Contagion on networks: Viral Marketing & Blogs

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
Plan for today

- Probabilistic models of network contagion
- How contagions diffuse in real-life:
  - Viral marketing
  - Blogs
  - Group membership
Models of Virus Propagation

- How do viruses/rumors propagate?
- Will a flu-like virus linger, or will it become extinct?
- (Virus) birth rate $\beta$:
  - probability than an infected neighbor attacks
- (Virus) death rate $\delta$:
  - probability that an infected node heals
General scheme for epidemic models:

- **S**...susceptible
- **E**...exposed
- **I**...infected
- **R**...recovered
- **Z**...immune
Assuming perfect mixing, i.e., a network is a complete graph

The model dynamics:

\[ \frac{dS}{dt} = -\beta IS \]
\[ \frac{dI}{dt} = \beta IS - \nu I \]
\[ \frac{dR}{dt} = \nu I \]
SIS model

- Susceptible-Infective-Susceptible (SIS) model
- Cured nodes immediately become susceptible
- Virus “strength”: $s = \frac{\beta}{\delta}$

Infected by neighbor with prob. $\beta$

Cured internally with prob. $\delta$
Assuming perfect mixing (complete graph):

\[ \frac{dS}{dt} = -\beta SI + \delta I \]

\[ \frac{dI}{dt} = \beta SI - \delta I \]
Representing SIS epidemic as an SIR model

SIS and SIR: Connection

- Representing SIS epidemic as an SIR model

Diagram showing the progression from S to I and then back to S, with steps labeled 0 to 4.

S → I → R
Epidemic threshold of a graph is a value of $t$, such that:

- If strength $s = \beta / \delta < t$ epidemic can not happen (it eventually dies out)

- Given a graph compute its epidemic threshold
What should $t$ depend on?

- Avg. degree? And/or highest degree?
- And/or variance of degree?
- And/or third moment of degree?
- And/or diameter?
We have no epidemic if:

\[
\frac{\beta}{\delta} < t = \frac{1}{\lambda_{1,A}}
\]

(Virus) Death rate

(Virus) Birth rate

Epidemic threshold

\[\lambda_{1,A}\] alone captures the property of the graph!
Experiments (AS graph)

10,900 nodes and 31,180 edges

- $\beta/\delta > \tau$ (above threshold)
- $\beta/\delta = \tau$ (at the threshold)
- $\beta/\delta < \tau$ (below threshold)
Experiments

- Does it matter how many people are initially infected?

(a) Below the threshold, $s=0.912$

(b) At the threshold, $s=1.003$

(c) Above the threshold, $s=1.1$
Human adoption curves

- Prob. of adoption depends on the number of friends who have adopted [Bass ‘69, Granovetter ’78]
  - What is the shape?
    - Distinction has consequences for models and algorithms

Diminishing returns?  Critical mass?
**Diffusion in Viral Marketing**

- Senders and followers of recommendations receive discounts on products
  - 10% credit
  - 10% off

- Data – Incentivized Viral Marketing program
  - 16 million recommendations
  - 4 million people
  - 500,000 products
Diffusion of Community Membership

- Use social networks where people belong to explicitly defined groups
- Each group defines a behavior that diffuses

Data – LiveJournal:
- On-line blogging community with friendship links and user-defined groups
- Over a million users update content each month
- Over 250,000 groups to join
Adoption curve: Validation

DVD recommendations
(8.2 million observations)
How do diffusion curves look like?

- LiveJournal community membership

![Graph showing probability of joining vs. number of friends in the community](image-url)
What are we really measuring?

- For viral marketing:
  - We see that node $v$ receiving the $i$-th recommendation and then purchased the product

- For communities:
  - At time $t$ we see the behavior of node $v$’s friends

- Questions:
  - When did $v$ become aware of recommendations or friends’ behavior?
  - When did it translate into a decision by $v$ to act?
  - How long after this decision did $v$ act?
Diffusion in Viral Marketing

- Large anonymous online retailer (June 2001 to May 2003)
  - 15,646,121 recommendations
  - 3,943,084 distinct customers
  - 548,523 products recommended
  - Products belonging to 4 product groups:
    - books
    - DVDs
    - music
    - VHS
Majority of recommendations do not cause purchases nor propagation
- Notice many star-like patterns
- Many disconnected components

• purchase following a recommendation
• customer recommending a product
• customer not buying a recommended product
Cascade formation process

- $t_1 < t_2 < \ldots < t_n$

Legend:
- Blue: bought bit
- Pink: buy edge
- Green: late recommendation
- Orange: received a discount
- Yellow: bought and received a discount
- Purple: bought but didn’t receive a discount
- White: received a recommendation but didn’t buy
### What role does the product category play?

