The Science of Social Media

Kristina Lerman
USC Information Sciences Institute

ML meetup, July 2011
What is a science?

- Explain observed phenomena
- Make verifiable predictions
- Help engineer systems with desired behavior
Why now? Data-driven network science

Availability of large-scale, time-resolved data from social media sites allows us to ask new questions about social behavior:

- How does social behavior arise from individual interactions?
- How far and how fast does information/behavior spread on networks?
- How do we measure network structure, find communities and influential people?
- How does the network structure affect information flow?
- What does the flow of information tell us about its quality?
Alternative approaches to the science of social media

• Data mining (CS) approach
  • Apply statistical regression to identify correlated sets of features in large data sets
  • Predict outcomes for new cases based on features
  • Cannot identify causal mechanisms

• Experimental approach
  • Controlled experiments to uncover causal mechanisms
  • Not practical for researchers – need access to large pool of study participants

• Empirical (physics-based) approach
  • Empirically grounded framework for discovering mechanistic models
  • Leverage statistical physics methodology to study social behavior
Roadmap

Use physics-based approach to study social media

1. Mathematical analysis of social dynamics
   • How do user actions (e.g., specified by the UI) give rise to collective behavior?
   • Mathematical models relate microscopic actions to collective dynamics
   • Explain and predict collective behavior

2. Dynamics of social contagion
   • How far does information spread on a network?
   • Empirical analysis and simulations to study how the microscopic mechanism affects macroscopic properties of spread
   • Fundamental differences between disease and information spread

3. Beyond PageRank
   • How to find important nodes in a network?
   • The role of diffusion in network analysis
Social News: Digg

Users submit or vote for (digg) news stories

Online social networks

• Users follow ‘friends’ to see
  • Stories friends submit
  • Stories friends vote for
• Shown in Friends Activity

Trending stories

• Digg promotes most popular stories to its Popular page

Data set*

• Votes on ~3K popular stories
• Follower graph w. ~280K users

Microblogging: Twitter

Users tweet short text posts
- Retweet posts of others
- Tweets may contain URLs

Online social networks
- Users follow ‘friends’ to see
  - Tweets by friends
  - Retweets by friends

Trending topics
- Twitter analyzes activity to identify popular trends

Data set
- Tweets mention ~70K URLs
- Follower graph w. ~700K users
Mathematics of social dynamics: Modeling collective behavior in social media
User interface

Collective behavior

FDA to be aggressive in tobacco regulation

counter-journal.com — If there is any doubt about how aggressive the federal Food and Drug Administration intends to be in regulating tobacco, take a look at a letter the agency sent out last week.

Can a Daily Pill Really Boost Your Brain Power?
guardian.co.uk — In America, university students are taking illegally obtained prescription drugs to make them more intelligent. Here, an investigation into the brave new world of neuro enhancement...

What happens when your Mom cancels your WoW account...

revisor.com - Parents just don’t understand. We don’t either - what WAS he trying to do with that remote control?

Three suspects arrested in U.S. terrorism probe

reuters.com — A Colorado man, his father and an accused accomplice in New York were arrested on Saturday and charged with lying to federal agents about a plot to blow up unspecified targets in the United States, the U.S. Department of Justice said.
User as a stochastic Markov process

- **browse front page**
  - **navigate**
  - **read**
  - **interesting?**

- **browse friends**
  - **read**
  - **interesting?**

- **browse upcoming**
  - **read**
  - **interesting?**
User as a stochastic Markov process

submitter

other

fans

fans

nonfans
individual behavior

collective behavior

stochastic modeling in one slide

model solutions

mathematical model
• Model solutions give dynamics of collective behavior
• Fully calibrated model
  • Interestingness \((r_S, r_F, r_N)\) the only adjustable parameters.
  • Max likelihood estimate (MLE) to find best \(r\) values that fit observed collective behavior
Popular submitter advantage

Less interesting (lower $r$) stories submitted by popular users (many followers) will be promoted to the front pages.
Predict popularity

- Estimate parameters \((r_S, r_F, r_N)\) based on early voting history
- Then solve model for later times to predict future dynamics (popularity)
Confidence of prediction

- Likelihood surface indicates how well data constrains r-values
- Compute 95% confidence bounds on model predictions
Confidence correlated with prediction error

- Predict at promotion time for ~90 stories
- Compute error 24hrs after promotion
  - Error = predicted votes - actual votes
Dynamics of social contagion: What stops social epidemics?
Information spread on network

A cascade is a sequence of activations generated by a contagion process, in which nodes cause their neighbors to be activated with some probability (transmissibility)

Underlying network

Two cascades spreading on the network*

*Nodes are labeled in the temporal order of activation
***Cf Spread of disease***

An epidemic is a contagion process that spreads to a fraction of all the nodes.

