The study of social phenomena via social media

Alessandro Flammini
Indiana University

James S. McDonnell Foundation

DARPA

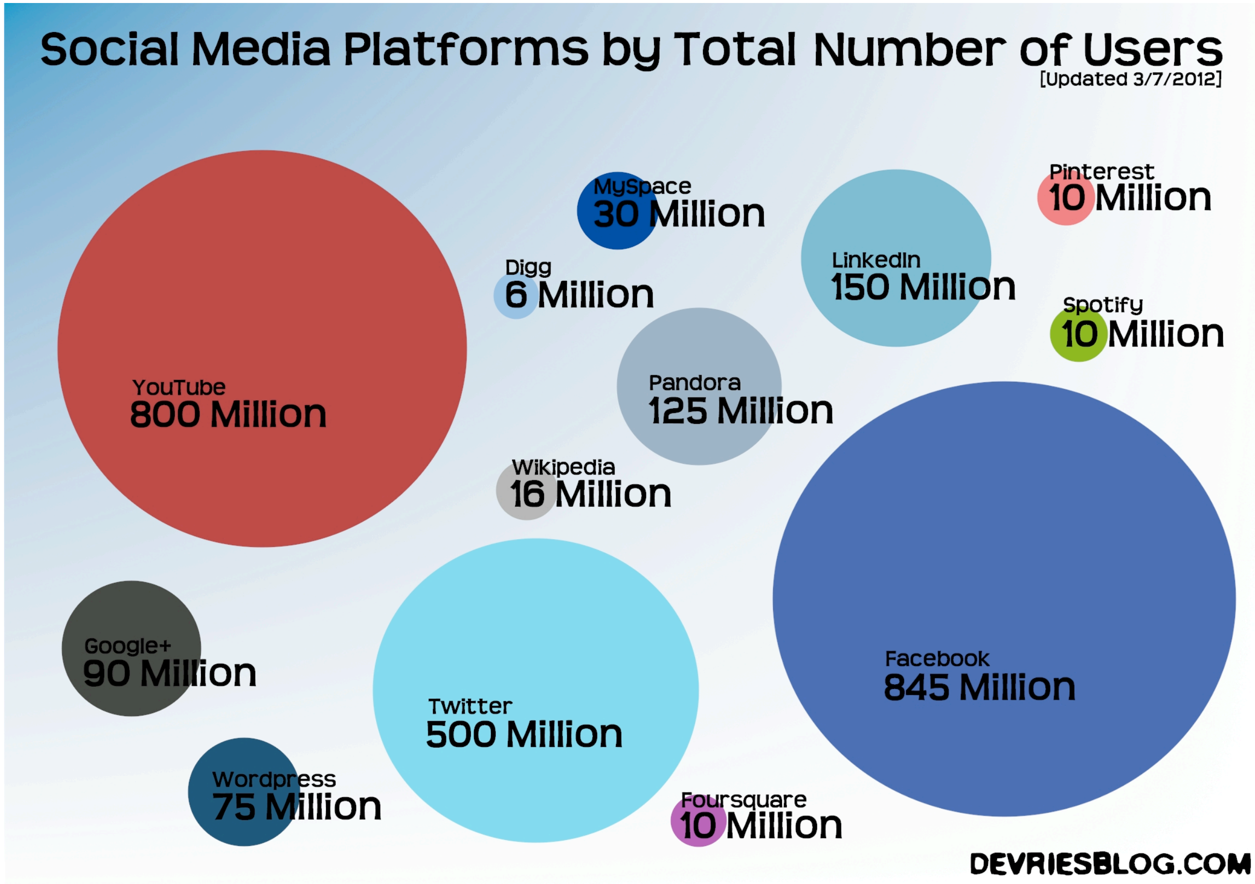
NSF

CNetS

Thanks to:

And to: James S. McDonnell Foundation, DARPA, NSF

Why do we care about Online Social Media?



Information shared on SM matters

People increasingly makes choices based on info share on SM:

- purchasing product
- political information
- health
- lifestyle
- eternal information consumed (TV shows, blogs, news, books, music ....)

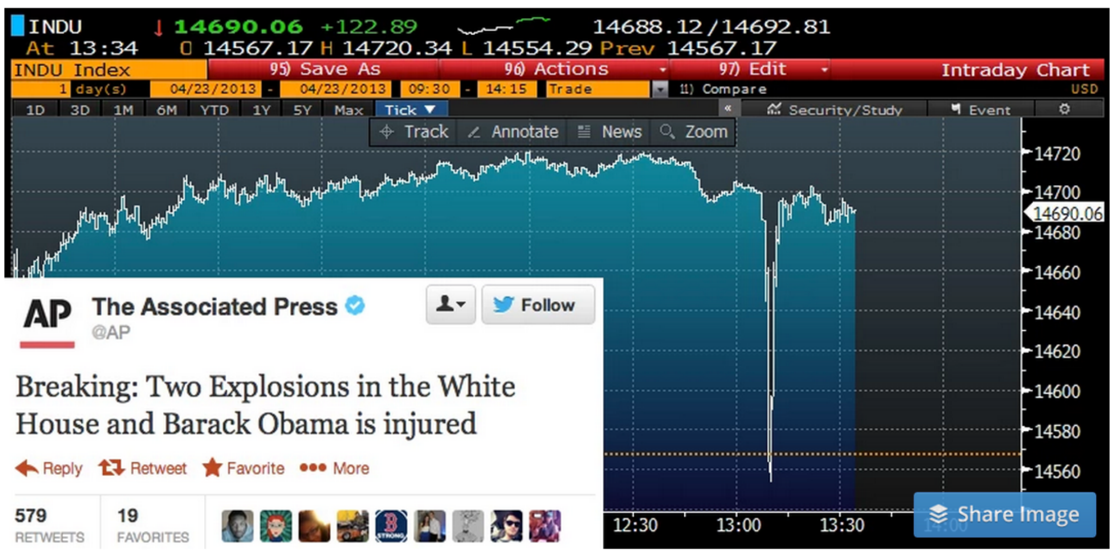
SM have been leveraged to:

- anticipate stock market
- forecast movie revenue
- probe consumer preferences
- coordinate action in crisis
- predict outcome of political elections

WorldViews

Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism?

