CS 6604: Data Mining Large Networks and Time-series

B. Aditya Prakash

Lecture #7: Epidemics: Probabilistic Models
Epidemics on random trees

- A patient meets $d$ other people
- With probability $q > 0$ infects each of them
- For which values of $d$ and $q$ does the epidemic run forever?

Depends on

$$\lim_{h \to \infty} P\left[\text{infected node at depth } h\right]$$

Run forever: 

Die out:
Branching processes

- \( p_h = \text{prob. there is an infected node at depth } h \)

\[
p_h = 1 - (1 - q \cdot p_{h-1})^d
\]

- \( \lim_{h \to \infty} p_h = ? \)

No parent from depth \( h-1 \) infects a child
Branching processes

- $p_h = \text{prob. there is an infected nodes at depth } h$

$$p_h = 1 - (1 - q \cdot p_{h-1})^d$$

- $\lim_{h \to \infty} p_h = \text{result of iterating}$

$$f(x) = 1 - (1 - q \cdot x)^d$$

Start at $x = 1$, why?
Fixed point of $f(x) = 1 - (1 - q \cdot x)^d$

Going to first fixed point

When is this going to 0?

What do we know about $f(x)$?

- $f(0) = 0$
- $f(1) = 1 - (1 - q)^d < 1$
- $f'(x) = q \cdot d (1 - qx)^{d-1}$
- $f'(0) = q \cdot d$ so $f'(x)$ is monotone decreasing on $[0,1]$!
When is the fixed point zero?

For the epidemic to die out we need $f(x)$ to be below $y=x$!

So: $f'(0) = q \cdot d < 1$

$$\lim_{h \to \infty} p_h = 0 \quad \text{when} \quad q \cdot d < 1$$

$q \cdot d = \text{expected # of people at we infect}$

Reproductive number $R_0 = q \cdot d$: There is an epidemic if $R_0 \geq 1$
Other models

- Till now we allowed nodes from healthy-infected on trees
- We can generalize nodes to alternate....
“SIR” model: life immunity (mumps)

- Each node in the graph is in one of three states
  - Susceptible (i.e. healthy)
  - Infected
  - Removed (i.e. can’t get infected again)

\[
\begin{align*}
  t = 1 & \quad \text{Prob. } \beta \\
  t = 2 & \quad \text{Prob. } \delta \\
  t = 3 & \\
\end{align*}
\]
Terminology: continued

- Other virus propagation models ("VPM")
  - SIS: susceptible-infected-susceptible, flu-like
  - SIRS: temporary immunity, like pertussis
  - SEIR: mumps-like, with virus incubation
    \[(E = Exposed)\]

- Underlying contact-network – ‘who-can-infect-whom’
SIR Model

- Assuming perfect mixing (cliques)

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

Number of new infections = $\beta \ast \text{# attacks}$

Number of infected nodes curing
Force of Infection

- $F = \lambda S$
  - $\lambda$ is number of infected people

- Kinds of transmission
  - Mass-action transmission $\lambda = \beta I$
  - Density-dependent transmission $\lambda = \beta I/N$
    - As the density increases so does transmission
Solving SIR on cliques

\[ S(t) = S(0)e^{-R_0(R(t)-R(0))} \]

\[ R_\infty = 1 - S(0)e^{-R_0(R_\infty-R(0))} \]

\[ R_0 = \frac{N\beta}{\delta} \]
SIR on cliques

Number of nodes

S(t)

I(t)

R(t)

time

Susceptible

Infected

Recovered

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

\[ \frac{dI}{dt} = \beta SI - \gamma I \]

\[ \frac{dR}{dt} = \gamma I \]
SIR on cliques: Measles in the Netherlands

Fig. 3: Solutions of the classic SIR epidemic model with contact number $\sigma = 3$ and average infectious period $1/\gamma = 3$ days.

Fig. 4: Reported number of measles cases in the Netherlands by week of onset and vaccination status during April 1999 to January 2000. Most of the unvaccinated cases were people belonging to a religious denomination that routinely does not accept vaccination. The 2,961 measles cases included 3 measles-related deaths. Reprinted from [52].
SIS model

- Assuming perfect mixing (cliques)

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

\[
\frac{dI}{dt} = \beta SI - \delta I
\]
Many many extensions

- With birth/death rates (‘vital dynamics’)
- Variable contact rates
- Age-structured models
- ........
- ........
General Scheme...

