Dynamics of information spread on networks

Kristina Lerman
USC Information Sciences Institute
Information spread in online social networks

Diffusion of activation on a graph, where each *infected* (activated) node infects neighbors with some probability
Why now? Data-driven network science

Availability of large-scale, time-resolved data on the Social Web allows us to ask new questions about social behavior

- How does information spread on networks?
  - How far and how fast does information flow?
  - How does the network structure affect information flow?
- What does this tell us about the quality of information?
- What is the structure of the network?
  - Who are the influential users and communities?
  - How does network structure affect the flow of information?
- What is the collective behavior of users?
  - How does individual behavior affect collective behavior?
Outline

Online social networks (OSN)

Empirical study of information cascades on networks
  • Statistical properties of OSN
  • Quantitative analysis of information cascades on networks
  • What limits the size of information cascades in OSN?

Network structure and dynamics
  • Classes of diffusion processes on networks
  • Diffusion and centrality equivalence
    • What is the appropriate centrality metric for a given network?
    • Compare performance of centrality metrics on OSN
  • Alpha-Centrality
    • Parametrized centrality for network analysis
    • Node Ranking and Community Detection
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 My News

Trending stories

• Digg promotes most popular stories to its Top News page
Microblogging: Twitter

Users tweet short text posts
- Retweet posts of others
- 140 characters long

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

Trending topics
- Twitter analyzes activity to identify popular trends
Comparative empirical analysis

Digg
- Voting on a single story
- one month
  ~3K stories
  ~280K users
  ~1M links

Twitter
- Use URLs as markers for tracking the flow of information
- three weeks
  ~70K stories
  ~700K users
  ~36M links

[Lerman and Ghosh, ICWSM 10]
Dynamics of story popularity

1: U.S. Government Asks Twitter to Stay Up for #IranElection
2: Western Corporations Helped Censor Iranian Internet
3: Iranian clerics defy ayatollah, join protests

1: U.S. Government Asks Twitter to Stay Up for #IranElection
2: Western Corporations Helped Censor Iranian Internet
3: Iranian clerics defy ayatollah, join protests
Distribution of popularity of stories

Digg (promoted stories)

- Aggregate over all stories to factor out influence of submitter and story quality
- “Inequality of popularity” – some stories much more popular than others
  *cf* social influence study of [Salganik, Dodds & Watts, 2006]

Twitter (all stories)
Network structure: Distribution of followers

Digg

Twitter

follower friend

number of fans per user

number of users

number of followers per user

number of users
Cascades on networks

Information spreads through cascades on networks

Underlying network

Two cascades spreading on the network*

*Nodes are labeled in the temporal order of activation
Analysis of information cascades

Figure 9: Common cascade shapes ordered by the frequency. Cascade with label $G_r$ has the frequency rank $r$.

(a) All cascades  (b) Star cascade  (c) Chain cascade

[Leskovec, McGlohon, & Faloutsos, Cascading Behavior in Large Blog Graphs, in SDM (2007)]
Cascade generating function
Quantitative framework for measuring the structure of evolving cascades
• Microscopic
• Macroscopic

Efficient compression
• Time sequence of real numbers carries information about cascade structure
• Allows to reconstruct the cascade

Fast and Scalable
• $O(kN)$ space
• $O(dkN)$ runtime complexity ($k=$#of seeds, $d=$max. degree, $N=$#of nodes)

Calculating cascade generating function

\[ \varphi(j, \alpha) = \sum_{i \in \text{friend}(j)} \alpha \varphi(i, \alpha) \]

\[ \varphi(6, \alpha) = \varphi(3, \alpha) + \alpha \varphi(1, \alpha) = (\alpha + \alpha^2)c_1 \]
Some common cascades

branching

chaining

community
‘Collision of cascades’
Cascade reconstruction: degeneracy
Digg case study: microscopic properties

Evolution of three largest cascades for two non-popular stories

Play Doctor On Yourself: 16 Things To Do Between Checkups

APOD: 2009 July 1 – Three Galaxies in Draco
Digg case study: microscopic properties

Evolution of three largest cascades for two popular stories

Infomercial King’ Billy
Mays Dead at 50

Bender’s back
Digg: macroscopic properties

No. of cascades

Diameter

Cascade size

Spread

Pr(X = x)

Pr(X = x)

Pr(X = x)

Pr(X = x)
Why are Digg cascades so small?
What limits cascade size?

