Patterns of Information Diffusion

David Liben-Nowell

dlibenno@carleton.edu

Carleton College, Department of Computer Science

Joint work with

Jon Kleinberg and Flavio Chierichetti

Central question:

How do ideas spread?

Ideas, disease, innovation, jokes spread continually via the global social network.

to [redacted] show details 7/2/05 Reply

this is actually kind of cool . . .

[redacted]
mobile [redacted]
fax [redacted]
email [redacted]
web [redacted]

-----Original Message-----
From: [redacted]
Sent: Friday, July 01, 2005 11:22 PM
To: [redacted]
Subject: Fw: (no subject)

----- Original Message -----
From: [redacted]
To: [redacted]
Sent: Tuesday, June 28, 2005 9:25 PM
Subject: (no subject)

I cdnuolt blveiee taht I cluod aulaclty uesdnatnrd waht I was rdanieg. The phaonmneal pweor of the hmuan mnid aoccdrnig to rscheearch taem at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the

believes in health care but is also super excited that she got the job!!!!!!!!!
11 minutes ago · Comment · Like

and am especially glad to have a job as my health benefits just ran out.
8 minutes ago

Congrats!
8 minutes ago

Congratulations! You go woman!
2 minutes ago

Write a comment...

No one should die because they cannot afford health care. No one should go broke because they get sick. No one should be denied medical care by their own insurer. If you agree, post as your status for today.
27 minutes ago · Comment · Like

believes that no one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.
28 minutes ago · Comment · Like

likes this.

Write a comment...

No one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.
35 minutes ago · Comment · Like

thinks that no one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.
42 minutes ago · Comment · Like

Central question:

How do ideas spread?

Ideas, disease, innovation, jokes spread continually via the global social network.

But *how?*

A known but typically unobserved process. How can we observe it?

A known but typically unobserved process. How can we observe it?

Pacific bluefin tuna transport Fukushima-derived radionuclides from Japan to California

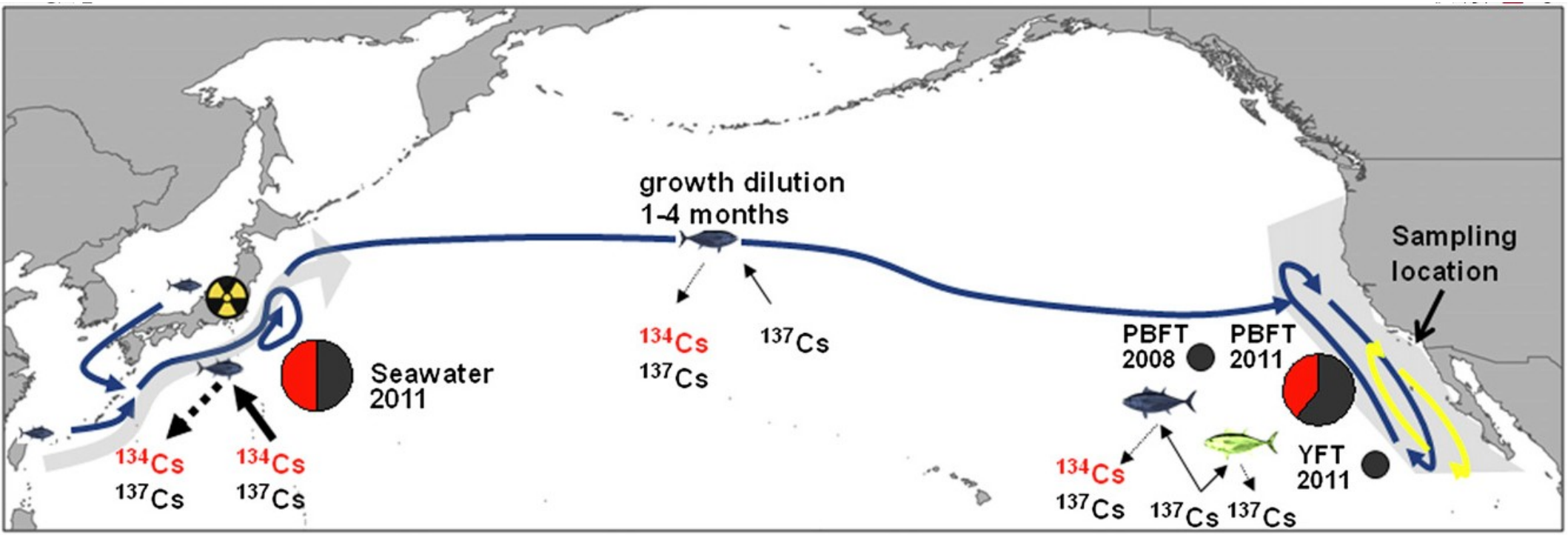
Daniel J. Madigan[a,1], Zofia Baumann[b], and Nicholas S. Fisher[b]

[a]Hopkins Marine Station, Stanford University, Pacific Grove, CA 93950; and [b]School of Marine and Atmospheric Sciences, Stony Brook University, Stony B
NY 11794

Edited by Karl K. Turekian, Yale University, North Haven, CT, and approved April 25, 2012 (received for review March 22, 2012)

