Searching for Influencers via Optimal Percolation: from Twitter to the Brain

MAIN QUESTION: Where are the superspreaders of information?

Hernán Makse, Physics Department, City College of New York

NetSci2016

Friday, June 3, 16
MAIN QUESTION: Where are the superspreaders of information?

Hernán Makse, Physics Department, City College of New York

Friday, June 3, 16
Searching for Influencers via Optimal Percolation: from Twitter to the Brain

MAIN QUESTION: Where are the superspreaders of information?

Answer: Optimal Percolation and the Non-Backtracking Matrix uncover the Collective Influencers.

Hernán Makse, Physics Department, City College of New York

Friday, June 3, 16
Longstanding Problem in Network Science

Find the minimum number of influencers/superspreaders to “infect” the entire network

- Viral marketing: $$$
- Predicting trends from social media
- Immunization epidemics
- Essential genes:
  - gene regulation
  - protein networks
- Essential species:
  - ecological networks
- Brain Networks:
  - neural correlates of consciousness
- Financial networks
  - “too big to fail” --> “too weak to fail”? 
- Predicting stock markets
How to become viral in social media?
EASY

1. Funny baby faces videos

50 million hits

Friday, June 3, 16
How to spread information in social media? EASY

2. Funny dog faces videos

50 million hits

lindos perritos!!
3. Most successful viral superspreading event in the history of humankind

Gangnam style video by Mr Psy

as of today: 2 billion hits (mainly teenagers)
3. Most successful viral superspreading event in the history of humankind

Gangnam style video by Mr Psy as of today: 2 billion hits (mainly teenagers)

-Can we predict the onset of collective behavior?
-More is different, Phil Anderson, Nobel ’77
1. COMPLEX NETWORK SCIENCE: superspreading of ideas

- Analogous to herding behavior in the stock market

Courtesy of Gene Stanley
1. COMPLEX NETWORK SCIENCE: superspreading of ideas
- Analogous to herding behavior in the stock market

Courtesy of Gene Stanley
1. COMPLEX NETWORK SCIENCE: superspreading of ideas

- Analogous to herding behavior in the stock market

“ Irrational exuberance”
- Chairman Greenspan (FED) on the dot-com bubble
- Keynes: “animal spirits” of excessive optimism
THE ANSWER:
Understanding BIG DATA
What is the Problem with Big Data?
What is the Problem with Big Data?

Eric Schmidt (CEO Google): “Every two years we create as much information (5 exabytes) as we did from the dawn of civilization until 2003”
THE QUESTION ARISES:
What happened in the last two years?
THE QUESTION ARISES:
What happened in the last two years?

Did we all get clever all of a sudden?
THE QUESTION ARISES:
What happened in the last two years?

Did we all get clever all of a sudden?
Most probably not
THE QUESTION ARISES:
What happened in the last two years?

Did we all get clever all of a sudden?
Most probably not

Evidence #1: The USA Presidential Primary Election Debates
THE QUESTION ARISES:
What happened in the last two years?

Did we all get clever all of a sudden?
Most probably not

Evidence #1: The USA Presidential Primary Election Debates

Evidence #2: Peter Thiel (founder PayPal)

“We wanted flying cars, instead we got 140 characters.”
—— Peter Thiel ——

“The rate of technological innovation is actually slowing”
WHY IS THAT?
WHY IS THAT?

THE AGE OF BIG DATA

OR
WHY IS THAT?

THE AGE OF BIG DATA

OR

THE AGE OF BIG DATA JUNK?
THE SOLUTION: REDUCE BIG DATA TO A FEW INFLUENCERS

MINIMAL INFLUENCERS

NP-hard Maximization of Influence Problem
Kempe, Kleinberg, Tardos (2003)
How to predict the most influentials in social networks

Typical approach: Brute-Force (heuristics)

“Attacking” the hubs (high degree)

Pres Obama: 55M followers
Lady Gaga: 45M followers
(very few: 1 percenters....)

Other heuristics: ranking the nodes by PageRank (Google), betweenness, eigenvector, closeness centralities, kcore, equal graph partitioning, etc...

Problems: heuristics do not maximize any function of influence
Our approach

Inspired by French mime Marcel Marceau:

“Making visible the invisible”

Collective optimization theory unravels the strength of “weak nodes”

Granovetter, 1973:

“The strength of weak ties”
Mapping to Optimal Percolation to find the minimal set of "influencers" to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with k-1 threshold

SIR model

$G_\infty$ $G_\infty$
Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

**Best spreaders** = minimize the inactive nodes = maximize giant component

**Optimal Percolation** = **Best immunizators** = best influencers

Linear-threshold model with k-1 threshold

SIR model

**Best spreaders** = minimize the inactive nodes = maximize giant component

**Best immunizators** = minimize the giant connected component
Optimal Percolation =
Best immunizators = best influencers

