Social and Technological Network Data Analytics

Lecture 11: Information Cascades and Epidemics Applications

Ack to D. Liben-Nowell and M. Cha for slides.

Prof Cecilia Mascolo
In this lecture we will show some more examples of applications of epidemics and information cascades in real networks.
Characterizing Social Cascades in Flickr

- Flickr social network (25%): WCC.
- Growing dataset over 100 days.
- 2M users.
- Favourite photo info used.
- 34,734,221 favorite markings over 11,267,320 distinct photos.
Questions Answered

• Does content in Flickr spread along links in the social network?
• What are the properties of content dissemination in Flickr (e.g., how long after being exposed to a piece of content do users tend to propagate it)?
• Can existing epidemiological models characterize the information dissemination observed in Flickr?
How did fans get to know this picture?

Fire Canoe #2

![Graph showing the number of fans over photo age (days)]
Mechanisms of Information Propagation

- Featuring (front page, hotlists)
- External links
- Search results
- Links between content
- Online social links
How to identify information flow through social links?

- Did a particular bookmark spread through social links?
- **No**: if a user bookmarks a photo and if *none* of his friends have previously bookmarked the photo

- **Yes**: if a user bookmarks a photo *after* one of his friends bookmarked the photo
Steady Increase

- 75% of bookmarks through social links

![Graph](image-url)

Found through social links
Through other mechanisms
Who is driving the increase in fan numbers?

the “social cascade” group accounts for over half of new fans

the dominance of the “social cascade” group over the “other” group switches during the two popularity surges exhibited by photo B
is also possible that some users have many contacts but only one social cascade happens only after several years of exposure to the content, which is possibly due to infrequent user activity. We observe that 50% of cascade events are influenced by a single infector; 20% of the events by two infectors; and the remaining 45% involve three or more potential infectors. For 10% of the events, the infectee had more than 10 contacts who had already marked the photo as a favorite and then counted the proportion of its connected nodes (i.e., the picture was introduced in an isolated part of the network) before the infection happens. To do so, we identify the earliest time when any contacts of the infectee marked the photo as a favorite himself. The dotted line in Figure 4 shows the cumulative distribution of the estimated time when any contacts of the infectee marked the photo as a favorite before marking it as a favorite himself. The solid line in Figure 5 shows the cumulative distribution (in log scale) of the number of infected contacts at the time a user marks a photo as a favorite. Based on these findings, we can assume that a picture will definitely be shared between two connected nodes. If we then define \( R_0 \) as the node degree, and then counting the proportion of its connected nodes (i.e., the natural duration of infection is equal to the lifetime of the user, which is very large in comparison to the timescale of the cascade. In this case, we can assume that a picture will definitely be shared in the Flickr network. However, 20% of the first cascade steps spread through the Flickr network. The solid line in Figure 6 compares the basic reproduction number \( R_0 \) attained directly from the trace and the estimated value can then estimate the reproductive number \( R_0 \) can be assessed by counting the number of nodes participating in each event from collapse to growth. Epidemiologists call this point the epidemic threshold. An empirical estimate of the transmission probability of a social contact in the reverse direction) that subsequent cascade events are influenced by a single infector; 20% of the infectees in a social cascade either result from a single infected individual in an entirely susceptible population or because it initially faced strong competition from other users. Although the concept of heterogeneity in the node degree distribution, \( \hat{R}_0 \) can be calculated by identifying an infected node and in public health the goal is to reduce this number. Figure 5 shows the cumulative distribution (in log scale) of the number of infected contacts at the time a user marks a photo as a favorite. Since we know the number of exposing the object before the infection happens. To do so, we identify the earliest time when any contacts of the infectee marked the photo as a favorite and then counting the proportion of its connected nodes; i.e., the natural duration of infection is equal to the lifetime of the user, which is very large in comparison to the timescale of the cascade. In this situation, we can assume that a picture will definitely be shared in the Flickr network. However, 20% of the first cascade steps spread through the Flickr network. The solid line in Figure 6 shows the cumulative distribution (in log scale) of the number of infected contacts at the time a user marks a photo as a favorite. Since we know the number of exposures to candidate infectors, we can then estimate the reproductive number \( R_0 \). It is defined as the expected number of secondary infections resulting from a single infected individual in an entirely susceptible population. If \( R_0 \) is below 1 in order to stop epidemics. HIV has an \( R_0 \) between 2 and 5; and measles has an \( R_0 \) between 12 and 18.

![Time in Cascades](image)

50% of first cascade steps happens in the first 3 days. 20% take longer than a month. 50% of cascade steps take 50 days.
35% of social cascade events are influenced by a single infector; 20% of the events by two infectors; and the remaining 45% involve three or more potential infectors. For 10% of the events, the infectee had more than 10 contacts who had already marked the same photo as a favorite.