The Fukushima Dai-ichi release of radionuclides into ocean waters caused significant local and global concern regarding the spread of radioactive material. We report unequivocal evidence that Pacific bluefin tuna, *Thunnus orientalis*, transported Fukushima-derived radionuclides across the entire North Pacific Ocean. We measured γ-emitting radionuclides in California-caught tunas and found ^{134}Cs (4.0 ± 1.4 Bq kg^{-1}) and elevated ^{137}Cs (6.3 ± 1.5 Bq kg^{-1}) in 15 Pacific

in their first year or early in their second (5). Thus, all blu
between years 1–2 (here, 2-y-old PBFT) caught during summ
the eastern Pacific must have migrated from the western Pa
within several months of capture. Waters north of the Kuro
Current (Fig. 1*A*) showed high radionuclide concentration
spring 2011 (3), and juveniles make extensive use of this re
before their eastward migration to the CCLME (6).



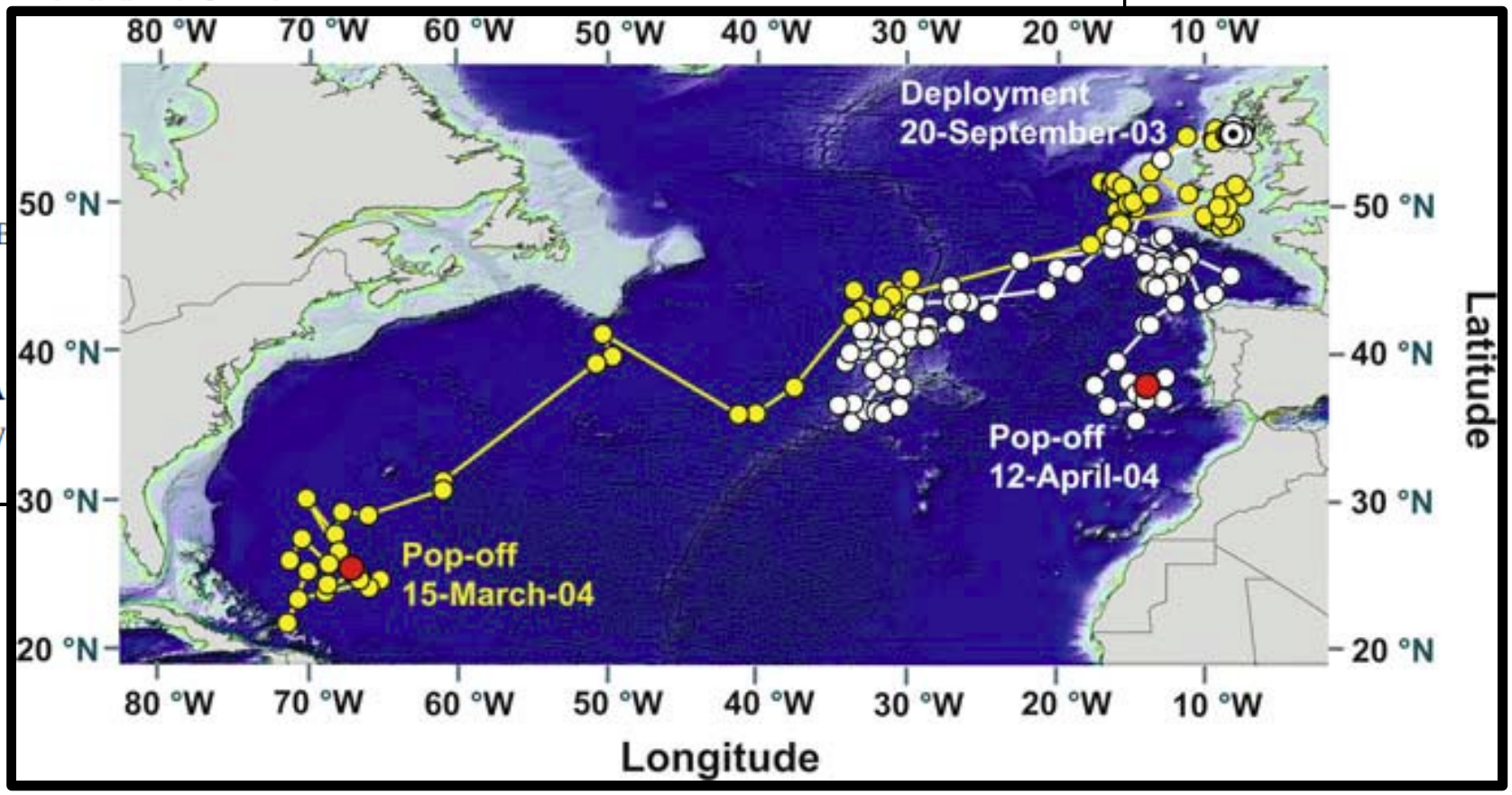
Hydrobiologia (2007) 582:91–97
DOI 10.1007/s10750-006-0552-y

FISH TELEMETRY

Results of satellite tagging of Atlantic bluefin tuna, *Thunnus thynnus*, off the coast of Ireland

Michael J. W. Stokesbury · Ronan Cosgrove · Andre Boustany · Daragh Browne · Steven L. H. Teo · Ronald K. O'Dor · Barbara A. Block

Abstract Pop-up
attached to six A
thynnus) off the w
2003 and 2004.