The graph shows the relationship between the fraction of nodes infected and the transmissibility, $\lambda$. The epidemic threshold, predicted for many cascade models, is indicated by the point where the fraction of nodes infected begins to significantly rise as $\lambda$ increases.
Size of cascades on social media

Digg (~280K users)

Twitter (~700K users)

Most cascades reach fewer than 1% of all users!
Why are these cascades so small?

Transmissibility of almost all Digg stories fall within width of this line?!

Standard model of epidemic growth
(Heterogeneous mean field theory, SIR model, same degree distribution as Digg)
Maybe graph structure is responsible?

→ clustering reduces epidemic threshold and cascade size, but not enough!
What about the spreading mechanism?

Independent Cascade Model implicit in many epidemic models: a node with $n$ infected friends has $n$ chances to be infected
How important are repeat exposures?

More than half exposed to a story more than once!
How do people respond to repeated exposure?

Not much.

We have similar results for Twitter.

Also noted by Romero, et al, WWW 2011.

→ repeated exposure has little effect on probability to become infected.
Big consequences for epidemic growth

1. Most people are exposed to a story more than once
2. Repeated exposures have little effect

Growth of epidemics is severely curtailed (especially compared to Ind. Cascade Model)
Weak response to repeated exposure

Take effect of repeat exposure into account:

Actual Digg cascades

Result of simulations

[Ver Steeg, Ghosh & Lerman, What stops social epidemics? in ICWSM 11]
Beyond PageRank: Dynamics and network structure
What are the important nodes in a network?

Centrality metrics examine structure of the network to determine relative importance of nodes

- **Degree Centrality**
  Measures number of node’s neighbors

- **Betweenness centrality [Freeman, 1979]**
  Fraction of all shortest paths passing through a given node

- **PageRank [Brin et al., 1998]**
  Probability random surfer ends up at a node

- **Alpha-centrality [Bonacich, 1987], Katz score [Katz, 1953]**
  Number of paths of any length, exponentially attenuated by their length, from a given node to all nodes

- ...
The Billion Dollar Algorithm: PageRank \textcolor{red}{(Brin\ et\ al\ 98)}

\[
\alpha^t_{PR}(i) = (1 - \alpha) \frac{1}{n} + \alpha \sum_{j \in \text{follows}(i)} \frac{r^t_{PR}(i, j)}{\alpha_{out}}
\]

After many iterations, converges to PageRank scores of nodes.
PageRank and the Random Surfer

Consider a web surfer who clicks on links on web pages at random with no regard to content

• With probability $\alpha$, the Surfer follows a hyperlink from a given web page
• With probability $(1-\alpha)$, the Surfer jumps to a random web page

After a long time ... the probability the Random Surfer visits a web page is given by that page's PageRank.
Random walk and conservative diffusion

• Random surfer executes a random walk (with restarts) on a web graph – a stochastic process obeying the following rules
  • With probability $\alpha$, walker jumps from a node to one of its neighbors
  • With probability $(1-\alpha)$, walker jumps to any node

• Random walk and diffusion
  • Random walk is a prototypical example of conservative diffusion process
  • Other examples are money transfer, used goods circulation, etc. where some quantity ($$, goods, probability) is conserved
Matrix formulation

Adjacency matrix of the network

\[ A = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 \\
\end{pmatrix} \]

Outdegree matrix

\[ D = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{pmatrix} \]
Matrix formulation of conservative diffusion

Weight vector gives the probability of finding the walker anywhere on the graph at time t. It is changed by diffusion:

$$w^C_t = (1 - \alpha)s + \alpha D^{-1} A w^C_{t-1}$$

where starting vector s specifies where a random jump ends up

Steady state solution: as $t \to \infty$

$$w^C_\infty = (1 - \alpha)s + (1 - \alpha)\alpha D^{-1} As + ... = \frac{(1 - \alpha)s}{(I - \alpha D^{-1} A)}$$

... same as PageRank when $s = 1/n$
Social phenomena are non-conservative

Random walk is not a good model of social phenomena, such as epidemics, adoption of innovation, or information spread in social media – broadcast-based diffusion

Broadcast-based diffusion is non-conservative; i.e., amount of the diffusing quantity (disease, information) changes over time.
Matrix formulation of non-conservative diffusion

Weight gives the amount of quantity on the graph at time $t$. Weight evolves according to rules of non-conservative diffusion

$$w_{t}^{NC} = s + \alpha A w_{t-1}^{NC} \quad \text{with starting vector } s$$

Steady state solution: as $t \to \infty$

$$w_{\infty}^{NC} = \sum_{k=0}^{\infty} \alpha^k A^k s = s(I - \alpha A)^{-1} \quad \text{Holds for } \alpha < 1/\lambda_1$$