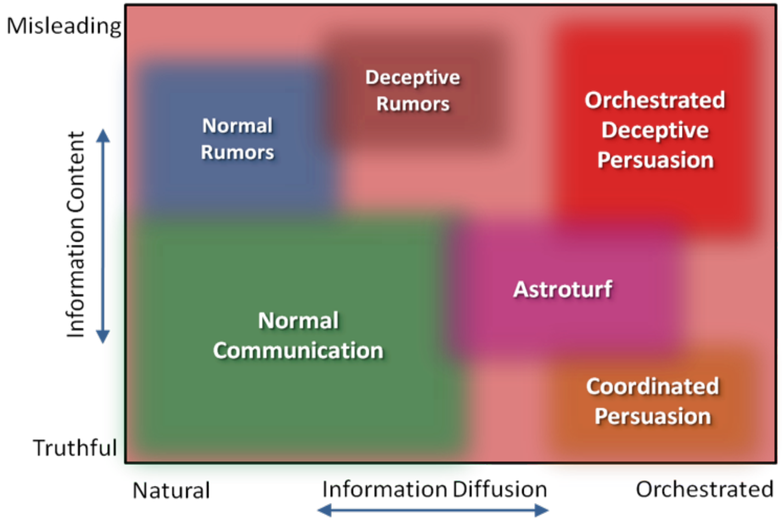
By Max Fisher April 23, 2013 Follow @Max_Fisher



