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

UNIVERSITY OF CAMBRIDGE

In This Lecture

- 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.

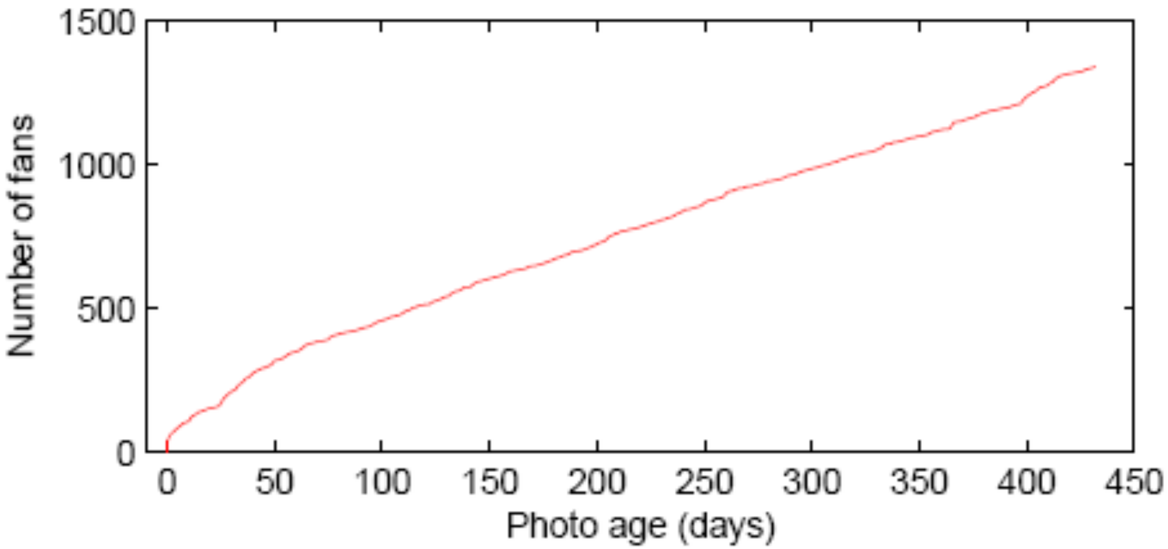
Questions Answered

- 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?

UNIVERSITY OF CAMBRIDGE

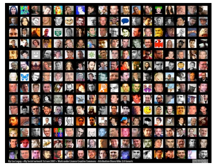
How did fans get to know this picture?

Fire Canoe #2



UNIVERSITY OF CAMBRIDGE

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?

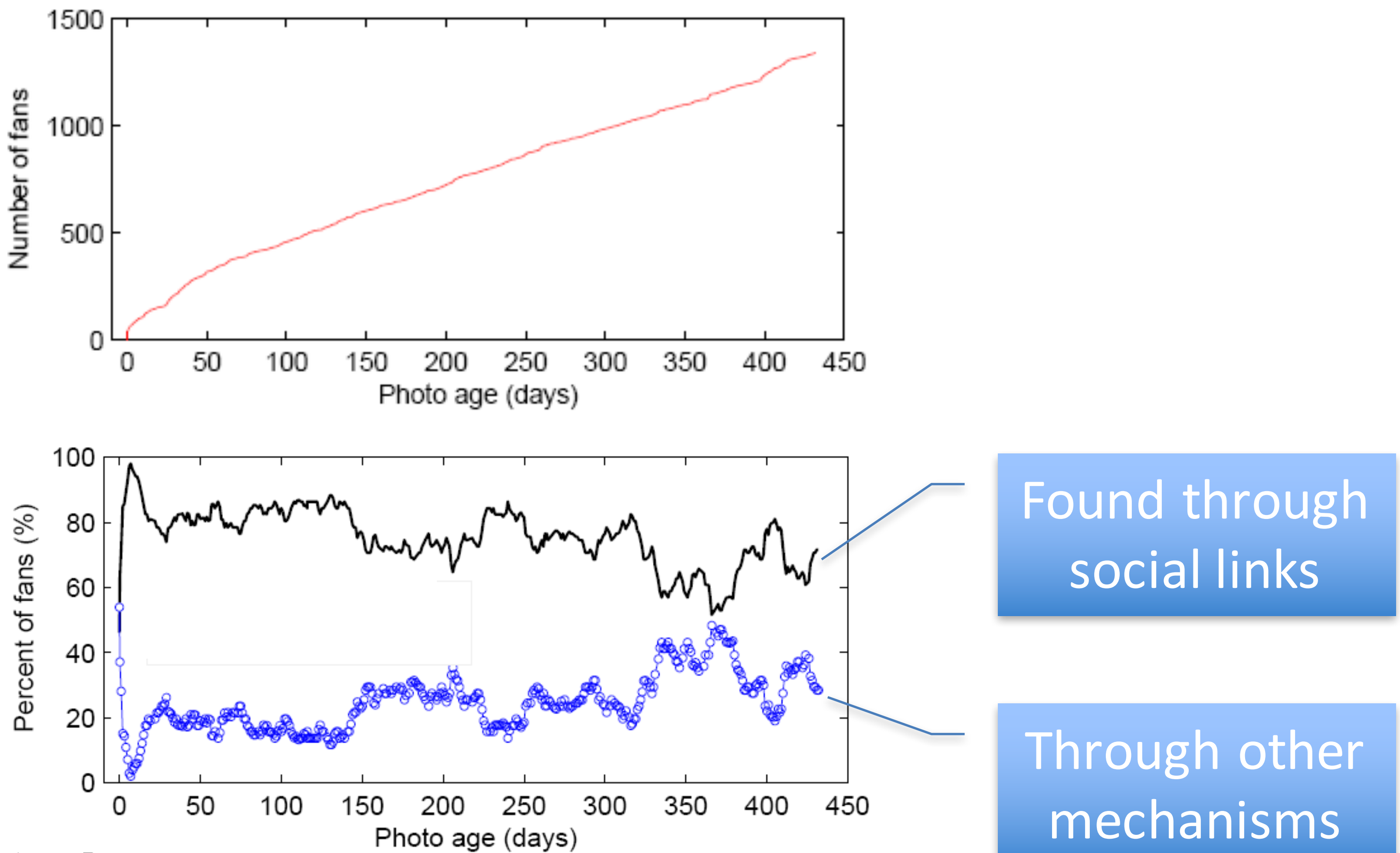
- Did a particular bookmark spread through social links?
- **No:** if a user bookmarks a photo and if *none* of his friends have previously bookmarked the photo

