The study of social phenomena via social media

Alessandro Flammini
Indiana University
Thanks to:

And to: James S. McDonnell Foundation
Why do we care about Online Social Media?

Social Media Platforms by Total Number of Users
[updated 3/7/2012]

- YouTube: 800 Million
- Facebook: 845 Million
- Twitter: 500 Million
- Wordpress: 75 Million
- Google+: 90 Million
- Foursquare: 10 Million
- Pinterest: 10 Million
- Linkedin: 150 Million
- Spotify: 10 Million
- Pandora: 125 Million
- Myspace: 30 Million
- Digg: 6 Million

DEVRIESBLOG.COM
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
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 affect information diffusion?

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

How do memes compete for our attention?

How does our finite attention affect the lifetime, popularity, and diversity of the information we consume?
Political discourse on Twitter & the issue of polarization

In Congress as Well as Public, the Center Increasingly Cannot Hold

Ideological scores of senators and representatives based on roll-call votes. Negative numbers represent liberal views and positive numbers conservative views.

Number of Senators
93rd Congress, 1973-74

Number of Representatives
93rd Congress, 1973-74

103rd Congress, 1993-94

112th Congress, 2011-12

Sources: Bryce Carroll, Jeff Lewis, James Lo, Nolan McCarty, Keith Poole and Howard Rosenthal, Votesview.com

PEW RESEARCH CENTER

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

<table>
<thead>
<tr>
<th></th>
<th>Left</th>
<th>Right</th>
<th>Und.</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweet</td>
<td>1.1%</td>
<td>93.4%</td>
<td>5.3%</td>
<td>7,115</td>
</tr>
<tr>
<td></td>
<td>80.1%</td>
<td>8.7%</td>
<td>11.1%</td>
<td>11,355</td>
</tr>
<tr>
<td>Mention</td>
<td>39.5%</td>
<td>52.2%</td>
<td>8.1%</td>
<td>7,021</td>
</tr>
</tbody>
</table>
Classification of users political leaning

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text (TF-IDF)</td>
<td>79%</td>
</tr>
<tr>
<td>Hashtags</td>
<td>91%</td>
</tr>
<tr>
<td>Retweet network</td>
<td>95%</td>
</tr>
<tr>
<td>Tags + Network</td>
<td>95%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Collective framing</th>
<th>Resource mobilization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table</strong></td>
<td><strong>Table</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Interstate</strong></th>
<th><strong>Intrastate</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>Ratio</td>
</tr>
<tr>
<td>wall</td>
<td>.590</td>
</tr>
<tr>
<td>nyc</td>
<td>.600</td>
</tr>
<tr>
<td>street</td>
<td>.699</td>
</tr>
<tr>
<td>news</td>
<td>.718</td>
</tr>
<tr>
<td>99%</td>
<td>.756</td>
</tr>
<tr>
<td>bank</td>
<td>.763</td>
</tr>
<tr>
<td>don’t</td>
<td>.782</td>
</tr>
<tr>
<td>media</td>
<td>.837</td>
</tr>
<tr>
<td>peaceful</td>
<td>.845</td>
</tr>
<tr>
<td>nypd</td>
<td>.847</td>
</tr>
</tbody>
</table>

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

**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

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 on what people says about the self or others.

Complement traditional approaches.