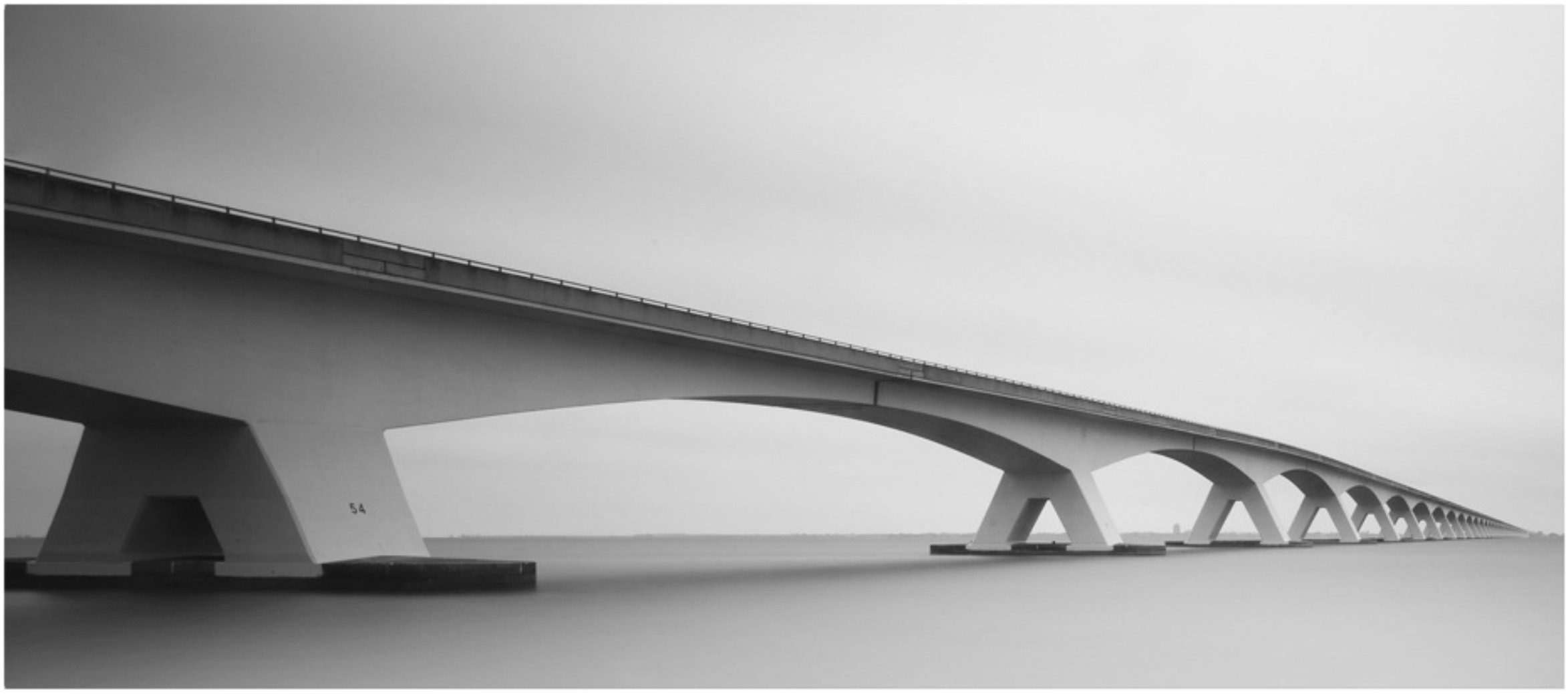


David Liben-Nowell and Jon Kleinberg

Tracing information flow on a global scale using internet chain-letter data

presented by Liu Tong
30/04/2014

Session Outline

- Data of internet chain-letters
- Structure of the dissemination tree
- Tree structure modelling
- Models based on asynchrony response time

How does information diffuse?

The basic models posit that information will diffuse from person to person in the style of an epidemic, expanding widely in a short number of steps according to "small-world" principles.

It has remained an open question whether the spreading of information truly proceeds with a rapid, epidemic-style fan-out or whether it follows a potentially more complex structure.

Chain-Letter

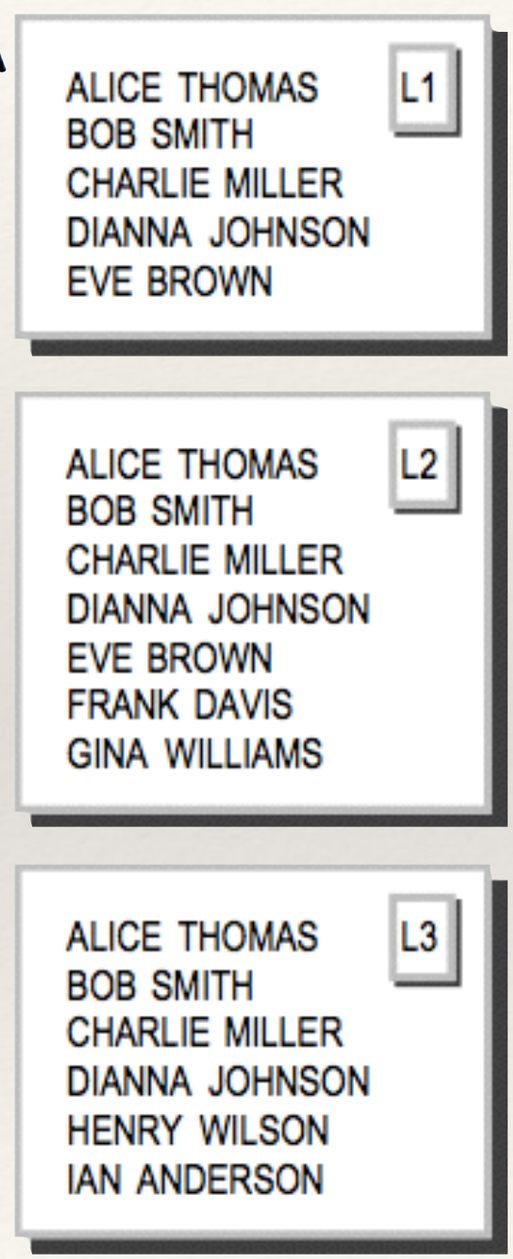
We observe the dissemination of petitions that circulated widely in chain-letter form on the Internet over the past several years. The petitions instruct each recipient to append his or her name to a copy of the letter and then forward it to friends. Each copy will thus contain a list of people, representing a particular sequence of forwardings of the message; and hence different copies will contain different but overlapping lists of people, reflecting the paths they followed to their respective current recipients.

year: 2002-2003
637 copies
20,000 distinct signatures

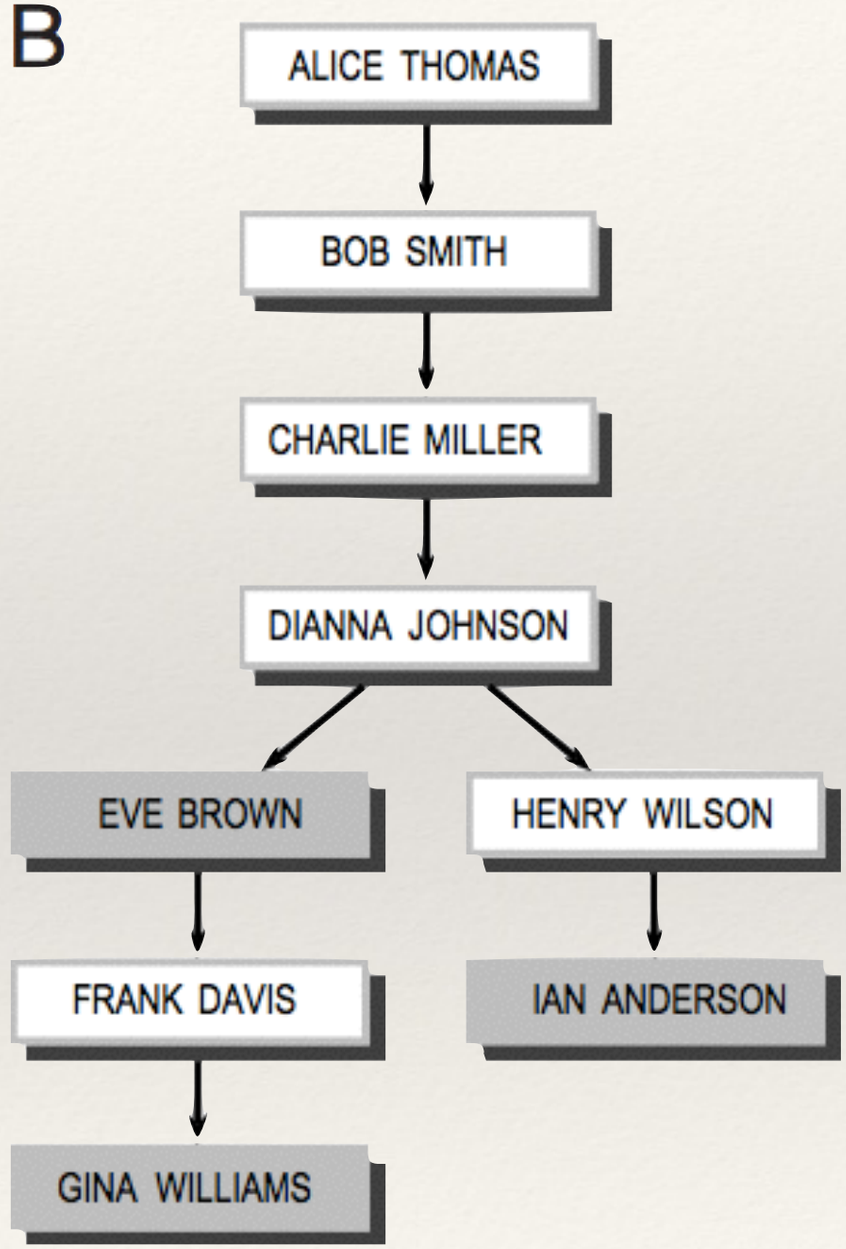
Other data from “funds petition for NBS, NPR”