<table>
<thead>
<tr>
<th></th>
<th>products</th>
<th>customers</th>
<th>recommendations</th>
<th>edges</th>
<th>buy + get discount</th>
<th>buy + no discount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>103,161</td>
<td>2,863,977</td>
<td>5,741,611</td>
<td>2,097,809</td>
<td>65,344</td>
<td>17,769</td>
</tr>
<tr>
<td>DVD</td>
<td>19,829</td>
<td>805,285</td>
<td>8,180,393</td>
<td>962,341</td>
<td>17,232</td>
<td>58,189</td>
</tr>
<tr>
<td>Music</td>
<td>393,598</td>
<td>794,148</td>
<td>1,443,847</td>
<td>585,738</td>
<td>7,837</td>
<td>2,739</td>
</tr>
<tr>
<td>Video</td>
<td>26,131</td>
<td>239,583</td>
<td>280,270</td>
<td>160,683</td>
<td>909</td>
<td>467</td>
</tr>
<tr>
<td>Full</td>
<td>542,719</td>
<td>3,943,084</td>
<td>15,646,121</td>
<td>3,153,676</td>
<td>91,322</td>
<td>79,164</td>
</tr>
</tbody>
</table>

**Diagram:**
- **high**
- **low**
- **people**
- **recommendations**
Observations on product groups

- There are relatively few DVD titles, but DVDs account for ~ 50% of recommendations.

- **Recommendations per person**
  - DVD: 10
  - books and music: 2
  - VHS: 1

- **Recommendations per purchase**
  - books: 69
  - DVDs: 108
  - music: 136
  - VHS: 203

- Overall there are 3.69 recommendations per node on 3.85 different products.

- Music recommendations reached about the same number of people as DVDs but used only 1/5 as many recommendations

- Book recommendations reached by far the most people – 2.8 million.

- All networks have a very small number of unique edges. For books, videos and music the number of unique edges is smaller than the number of nodes – the networks are highly disconnected
More subtle features

- Does sending more recommendations influence more purchases?

![Graph showing the relationship between outgoing recommendations and number of purchases for books and DVDs.](chart.png)
What is the effectiveness of subsequent recommendations?
Product characteristics

- Consider successful recommendations in terms of
  - av. # senders of recommendations per book category
  - av. # of recommendations accepted
- Books overall have a 3% success rate
  - (2% with discount, 1% without)
- Lower than average success rate (significant at p=0.01 level)
  - Fiction
    - Romance (1.78), horror (1.81)
    - Teen (1.94), children’s books (2.06)
    - Comics (2.30), sci-fi (2.34), mystery and thrillers (2.40)
  - Nonfiction
    - Sports (2.26)
    - Home & garden (2.26)
    - Travel (2.39)
- Higher than average success rate (statistically significant)
  - Professional & technical
    - Medicine (5.68)
    - Professional & technical (4.54)
    - Engineering (4.10), science (3.90), computers & internet (3.61)
    - Law (3.66), business & investing (3.62)
Anime DVDs

- 47,000 customers responsible for the 2.5 out of 16 million recommendations in the system
- 29% success rate per recommender of an anime DVD
- Giant component covers 19% of the nodes
- Overall, recommendations for DVDs are more likely to result in a purchase (7%), but the anime community stands out
## Predicting recommendation success

<table>
<thead>
<tr>
<th>Variable</th>
<th>Transformation</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td></td>
<td>-0.940 ***</td>
</tr>
<tr>
<td># recommendations</td>
<td>ln(r)</td>
<td>0.426 ***</td>
</tr>
<tr>
<td># senders</td>
<td>ln(n_s)</td>
<td>-0.782 ***</td>
</tr>
<tr>
<td># recipients</td>
<td>ln(n_r)</td>
<td>-1.307 ***</td>
</tr>
<tr>
<td>product price</td>
<td>ln(p)</td>
<td>0.128 ***</td>
</tr>
<tr>
<td># reviews</td>
<td>ln(v)</td>
<td>-0.011 ***</td>
</tr>
<tr>
<td>avg. rating</td>
<td>ln(t)</td>
<td>-0.027 *</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.74</td>
</tr>
</tbody>
</table>

Significance at the 0.01 (***) , 0.05 (**) and 0.1 (*) levels
Adoption of viral marketing program & social network connectivity

![Graph showing the size of the giant component as a function of the number of nodes and the month.](graph.png)

- Size of giant component vs. number of nodes and month.
- Quadratic fit: $1.7 \times 10^6 m$.
- Number of nodes: $\leq 2$. 