This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

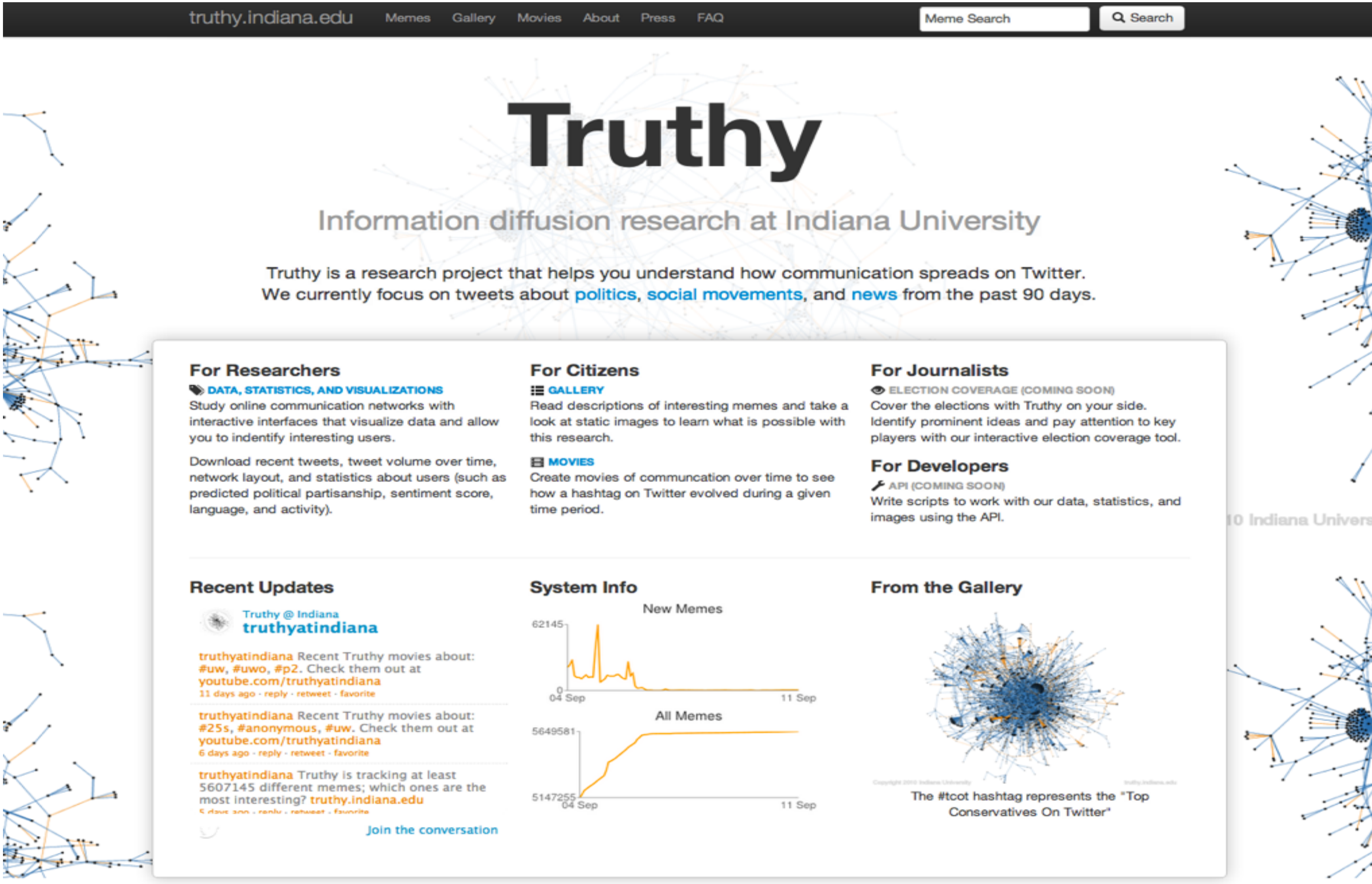
Social Media: many actors, many agendas

- Detecting persuasion campaigns on Twitter

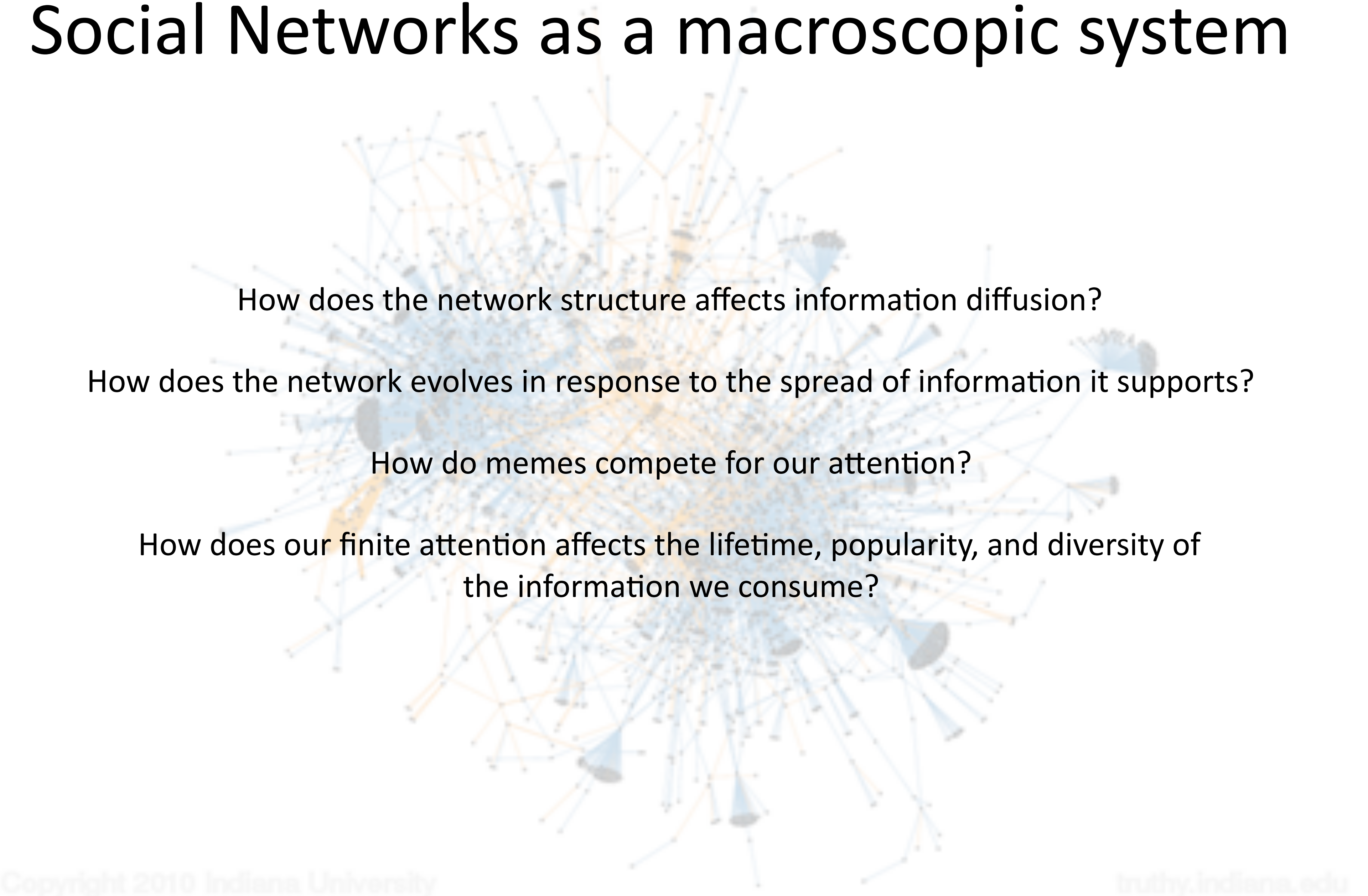


- Automatic detection of artificial users

Visit us @truthy.indiana.edu



Social Networks as a macroscopic system



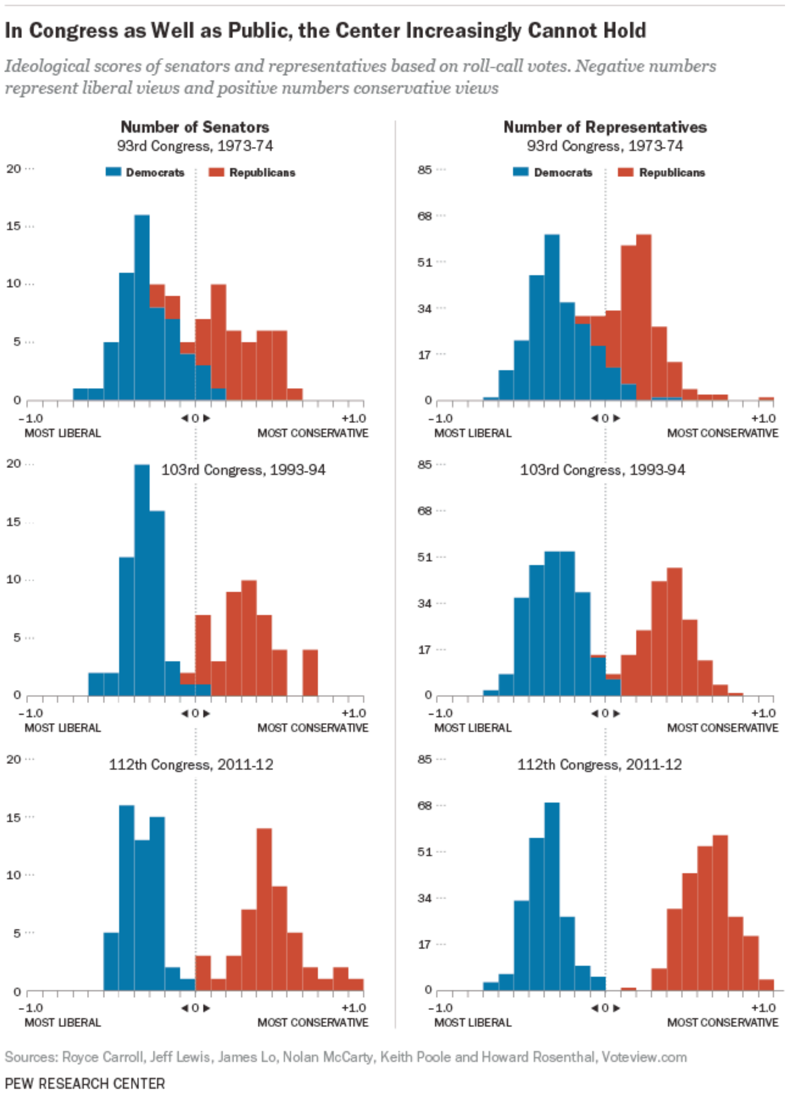
How does the network structure affects information diffusion?