Letter data

A



B

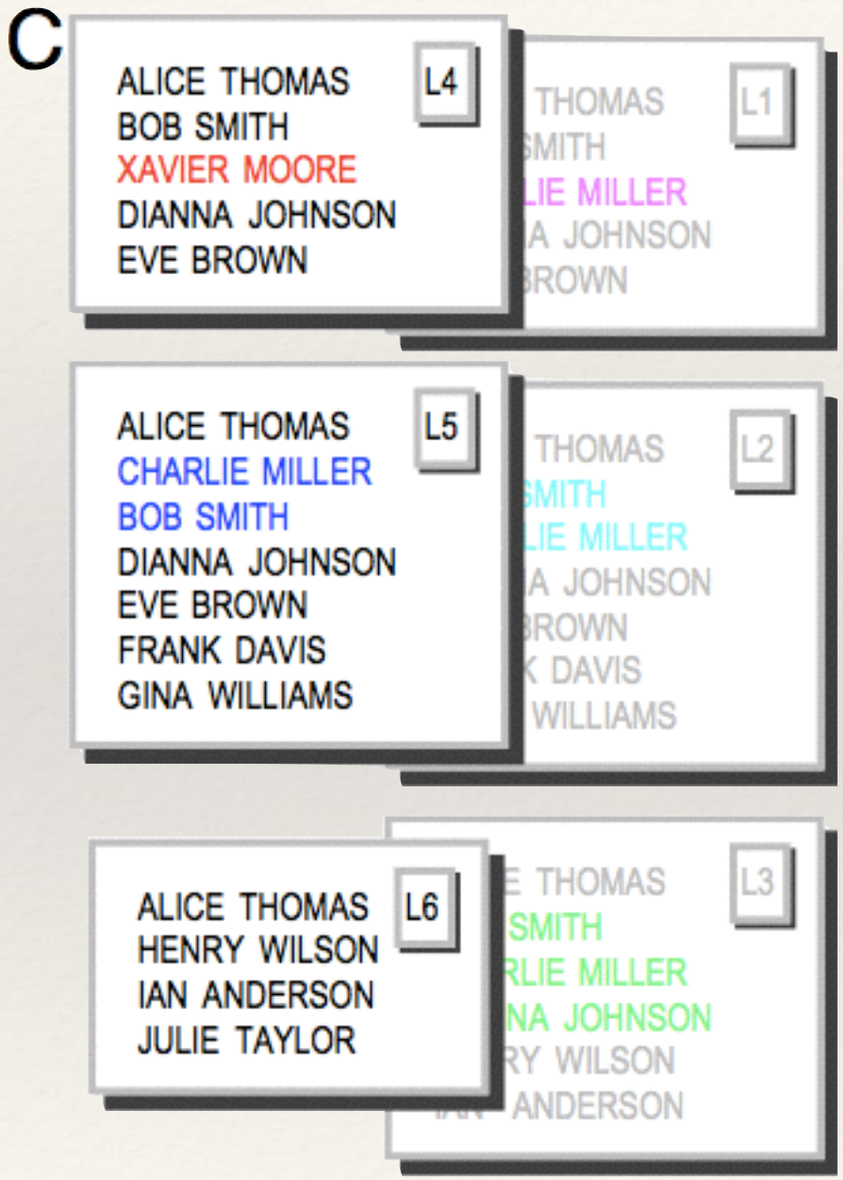


ERCW?

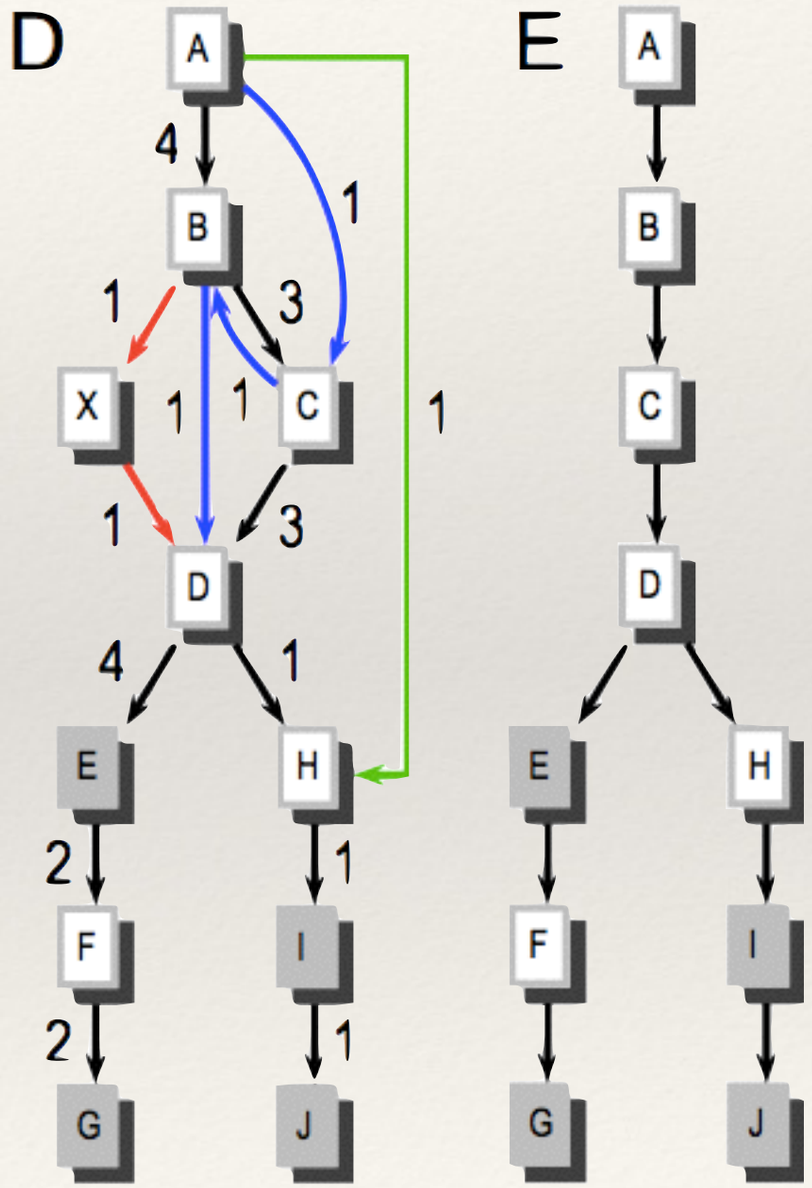
Data Cleaning

Heuristic based on sequence alignment to declare two names with a common list predecessor and very small typographical variation to be equivalent.