Network structure

• Clustering
• Degree heterogeneity

Dynamics

• Social contagion mechanism
• Change in transmissibility (e.g., novelty decay)

[Ver Steeg, Ghosh & Lerman, What stops social epidemics? in ICWSM 11]
Effect of network structure

Study the effect of clustering by simulating cascades on synthetic and real graphs

Digg graph

- Power law degree distribution with exponent -2

Synthetic graph

- Constructed using directed configuration Model [Newman et al.]
- Preserves degree distribution but destroys clustering and degree correlation

Cascade simulations

- Independent Cascade Model (ICM)
  - Widely used to model epidemics, viral marketing campaigns, etc.
  - Start with *infected* seed (submitter)
  - *Susceptible* fans are infected with probability $\lambda$ (transmissibility)
    - If user has n contagious friends, she has n chances to be infected
  - Users can infect friends during a single round, then they are removed
  - When cascade stops, measure its size (# of infected nodes)
Cascade size vs transmissibility

→ clustering reduces epidemic threshold and cascade size, but not enough!
Friend Saturation Model

Perhaps we got the contagion mechanism wrong?

- ICM: each friend has probability $\lambda$ to infect the node; therefore,
  
  $p_{ICM}(\text{vote} | \text{n friends voted}) = 1 - (1 - \lambda)^n$

- On Digg, empirically
  
  $p(\text{vote} | \text{n friends voted}) \sim \lambda$

Friend Saturation Model (FSM)

- Repeated exposure to a story does not make the user much more likely to vote for it
  
  - Cf, market saturation effect, decreasing cascade model [Kempe et al, 2005]
Simulations of FSM Model on Digg graph

Actual and simulated (FSM) cascades on the Digg graph

→ FSM mechanism drastically reduces cascade size
Outline

Online social networks (OSN)
Empirical study of information cascades on networks
  • Online social networks (OSN) and their properties
  • Quantitative analysis of information cascades on networks
  • What limits the size of information cascades in OSN?

→ Network structure and dynamics
  • Classes of diffusion processes on networks
  • Diffusion and centrality equivalence
    • What is the appropriate centrality metric for a given network?
    • Compare performance of centrality metrics on OSN
  • Alpha-Centrality
    • Parametrized centrality for network analysis
    • Node Ranking and Community Detection
Two classes of diffusion on networks

**Conservative diffusion**: models money transfer, conversations, etc. where some quantity (\$, attention) is conserved

**Non-conservative diffusion**: models epidemics, spread of information, etc., where quantity (viruses) is not conserved.

Mathematical formulation of two types of diffusion

- Equivalence of steady state solutions and centrality metrics
- Unifies social network analysis and epidemic models
  - Existence of threshold for epidemic processes
  - Location of the epidemic threshold

[Ghosh and Lerman, Predicting Influential Users in Online Social Networks. SNAKDD10]
Conservative diffusion

\(\alpha \Delta_2/2 + \alpha \Delta_3/2\)

\(w_1 - \alpha \Delta_1\)

\(\alpha \Delta_1\)

\(\alpha \Delta_2/2\)

\(w_2 - \alpha \Delta_2\)

\(\alpha \Delta_2/2\)

\(\alpha \Delta_3/2\)

\(w_3 - \alpha \Delta_3\)

\(\alpha \Delta_3/2 + \alpha \Delta_4\)

\(w_4 - \alpha \Delta_4\)

\(\alpha \Delta_4\)

\(w_5 - \alpha \Delta_5\)

\(\alpha \Delta_5\)

\(\alpha \Delta_2/2\)

\(\alpha \Delta_1 + \alpha \Delta_5\)
Conservative diffusion

\[ w_4 = \alpha \Delta_4 + \alpha \Delta_2 / 2 + \alpha \Delta_3 / 2 \]

\[ w_1 = \alpha \Delta_1 \]

\[ w_2 = \alpha \Delta_2 + \alpha \Delta_1 + \alpha \Delta_5 \]

\[ w_3 = \alpha \Delta_3 + \alpha \Delta_2 / 2 \]

\[ w_5 = \alpha \Delta_5 + \alpha \Delta_4 + \alpha \Delta_3 / 2 \]

\[ \alpha \Delta_2 / 2 + \alpha \Delta_3 / 2 \]

\[ \alpha \Delta_4 + \alpha \Delta_3 / 2 \]
Matrix formulation

Adjacency matrix of the network

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

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

\[ w_t^C = (1 - \alpha)s + \alpha T w_{t-1}^C \]

with starting vector \( s \) and transfer matrix \( T = D^{-1}A \)