How does the network evolves in response to the spread of information it supports?

How do memes compete for our attention?

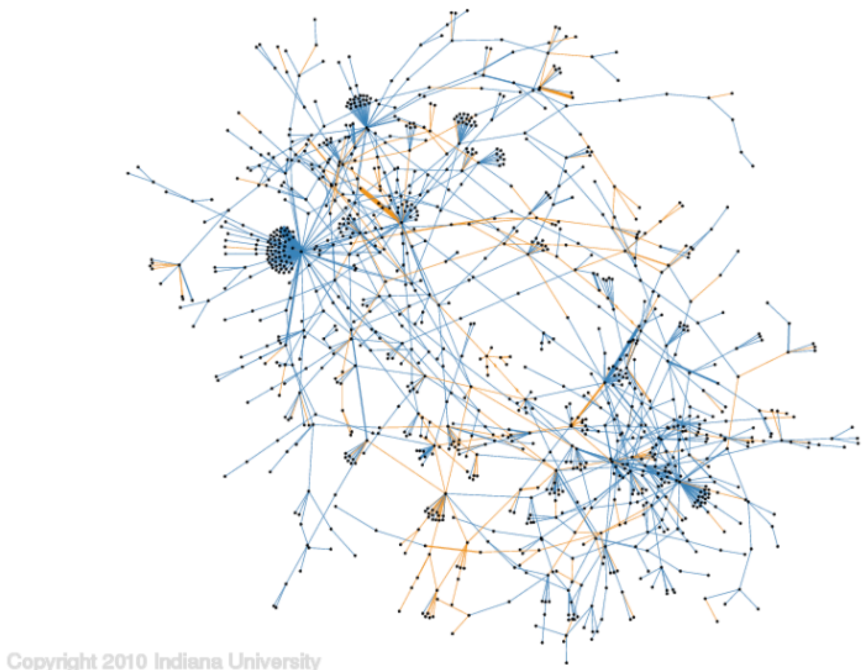
How does our finite attention affects the lifetime, popularity, and diversity of the information we consume?

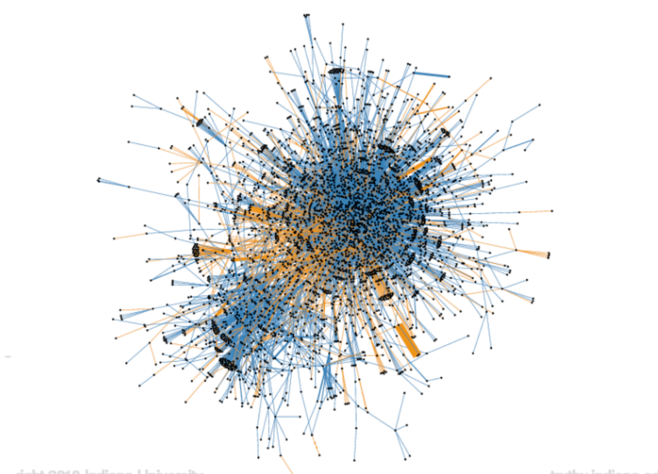
Political discourse on Twitter & the issue of polarization



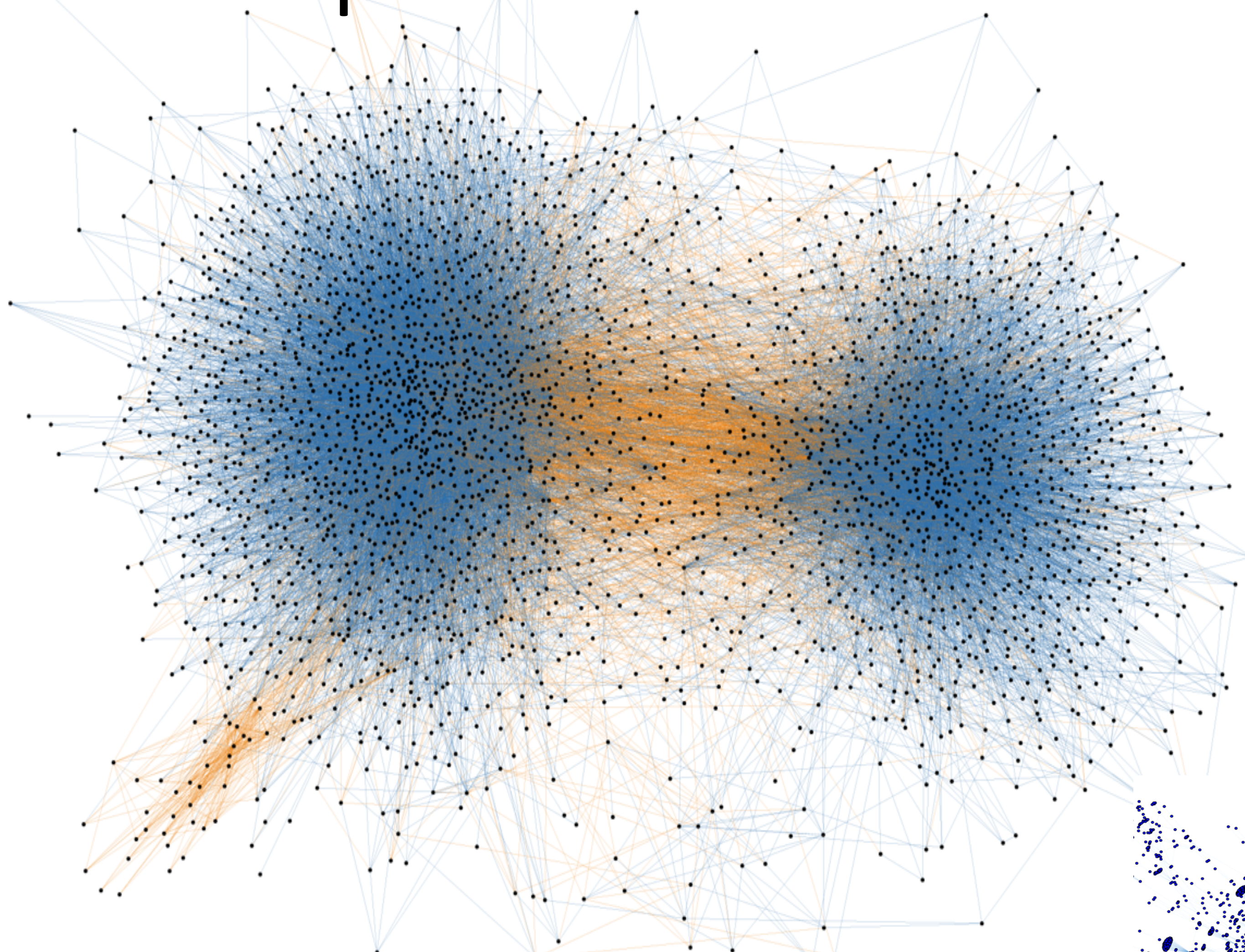
Is people polarized as their congressmen?