Number of infected contacts when user marks the picture as favourite.

The number of potential infectors

max=236
Model of Spreading

- Let’s recall the definition of $R_0$ in epidemic models
- If $R_0 > 1$ spreads
- If $R_0 < 1$ dies out
- $R_0 = 1$ epidemic threshold
- $R_0$ empirical calculation:
  - For each fan, count how many friends further bookmark the same photo. Average the count.

$$R_0 = \left( \rho_0 \langle k^2 \rangle / \langle k \rangle^2 \right)$$

\[
\langle k \rangle = 14.7, \quad \langle k^2 \rangle / \langle k \rangle^2 = 48.0.
\]
1. Formula based estimation of $R_0$:
   - Estimating $p$: Given an infected node count the neighbours subsequently infected and average.
   - This allows to derive a general $R^0$ from the equation

2. Empirical estimation of $R_0$:
   - Given start node of cascade, count the number of directly infected nodes
R0 correlation across all photos

corrcoef = 0.9765

y = x
Predicting spreading

• The correlation means that by using the social network properties and some simple observation over a short time series of user activity we can predict the popularity of photos.
Discussion

• Social Cascades occur in Flickr
• The basic reproduction number of popular photos is between 1 and 190. This is much higher than very infectious diseases like measles, indicating that social networks are efficient transmission mediums and online content can be very infectious.
• Given the expected spread and the node degree they can predict the expected spread on various networks (knowing $<k>$).
Another study on cascades

- Tracing information flow on a global scale using Internet chain-letter data.
- Iraq Petition Example:

  **UN Petition for Peace**
  
  Non-essential personnel are now evacuating from the US embassies in the middle east. War is about to start. It takes is 20% of us to cry out for "NO WAR" to induce further diplomacy, but they say our numbers are more like 2%. US Congress has authorized the President of the US to go to war against Iraq. Please consider this an urgent request. UN Petition for Peace, Stand for Peace. Islam is not the Enemy. War is NOT the Answer. Speak against a THIRD WORLD WAR. The UN is gathering signatures in an effort to avoid a tragic world event.

  Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know. If you receive this list with more than 500 names signed, please send a copy of the message to:

  usa@un.int
  and president@whitehouse.gov

  Even if you decide not to sign, please consider forwarding the petition on instead of eliminating it.
Data Cleaning and Gathering

- Query search engine to find copies of petitions.
  - (~650 distinct copies found.)
  - (~20K distinct names.)
- compute propagation tree from these copies
  - (x \rightarrow y \text{ if there is a copy where } x \text{ immediately precedes } y.)
Building a propagation tree
Is this really a tree?

- No. some responded twice (have 2 parents)
- Typographical changes are frequent
  
  | John Smith Santa Monica Calif | John Smith Santa Monica USA | John Smith Santa Monica Calif USA |

- List rearrangements are common
Propagation tree

http://petitions.cs.carleton.edu
Solution

- Use the graph

weight($x \rightarrow y$) := # copies s.t. $x$ immediately precedes $y$.

- run max-weight spanning arborescence algorithm to produce a tree from G. [Edmonds 1967]
- prune tree to eliminate any nodes that have no poster nodes beneath them.
Graph to Tree

Nodes who made the list public are in gray.

http://petitions.cs.carleton.edu
Expectations

• The petition is flooding the social network.
• Small world $\Rightarrow$ the tree’s depth will be small. High branching: people have many friends (10’s or 100’s).
• So the propagation tree should be shallow and wide.
• (unless it dies out quickly.)
The tree looks like this

- process doesn’t die out quickly.
  - 20K nodes in posted copies.
- tree is very deep.
  - median node depth ≈ 288.
- tree is very narrow.
  - over 94% of nodes have only one child.
Modelling

• Let us try to find a model that reproduces this:
  • Deep tree
  • Small width
  • Large single child fraction

• Iraq tree (18k nodes)
  • depth 288, width 82, single-child fraction 94%
Tried simulation on LiveJournal

- simulate on real social network (LiveJournal, 4.4M nodes). Randomly choose an initiator node (= root).
- each recipient discards with prob $\delta$, forwards with prob $1 - \delta$. $\delta := 0.65$ [Dodds Muhamad Watts 2003]
- a non-discarding recipient posts his copy with prob $\pi$ (a posted copy ‘lights up’ the root-to-poster path.)
- tree propagates from root until either (i) the process dies out (‘fizzles’) or (ii) observable portion of tree reaches size of Iraq tree.
Epidemic Model

- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x discards with probability $\delta = 0.65$; otherwise:
  - x appends x to
  - x forwards to all neighbours (who act at time $t + 1$)
  - x posts with probability $\pi$.