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

UNIVERSITY OF CAMBRIDGE

Steady Increase

- **75%** of bookmarks through social links



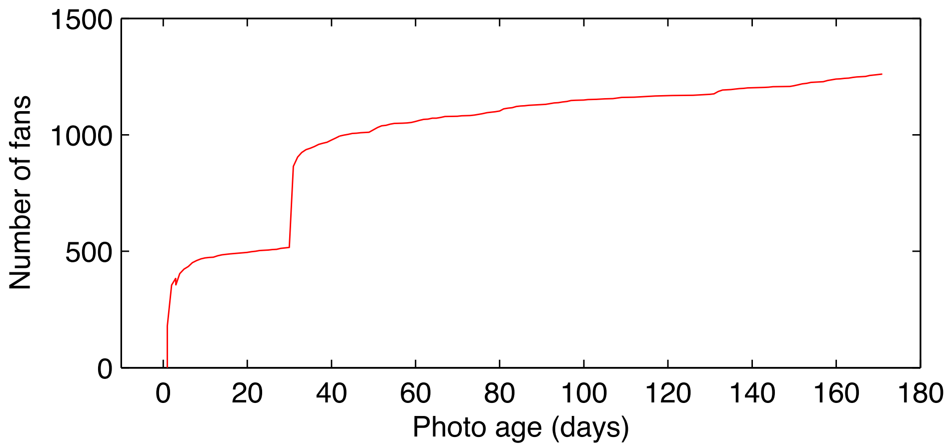
UNIVERSI CAMBR

Fire Canoe #2 by Peter Bowers

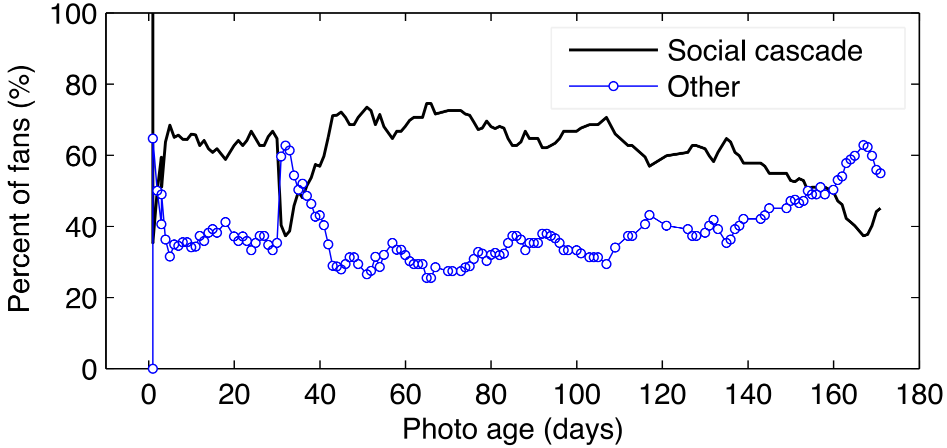
Who is driving the increase in fan numbers?

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

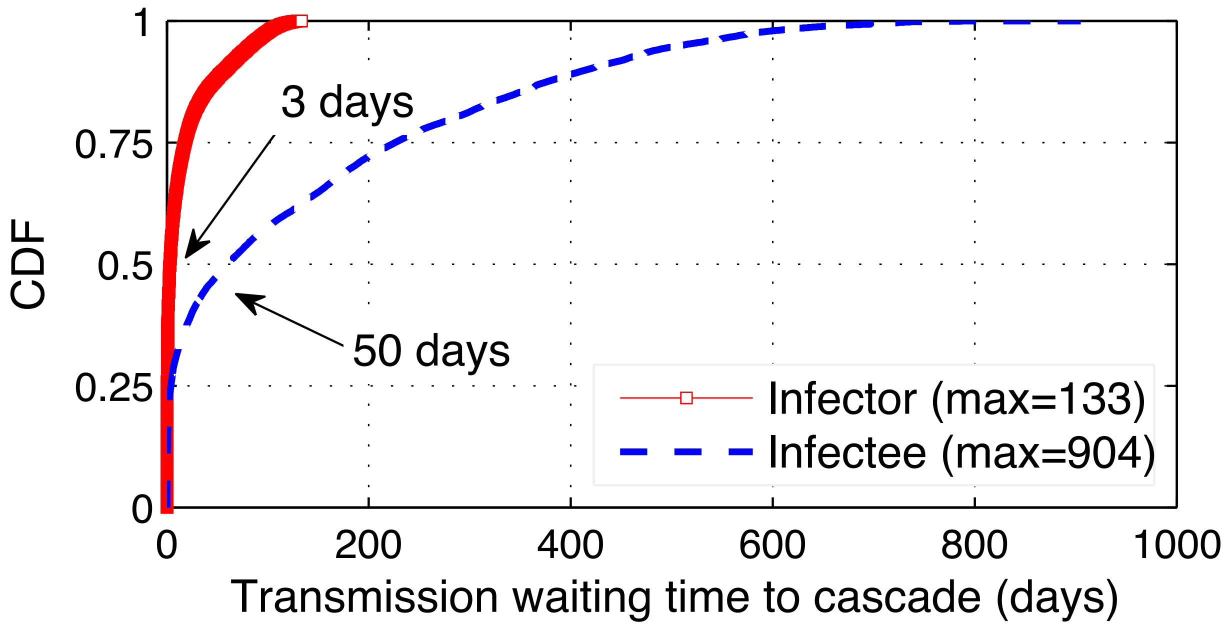


(b) Growth of fans, photo B



UNIVERSITY OF CAMBRIDGE

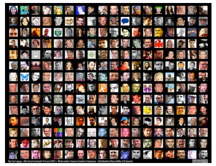
Time in Cascades



50% of first cascade steps happens in the first 3 days.
20% take longer than a month.
50% of cascade steps take 50 days

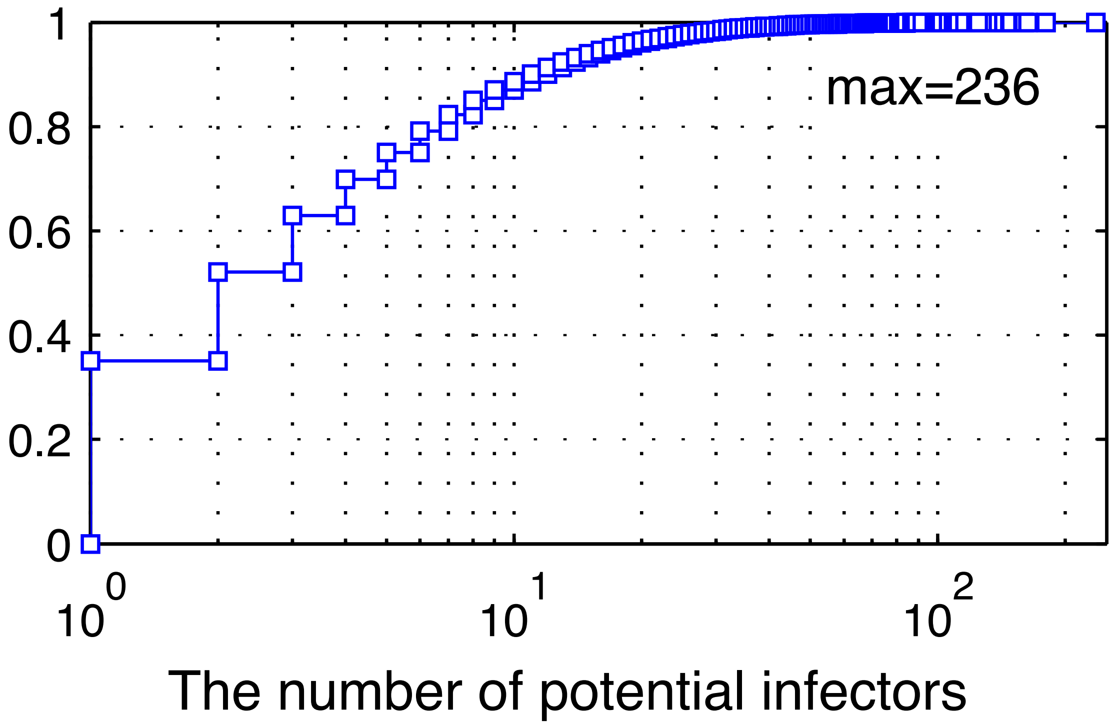
UNIVERSITY OF CAMBRIDGE

Potential Infectors



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.