Noncentral question:

How do tuna spread?

Classic version:

Perform detailed study of diffusion among small group of individuals.

Alternative version:

Make use of an unusual event that makes the typically invisible patterns visible.

Central question:

How do ideas spread?

Classic version:

Perform detailed study of diffusion among small group of individuals.

Alternative version:

Make use of an unusual event that makes the typically invisible patterns visible.

Central question:

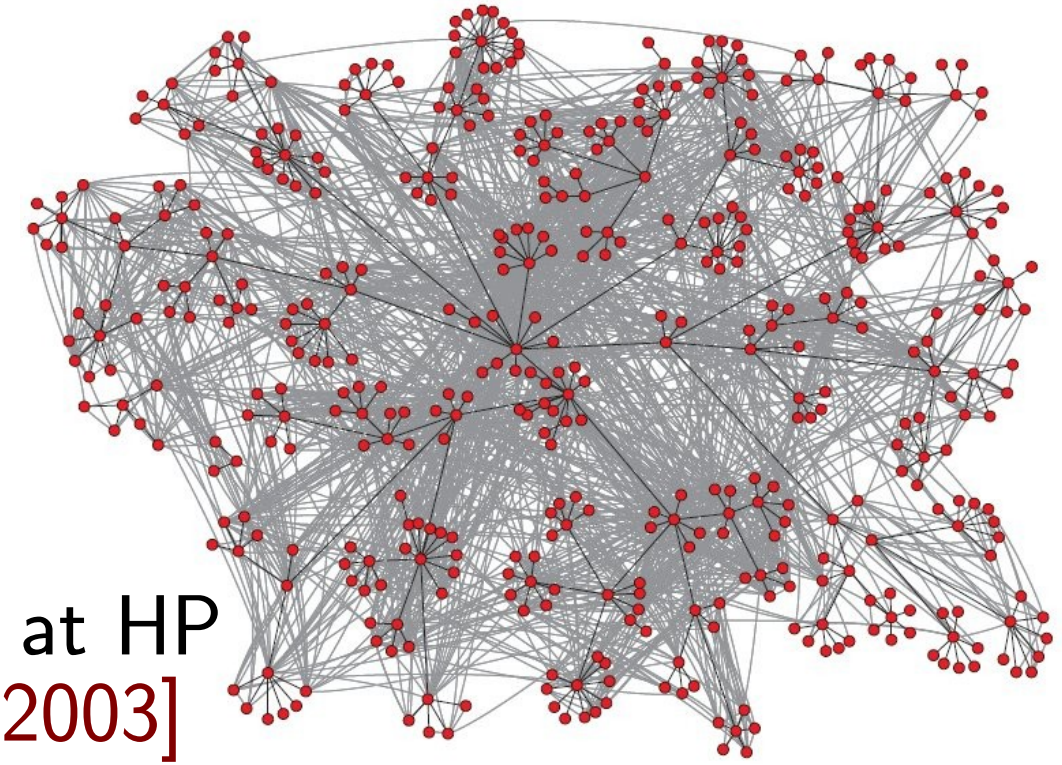
How do ideas spread?

Ideas, disease, innovation, jokes spread continually via the global social network.

But *how?*

What mechanisms?