Weight vector is not conserved: $\|w_{t}^{NC}\| \neq \|w_{t-1}^{NC}\|$
Alpha-Centrality [Bonacich, 87]

Number of paths of any length, attenuated by their length with $\alpha$

$$r_t^{\text{Alpha}} = eA + \alpha eA^2 + \ldots + \alpha^t eA^{t+1}$$
Alpha-Centrality [Bonacich, 87]

Number of paths of any length, attenuated by their length with $\alpha$

$$r_t^{\text{Alpha}} = eA + \alpha eA^2 + \ldots + \alpha^t eA^{t+1}$$
Alpha-Centrality [Bonacich, 87]

Number of paths of any length, attenuated by their length with $\alpha$

$t \rightarrow \infty$

$$r_{\infty}^{\text{Alpha}} = \frac{eA}{(I - \alpha A)}$$

Holds for $\alpha < 1/\lambda_1$
**Alpha-Centrality [Bonacich, 87]**

Number of paths of any length, attenuated by their length with $\alpha$

\[ t \to \infty \]

\[ r^\text{Alpha}_\infty = \frac{eA}{(I - \alpha A)} \]

Holds for $\alpha < 1/\lambda_1$

\[ w^{NC}_\infty = \frac{s}{(I - \alpha A)} \]
Two classes of dynamic processes on networks

Conservative diffusion

Non-conservative diffusion

- Mathematical formulation of two types of diffusion
  - Equivalence of steady state solutions and centrality metrics
  - Unifies social network analysis and epidemic models
- Implications
  - Location of the epidemic threshold
  - Social Network Analysis should consider not only the network topology, but type of dynamics also
    - I.e., which centrality metric?

[Ghosh and Lerman, Predicting Influential Users in Online Social Networks. SNAKDD10]
Implications: epidemic threshold

Non-conservative diffusion has a threshold, given by \( 1/\lambda_1 \), (largest eigenvalue of A) [Wang et al., 2003]

- For \( \lambda < 1/\lambda_1 \), process (epidemic) dies out
- For \( \lambda > 1/\lambda_1 \), process (epidemic) reaches many nodes

\[ c = 0.006 \]
\[ c = 0.009 \]
Implications: Social Network Analysis

Which centrality metric best predicts important nodes?

PageRank

Alpha-Centrality
Implications: Social Network Analysis

Which centrality metric best predicts important nodes?

PageRank

Alpha-Centrality
Which metric is right?

How can we evaluate centrality metrics?

• User activity in social media provides an independent measure of importance/influence
  • Serves as ground truth for evaluating centrality metrics

Evaluation methodology

• Define an empirical measure of influence (ground truth)
• Compare centrality metrics with the ground truth
Information flow on Digg

Information spread on Digg is non-conservative
→ Non-conservative metric will best predict influential users
Empirical estimate of influence

1. Average follower votes
   - Likelihood a follower votes for the story
     - Influence of submitter
     - Quality of the story
   - Story quality
     - Random variable
     - Average out by aggregating fan votes over all stories submitted by the same submitter
       - 289 users submitting at least 2 stories

2. Average Cascade size
   - How far does the story spread into the network

Ground truth(s): Rank users according to each estimate
Evaluation of importance prediction

Correlation between the rankings produced by the empirical measures of influence and those predicted by Alpha-Centrality and PageRank

(1) Avg. # of follower votes
(2) Avg. cascade size

→ Non-conservative Alpha-Centrality best predicts influence rankings
Conclusion

Physics-inspired analysis of large-scale, time-resolved data about social behavior online leads to new understanding of social dynamics and social networks

- Mathematical modeling links details of individual actions with collective social behavior
  - Explain phenomenology of Digg
  - Predict future popularity of new content
  - Design better social systems

- Dynamics of information spread on networks
  - What limits the size of information cascades?
  - Fundamental differences between social and disease contagion

- Understanding network structure
  - Topology alone does not explain structure, e.g., important nodes. Also need to consider the type of diffusion process
  - Centrality metric based on epidemic models better predicts influential users of Digg.
Readings

• **Mathematical modeling**

• **Dynamics of information spread**

• **Diffusion and network structure**
Thanks

Collaborators

- Rumi Ghosh (USC)
- Greg Ver Steeg (USC)
- Tad Hogg (IMM)
- Rob Zinkov (USC)
- Tawan Surachawala (USC)

Funding agencies

- NSF
- AFOSR
- AFRL