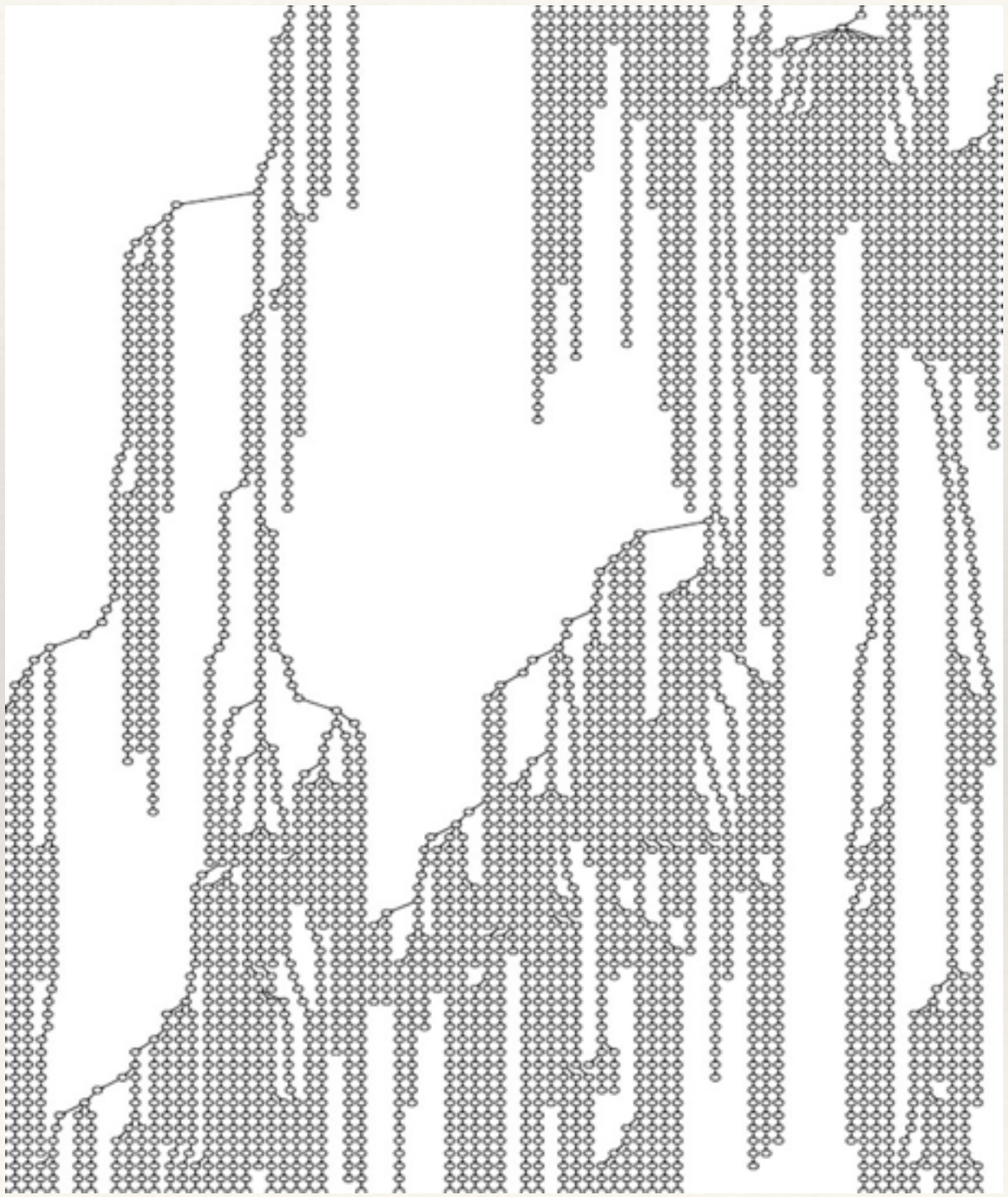
Using “maximum spanning tree” to construct the tree presentation of the direct graph



Direct Graph



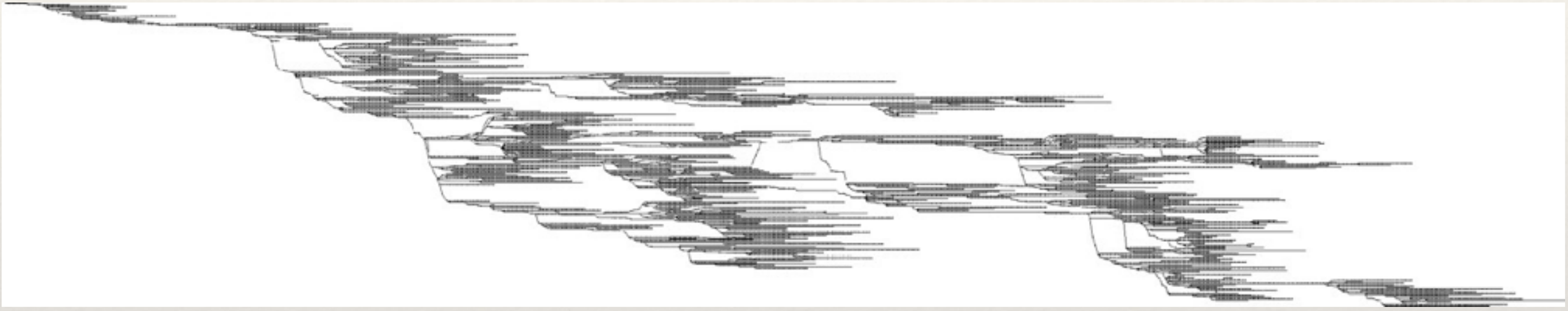
A portion of the tree



Metrics:

- Depth: Distance from the root
- Width: maximum size of a set of nodes that all possess the same depth
- Median node depth
- Fraction of nodes with exactly one child.

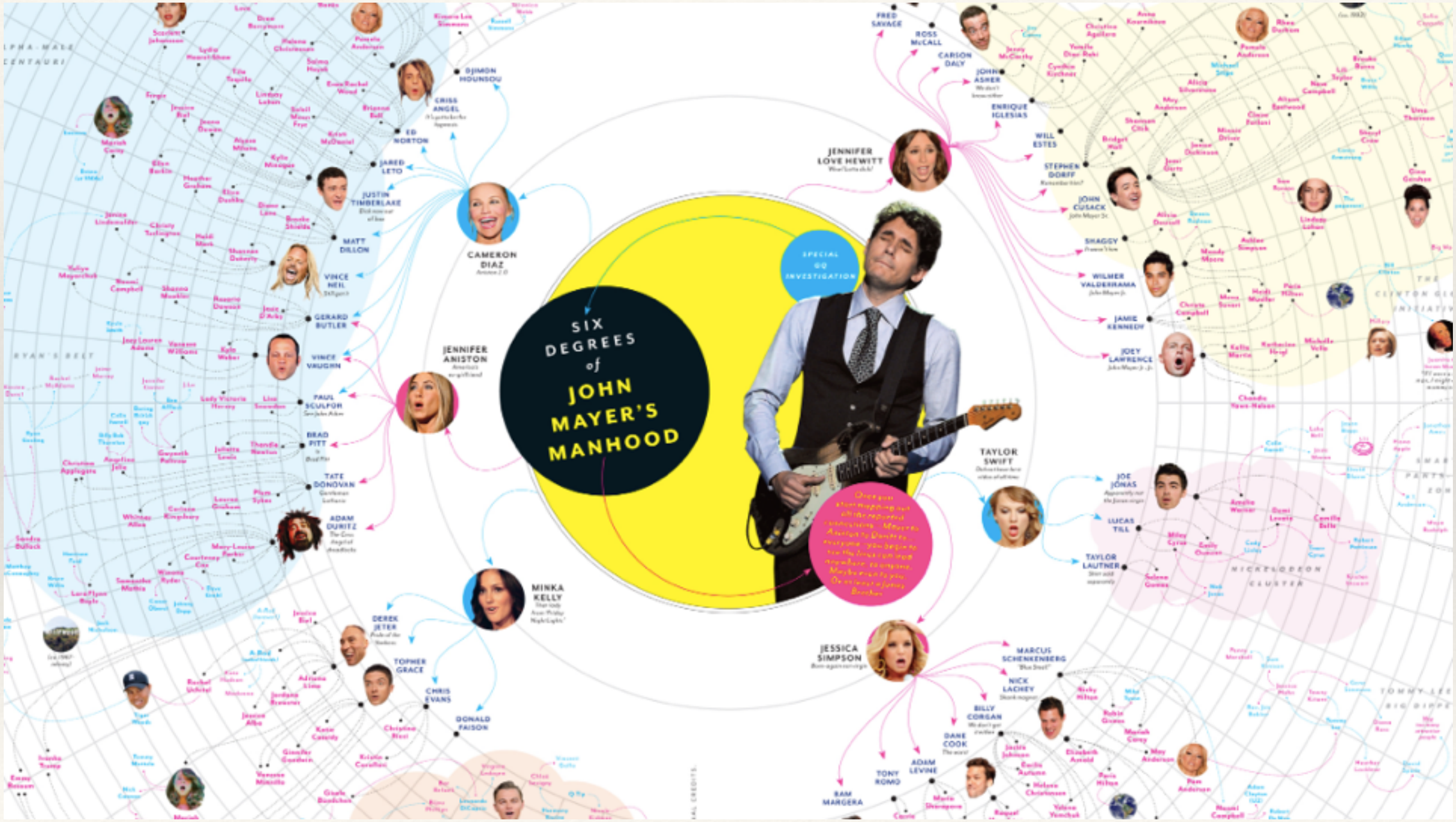
Tree's global view



This tree has 18,119 nodes, of which,

- Width of the tree: 82.
- Median node depth: 288
- Nodes have 1 child: 17,079 (94.26%)

Small World Property?