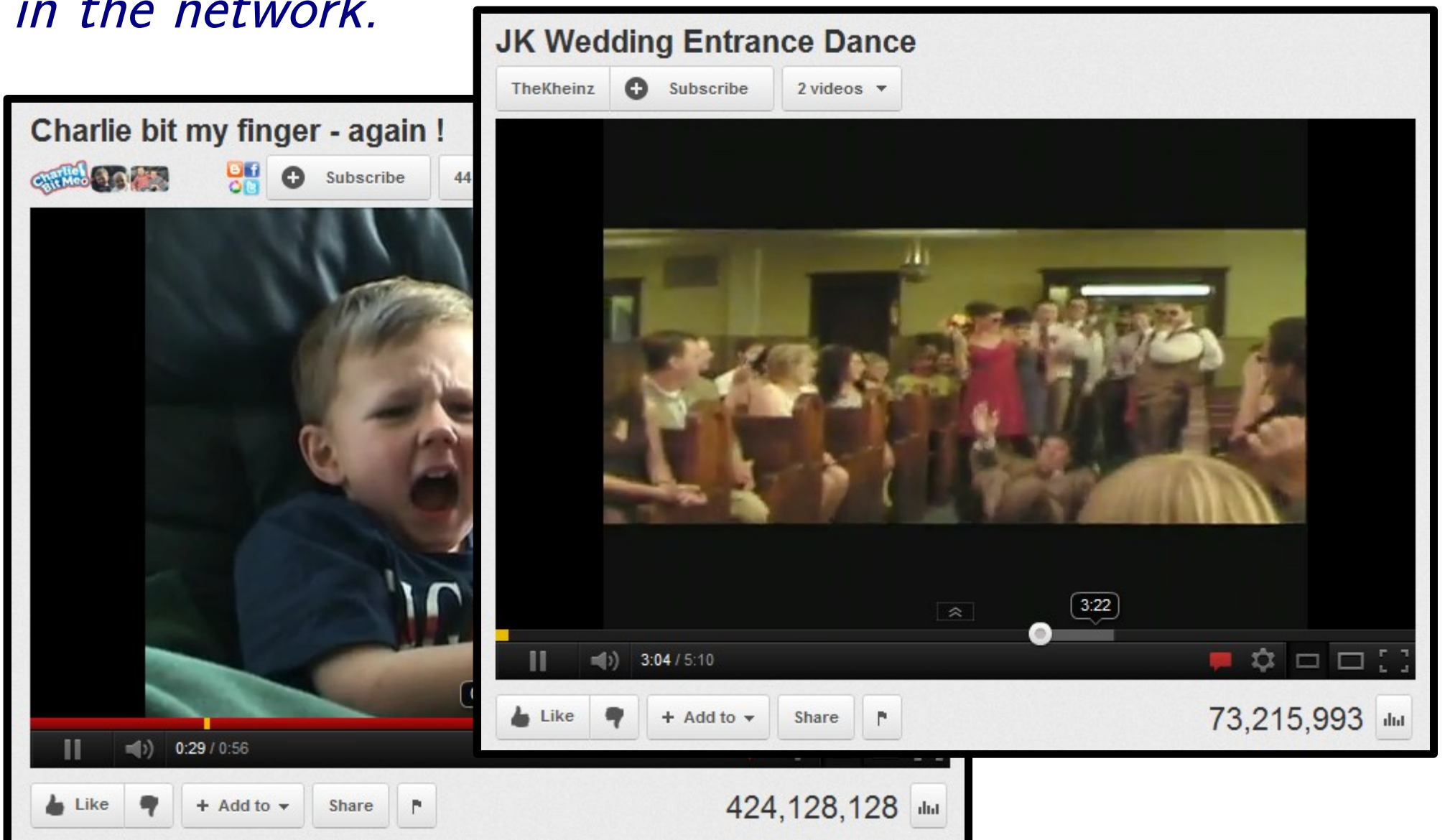
From whence the data?



Email network at HP
[Adamic Adar 2003]

Intuition: *going viral!*

Some people are susceptible to the meme; it spreads exponentially from "patient 0" through "susceptibles" in the network.



Intuition: *the small-world phenomenon!*

Two people chosen arbitrarily from the social network are connected by a small number of intermediate friends.

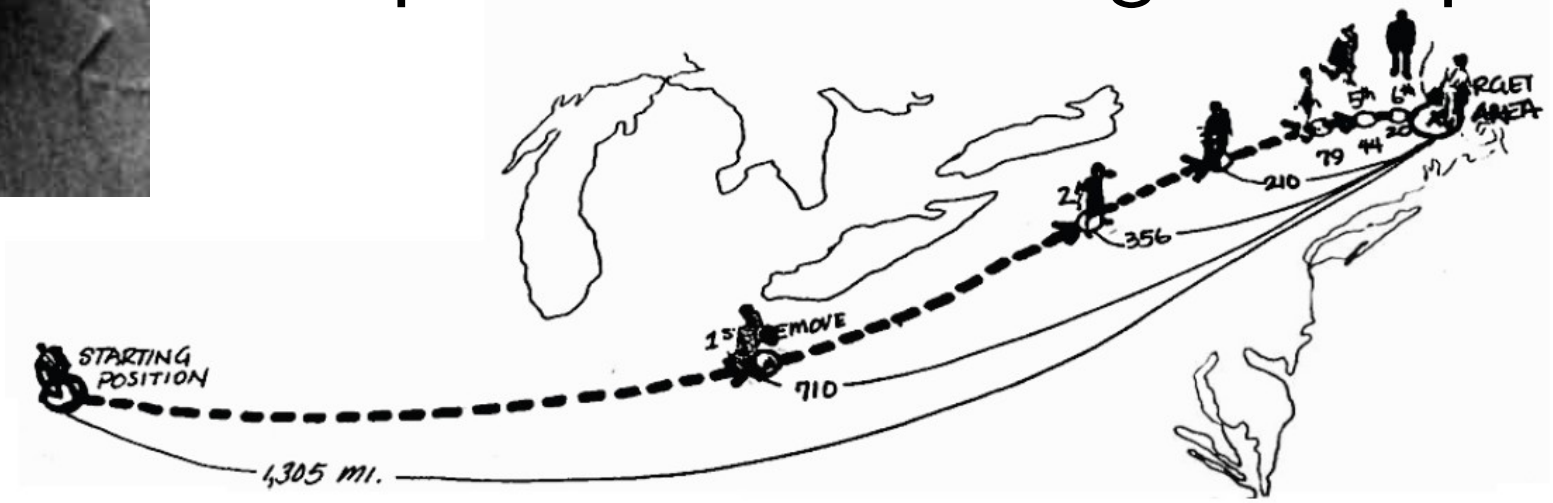
Stanley Milgram [1967]

Participants asked to forward letter to one friend.

source: resident of Omaha, NE

target: stockbroker near Boston

Completed chains averaged 6 hops!