- Iraq tree (18K nodes): depth 288, width 82, single-child 94%
- Epidemic tree: depth 5, width 9625, single-child 19%
Social networks have lots of “cliquey” communities.
⇒ high degrees in epidemic tree (*not true in Iraq*).

One reason to think that cliques can be “serialized”:
BCDE don’t react synchronously at time $t = 1$.

A mails BCDE at $t = 0$
Each receives message at first email check after $t = 0$.
B responds first ⇒ C gets new copy from B.
Asynchronous Model

- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x chooses a delay $\tau$, where $\Pr[\tau] = ???$. At time $t + \tau$:
  - x discards with probability $\delta = 0.65$; otherwise:
    - x appends x to longest list x received (in $[t, t + \tau]$)
    - x forwards to all x’s neighbours.
  - x posts with probability $\pi$. 
You receive a message at time $t$; you respond at time $t + \tau$.

What does $\tau$ look like?

Letters from Darwin and Einstein:

We use $\Pr[\tau] \propto \tau^{-3/2}$.

(though the precise exponent actually doesn’t matter much.)
Epidemic Model

• randomly choose an initiator node (= root)
• for each $x$ who first receives a list at time $t$: $x$ chooses a delay $\tau$, where $\Pr[\tau] \propto \tau^{-3/2}$. At time $t + \tau$:
  - $x$ discards with probability $\delta = 0.65$; otherwise:
  - $x$ appends $x$ to longest list $x$ received (in $[t, t + \tau]$),
  - $x$ forwards to all $x$’s neighbours $x$,
  - $x$ posts with probability $\pi$. 
Epidemic Model

The Asynchronous Model

Randomly choose an initiator node (= root) for each \( x \) who first receives a list at time \( t \):

\( x \) chooses a delay \( \tau \), where \( P_r[\tau] \propto \tau^{-3/2} \).

At time \( t + \tau \):

- \( x \) discards with probability \( \delta = 0.65 \); otherwise:
  - \( x \) appends \( x \) to the longest list \( x \) received (in \( [t, t + \tau] \))
  - \( x \) forwards to all \( x \)'s neighbors
  - \( x \) posts with probability \( \pi \).

http://petitions.cs.carleton.edu
Asynchronicity has serialized cliques, but we need more. (e.g., social networks are “cliquey” but not just cliques.)

When \( x \) receives a list, it can either

— forward that list to all of \( x \)’s friends, OR

— reply-to-all to all of \( x \)’s corecipients on the message.
The Asynchronous Model

- randomly choose an initiator node (= root)
- for each $x$ who first receives a list at time $t$: $x$ chooses a delay $\tau$, where $\Pr[\tau] \propto \tau^{-3/2}$. At time $t + \tau$:
  - $x$ discards with probability $\delta = 0.65$; otherwise:
  - $x$ appends $x$ to longest list $x$ received (in $[t, t + \tau]$)
  - $x$ forwards to all $x$’s neighbors.
  - $x$ posts with probability $\pi$. 
The Full Model

- randomly choose an initiator node (= root)
- for each $x$ who first receives a list at time $t$: $x$ chooses a delay $\tau$, where $\Pr[\tau] \propto \tau^{-3/2}$. At time $t + \tau$:
  - $x$ discards with probability $\delta = 0.65$; otherwise:
  - $x$ appends $x$ to longest list $x$ received (in $[t, t + \tau]$)
  - with prob $\beta$, $x$ replies to all of $x$’s corecipients;
  - with prob $1 - \beta$, $x$ forwards to all of $x$’s neighbors.
- $x$ posts with probability $\pi$.

The asynchronous model = full model with $\beta=0$
Studying Single Child Proportion

Iraq: 94%

http://petitions.cs.carleton.edu

single-child node fraction

back rate

post rate

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Tree Depth

Iraq: 288
Tree Width

Iraq: 82

Diagram showing a heatmap with axes labeled 'back rate' on the x-axis and 'post rate' on the y-axis, and 'width' on the z-axis. The heatmap ranges from 0 to 1 on both axes, with colors transitioning from red to blue to indicate varying values. The label 'http://petitions.cs.carleton.edu' is present in the image.
It matches!

Simulations: $\beta = 0.950, \pi = 0.22$
Discussion

• A model with asynchronicity and group-reply was a good initial approximation

• More data needed to understand what’s happening
Summary

• We have shown examples of application of cascades and epidemic models to real data

• Real data is challenging and often processes do not match exact models and need tweaking.
References