Simulation Data Source

- Live Journal
- DBLP
- Wikipedia

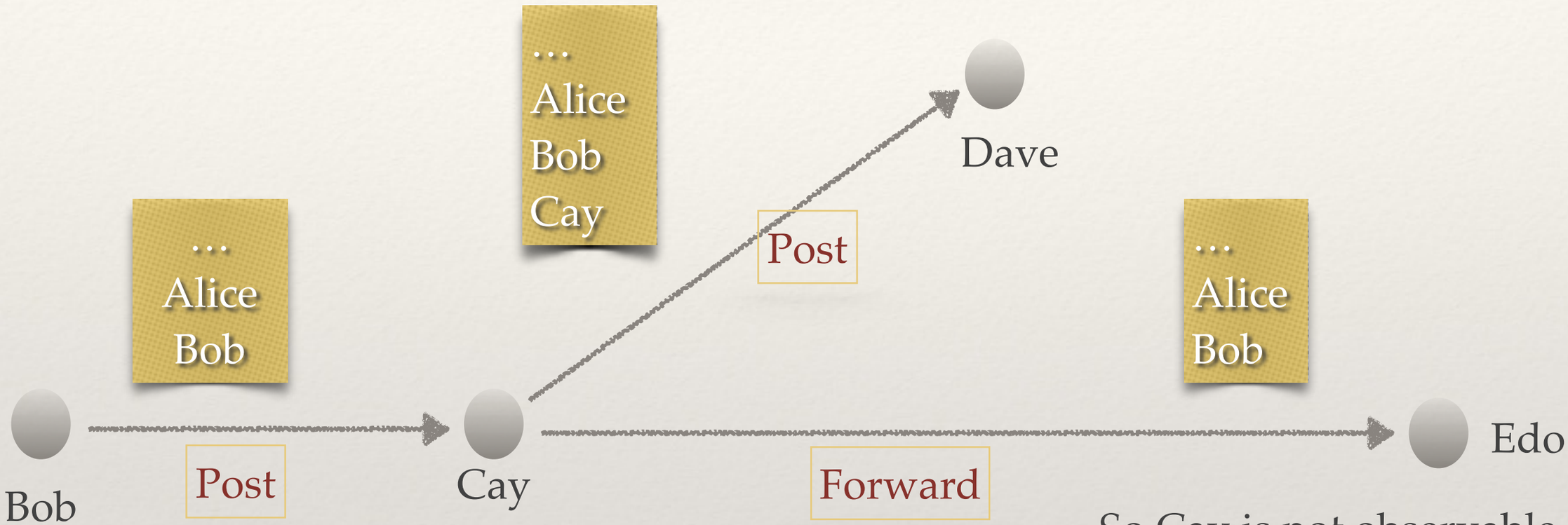
Relations of 4,400,000 individuals gathered from LJ

WIKIPEDIA Editors

Coauthor
400,000

dblp
computer science bibliography

Simulation One



Two principles:

Many recipients may choose not to forward the letter at all

only a few recipients will choose to post the letter publicly.

Discard rate d=0.65

Posting rate π=0.10

Simulation One

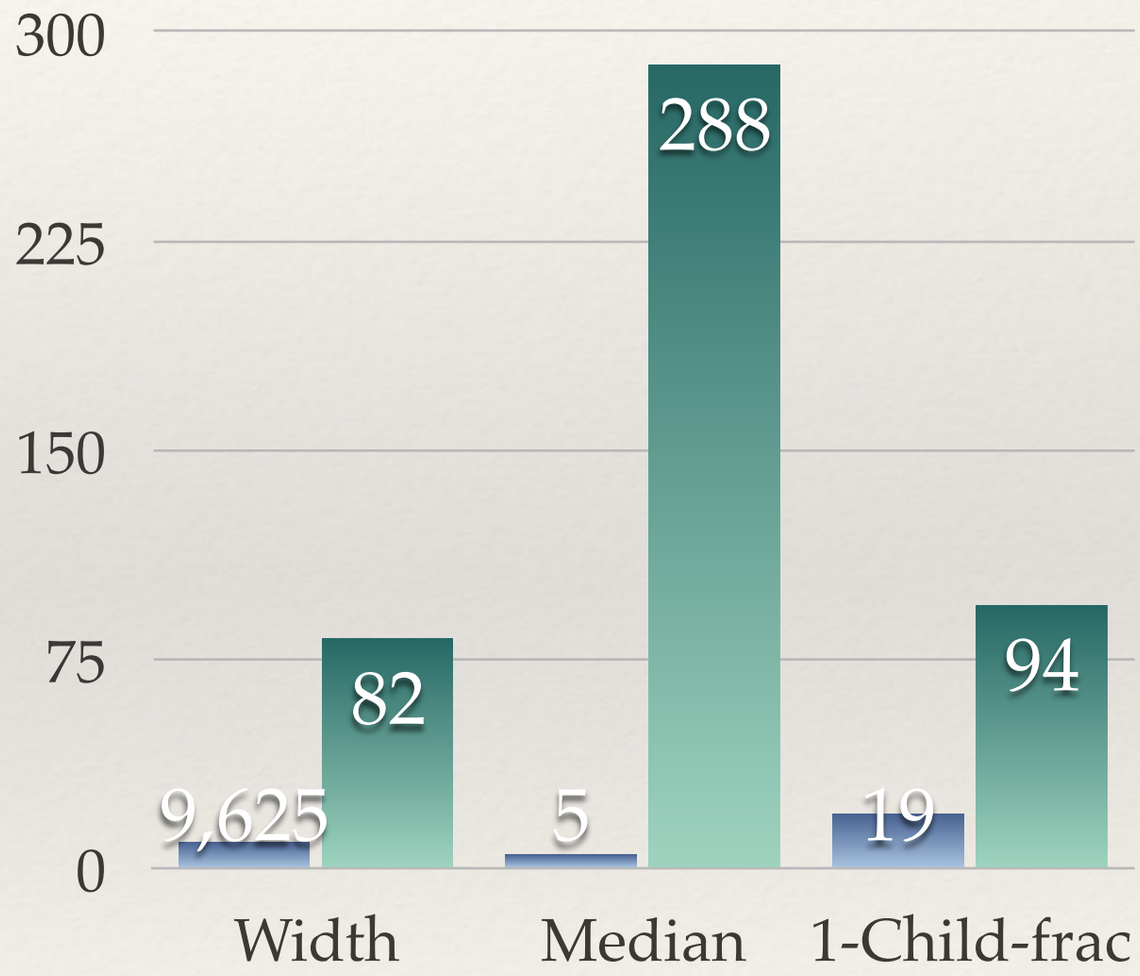
Observable portion of the tree has:

Width: 9,625

Median depth: 5.0

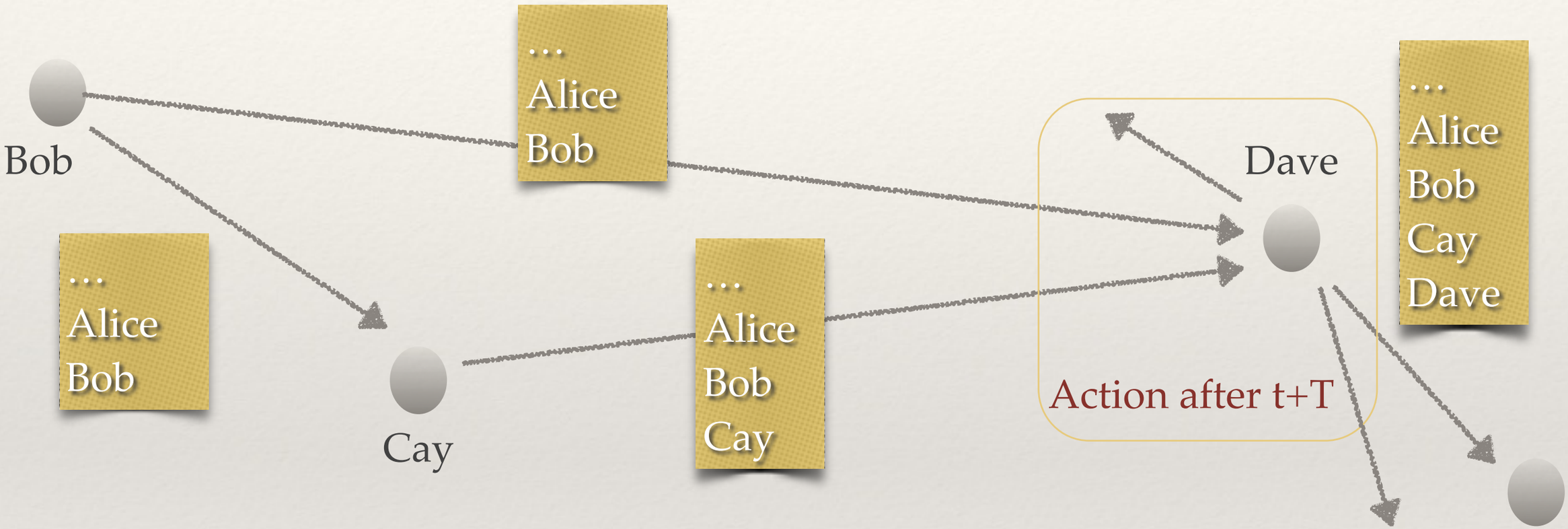
Single-child fraction: 19.04% (averaged over 10 independent runs)

Forwarding to a random subset of 4 or 5 of his or her neighbours, then the width remains in the thousands, the median depth remains <50, and the single-child fraction remains <70%.



