisis Communications

- During crises people have limited time
- Texts that are hard to read require more time
- Texts that are hard to read can be misleading

Data: 15 crises

- 15 events from CrisisLexT26 in countries with majority of native English speakers
- “Informative” tweetes from media+gov. +NGOs

Crisis	Country
2013 Alberta floods	Canada
2013 Australia bushfires	Australia
2013 Bohol earthquake	Philippines
2013 Boston bombings	USA
2013 Colorado Floods	USA
2013 Glasgow helicopter crash	UK
2013 Los Angeles airport shooting	USA
2013 Lac Mégantic train crash	Canada
2013 Manila floods	Philippines
2013 New York train crash	USA
2013 Queensland floods	Australia
2013 Savar building collapse	Bangladesh
2013 Singapore haze	Singapore
2013 Typhoon Yolanda	Philippines
2013 West Texas explosion	USA

Data Annotation

- Used CrowdFlower
  - Annotators in AU, CA, NZ, UK, USA
  - 5 annotators/tweet
  - Instructions and quiz before starting
- Annotated 500 tweets
- Pre-processing: Removed “RT @user:”
- Only tweets with a weighted measure of agreement θ >= 0.66 selected

#SGhaze update: 3-hour PSI at 5pm is 73, in ’moderate’ range, 24-hr PSI is 52-65. @NEAsg
(Posted during the 2013 Singapore haze)

This tweet:

- ○ Is very CLEAR - easy to understand
- ○ Needs slights IMPROVEMENT to be clear
- ○ Is very UNCLEAR - hard to understand

How would you improve this tweet?

Free text, optional

Feel free to re-write the tweet completely.

All tweets with confidence $>= \theta$	301	100.0%
Is very CLEAR - easy to understand	247	82.1%
Needs slight IMPROVEMENT to be clear	36	12.0%
Is very UNCLEAR - hard to understand	18	6.0%

Very Unclear

Tweet	Crisis	Source
[最新] 截至 9 点，本地空气污染指数狂飙，达 290 点，属于非常不健康水平！公众请多留意！ #sghaze	2013 Singapore haze	Media
//t.co/Lti7AeKB8a or call 1-800-621-FEMA Plz RT	2013 Colorado floods	Government
NDRRMC Update SitRep No. 26 re Effects of Typhoon PABLO (BOPHA) as of 13 December 2012. 10:00AM. http://t.co/G8MHAWrq	2013 Typhoon Pablo	Government

Needs Improvement

Tweet	How to improve?	Crisis	Source
#SGHaze: PSI now at 155 as of 10pm. Here's the health advisory from @NEAsg http://t.co/tvG4bIYZYO	Singapore Haze update: Pressure per square inch now at 155 as of 10pm. Here's the health advisory from @NEAsg #SGHaze Pollutant standard index PSI now at 155 as of 10pm. Here's the health advisory from @NEAsg http://t.co/tvG4bIYZYO #SGHaze	2013 Singapore haze	Media
Office of Civil Defense-NCR: Per MMDA flood control info, 50-60% of Metro Manila flooded.	Office of Civil Defense - 50-60% of Metro Manila flooded Office of Civil Defense- National Capital Region: Per Metropolitan Manila Development Authority flood control info, 50-60% of Metro Manila flooded.	2013 Manila floods	Media

Very Clear

Tweet	Crisis	Source
Deadly quake hits Philippines http://t.co/ERb2CjSwzf	2013 Bohol earthquake	Media
Breaking: Flood maps for Brisbane River are now available http://t.co/2ExK39rY #bigwet	2013 Queensland floods	Media
Colorado Springs POLICE are closing PALMER PARK as a PRECAUTION ONLY!!!!! #WaldoCanyonFire	2013 Colorado fires	Government

Statistics

Characteristics of selected tweets in our dataset.

"Unclear" means "Needs Slight Improvement" or "Very Unclear".

	Clear	Unclear	
Average length	108.6	93.1	***
Average num. of words	15.5	14.0	**
Average num. of English words	12.0	7.7	***
Average word length	6.3	6.1	
Average number of acronyms	0.3	0.7	***
Average number of mentions	0.3	0.5	*
Average number of hashtags	1.1	1.2	
Fraction with acronyms	25.5%	64.8%	***
Fraction with mentions	23.5%	38.9%	**
Fraction with URLs	56.3%	22.2%	***
Fraction with URLs in the middle	29.2%	11.1%	***
Fraction with ellipsis	17.8%	14.8%	
Fraction with hashtags (#)	68.8%	87.0%	***
Fraction with # at the beginning	6.1%	37.0%	***
Fraction with # in the middle	31.6%	35.2%	
Fraction with # at the end	37.3%	25.9%	*

*** $p<0.01$
** $p<0.05$
* $p< 0.1$

Readability observations

Tweets should be short, but not shorter than necessary:

- Include a maximum of 1 or 2 main points per tweet
- Use abbreviations and acronyms with care (e.g. PSI in Singapore), simple and familiar words.
- **Bad strategies for shortening tweets can render them unreadable!**
- **Write brief, concise sentences, but avoid incomplete sentences.**

Use Twitter-specific syntax with care:

- Hashtags at the beginning of tweets make them less readable!
- **Include at most 1 or 2 hashtags, and only at the end of the tweet.**
- **Avoid user mentions (i.e. "@user") when possible.**

Next steps: automation?