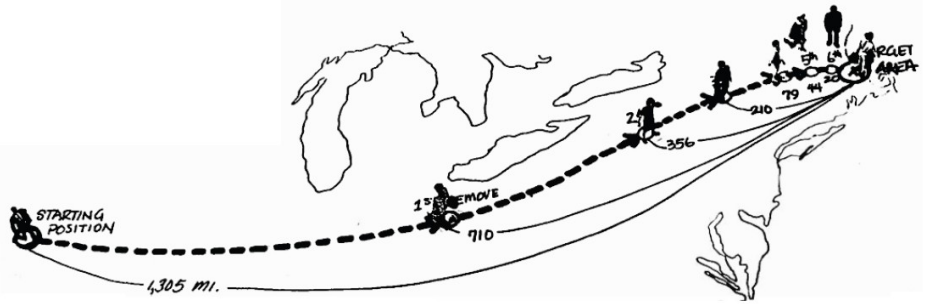


Central question:

How do ideas spread?

Intuition:

- exponential growth ("going viral")
- short chains ("small-world phenomenon")



Central question:

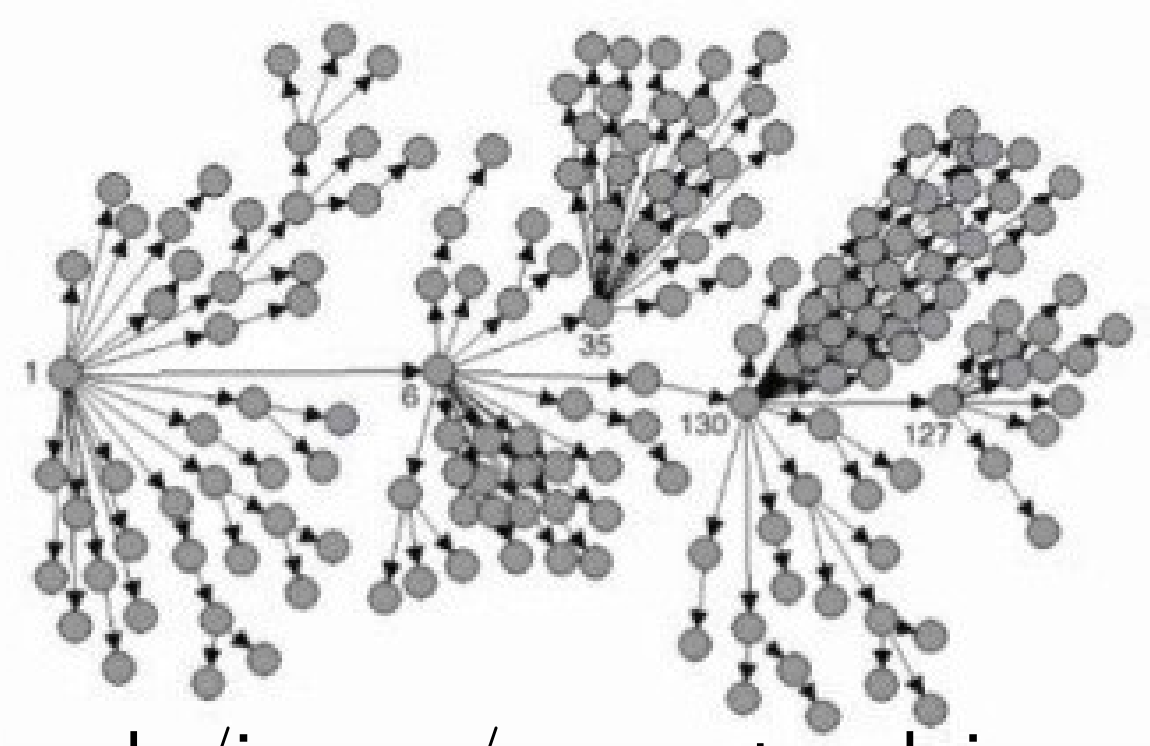
How do ideas spread?

Intuition: exponential growth, short chains

How do we test the intuition?
Where's data on the spread of one idea?

Generally hard to get genuine, large-scale data on a single entity's diffusion.