Simulation Two

based on Asynchronous Response Time

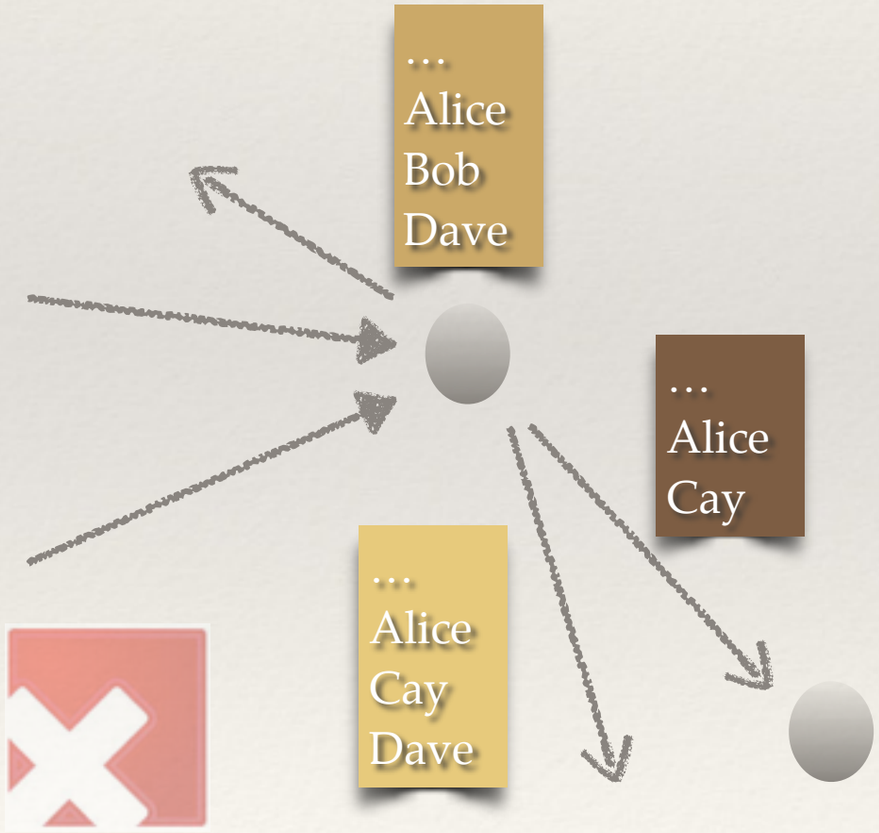


when a given node *w* in the network first receives a copy of the letter, at time *t*, it first decides whether to participate in the process at all, choosing to do so with probability 1 -d. If *w* chooses to participate, it then chooses a random waiting time T. Between times *t* and *t* +*T*, node *w* may receive multiple copies of the letter (including the initial one it received at time *t*). At time *t* +*T*, node *w* selects the one with the longest list of names, forwards it to ALL its neighbours, and publicly posts this copy with probability π.

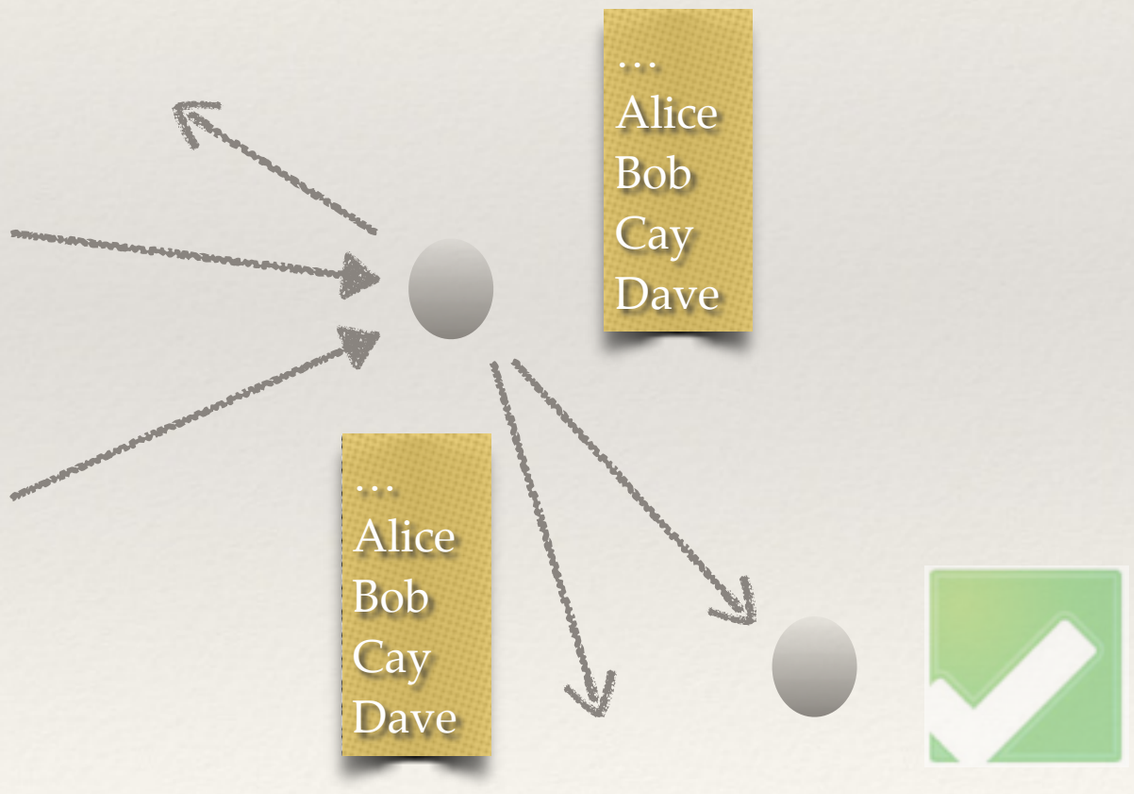
Simulation Two

This asynchronous pattern of response has a “serialising” effect in networks with large clustering coefficient!

If the neighbours of a forwarding node are mutually connected, then they will forward the letter to each other as they act on it in order, producing a single long list with all of their names rather than many distinct shorter lists. So it begins to produce the “Deeper Run”.