Steady state solution: as \( t \to \infty \)

\[ w_\infty^C = (1 - \alpha)s + (1 - \alpha)\alpha Ts + ... = \frac{(1 - \alpha)s}{(I - \alpha T)} \]

Weight vector is conserved \( \rightarrow \) conservative diffusion

\[ \| w_t^C \| = \| w_{t-1}^C \| \]
Non-conservative diffusion

\[
\begin{align*}
\alpha \Delta_2 + \alpha \Delta_3 & \\
w_4 - \alpha \Delta_4 & \\
w_1 - \alpha \Delta_1 & \\
\alpha \Delta_2 & \\
\alpha \Delta_1 & \\
\alpha \Delta_1 + \alpha \Delta_5 & \\
w_2 - \alpha \Delta_2 & \\
\alpha \Delta_2 & \\
\alpha \Delta_5 & \\
w_5 - \alpha \Delta_5 & \\
\alpha \Delta_3 + \alpha \Delta_4 & \\
w_3 - \alpha \Delta_3 & \\
\alpha \Delta_2 & 
\end{align*}
\]
Non-conservative diffusion

\[ w_1 - \alpha \Delta_1 \]

\[ \alpha \Delta_2 + \alpha \Delta_3 \]

\[ w_2 - \alpha \Delta_2 + \alpha \Delta_1 + \alpha \Delta_5 \]

\[ w_3 - \alpha \Delta_3 + \alpha \Delta_2 \]

\[ w_4 - \alpha \Delta_4 + \alpha \Delta_2 + \alpha \Delta_3 \]

\[ w_5 - \alpha \Delta_5 + \alpha \Delta_3 + \alpha \Delta_4 \]
Matrix formulation of non-conservative diffusion

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

Steady state solution: as \( t \to \infty \)

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

Weight vector is not conserved \( \Rightarrow \) non-conservative diffusion

\[ \| w_{NC}^t \| \neq \| w_{NC}^{t-1} \| \]
Epidemics as non-conservative diffusion

$time = t-1$
Epidemics as non-conservative diffusion

\[ t = t \]

\[ P_t = \lambda A P_{t-1} = (\lambda A)^t P_0 \]

Transmissibility \( \lambda \)

[Wang et al. 2003]
Epidemics as non-conservative diffusion

Expected # of virus up to time t

\[ P_{t}^{cum} = \sum_{k=0}^{t} P_{k} = \sum_{k=0}^{t} (\lambda A)^{k} P_{0} \]
Epidemics as non-conservative diffusion

\[ P_{t}^{\text{cum}} = \sum_{k=0}^{t} P_k = \sum_{k=0}^{t} (\lambda A)^k P_0 \]

Expected # of virus up to time \( t \)

\[ w = \sum_{k=0}^{\infty} (\alpha A)^k s \]
Implications: Epidemic threshold

In spreading epidemics there exists a threshold below which epidemics die out and above which they reach significant fraction of nodes

- Epidemic threshold is given by $\frac{1}{\lambda_1}$, inverse of the largest eigenvalue of $A$ [Wang et al., 2003]
  - For $\lambda > \frac{1}{\lambda_1}$ epidemics reach many nodes
  - Agrees with empirically observed threshold on Digg
- Threshold also exists in non-conservative diffusion
Implications for network analysis

Who is important in a network?

Geodesic path-based ranking measures
  • Betweenness centrality [Freeman, 1979]

Topological ranking measures
  • Page Rank [Brin et al., 1998]
  • Degree Centrality

Path-Based Ranking Measures
  • Alpha-centrality [Bonacich, 1987]
  • Katz score [Katz, 1953]
  • EigenVector centrality [Bonacich, 2001]

[Node size = importance]
Diffusion and centrality

Current approaches take into account topology only and may lead to conflicting answers.