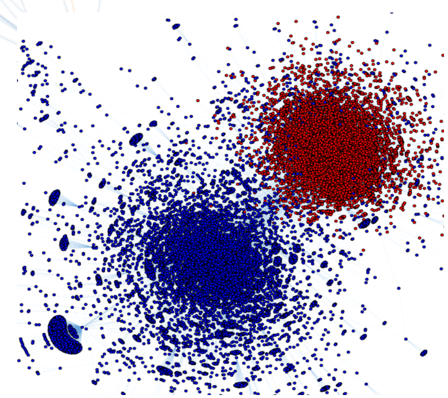
Data and Networks





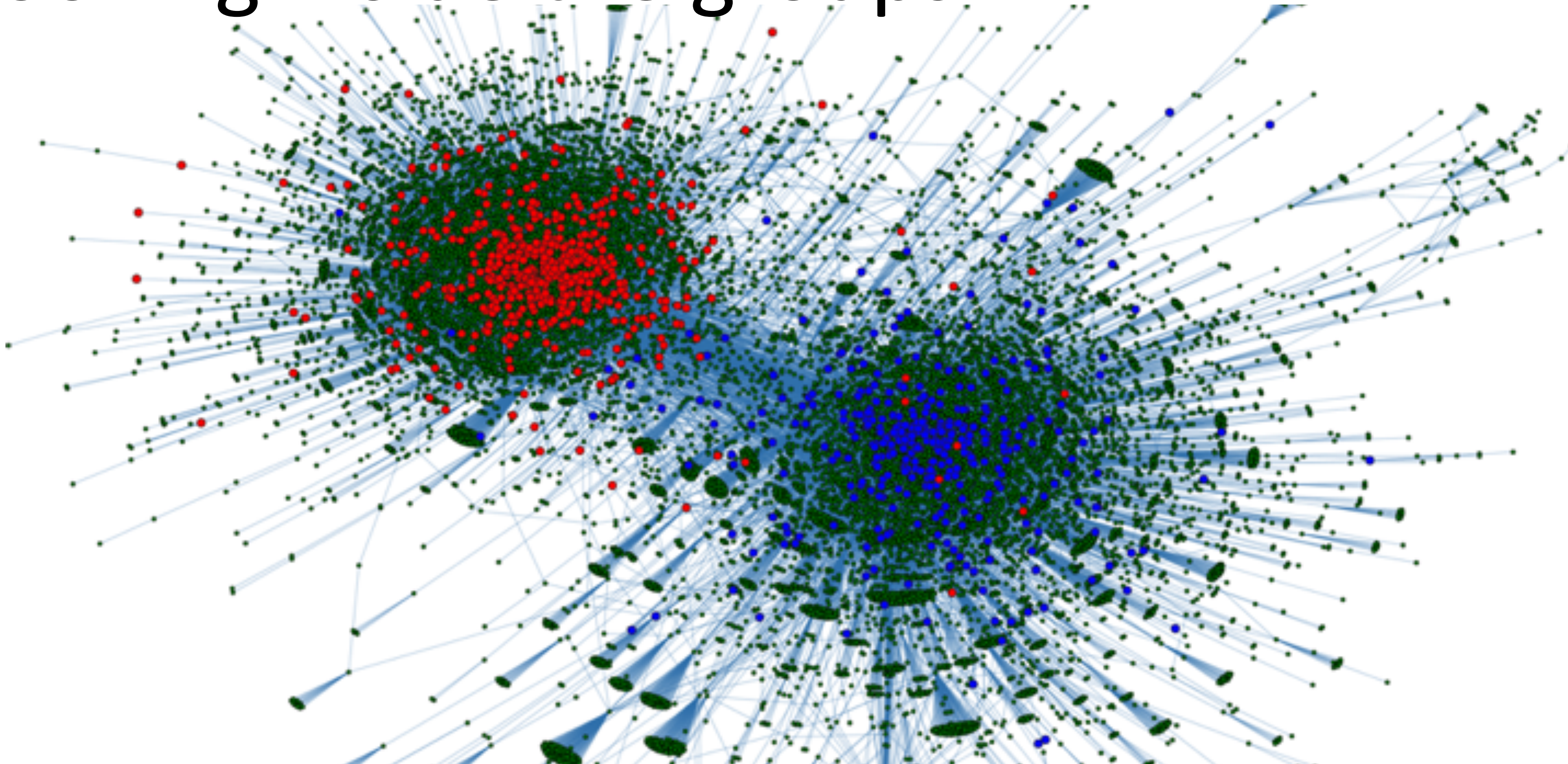
How the political network looks like

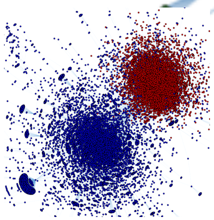


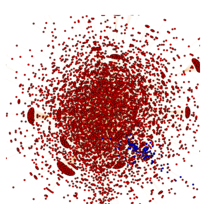


3 months of data. ~300k tweets. ~20k users.
On the right the result of label propagation clustering algorithm