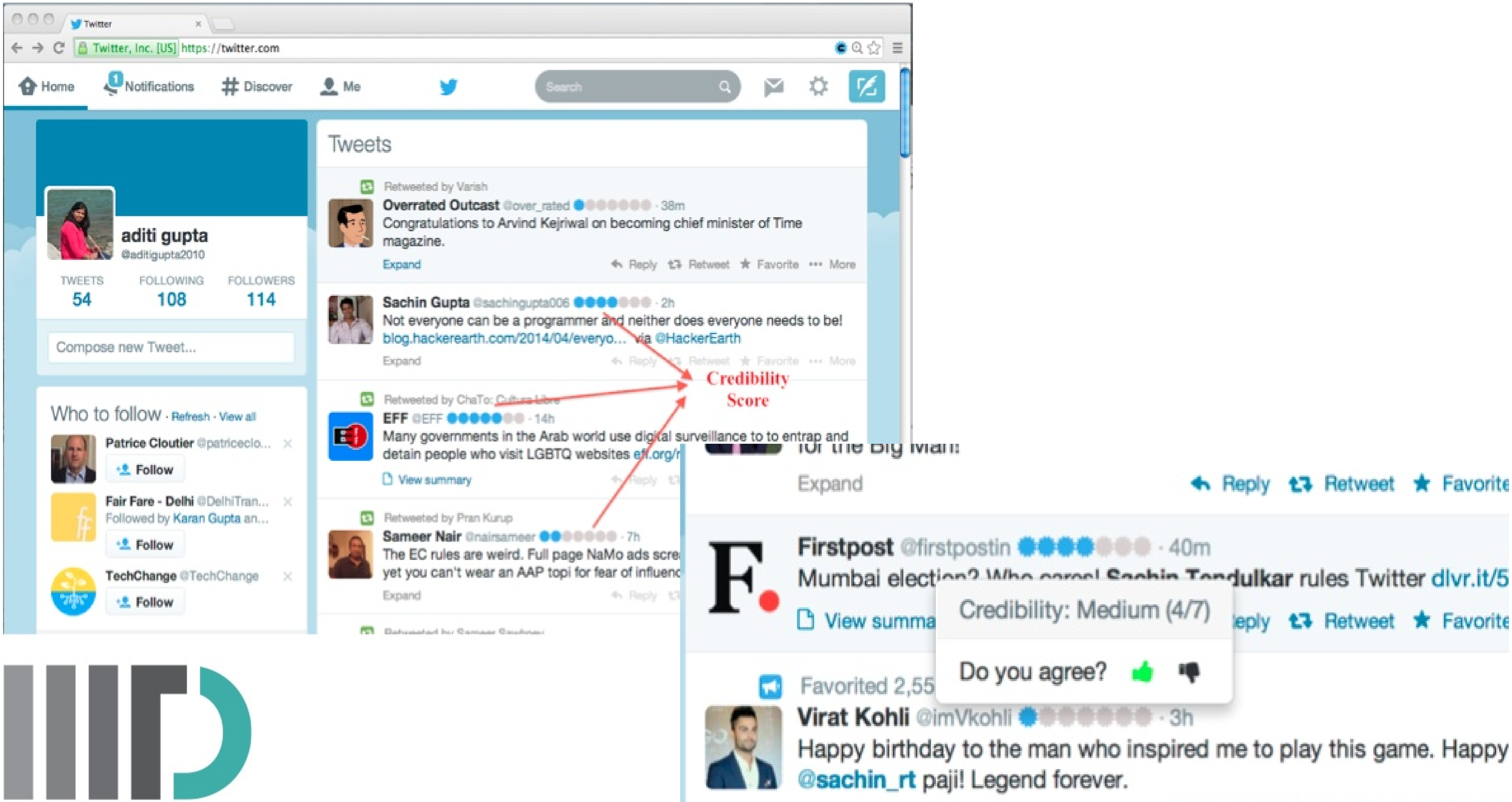
Example 2/2: Credibility

- Another perceived characteristic
- Can be approximated quite well with content-based, user-based, and propagation-based features