UNIVERSITY OF CAMBRIDGE

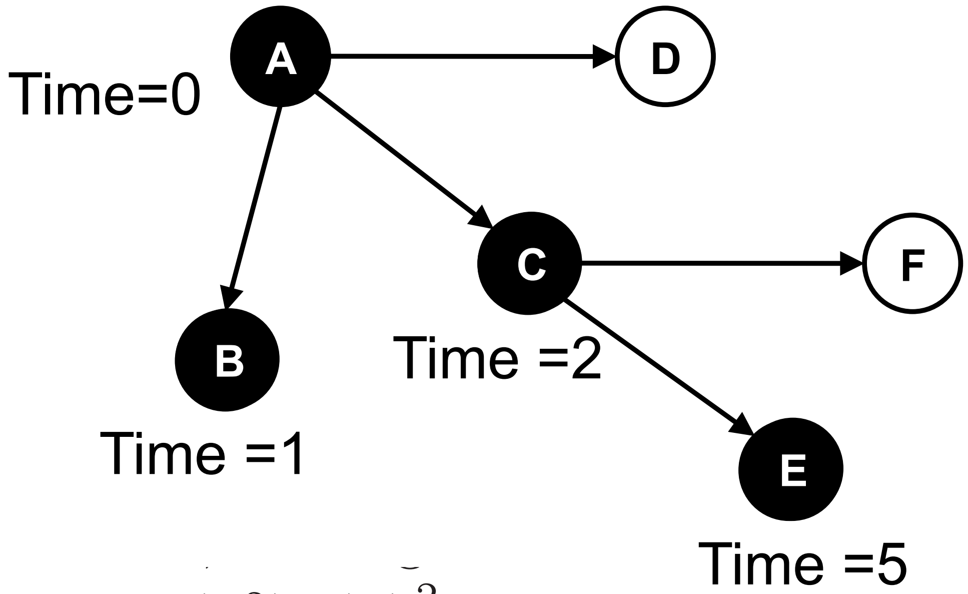
Model of Spreading

- Let's recall the definition of R0 in epidemic models

$$R_0 = \rho_0 \langle k^2 \rangle / \langle k \rangle^2$$

- If R0>1 spreads
- If R0<1 dies out
- R0=1 epidemic threshold

- ro0 empirical calculation:
  - For each fan, count how many friends further bookmark the same photo. Average the count.



UNIVERSITY OF CAMBRIDGE

$$\langle k \rangle = 14.7, \quad \langle k^2 \rangle / \langle k \rangle^2 = 48.0.$$

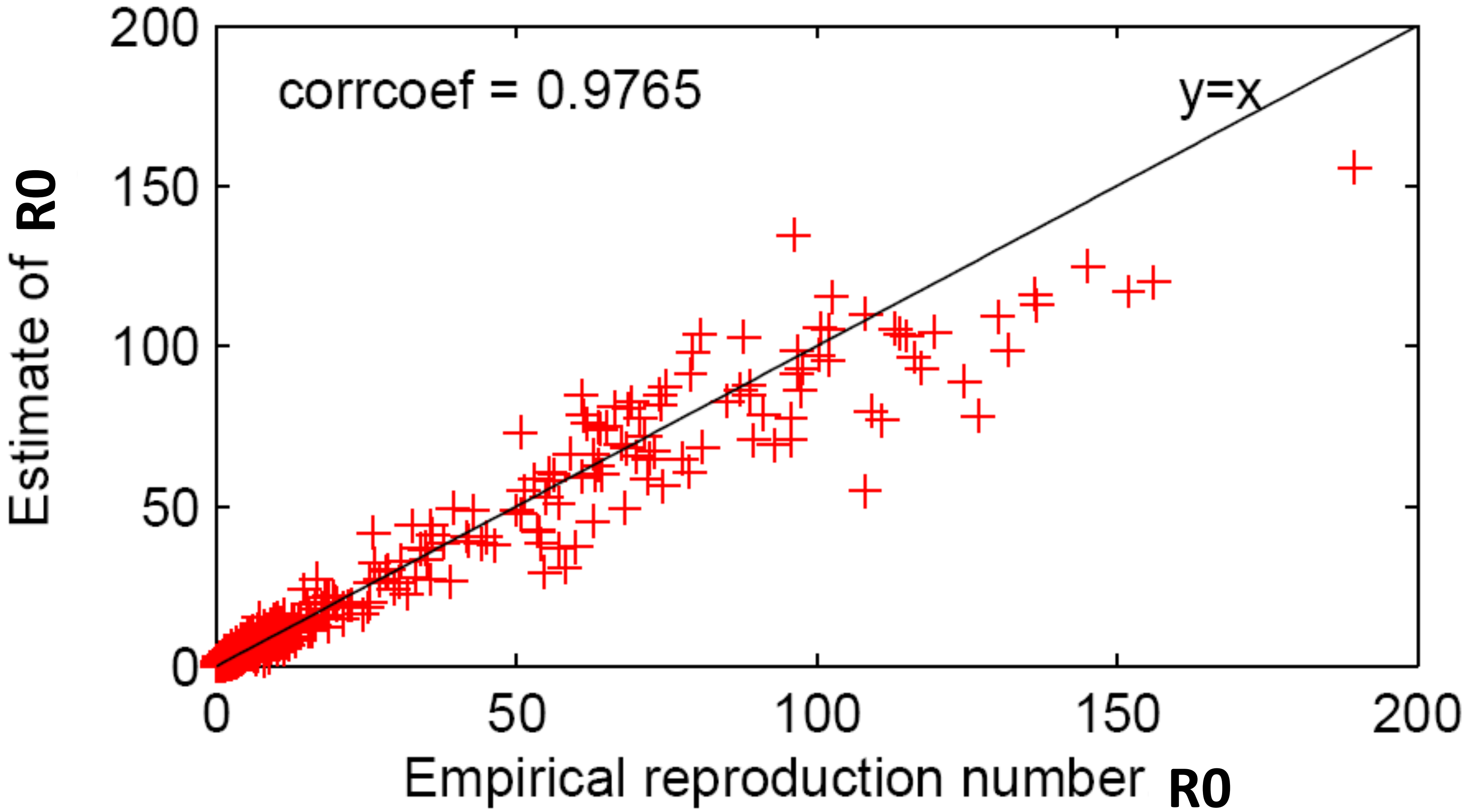
Estimations

1. Formula based estimation of R0:
   – 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 R0:
   – Given start node of cascade, count the number of directly infected nodes

UNIVERSITY OF CAMBRIDGE

R0 correlation across all photos



UNIVERSITY OF CAMBRIDGE

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>).

UNIVERSITY OF CAMBRIDGE

Another study on cascades

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

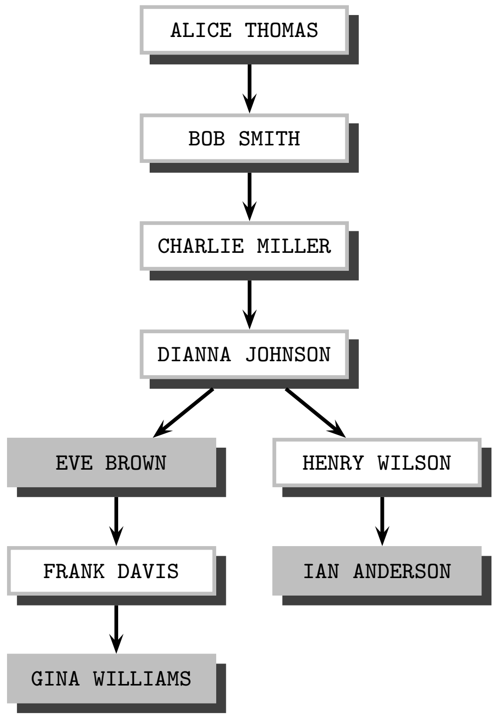
```
Date:  Mon, 17 Mar 2003 16:39:51 -0600
From:  XXXX <XXXX@mac.com>
To:  usa@un.int, president@whitehouse.gov
Subject:  UN Petition

UN Petition for Peace

Non-essential personnel are now evacuating from the US embassies in
the middle east.  Was 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
```

```
1) Alice Thomas
2) Bob Smith
3) Charlie Miller
4) Dianna Johnson
5) Eve Brown
6) Frank Davis
7) Gina Williams
[...]
```