Looking inside the groups



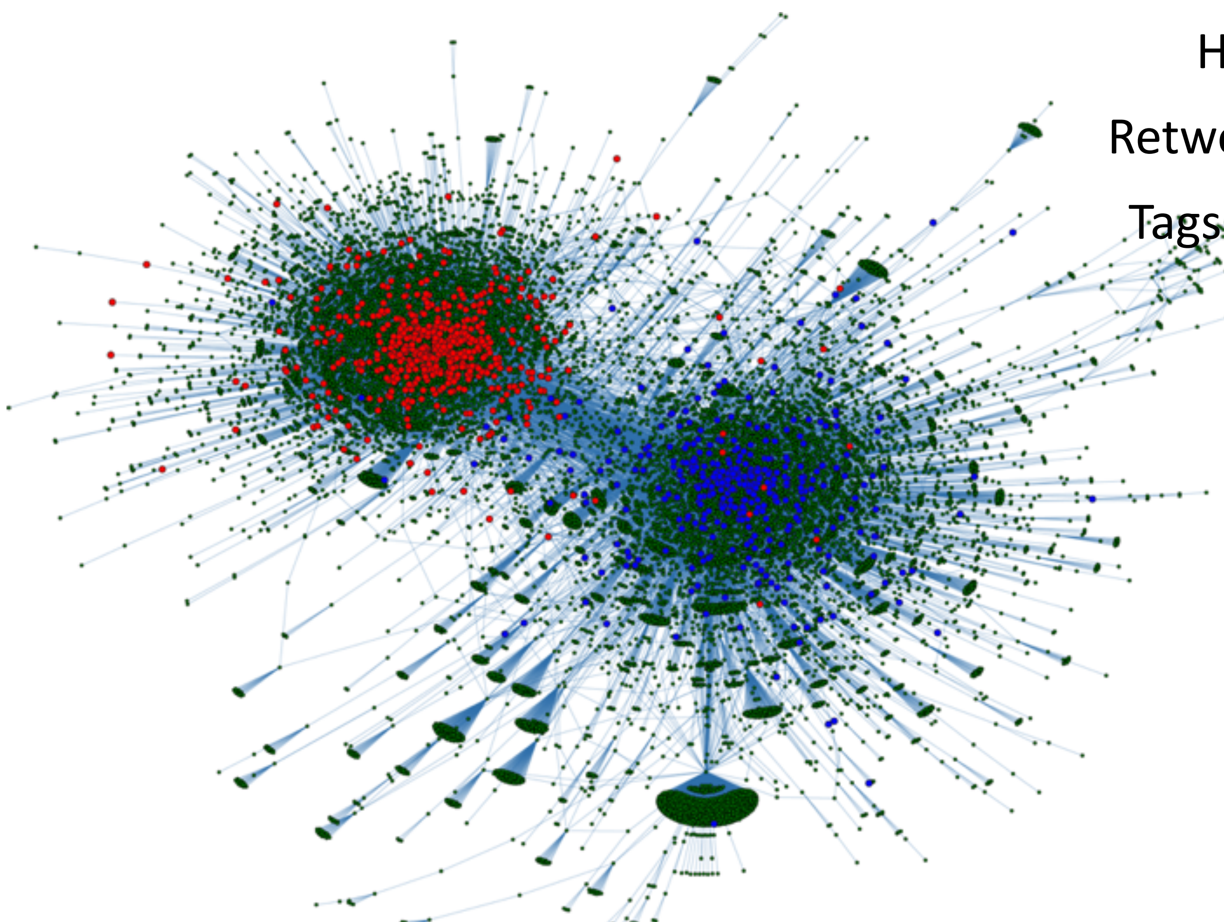




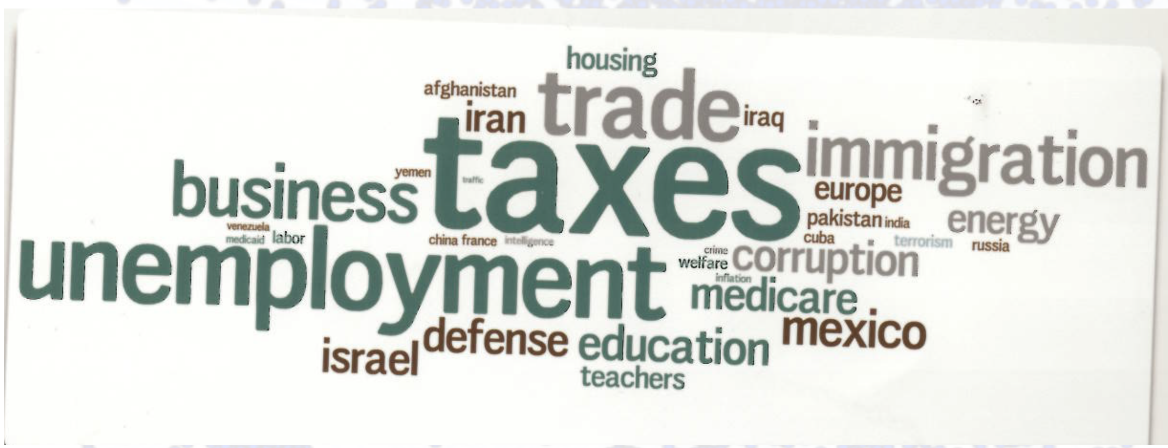
	Left	Right	Und.	Nodes
Retweet	1.1%	93.4%	5.3%	7,115
	80.1%	8.7%	11.1%	11,355
Mention	39.5%	52.2%	8.1%	7,021

Classification of users political leaning

Features	Accuracy
Text (TF-IDF)	79%
Hashtags	91%
Retweet network	95%
Tags + Network	95%