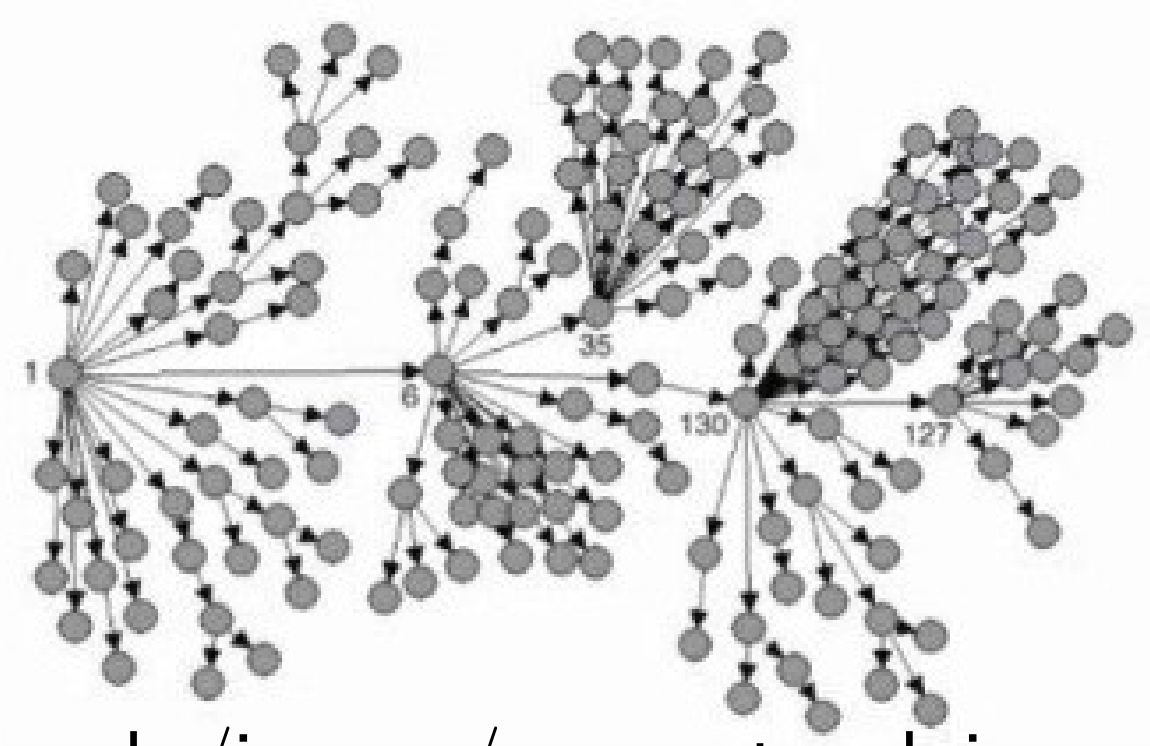
Diffusion of innovation
Contact tracing (epidemiology)
Folklore



http://web.mit.edu/networks/images/sars_network.jpg

Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore



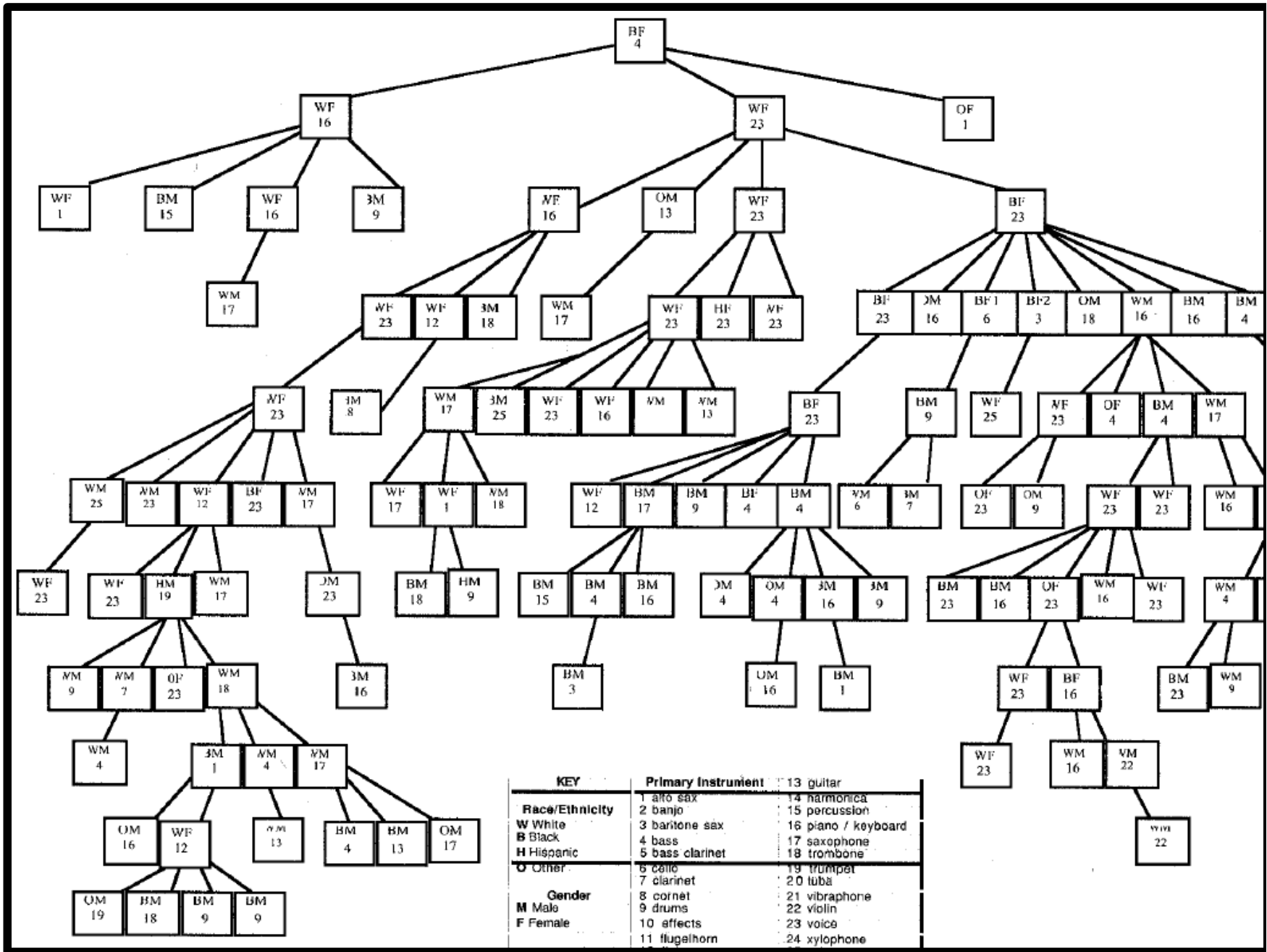
http://web.mit.edu/networks/images/sars_network.jpg

Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore
Sampling hidden populations
- [Goodman 1961]
- [Heckathorn 1997]
- [Heckathorn Jeffri 2003]
- ...

Hidden populations: Jazz Musicians in NYC

[Heckathorn Jeffri 2003]



Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore
Sampling hidden populations
Inferred spread of topics among blogs/email/Twitter

[Wu Huberman Adamic Tyler 2003]
[Adar Zhang Adamic Lukose 2004] [Adamic Adar 2005]
[Gruhl Guha DLN Tomkins 2004]
[Leskovec McGlohon Faloustos Glance Hurst 2007]
[Kumar Mahdian McGlohon 2010]
[Gomez-Rodriguez Leskovec Krause 2010] ← *come back tomorrow!*
...

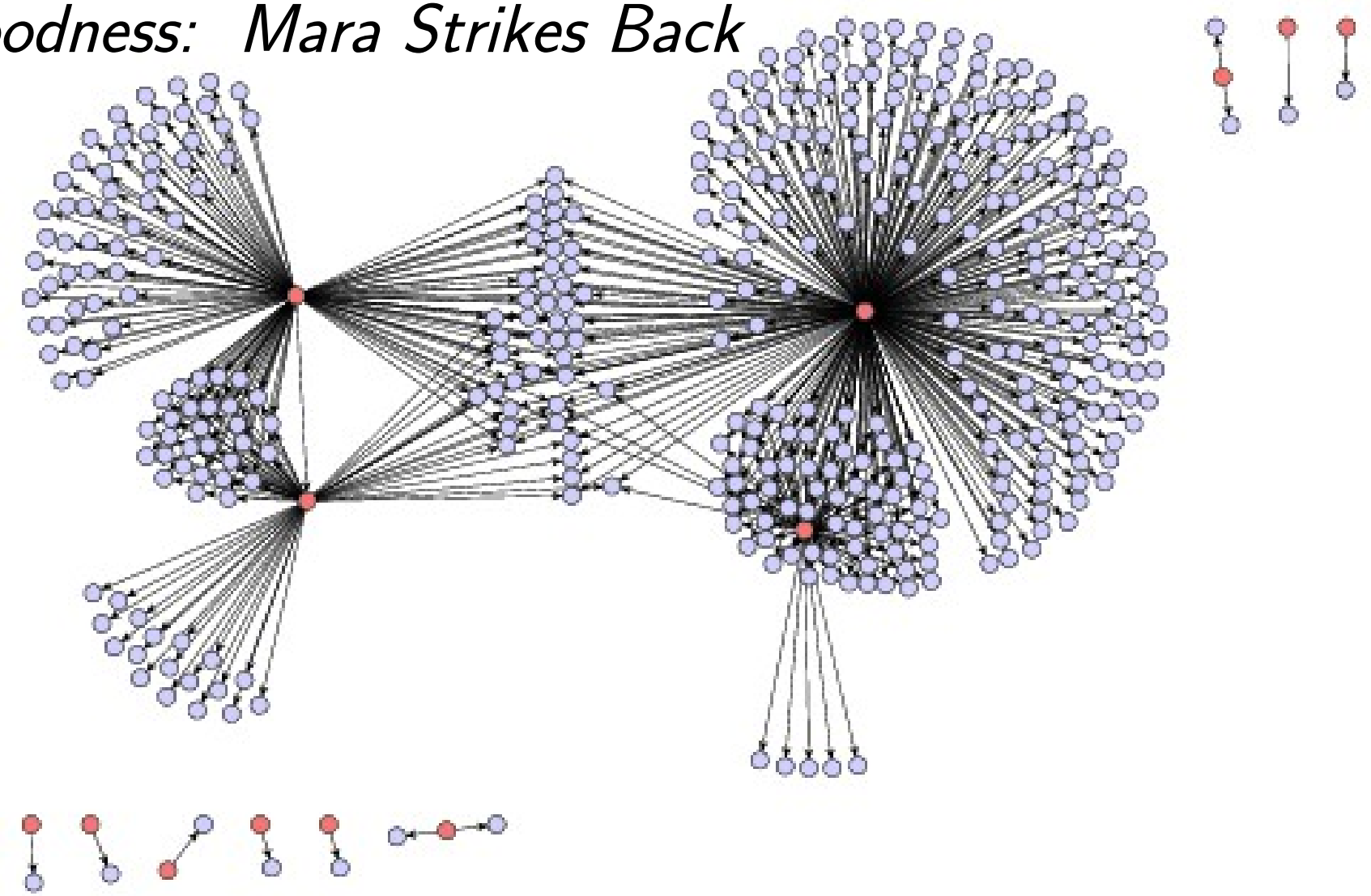
Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore
Sampling hidden populations
Inferred spread of topics among blogs/email/Twitter
Word of mouth/viral marketing
- [Goldenberg Libai Muller 2001]
- [Leskovec Adamic Huberman 2006]: product recommendations
- [Iribaren Moro 2009]
- [Sun Rosenn Marlow Lento 2009]: adoption in Facebook feeds
- ...