UNIVERSITY OF CAMBRIDGE

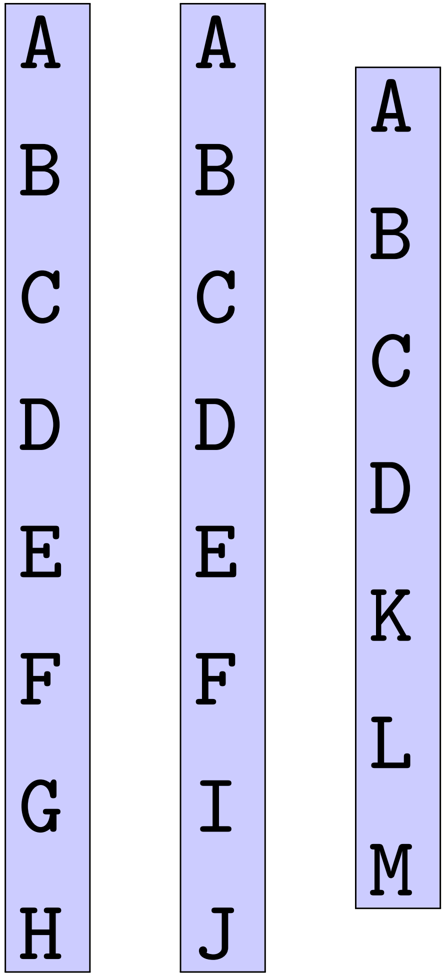
Data Cleaning and Gathering

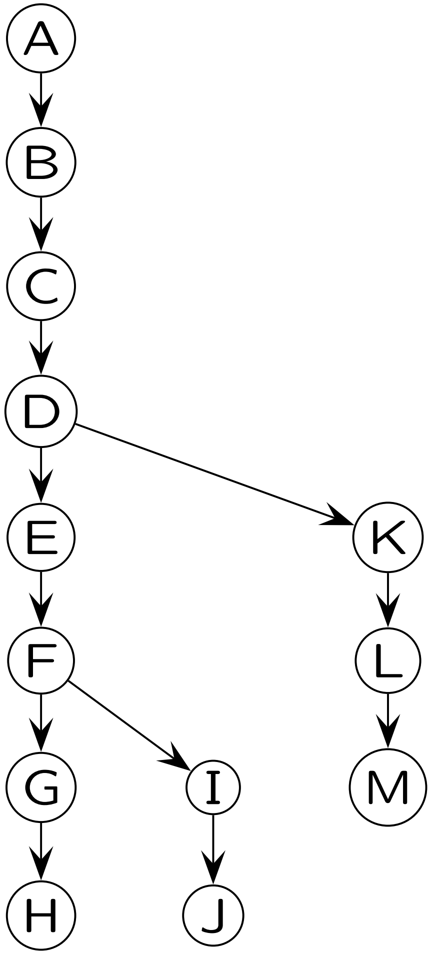
- Query search engine to find copies of petitions.
  - (~650 distinct copies found.)
  - (~20K distinct names.)
- compute propagation tree from these copies
  - (x → y if there is a copy where x immediately precedes y.)



UNIVERSITY OF CAMBRIDGE

Building a propagation tree





UNIVERSITY OF CAMBRIDGE

Is this really a tree?

- No. some responded twice (have 2 parents)
- Typographical changes are frequent

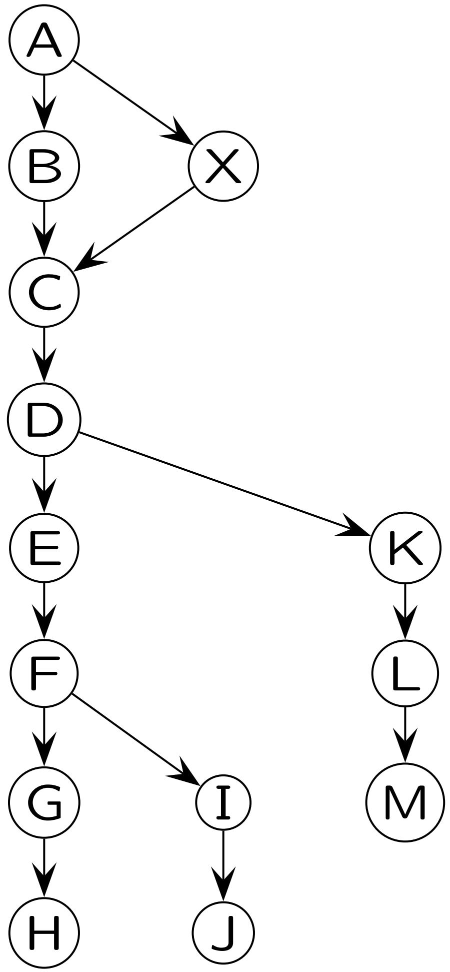
`John Smith Santa Monica Calif` `John Smith Santa Monica USA` `John Smith Santa Monica Calif USA`