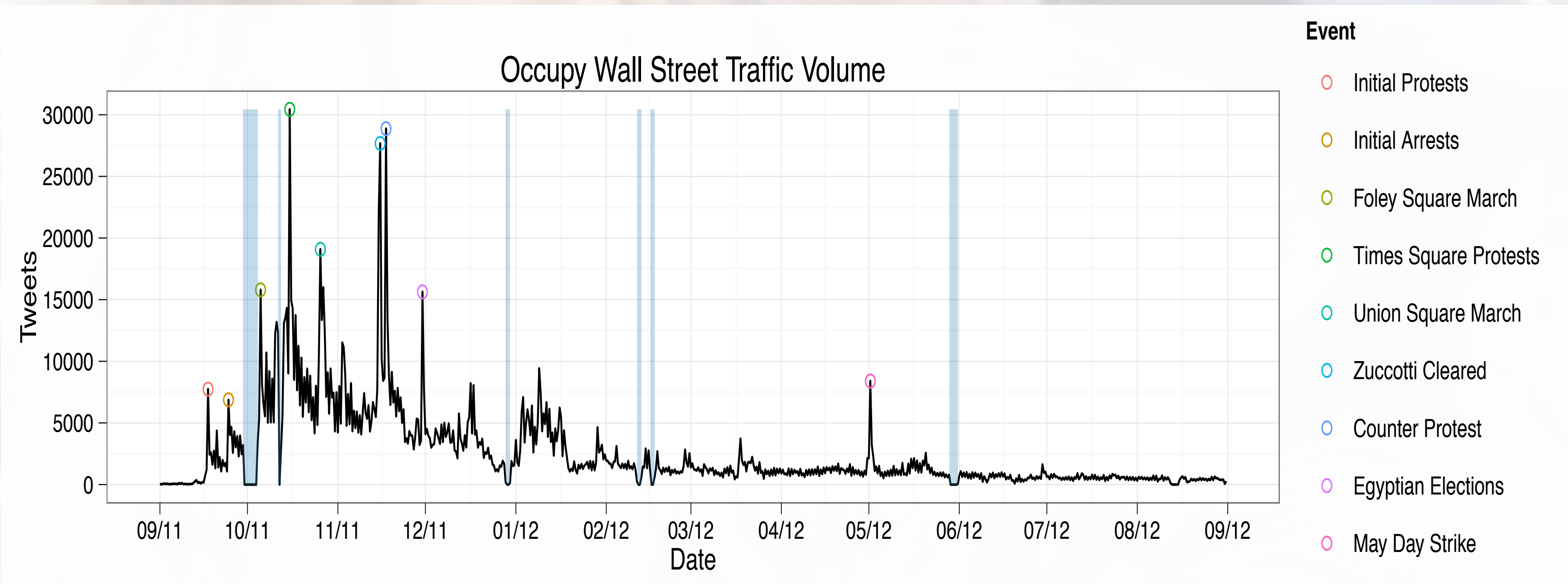
Standard machine learning classification with different classes of features.
Text is in bag of words + TF-IDF representation



Words were ranked according to their Term Frequency – Inverse Document Frequency, Represented here in Log-scale

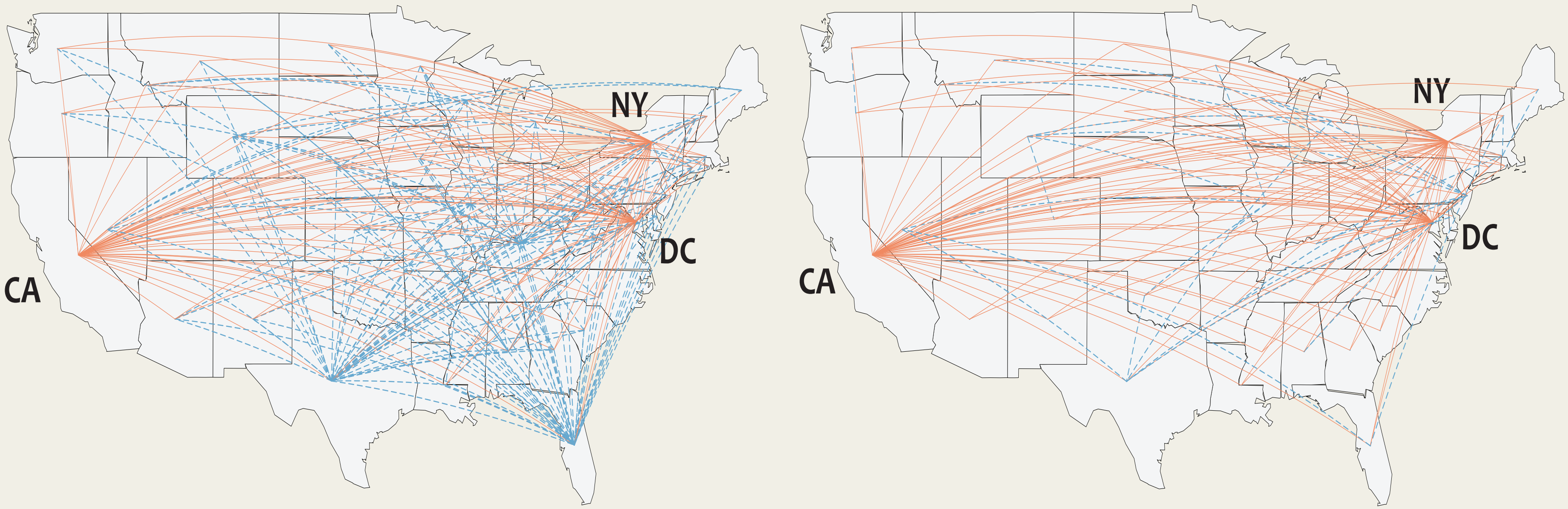
Occupy Wall Street

Twitter traffic & timeline



On-the-ground events (circles).
Twitter data-stream outages (blue bands) are highlighted. Bins have 12-hours length.

The geography of online OWS



Multi-scale backbone extraction – confidence level a = 0.15
Occupy-related discourse (right) shows a prominent hub-and-spoke structure with features distinct from conversations on domestic politics (left).