The general transfer diagram for the MSEIR model with the passively immune class M, the susceptible class S, the exposed class E, the infective class I, and the recovered class R.

- **Births with passive immunity** to an infection. The class M contains these infants with passive immunity. After the maternal antibodies disappear from the body, the infant moves to the susceptible class S.
- **Births without passive immunity** also enter the class S of susceptible individuals; that is, those who can become infected.
- When there is an adequate contact of a susceptible with an infective so that transmission occurs, then the susceptible enters the exposed class E of those in the latent period, who are infected but not yet infectious.
- After the latent period ends, the individual enters the class I of infectives, who are infectious in the sense that they are capable of transmitting the infection.
- When the infectious period ends, the individual enters the recovered class R consisting of those with permanent infection-acquired immunity.

The choice of which compartments to include in a model depends on the characteristics of the particular disease being modeled and the purpose of the model. The passively immune class M and the latent period class E are often omitted, because they are not crucial for the susceptible-infective interaction. Acronyms for epidemiology models are often based on the flow patterns between the compartments such as MSEIR, MSEIRS, SEIR, SEIRS, SIR, SIRS, SEI, SEIS, SI, and SIS. For example, in the MSEIR model shown in Figure 1, passively immune newborns first become susceptible, then exposed in the latent period, then infectious, and then removed with permanent immunity. An MSEIRS model would be similar, but the immunity in the R class would be temporary, so that individuals would regain their susceptibility when the temporary immunity ended.

The threshold for many epidemiology models is the basic reproduction number $R_0$, which is defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible [61]. For many deterministic epidemiology models, an infection can get started in a fully susceptible population if and only if $R_0 > 1$. Thus the basic reproduction number $R_0$ is often considered as the threshold quantity that determines when an infection can invade and persist in a new host population. Section 2 introduces epidemiology modeling by formulating and analyzing two classic deterministic models. The role of $R_0$ is demonstrated for the classic SIR endemic model in section 2.4. Then thresholds are estimated from data on several diseases and the implications of the estimates are considered for diseases such as smallpox, polio, measles, rubella, chickenpox, and influenza. An MSEIR endemic model in a population without age structure but with exponentially changing population size is formulated and analyzed in section 3. This model demonstrates how exponential population growth affects the basic reproduction number $R_0$.

Realistic infectious disease models include both time $t$ and age $a$ as independent variables, because age groups mix heterogeneously, the recovered fraction usually increases with age, risks from an infection may be related to age, vaccination pro-
What happens on other graphs?

- Power-Law random graphs
- Arbitrary graphs?
- ..... 

- Next class...
INFORMATION SPREAD

Prakash 2015
CS 6604: DM Large Networks & Time-Series
Independent Cascade Model

- Each edge \((u,v)\) has weight \(p_{uv}\)

- When node becomes active
  - It activates each out-neighbor with prob. on the edge
IC Model

- **Pros:**
  - Intuitive
  - Has some nice properties

- **Cons:**
  - Too many parameters
Cascades on the Blogosphere

Blogosphere blogs + posts

Blog network links among blogs

Post network links among posts

Cascade is graph induced by a time ordered propagation of information (edges)
Blog data

- 45,000 blogs participating in cascades
- All their posts for 3 months (Aug-Sept ‘05)
- 2.4 million posts
- ~5 million links (245,404 inside the dataset)
Post popularity drops-off – exponentially?
Popularity over time

Post popularity drops-off – exponentially? POWER LAW!
Exponent?
Popularity over time

Post popularity drops-off – exponentially? POWER LAW!
Exponent? -1.6
• close to -1.5: Barabasi’s stack model
• and like the zero-crossings of a random walk
-1.5 slope

Topological Observations

*How do we measure how information flows through the network?*

Common cascade shapes extracted using algorithms in [Leskovec, Singh, Kleinberg; PAKDD 2006].
Topological Observations

What graph properties do cascades exhibit?

Cascade size distributions also follow power law.

Observation 2: The probability of observing a cascade on n nodes follows a Zipf distribution:

\[ p(n) \propto n^{-2} \]
Topological Observations

What graph properties do cascades exhibit?