- List rearrangements are common

UNIVERSITY OF CAMBRIDGE

Propagation tree

A	A	A	A
B	B	B	X
C	C	C	C
D	D	D	D
E	E	K	E
F	F	L	F
G	I	M	G
H	J		H



UNIVERSITY OF CAMBRIDGE

Solution

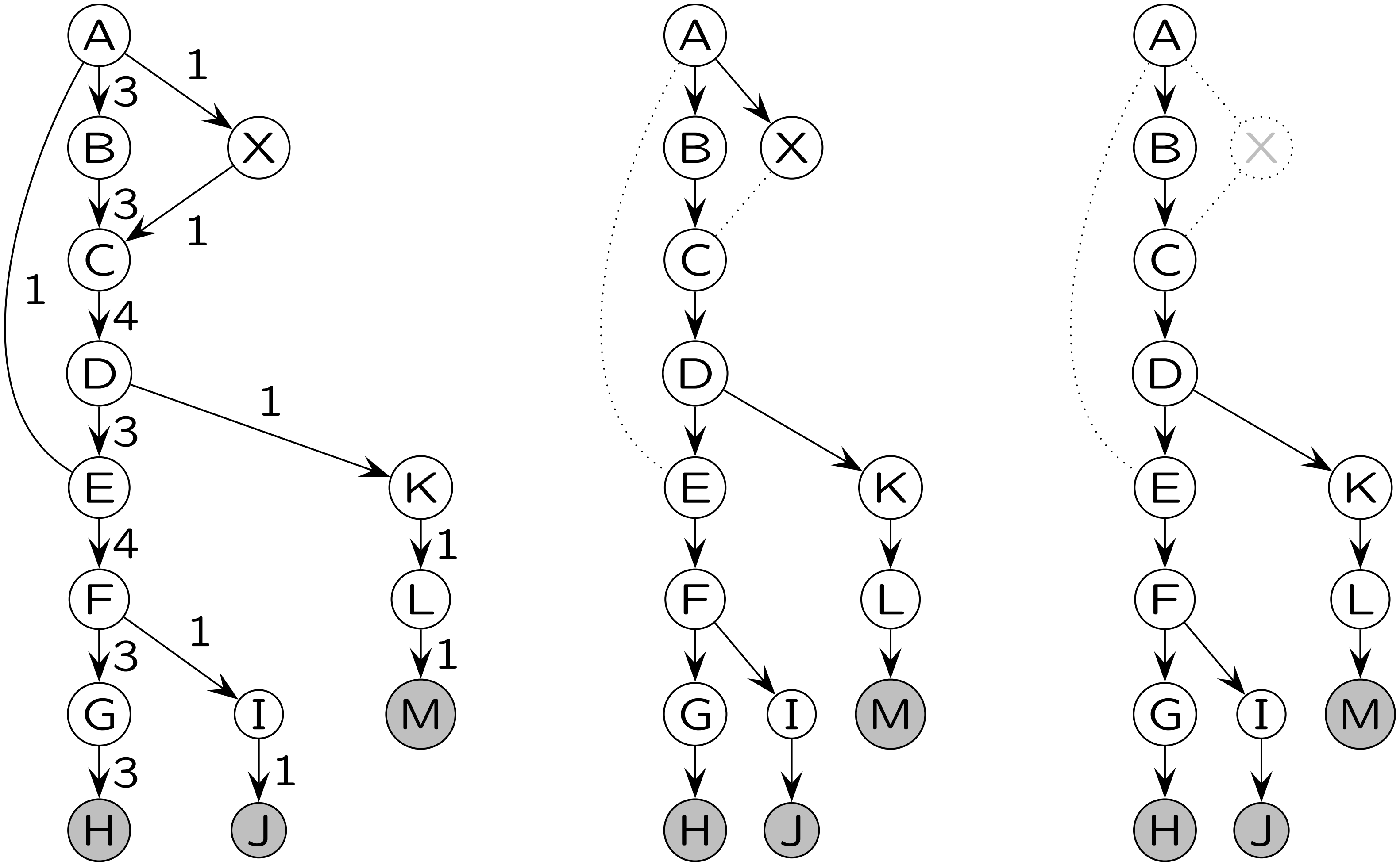
- Use the graph

$$\text{weight}(x \to y) := \#\ \text{copies s.t. } x \text{ immediately precedes } y.$$

- 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.

UNIVERSITY OF CAMBRIDGE

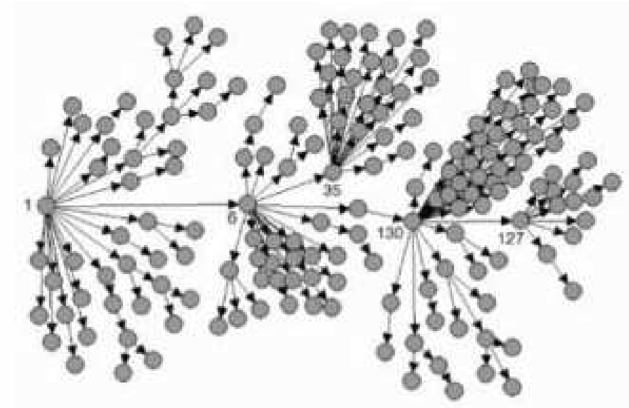
Graph to Tree

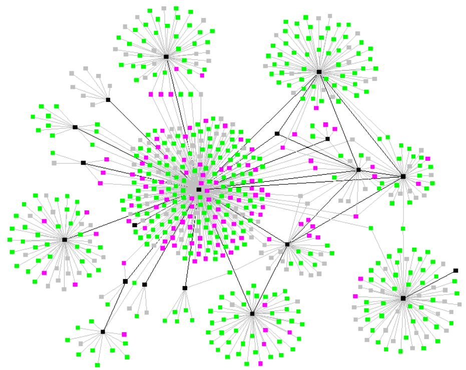


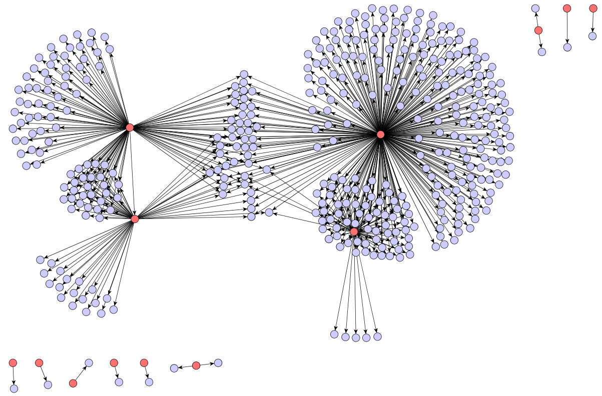
UNIVERSITY OF CAMBRIDGE

Nodes who made the list public are in gray

Expectations





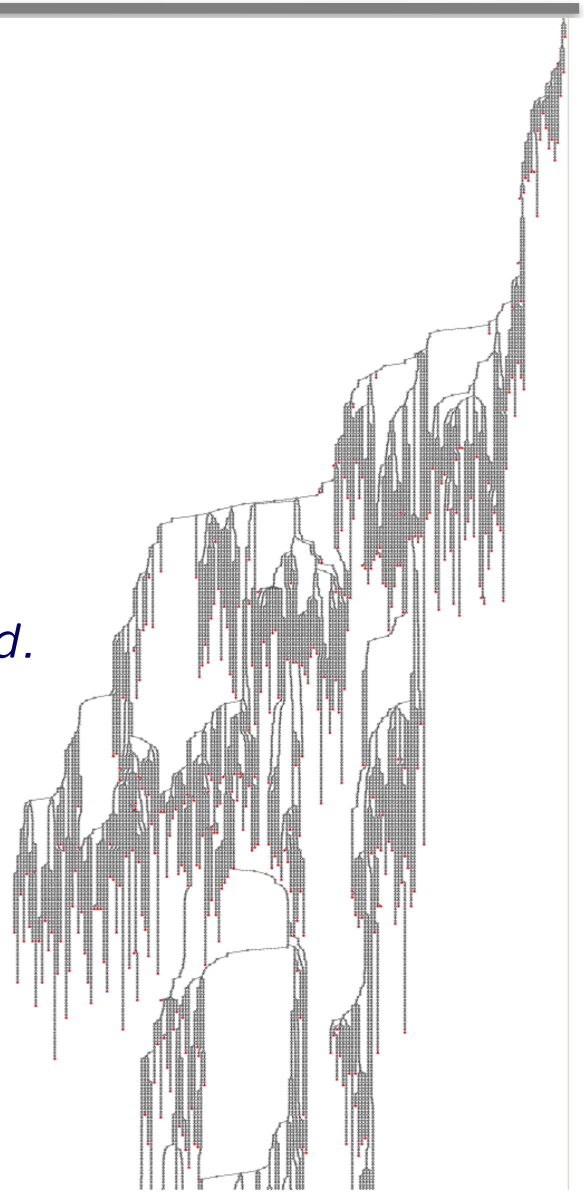


- The petition is flooding the social network.
- Small world $\Rightarrow$ the tree's depth will be small. High branching: people have many friends (10's or 100's).
- So the propagation tree should be shallow and wide.
- (unless it dies out quickly.)

UNIVERSITY OF CAMBRIDGE

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.



UNIVERSITY OF CAMBRIDGE

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%

UNIVERSITY OF CAMBRIDGE

Tried simulation on LiveJournal

- simulate on real social network (LiveJournal, 4.4M nodes). Randomly choose an initiator node (= root).
- each recipient discards with prob δ, forwards with prob $1 - \delta$. $\delta := 0.65$ [Dodds Muhamad Watts 2003]
- a non-discarding recipient posts his copy with prob π (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.

UNIVERSITY OF CAMBRIDGE

Epidemic Model

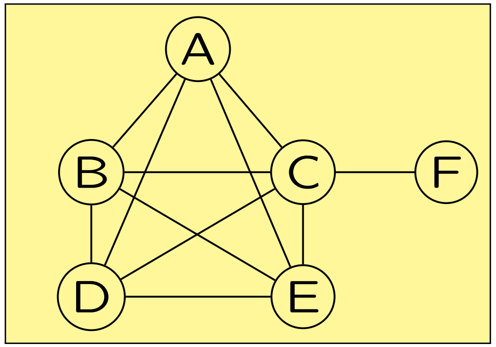
- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x discards with probability δ = 0.65; otherwise:
  - x appends x to
  - x forwards to all neighbours (who act at time t + 1)
  - x posts with probability π.
  - Iraq tree (18K nodes): depth 288, width 82, single-child 94%
  - Epidemic tree: depth 5, width 9625, single-child 19%

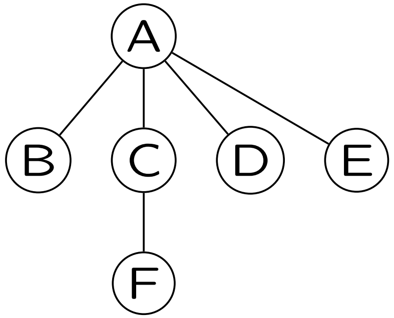
UNIVERSITY OF CAMBRIDGE

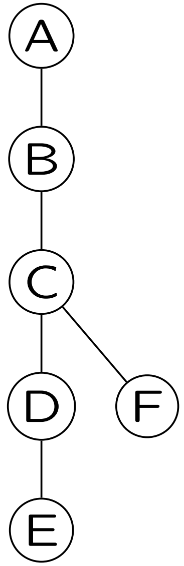
Why?