Credibility evaluation: TweetCred

- Real-time web-based service
- Used as a Chrome extension
- Annotates Twitter's timeline with credibility scores

http://twitdigest.iiitd.edu.in/TweetCred/



http://aidr.qcri.org/

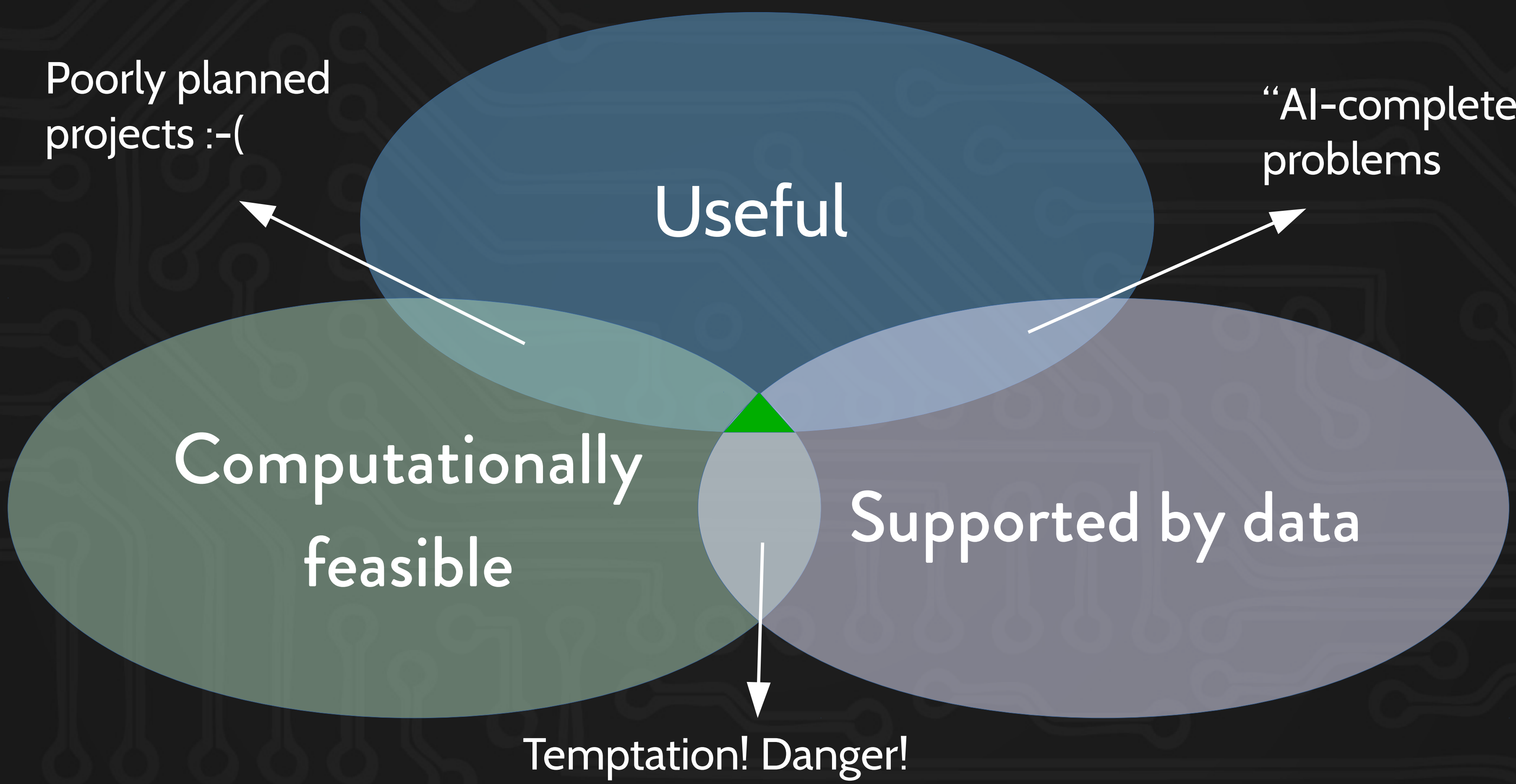
Social Media Mining and Retrieval

Conclusions

Some ethical aspects

- Disclosing private information is not a consent for any usage of this information in any context
- Authenticity, edited self and social anxiety
- Reducing/increasing inequality (gender, race, social class)
- Values embedded in social media platforms
  - Business thrive on disclosure and frame it as a value
  - Marketing strategies used by individuals: what is exactly the product and what is its price?

Finding an Interesting Problem



Things to remember

- Social media is beautifuly chaotic
- Validity vs hype of social media mining
  - Interdisciplinary research is hard but rewarding
- Lots of interesting topics to work on
  - Some of them are also useful
- Happiness is contagious!

Further references

- Tutorial: Twitter and the real world [Weber and Mejova 2013]
  - https://sites.google.com/site/twitterandtherealworld/home
- Social media mining [Zafarani, Abbasi and Liu 2014]
  - http://dmml.asu.edu/smm/book/
- Information and Influence Propagation in Social Networks [Chen, Lakshmanan and Castillo 2013]
  - http://www.morganclaypool.com/doi/abs/10.2200/S00527ED1V01Y201308DTM037