Collective Framing vs Resource mobilization

Collective framing

Interstate		Intrastate	
Token	Ratio	Token	Ratio
wall	.590	city	2.254
nyc	.600	tonight	1.737
street	.699	march	1.669
news	.718	join	1.494
99%	.756	solidarity	1.387
bank	.763	day	1.354
don't	.782	square	1.333
media	.837	please	1.243
peaceful	.845	park	1.220
nypd	.847	now	1.179

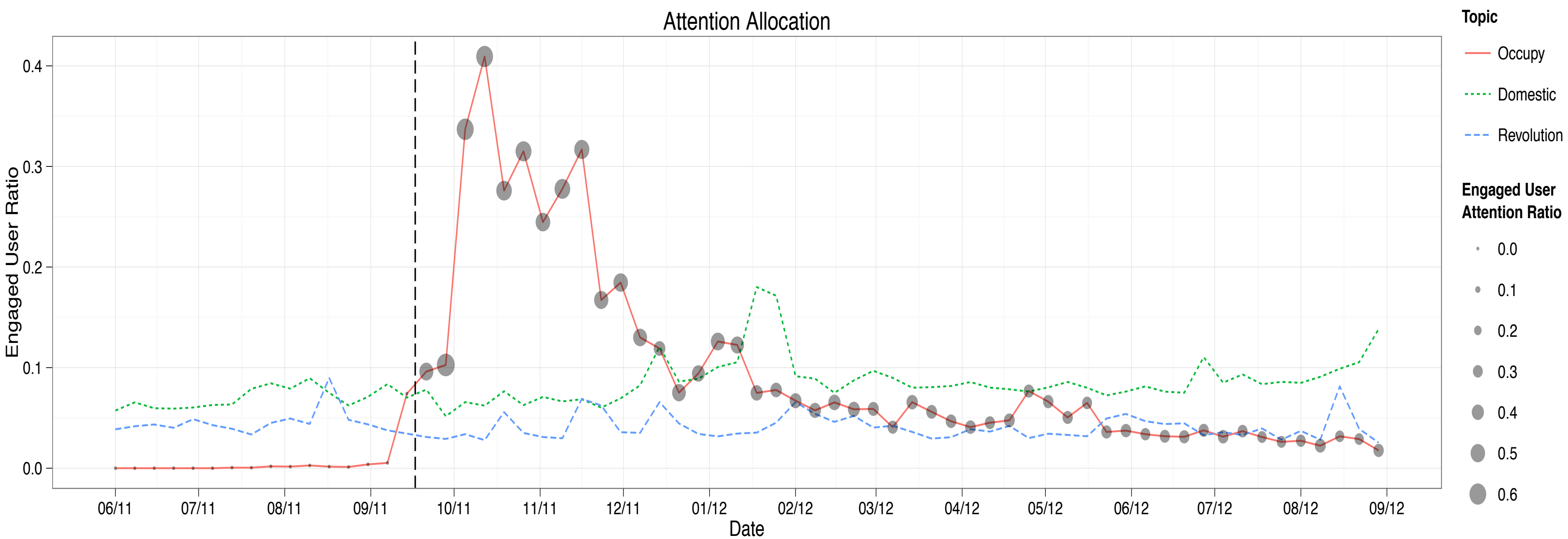
Resource mobilization

Ratio = $\frac{P(Token|Intrastate)}{P(Token|Interstate)}$

Collective framing: the social processes whereby participants negotiate the shared language and narrative frames to define the movement's identity and goals.

Resource mobilization: the work to marshal the physical and technological infrastructure, human resources, and financial capital to sustain ongoing activity.

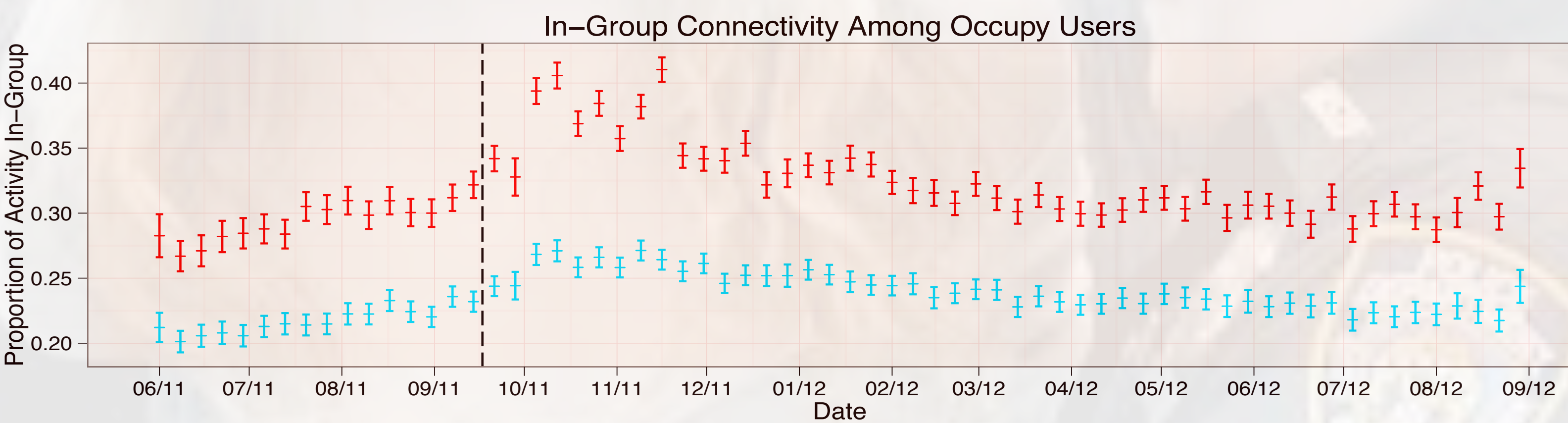
People engagement



Point size is proportional to user engagement, measured as average (across users) ratio of OWS-related tweets over entire production

At the core of the protest

Domestic Politics	Social Movements
#tcot	#syria
#p2	#bahrain
#teaparty	#egypt
#gop	#yemen
#anonymous	#libya
#obama	#tahrir
#tlot	#wiunion
#jobs	#iranelection
#ronpaul	#assange
#romney	#wikileaks
#sopa	#jan25
#ndaa	#14feb
#obama2012	#assad
#ocra	#greece
#twisters	#damascus
#sgp	#gaddafi
#politics	#feb14
#solidarity	#scaf
#gop2012	#antisec
#p21	#arabspring
#topprog	#tunisia
#obamacare	#noscaf
#mapoli	#syrian



Longitudinal connectivity level of 25k users – reported as mean and 95% confidence intervals

Conclusions

SM data can be leveraged to answer questions relevant to social science

Technical issues I carefully hide: data collection, filtering, archiving, retrieval, analysis

May possible biases

Opportunity to observe in the “wild”,
without relying o what people says about the self or others

Complement traditional approaches