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with k-1 threshold

SIR model

\( G_\infty \)

\( G_\infty \)
Optimal Percolation =

Best immunizators = best influencers

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with \( k-1 \) threshold

\( G_\infty \)

SIR model

\( G_\infty \)
Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Optimal Percolation = Best immunizators = best influencers

Linear-threshold model with k-1 threshold

SIR model

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Optimal Percolation = Best immunizators = best influencers

Linear-threshold model with $k-1$ threshold

SIR model
Optimal Percolation =
Best immunizators = best influencers

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with k-1 threshold

SIR model

Friday, June 3, 16
Optimal Percolation =
Best immunizators = best influencers

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with k-1 threshold

SIR model

\[ G_\infty \]

\[ G_\infty \]
Optimal Percolation =
Best immunizators = best influencers

Mapping to Optimal Percolation to find the minimal set of “influencers” to fragment the network

Best spreaders = minimize the inactive nodes = maximize giant component

Best immunizators = minimize the giant connected component

Linear-threshold model with k-1 threshold

SIR model

**Optimal Percolation** = minimize $q_c$

$q_c$ : minimal fraction of influencers to destroy the network

$G_\infty$ : stability problem

Transform the problem into a stability problem

Random percolation $q_{\text{rand}} = \frac{\kappa - 2}{\kappa - 1}$

Strategy = minimize the giant component

or

Minimize $q_c = $ find the minimal influencers until the solution

$G_\infty = 0 \rightarrow \{\nu_{i\rightarrow j} = 0\}$ becomes unstable
How to calculate the stability matrix

Order parameter = Prob to belong to giant component: \( G_{\infty} \)

Message passing equation in a locally tree-like network

\[
\nu_{i \to j} = n_i \left[ 1 - \prod_{k \in \partial i \setminus j} (1 - \nu_{k \to i}) \right]
\]

\( n_i = 0 \) removed node
\( n_i = 1 \) other nodes
Stability of \( G_\infty = 0 \) is given by largest eigenvalue of non-backtracking matrix

\[
\frac{\partial \nu_{i \rightarrow j}}{\partial \nu_{k \rightarrow \ell}} \bigg|_{\nu_{i \rightarrow j} = 0} \equiv n_k B_{k \rightarrow \ell, i \rightarrow j}
\]

Solution is stable when the largest eigenvalue of the modified NB matrix:

\[
\lambda_{\text{max}}(n, q) \leq 1
\]

Finding the superspreaders = finding the nodes that minimizes the largest eigenvalue of the modified NB matrix:

\[
\min_n \lambda_{\text{max}}(n, q_c) = 1
\]
The superspreaders are the best non-backtracking walkers

Non-backtracking walk of length $\ell$: a walker that cannot immediately come back

- Second largest eigenvalue of NB is also optimal for community detection. Krzakala PNAS (2013), Newman PRE (2013)
- In contrast, most heuristics are based on the adjacency matrix which measures only regular random walks: PageRank largest eigenvalue $A_{ij}$
Many-body problem: Power method

“Feynman-like” diagrams for $\lambda_{\text{max}}(n^*, q_c)$

Non-backtracking walk

\(\ell\)-level = steps

Diagramatic expansion contains all physical interactions. Exact solution for random graphs (not for real graphs)
For a given level $\ell$

1. Start with all nodes occupied: $n_i=1$

2. Assigned to every node the collective influence index:

$$CI_\ell = (k_i - 1) \sum_{j \in \partial Ball(\ell)} (k_j - 1)$$

3. Remove (attack) the highest CI node: $n_i^*=0$

4. Recalculate CI and repeat until giant component is zero.

Cl scales as:

$$O(N) \sim N \log N$$
Test: Exact optimal solution and best approximation with CI in Erdos-Renyi network

CI outperforms heuristic centralities and approximates well the exact optimal solution

The best possible attack is to remove the loops to get a tree at $q_c$

Others: hub removal (HD, high degree and adaptive HDA), PageRank, Closeness Centrality, Eigenvector Centrality, k-core, EGP
Why CI outperforms other heuristics?

Collective Influence identifies a new class of influencer neglected by previous rankings.

Weak node: a low degree node surrounded by hierarchical coronas of hubs at level $\ell$.

Related to Granovetter theory.

“Strength of weak ties” (1973).

\[ CI_{\ell} = (k_i - 1) \sum_{j \in \partial Ball(\ell)} (k_j - 1) \]

Weak node
Validation of CI in Big Data: Twitter

CI is valid in locally tree-like networks: what about real networks?

- CI identifies 40% less influencers than high degree

A broker node
Next, we address the question that will doubtless arise:
Next, we address the question that will doubtless arise:

Is all this mathematical gibberish of any real use?