Simulation One



Simulation Two

Simulation Three

group reply with rate ß

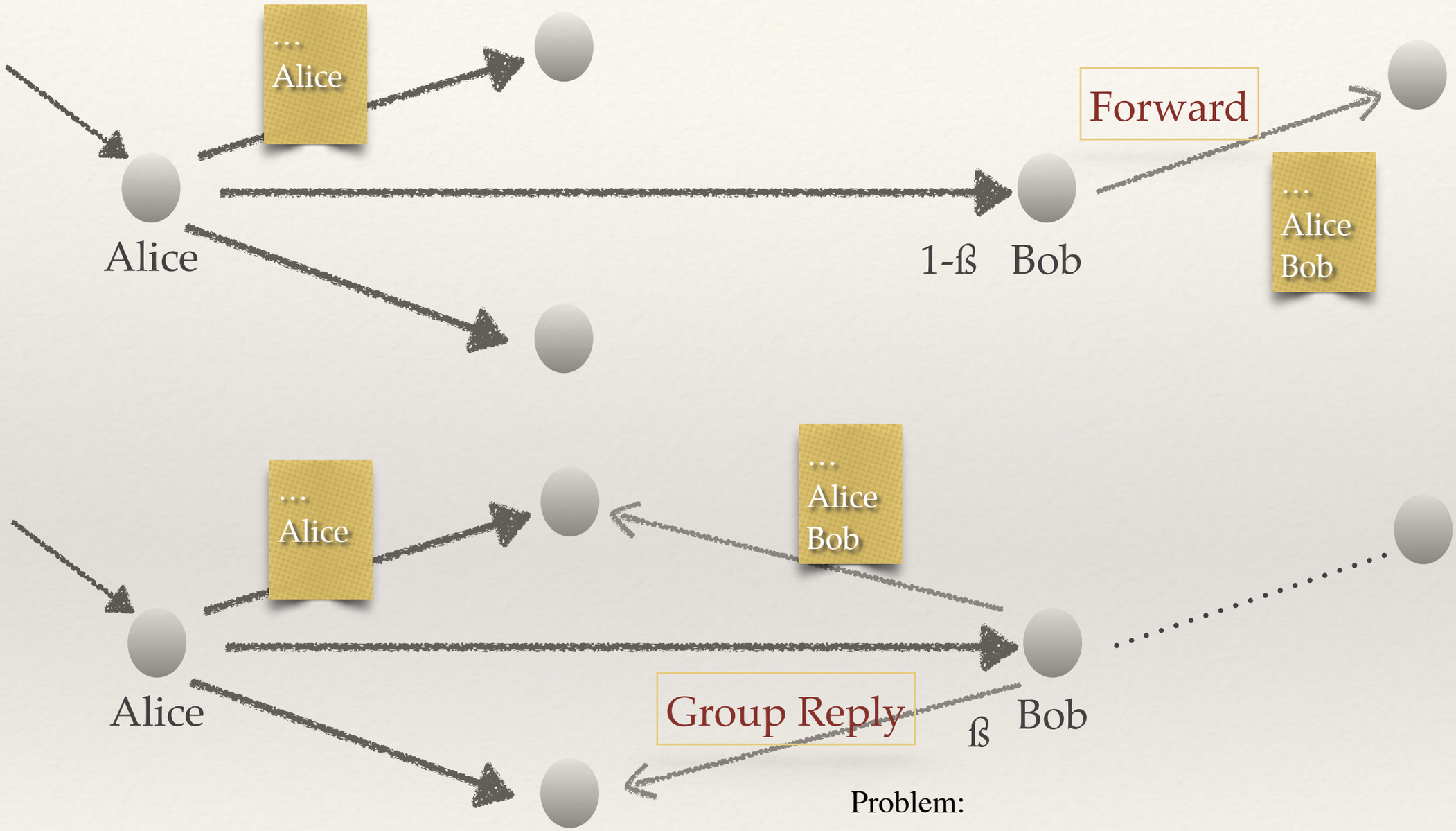
Asynchronous Response Time is a step toward trees with correct structure, but it is not enough by itself

The second extension is based on the fact that recipients actually have two natural ways of reacting to the message other than discarding it:

1) They can forward it to their neighbours in the network, as before, or
2) They can group-reply to the set of corecipients on the e-mail message they receive, in the case, these corecipients each receive a copy of the letter with the recipient's name appended

Simulation Three

group reply with rate ß



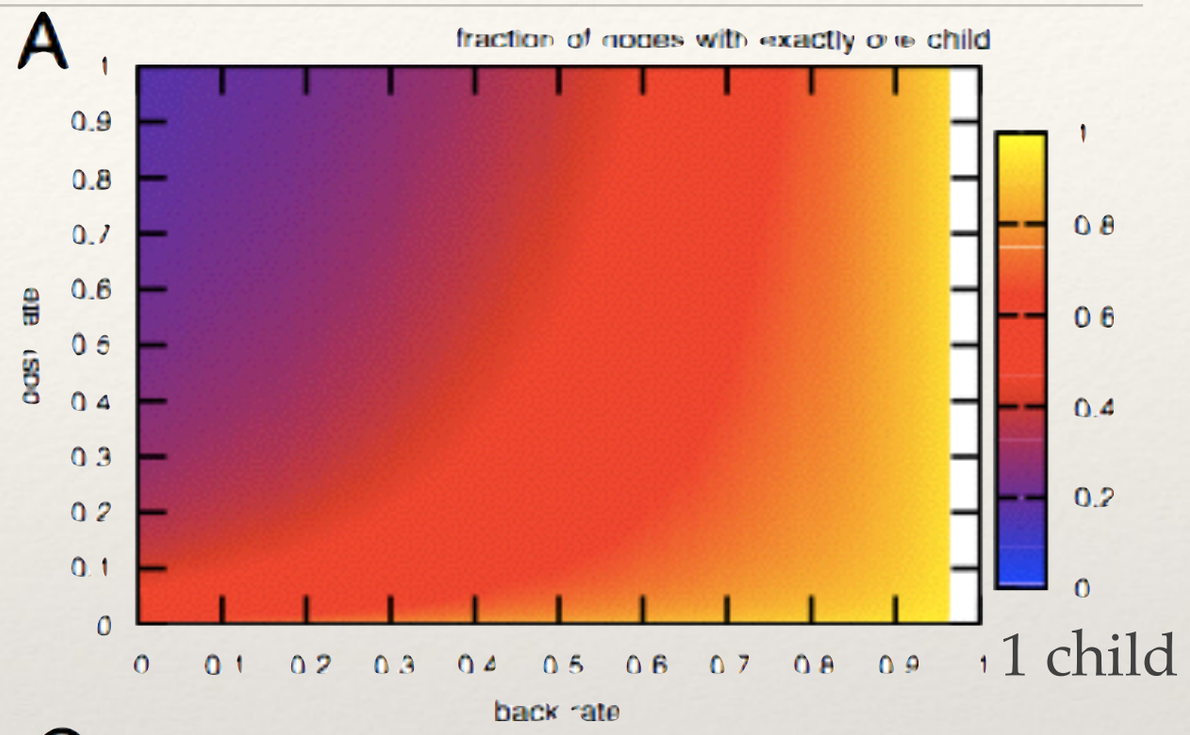
Problem:

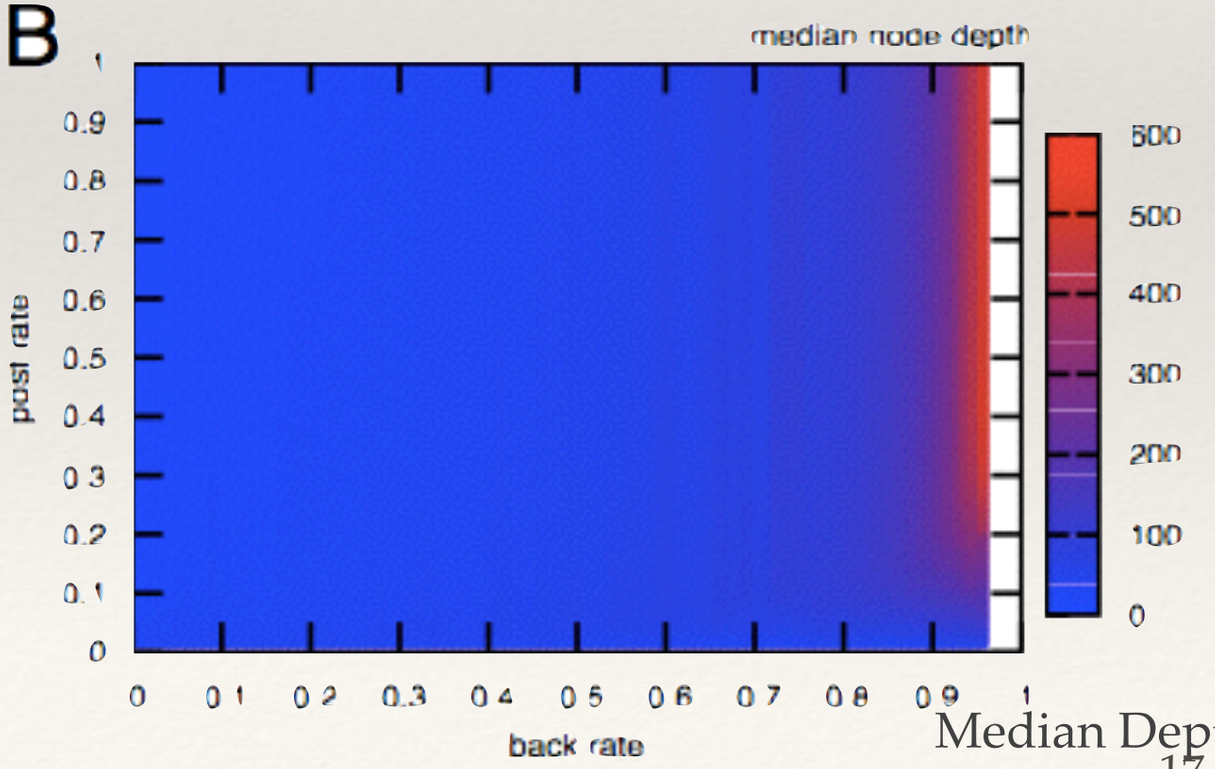
With a high back-rate, the letter is thus less likely to ever reach a large set of nodes. Thus, it becomes natural to study trade-offs in the tree structure.