Stars and chains also follow a power law, with different exponents (star $-3.1$, chain $-8.5$).
Blogs and structure

- Cascades take on different shapes (sorted by frequency):

How can we use cascades to identify communities?
PCA on cascade types

- Perform PCA on sparse matrix.
- Use log(count+1)
- Project onto 2 PC...

\[ \begin{array}{ccc}
\text{slashdot} & 4.6 & 2.1 & .09 \\
\text{boingboing} & 3.2 & 1.1 & 3.4 & .07 \\
\vdots & 4.2 & & \\
\vdots & 5.1 & & \\
\vdots & 2.1 & 1.1 & \\
\vdots & .67 & .07 & \\
\vdots & .01 & & \\
\end{array} \]

~9,000 cascade types

\~44,000 blogs

Prakash 2015
PCA on cascade types

• Observation: Content of blogs and cascade behavior are often related.

• Distinct clusters for “conservative” and “humorous” blogs (hand-labeling).

PCA on cascade types

• Observation: Content of blogs and cascade behavior are often related.

• Distinct clusters for “conservative” and “humorous” blogs (hand-labeling).

Complex Contagion

- From exposure to adoptions
  - Node’s neighbor exposes the node to contagion
  - The node acts on the contagion
Exposure Curves [Romero+2010]

- Probability of adopting a new behavior depends on the number of friends who have already adopted.

\[ k = \text{number of friends adopting} \]

Diminishing returns: Viruses, Information
Exposure Curves [Romero+2010]

- Probability of adopting a new behavior depends on the number of friends who have already adopted.

\[ k = \text{number of friends adopting} \]

**Diminishing returns:** Viruses, Information

**Critical mass:** Decision making
Exposure curves

- **Exposure**: Node's neighbor exposes the node to information.
- **Adoption**: The node acts on the information.
- **Adoption curve**: The probability of infection ever increases.
- **Nodes build resistance**: Probability decreases over time.
Example Application

- Marketing Agency would like you to buy their product
- They estimate the adoption curve

3
Exposure: Validation: DVDs on Amazon [Leskovec+, 07]
LiveJournal Group Membership
[Backstrom+ 06]

Prob. of joining vs. k (number of friends in the group)
Exposure Curve: Twitter [Romero et. al. ’11]

- Aug’09 to Jan’10: 3B tweets, 60M users

- Avg. exposure curve for the top 500 hts
- How to characterize the curves?
Modeling the Shape of the curve

- Persistence of $P$ is the ratio of the area under the curve $P$ and the area of the rectangle of length $\max(P)$, width $D(P)$

Persistence measures the decay of exposure curves
Stickiness of P is max(P)

– Probability of usage at the most effective exposure
Exposure Curves: Persistence

- Manually identify 8 broad categories with at least 20 hashtags each.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celebrity</td>
<td>mj, brazilwantsjb, regis, iwantpeterfacinelli</td>
</tr>
<tr>
<td>Music</td>
<td>thisiswar, mj, musicmonday, pandora</td>
</tr>
<tr>
<td>Games</td>
<td>mafiawars, spymaster, mw2, zyngapirates</td>
</tr>
<tr>
<td>Political</td>
<td>tcot, glennbeck, obama, hcr</td>
</tr>
<tr>
<td>Idiom</td>
<td>cantlivewithout, dontyouhate, musicmonday</td>
</tr>
<tr>
<td>Sports</td>
<td>golf, yankees, nhl, cricket</td>
</tr>
<tr>
<td>Movies/TV</td>
<td>lost, glennbeck, bones, newmoon</td>
</tr>
<tr>
<td>Technology</td>
<td>digg, iphone, jquery, photoshop</td>
</tr>
</tbody>
</table>
Exposure Curve: Persistence

- Idioms and Music have lower persistence than that of a random subset of hashtags of same size.
- Politics and Sports have higher persistence than that of a random subset of same size.
Exposure Curve: Stickiness

- Tech. and Movies have lower stickiness than random
- Music has higher stickiness than random
Other structural models

- Ugander+, 2012

A
Friendship neighborhood:

Contact neighborhood:

B
Connected components
Components of size ≥ 3
Components in 2-core
Components in 1-brace
Structural Models

C

D

E

F

[Ugander+, 2012]

Prakash 2015

CS 6604: DM Large Networks & Time-Series
Haven’t covered

- Decision-based Models
- Hybrid Models
- ...

Prakash 2015