Social networks have lots of “cliquey” communities.

$\Rightarrow$ 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 $\Rightarrow$ C gets new copy from B.

UNIVERSITY OF CAMBRIDGE

Asynchronous Model

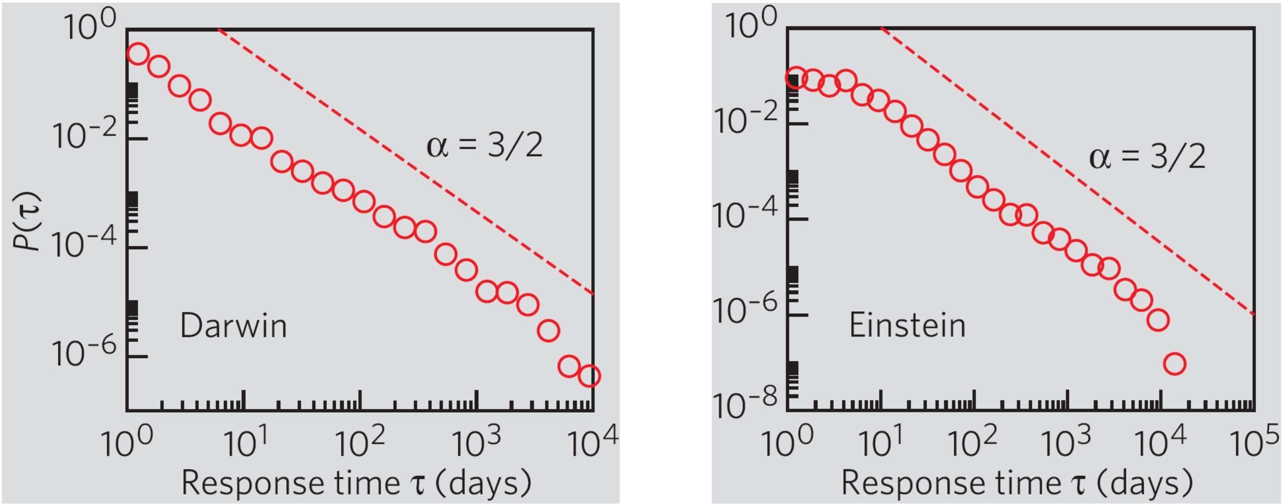
- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x chooses a delay τ, where Pr[τ] =???. At time t + τ:
  – x discards with probability δ = 0.65; otherwise:
  – x appends x to longest list x received (in [t, t + τ])
  – x forwards to all x’s neighbours.
  – x posts with probability π.

Delay Distribution

You receive a message at time t; you respond at time $t + \tau$.

What does τ look like?

Letters from Darwin and Einstein: [Oliveira Barabasi 2005]



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

(though the precise exponent actually doesn't matter much.)

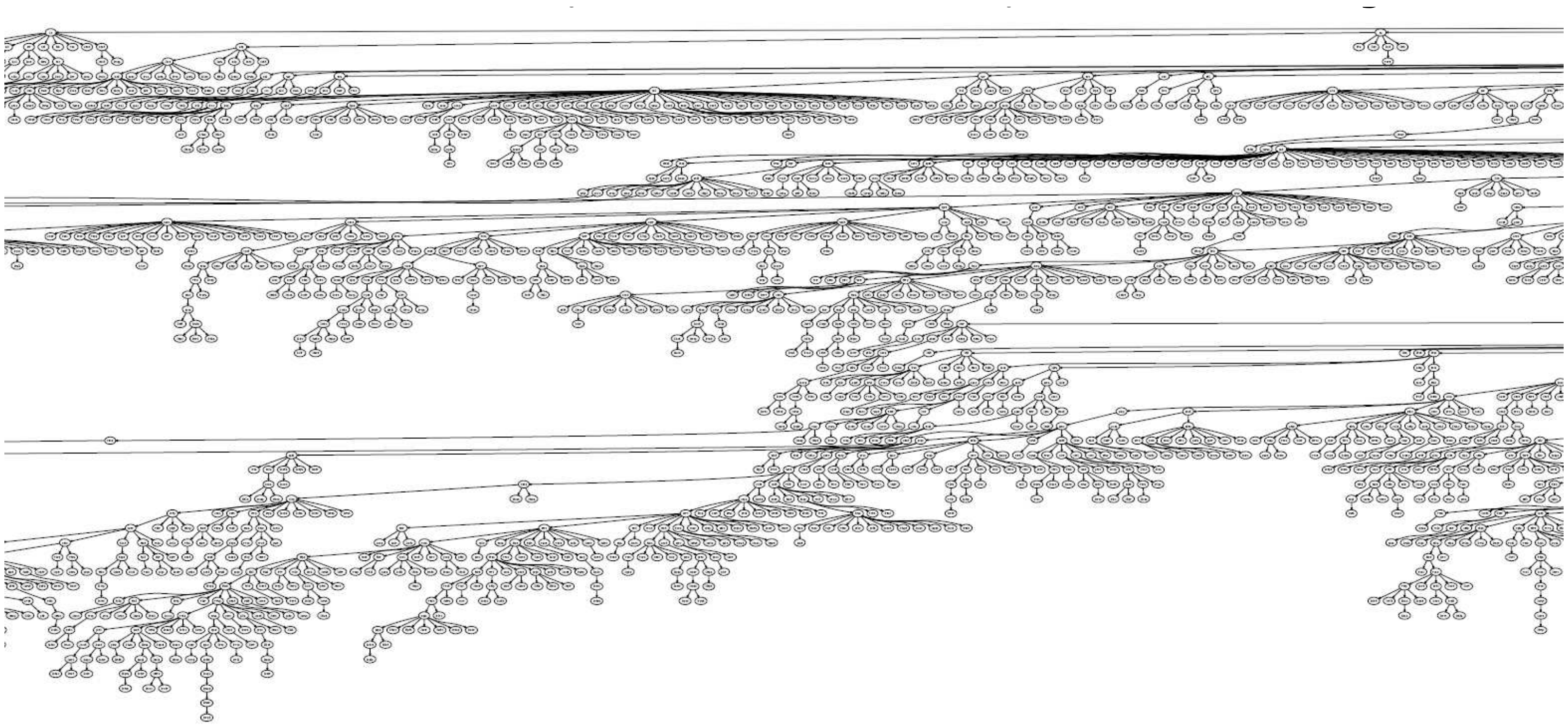
UNIVERSITY OF CAMBRIDGE

Epidemic Model

- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: **x chooses a delay τ, 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 π.

UNIVERSITY OF CAMBRIDGE

Epidemic Model



One More Ingredient

(18K nodes)	**depth**	**width**	**single-child %**
Iraq	288	82	94%
Epidemic	5	9625	19%
Asynchronous	42	505	55%

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.

UNIVERSITY OF CAMBRIDGE

The Asynchronous Model

- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x chooses a delay τ, 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 π.