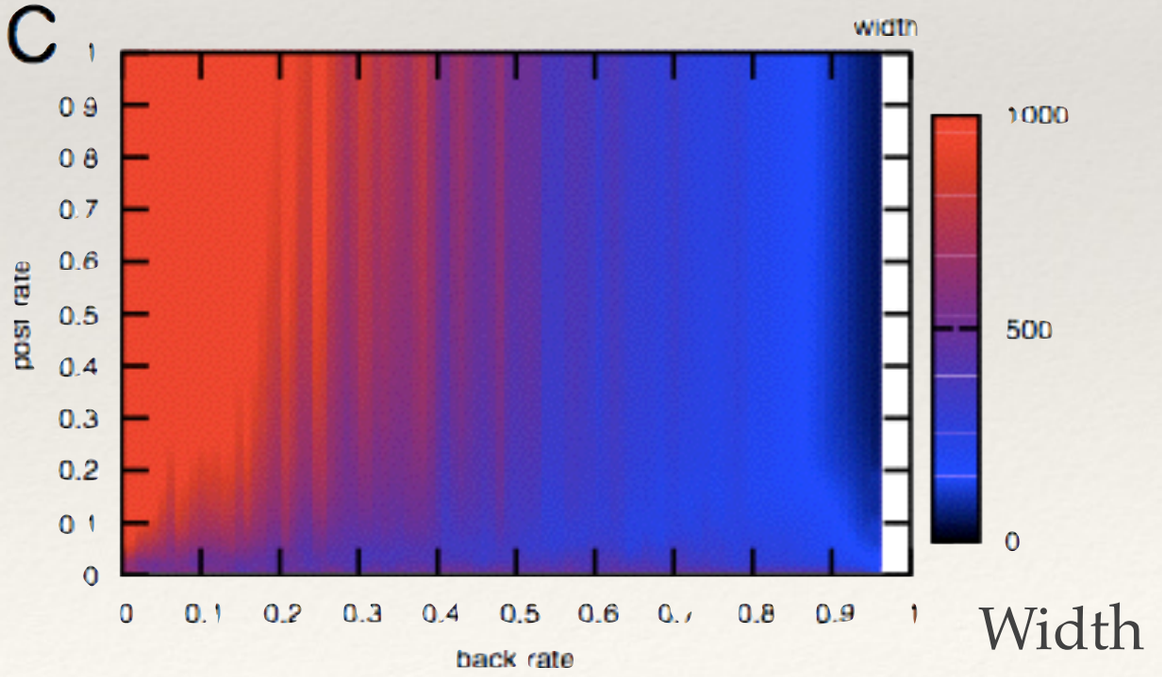
Trade-off of parameters

In Iraq chain-letter we had,

- Width of the tree: 82.
- Median node depth: 288
- Nodes have 1 child: 17,079 (94.26%)







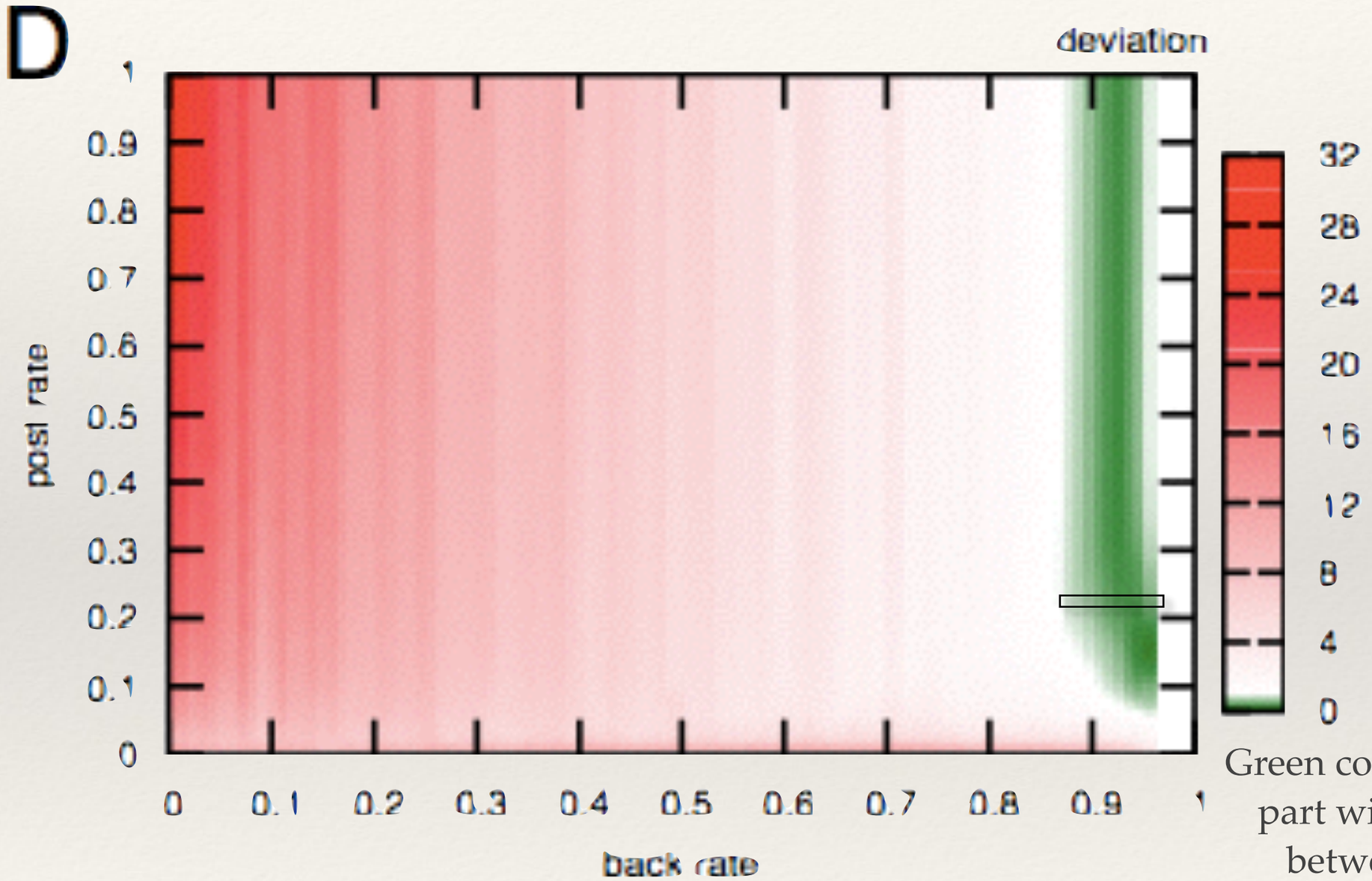
Median Deviation

The maximum over the three metrics of the ration

$$Deviation = \frac{|x-y|}{\min(x,y)}$$

Where **X** is the value of the metric on the **simulated tree** and **Y** is the value of the metric on the **real chain letter**