Viral Marketing: Not spreading virally

- 94% of users make first recommendation without having received one previously

- Size of giant connected component increases from 1% to 2.5% of the network (100,420 users) – small!

- Some sub-communities are better connected
  - 24% out of 18,000 users for westerns on DVD
  - 26% of 25,000 for classics on DVD
  - 19% of 47,000 for anime (Japanese animated film) on DVD

- Others are just as disconnected
  - 3% of 180,000 home and gardening
  - 2-7% for children’s and fitness DVDs
Viral Marketing: Consequences

Products suited for Viral Marketing:
- small and tightly knit community
  - few reviews, senders, and recipients
  - but sending more recommendations helps
- pricey products
- rating doesn’t play as much of a role

Observations for future diffusion models:
- purchase decision more complex than threshold or simple infection
- influence saturates as the number of contacts expands
- links user effectiveness if they are overused

Conditions for successful recommendations:
- professional and organizational contexts
- discounts on expensive items
- small, tightly knit communities
How Do Cascades Look Like?

- How big are cascades?
- What are the building blocks of cascades?
Cascades as Graphs

- Given a (social) network
- A process by spreading over the network creates a graph (a tree)

Social network

Cascade
(propagation graph)

Let’s count cascades
Frequent cascade subgraphs

- General observations:
  - DVDs have the richest cascades (most recommendations, most densely linked)
  - Books have small cascades
  - Music is 3 times larger than video but does not have much variety in cascades

<table>
<thead>
<tr>
<th></th>
<th>cascades</th>
<th>different</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>122,657</td>
<td>959</td>
</tr>
<tr>
<td>DVD</td>
<td>289,055</td>
<td>87,614</td>
</tr>
<tr>
<td>Music</td>
<td>13,330</td>
<td>158</td>
</tr>
<tr>
<td>Video</td>
<td>1,928</td>
<td>109</td>
</tr>
</tbody>
</table>

number of all “words” | vocabulary size
Viral Marketing: Frequent Cascades

- •→• is the most common cascade subgraph
- It accounts for ~75% cascades in books, CD and VHS, only 12% of DVD cascades
- •→••→•• is 6 (1.2 for DVD) times more frequent than •→••→•
- For DVDs •→••→•• is more frequent than •→••→•→•
- Chains (•→•→•) are more frequent than •→•→•→•
- •→••→• is more frequent than a collision (•→•→•) (but collision has less edges)
- Late split (•→•→•) is more frequent than •→•→••→•
Viral Marketing Cascades

- Stars ("no propagation")

- Bipartite cores ("common friends")

- Nodes having same friends
- Delete late recommendations
- Count how many people are in a single cascade
- Exclude nodes that did not buy

\[ 10^6 = 1.8 \times 10^{-4.98} \]

Steep drop-off

Very few large cascades

Cascade Size: DVDs

- DVD cascades can grow large
- Possibly as a result of websites where people sign up to exchange recommendations

\[ x \sim x^{-1.56} \]

Shallow drop off – fat tail

a number of large cascades
**Data – Blogs:**

- We crawled 45,000 blogs for 1 year
- 10 million posts and 350,000 cascades
Cascade shapes

The probability of observing a cascade on $n$ nodes follows a Zipf distribution:

$$p(n) \sim n^{-2}$$
Properties of Cascades

- Most of cascades are trees

Number of cascades per node also follows power-law distribution.
Cascade sizes follow a heavy-tailed distribution

- Viral marketing:
  - Books: steep drop-off: power-law exponent -5
  - DVDs: larger cascades: exponent -1.5

- Blogs:
  - Power-law exponent -2

What’s a good model?

- What role does the underlying social network play?
- Can make a step towards more realistic cascade generation (propagation) model?
Towards a Better Cascade Model

1) Randomly pick blog to infect, add to cascade.

2) Infect each in-linked neighbor with probability $\beta$.

3) Add infected neighbors to cascade.

4) Set node infected in (i) to uninfected.
Cascade Model: Results

Generative model produces realistic cascades

\[ \beta = 0.025 \]

Most frequent cascades

![Graphs of cascade model results](http://cs224w.stanford.edu)
Next: Creating big cascades

- Blogs – information epidemics
  - Which are the influential/infectious blogs?

- Viral marketing
  - Who are the trendsetters?
  - Influential people?

- Disease spreading
  - Where to place monitoring stations to detect epidemics?