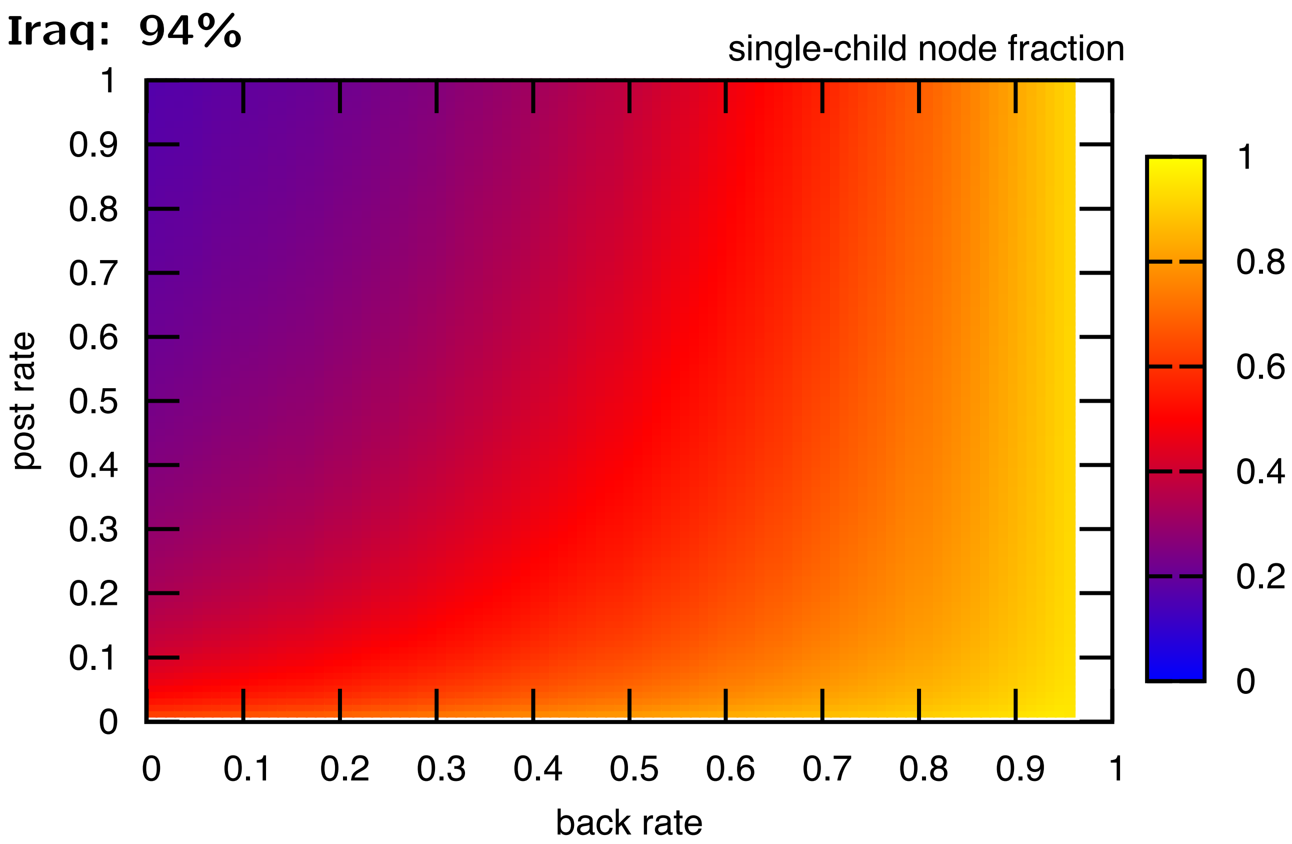
UNIVERSITY OF CAMBRIDGE

The Full Model

- randomly choose an initiator node (= root)
- for each x who first receives a list at time t: x chooses a delay τ, 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 β, x replies to all of x's corecipients; with prob $1 - \beta$, x forwards to all of x's neighbors.**
  - x posts with probability π.

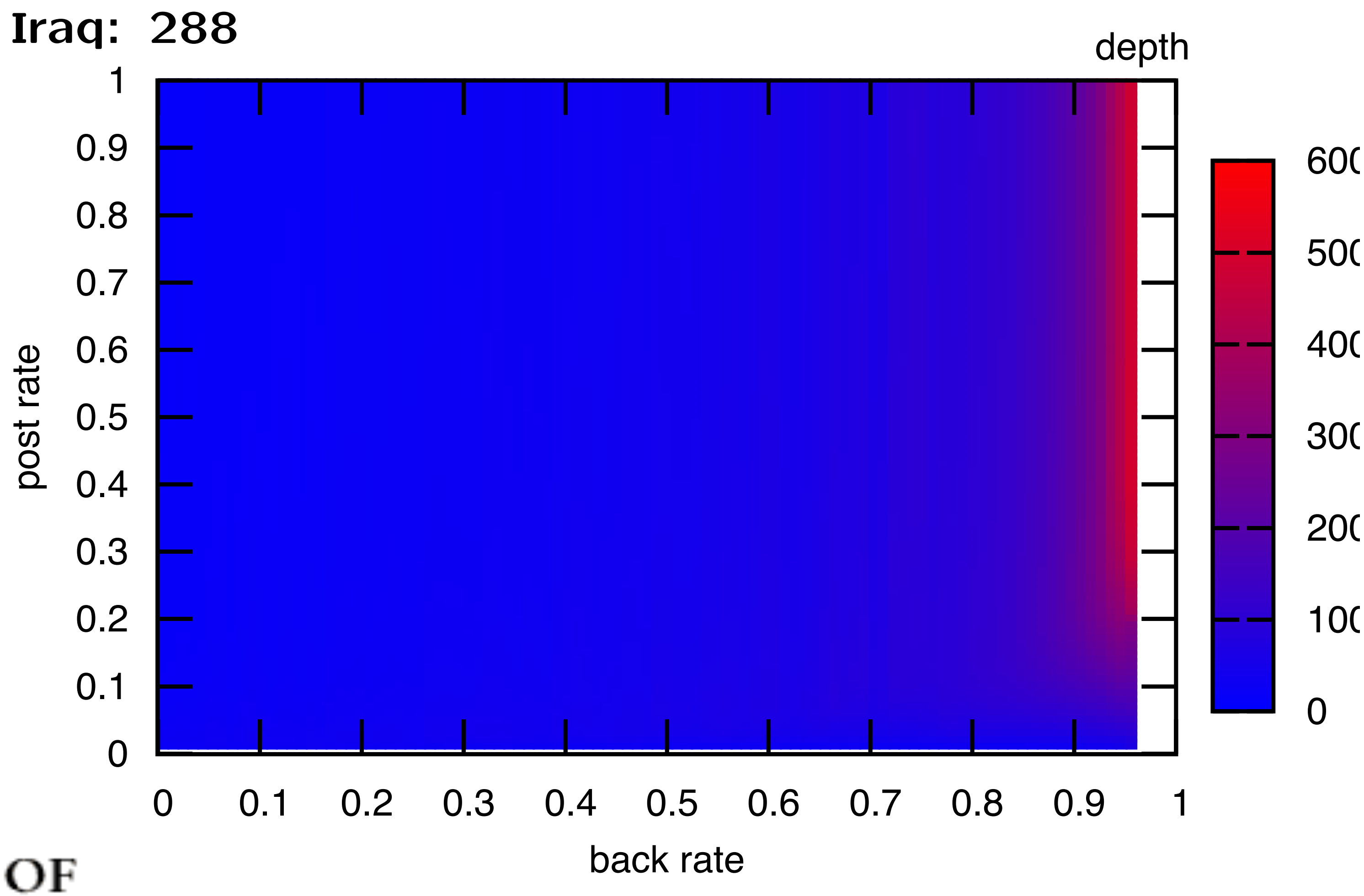
The asynchronous model = full model with $\beta=0$