Diffusion and centrality

- Random walk \(\rightarrow\) Conservative diffusion \(\rightarrow\) PageRank
- Information spread \(\rightarrow\) Non-conservative diffusion \(\rightarrow\) Alpha-Centrality

What is the appropriate metric for the given network?
**PageRank [Page & Brin, 98]**

Web surfer: with probability $\alpha$ navigates to a neighboring node

.... with probability $(1 - \alpha)$ jumps to a random node

\[
\begin{align*}
    r_t^{PR} &= (1 - \alpha)r_0^{PR} + (1 - \alpha)\alpha D^{-1} A r_0^{PR} + K + \left(\alpha D^{-1} A \right)^t r_0^{PR}
\end{align*}
\]
PageRank [Page & Brin, 98]

Web surfer: with probability $\alpha$ navigates to a neighboring node

.... with probability $(1 - \alpha)$ jumps to a random node

$$\lim_{t \to \infty} r_{PR}^t = (1 - \alpha) r_{0}^{PR} + K + \left(\alpha D^{-1} A\right)^t r_{0}^{PR} + ... = \frac{(1 - \alpha) r_{0}^{PR}}{(I - \alpha D^{-1} A)}$$
PageRank [Page & Brin, 98]

Web surfer: with probability $\alpha$ navigates to a neighboring node
.... with probability $(1 - \alpha)$ jumps to a random node

$$r_{\infty}^{PR} = \frac{(1 - \alpha)r_0^{PR}}{(I - \alpha D^{-1} A)}$$

$$w_{\infty}^C = \frac{(1 - \alpha)s}{(I - \alpha T)}$$
**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 $\alpha$

$$t \rightarrow \alpha$$

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

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

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

$t \rightarrow \infty$

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

Holds for $\alpha < 1/\lambda_1$

\[
\mathbf{w}^{NC}_\infty = \frac{s}{(I - \alpha A)}
\]

non-conservative diffusion
Ranking nodes by centrality

PageRank

Alpha-Centrality
Ranking nodes by centrality

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
**Statistical significance estimate 1**

**URN MODEL**

\[
P(X = k \mid K, N, n) = \frac{\binom{K}{k} \binom{N - k}{n - k}}{\binom{N}{n}}
\]

(Hypergeometric Dist.)
Statistical significance results

Avg # follower votes received by stories within the first 100 votes vs # of submitter’s followers.

Probability of the expected number of fan votes being generated purely by chance.
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
Time check?
**Alpha-Centrality** [Bonacich, 87]

\[ r^{\text{Alpha}}(\alpha) = eA \sum_{k=0}^{\infty} \alpha^k A^k = \frac{eA}{(I - \alpha A)} \]

- Measures the number of paths between nodes, each path attenuated by its length with parameter \( \alpha \)
  - \( \alpha \) sets the length scale of interactions
    - \( \alpha = 0 \) local interactions only \( \rightarrow \) degree centrality
    - As \( \alpha \) grows, increasingly longer distances become important
- But, condition must hold for convergence
  - \( \alpha < 1/|\lambda_1| \)
  where \( \lambda_1 \) is largest eigenvalue of \( A \)
Normalized Alpha-Centrality

\[ \overline{r}^{\text{Alpha}} (\alpha) = \frac{eA \sum_{k=0}^{\infty} \alpha^k A^k}{\sum_{i,j}^{n} \sum_{k=0}^{\infty} \alpha^k A^k} \]

- No longer bounded by eigenvalue: holds for \( 0 \leq \alpha \leq 1 \)
- Parameterized centrality metric
  \( \alpha \) sets the length scale of interactions
    - Local: For \( \alpha = 0 \) leads to the same rankings as degree centrality
    - Meso: As \( \alpha \) grows, increasingly longer distances become important
    - Global: As \( \alpha \to 1/\lambda_1 \), leads to the same rankings as eigenvector centrality
- Relation to Alpha-Centrality
  - For \( 0 \leq \alpha < 1/\lambda_1 \) \( \to \) leads to the same rankings as Alpha-Centrality
  - For \( \alpha > 1/\lambda_1 \), rankings independent of \( \alpha \)

[Ghosh and Lerman, Parameterized Metric for Network Analysis, Physical Review 2011]
Applications to network analysis