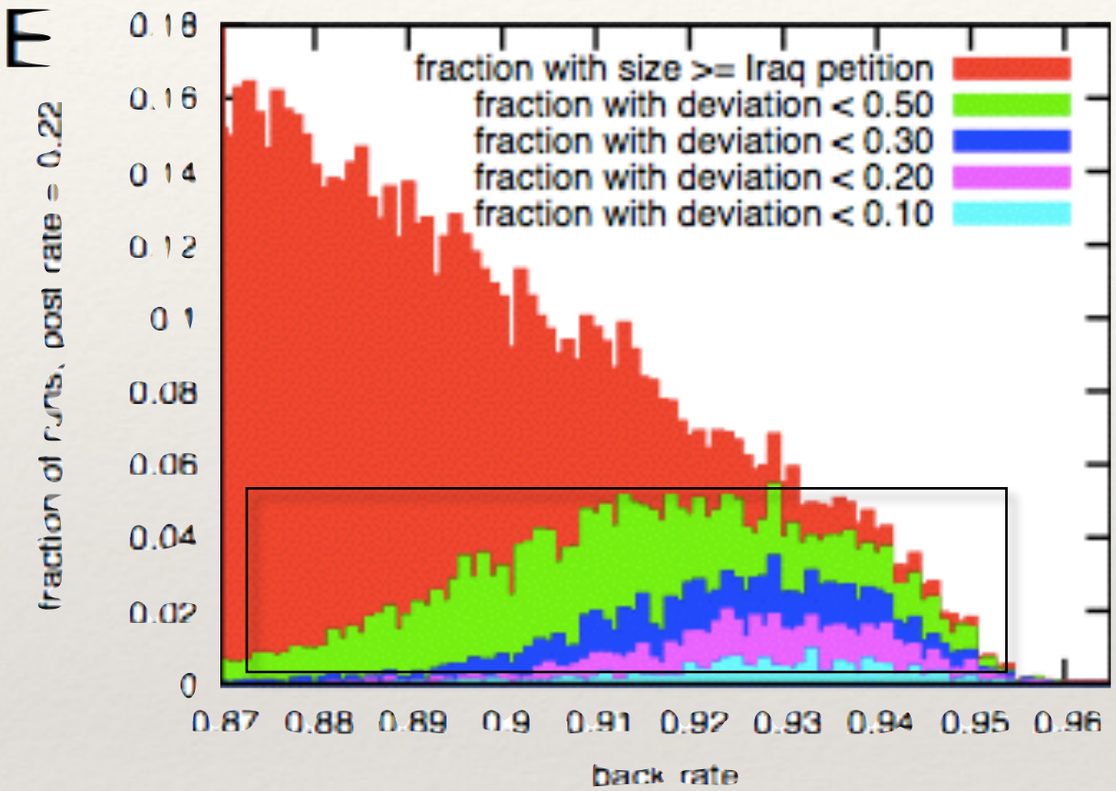
Trade-off of parameters



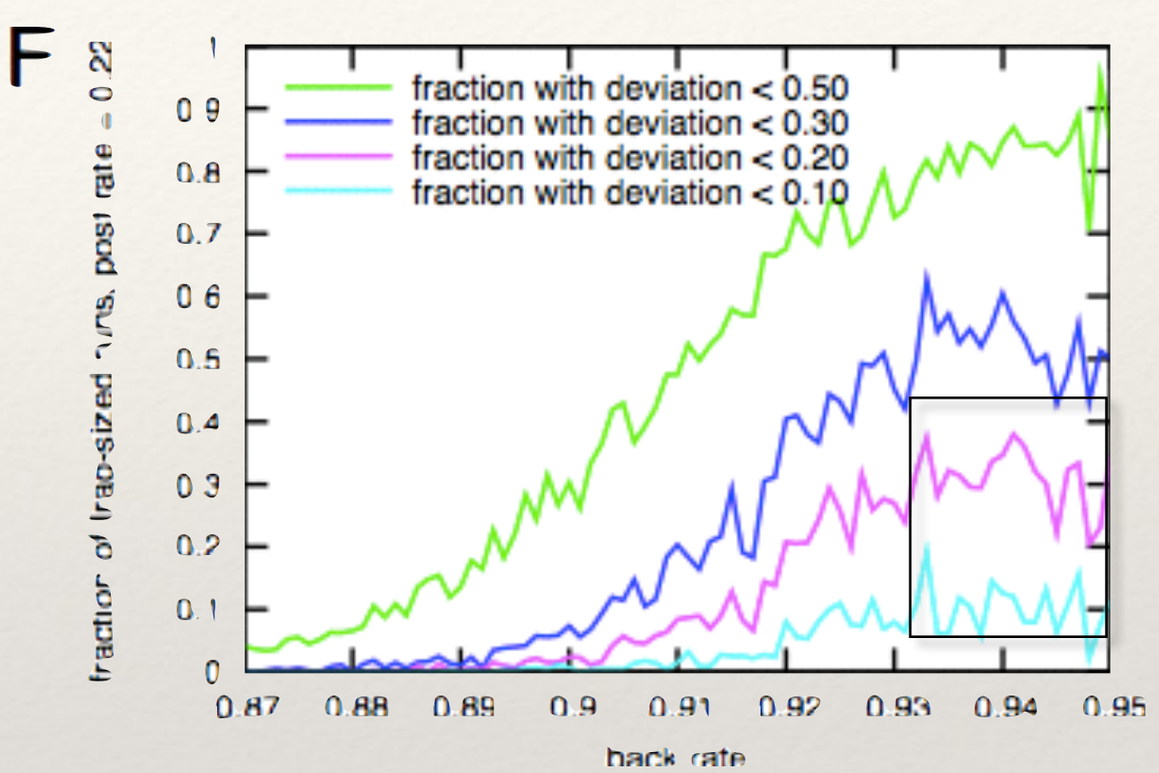
Deviation as a function of ß-back rate and post rate

Green colour shows the part with Deviation between 0 and 1

Simulation Three



Fraction of runs with post rate=0.22



Fraction of Iraq-sized runs with post rate=0.22

For high back-rates around 0.95, combined with low post-rates around 0.22, we obtain trees that approximately match the propagation tree of the real chain letter in all three metrics.

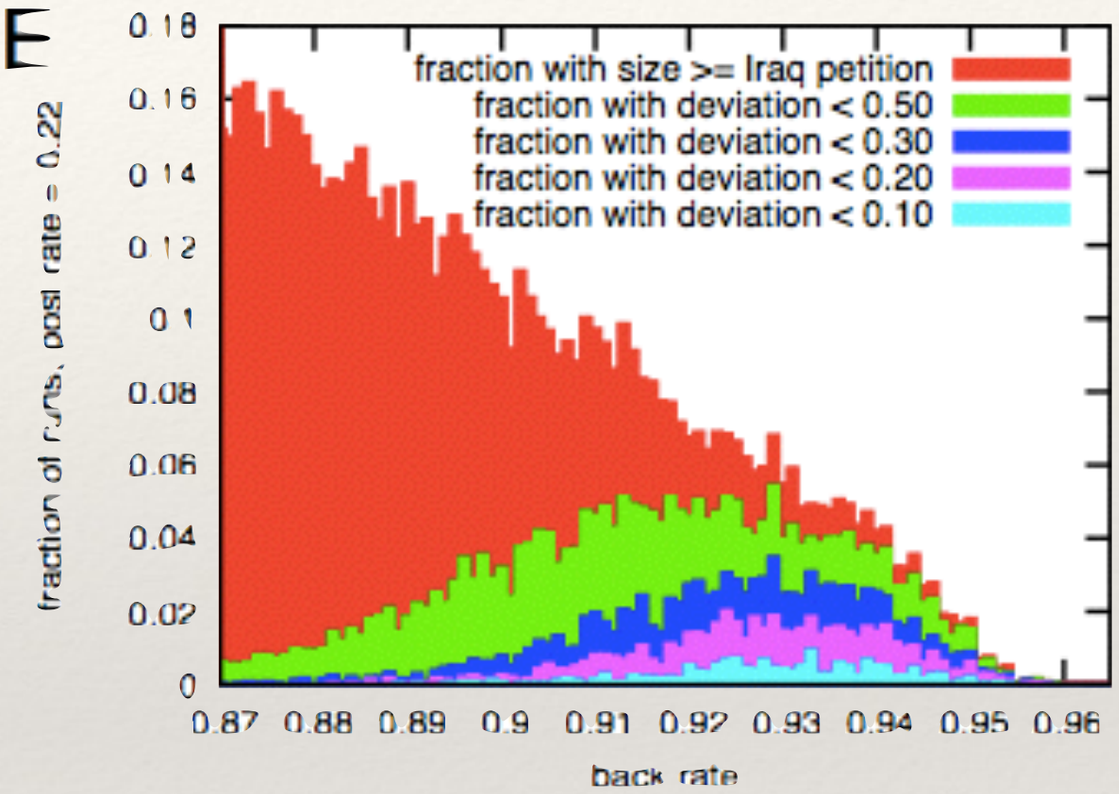
Observations

We involved 2 ingredients for our patterns to mach the observed chain-letter tree:

1) asynchronous response times

2) back-rate parameter(msg moves laterally)

both have effect of producing long narrow chain, which is the real property real dissemination tree structure.



for the parameters at which the closest approximations to the real tree are obtained, an extremely small fraction of the simulation runs produce trees as large as the real chain-letter tree before dying out. In other words, the structure of the real tree corresponds to a portion of the parameter space in which large trees are rare events—as they are in real life as well.

Conclusions

The resulting analysis has exposed several themes.

1) Accurately reconstructing the paths followed by the information is a **computational challenge** in itself, given the extensive ways in which the data are **mutated** as they spread.

3) The spreading patterns of the real chain letters are strongly at odds with the predictions of simpler theoretical models, which posit processes that reach many more people in radically fewer steps.

3) Simple **probabilistic** models incorporating the **speed** with which individuals respond to information can produce synthetic spreading patterns that closely resemble the ones we observe in real life.

Conclusions

The fact that the observed diffusion occurs along trees that are so deep and narrow suggests that the paths traversed by information through social networks can be more **complex** than might have been supposed, with the large number of steps giving the diffusion a certain **fragility** and presenting greater opportunities for the information to be **altered** or **lost** as it spreads.

Thank you all for your attention!

–Liu Tong