Studying Single Child Proportion



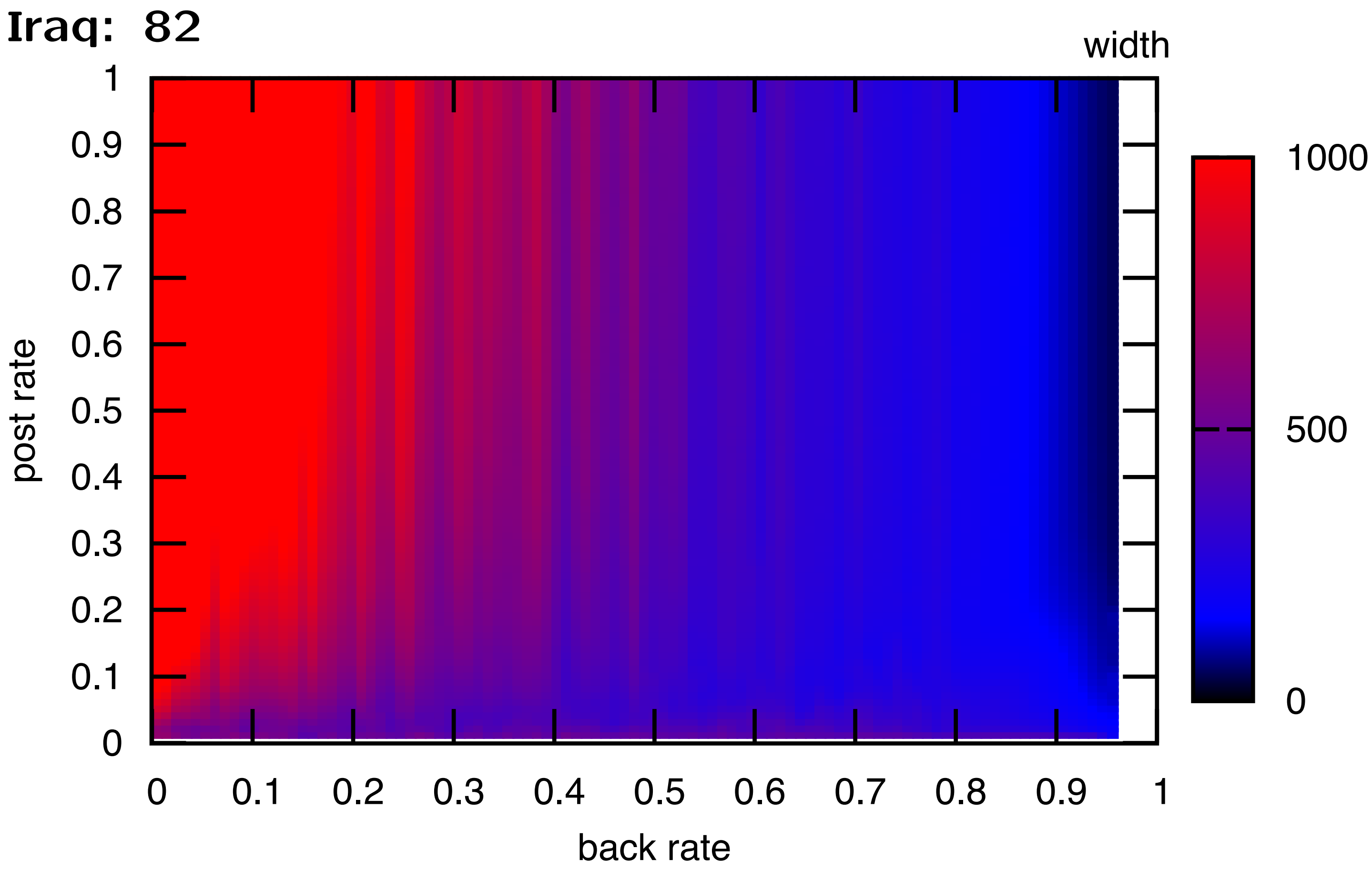
UNIVERSITY OF CAMBRIDGE

Tree Depth



UNIVERSITY OF CAMBRIDGE

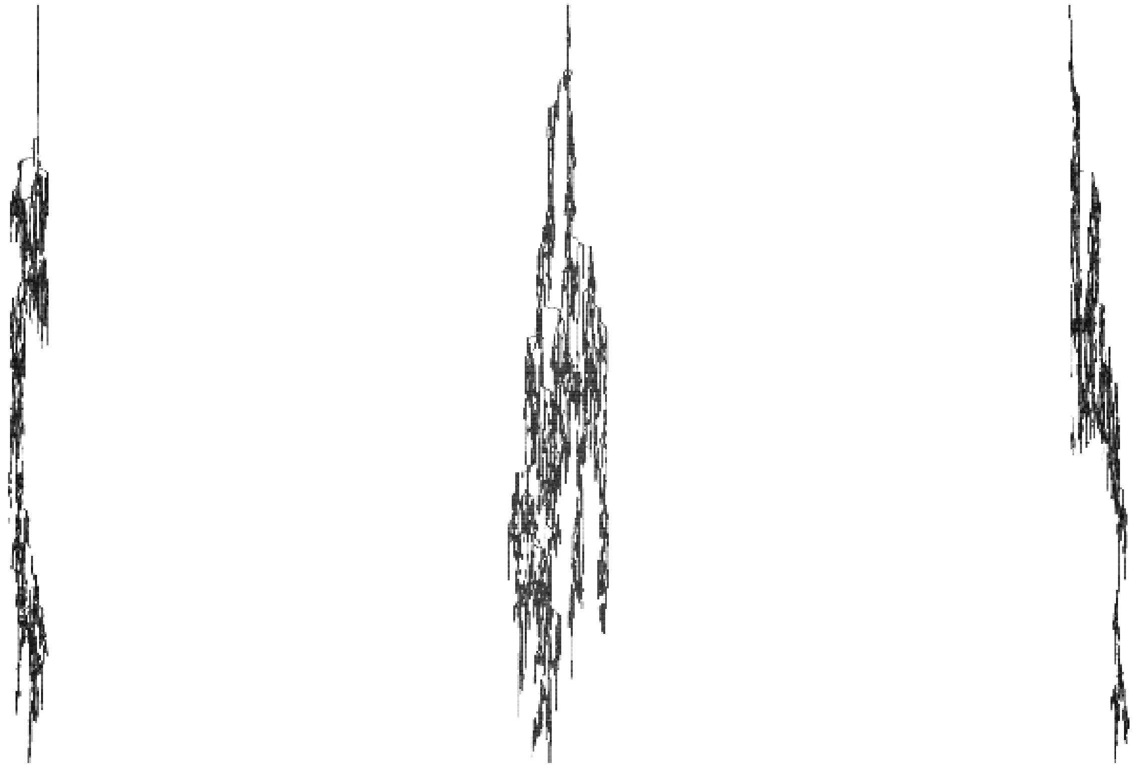
Tree Width



UNIVERSITY OF CAMBRIDGE

It matches!

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



UNIVERSITY OF CAMBRIDGE

Discussion

- A model with asynchronicity and group-reply was a good initial approximation
- More data needed to understand what’s happening

UNIVERSITY OF CAMBRIDGE

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.

UNIVERSITY OF CAMBRIDGE

References

- Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P. Gummadi. 2008. **Characterizing social cascades in flickr.** In *Proceedings of the first workshop on Online social networks* (WOSN '08). ACM, New York, NY, USA, 13-18.
- D. Liben-Nowell and J. Kleinberg. **Tracing information flow on a global scale using Internet chain-letter data.** *PNAS March 25, 2008 vol. 105 no. 12 pp. 4633-4638.*

UNIVERSITY OF CAMBRIDGE