• Ranking
  
  Change in ranking with a

  • Leaders: individuals with high importance score
  • Bridges: individuals whose importance score grows with a
    Mediate communication between communities
  • Peripherals: low importance score

• Community Detection

  Modularity maximization [Newman & Girvan, 04]

  • Modularity Q=(actual connectivity)-(expected connectivity)
  • Connectivity measured using normalized Alpha-Centrality
Advantages

Multi-scale analysis of networks

• Parameter sets the length scale of interaction

Connect the rankings produced by well-known local and global centrality metrics

• Degree Centrality
• Eigenvector Centrality

Can differentiate between locally and globally connected nodes and structures

• Leaders and bridges
• Local and global communities
Zachary’s karate club [Zachary, 77]

Centrality scores of nodes vs. $\alpha$
Communities in Zachary’s karate club network

$\alpha = 0$

$0 < \alpha < 0.14$

$\alpha \geq 0.14$

Newman’s modularity

$(1/\lambda_1 = 0.1487)$
Related Work

Empirical Investigation of Online Social Networks

• Network inferred from the observed links [Wu,04; Gruhl,04; Leskovec,07; Rodriguez,10]
  – Forwarding chains long narrow rather than bushy and wide trees
  – Networks are extracted independent of the spread of data.
• Diffusion terminates in few steps [Wu,04; Leskovec,08;] versus Influence spreads easily [Bakshy,09]
  – decay of similarity
  – Reach of spread does not depend on edge-based similarity
  – Left open the question whether OSN effective for spreading information rather than purchasing product
  – Information such as news reach many individuals
• Enumeration of shapes on local cascades [Leskovec,05; Leskovec,07]
  – As cascades grow in size, the number of possible shapes increases exponentially and such enumeration becomes infeasible
Related work: Modeling social epidemics

- **Epidemic Models**
  - Homogenous SIS, SIR Models [Bailey, 1975],
  - segregate heterogeneous population in homogenous subgroups [Hethcote, 1978],
  - Heterogeneous Mean Field (HMF) models [Moreno, 2002]
  - Structure of the underlying network [Wang et al, 2003], regardless of virus propagation mechanisms [Prakash et al., 2010]
  - Effect of adding `stifflers’ [Barrat, 08]

- **Spreading mechanism**
  - Decreasing cascade model [Kempe 03; Leskovec, 07; Kossinets, 06]
  - Complex contagion on Twitter [Romero et al. 2011]
  - Viral cascades as branching processes [Iribarren, 09]
Related work: Centrality and diffusion

- Network structure and diffusion processes
  - Epidemic models and spectral radius [Wang,03]
  - Random walk and Laplacian [Chung,97]
- Most centrality metrics make implicit assumptions about network flow [Borgatti, 05]
  - We advocate a simpler classification scheme
- Ranking of Twitter users
  - Comparison with centrality metrics [Cha,10; Lee,10]
- Community detection
  - We extend the edge-based modularity maximization [Newman,04] to take path-based connectivity into account
Conclusion

Empirical Investigations of Networks

• Leverage the power of OSN
  • Studied Digg and Twitter

• Structure
  • Degree distribution

• Dynamics-Information Propagation
  • How stories evolve?
  • Metric to Quantify Cascades
    − Macroscopic and Microscopic Properties
  • Puzzle: Why are cascades so small?
    − Analysis and Simulations
    − Clustering
    − Contagion Mechanism-FSM
Conclusion

Modeling Networks

• Structure and Functionality
• Dynamic Processes
  – Conservative and Non-Conservative
• Prediction of dynamics uses structure (Spectral radius)
  – Epidemic Models
    – Non-Conservative
•Prediction of structure uses dynamics
  – Importance and Centrality Metrics
    – Conservative and non-conservative
    – How to choose centrality metrics?
    – Evaluation of centrality metrics
  – Alpha centrality better models information diffusion
    – Parameterized and useful for network analysis
    – Normalized Alpha Centrality
    – Communities, leaders and bridges
Thanks

Collaborators

• Rumi Ghosh (USC)
• Greg Ver Steeg (USC)
• Tad Hogg (IMM)
• Tawan Surachawala (USC)

Funding agencies

• NSF
• AFOSR
• AFRL