Three applications:
1. Marketing campaign in Mexico
2. Predicting trends in Twitter: Trump vs Cruz
3. Finding superspreaders in the brain
1. Validating CI in a Real Marketing Campaign using Mobile Phone Networks

Market campaign targeting the high CI people, AKA influencers

Mexico:
110 million users calling over 3 months + banking data

Team up with GranData.com (an Argentinian Corporation)
Communication patterns of rich and poor

Top 1% richest people

Bottom 20% poorest people

Top 1%

HIGH CI

Bottom 20%

(same degree)

LOW CI

Bottom 20% poorest people

Friday, June 3, 16
Improve response rate = $$$ by five-fold between low CI and high CI targeting

Targeting 60,000 people according to their Collective Influence in the network of entire Mexico

Low CI: poorest people

High CI: wealthy people
2. Predicting Trump Sentiment from Twitter

Friday, June 3, 16
2. Predicting Trump Sentiment from Twitter

ALL TWEETS

INFLUENCERS

Trump wins

Cruz wins

Trump camp

Cruz camp

Friday, June 3, 16
Awake brain surgery to dissect tumor with no clear boundaries

Neurosurgeon stimulates areas around the tumor with electrodes to locate the essential functional areas.

Functional areas (e.g., language, motor) are located by asking the patient to talk, move, etc. Remove as much tumor as possible avoiding the essential areas.

GOAL: predict the essential areas of the brain with NoN theory

Collaboration with Andrei Holodny, MSKCC

Friday, June 3, 16
3. Finally: Influencers in a Brain Network of Networks

Important for any biological system: genetic networks, proteome, metabolome, etc

Finding the essential minimal nodes in the brain
- neural correlates of consciousness -

We extract the Network from fMRI/DTI experiments on dual tasking: auditory and visual

Gallos, Makse, Sigman, PNAS (2012)
Reis, Andrade, Sigman, Canals, Makse, Nat Phys (2014)

Friday, June 3, 16
Influencers in a Network of Networks

Influencers are the best non-backtracking walkers walking along two types of links. Information flows via two types of messages: intra-modular and inter-modular.

\[ \mathcal{M}_{k \to \ell, i \to j}(n_i) = \frac{\partial \rho_{i \to j}}{\partial \rho_{k \to \ell}} \bigg|_{\rho=0} \]

\[ \hat{\mathbf{M}} \equiv \left( \begin{array}{cc} \hat{\mathbf{M}}_{\rho \rho} & \hat{\mathbf{M}}_{\rho \varphi} \\ \hat{\mathbf{M}}_{\varphi \rho} & \hat{\mathbf{M}}_{\varphi \varphi} \end{array} \right) \bigg|_{G=0} \equiv \left( \begin{array}{cccc} \frac{\partial \rho_{k A \to l A}}{\partial \rho_{i A \to j A}} & \frac{\partial \rho_{k B \to l B}}{\partial \rho_{i A \to j A}} & \frac{\partial \varphi_{k A \to l B}}{\partial \rho_{i A \to j A}} & \frac{\partial \varphi_{k B \to l A}}{\partial \rho_{i A \to j A}} \\ \frac{\partial \rho_{k A \to l A}}{\partial \varphi_{i A \to j A}} & \frac{\partial \rho_{k B \to l B}}{\partial \varphi_{i A \to j A}} & \frac{\partial \varphi_{k A \to l B}}{\partial \varphi_{i A \to j A}} & \frac{\partial \varphi_{k B \to l A}}{\partial \varphi_{i A \to j A}} \\ \frac{\partial \rho_{k A \to l B}}{\partial \rho_{i B \to j B}} & \frac{\partial \rho_{k B \to l B}}{\partial \rho_{i B \to j B}} & \frac{\partial \varphi_{k A \to l B}}{\partial \rho_{i B \to j B}} & \frac{\partial \varphi_{k B \to l A}}{\partial \rho_{i B \to j B}} \\ \frac{\partial \rho_{k A \to l B}}{\partial \varphi_{i B \to j B}} & \frac{\partial \rho_{k B \to l B}}{\partial \varphi_{i B \to j B}} & \frac{\partial \varphi_{k A \to l B}}{\partial \varphi_{i B \to j B}} & \frac{\partial \varphi_{k B \to l A}}{\partial \varphi_{i B \to j B}} \end{array} \right) \bigg|_{G=0} \]
Collective Influence Map of the Human Brain: reducing the brain to its essential nodes

Inter-modular links
Intra-modular links
Complexity Reduction

High CI

G(q)
q (Frac. of zero inputs)
Giant active component

CI - Rare inputs
Random - Typical inputs
Data Analytics at the Cutting-Edge for Free!
kcore-analytics.com

Kcore DREAM TEAM:

Search Engine for Influencers

Flaviano Morone
George Furbish
Andrea Morone
Alex Bovet
Kevin Roth
Francesca Lucini
Hernan Makse

Friday, June 3, 16