Viral marketing ("large online retailer")

[Leskovec Adamic Huberman 2006]

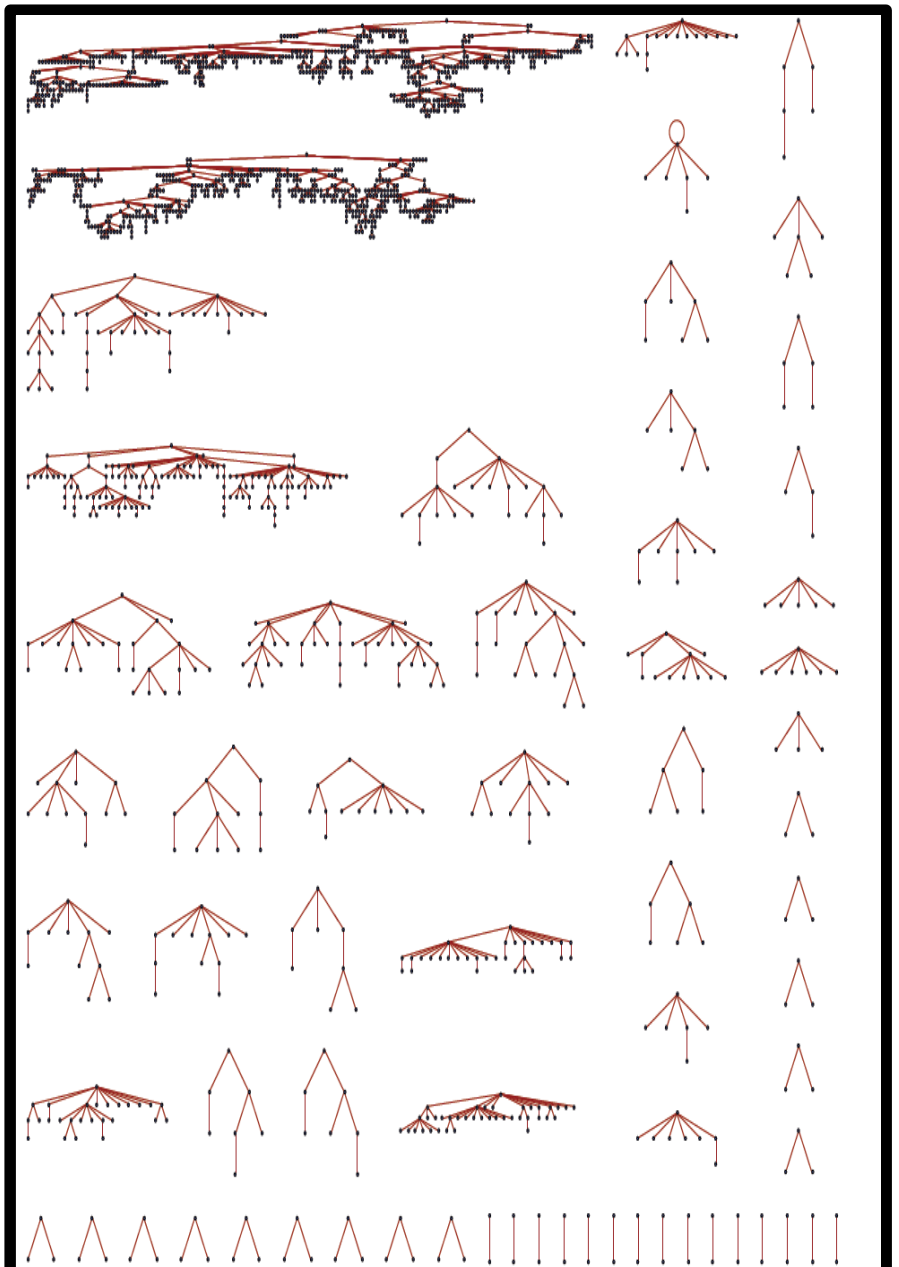
Friends' recommendations for
Oh My Goodness: Mara Strikes Back



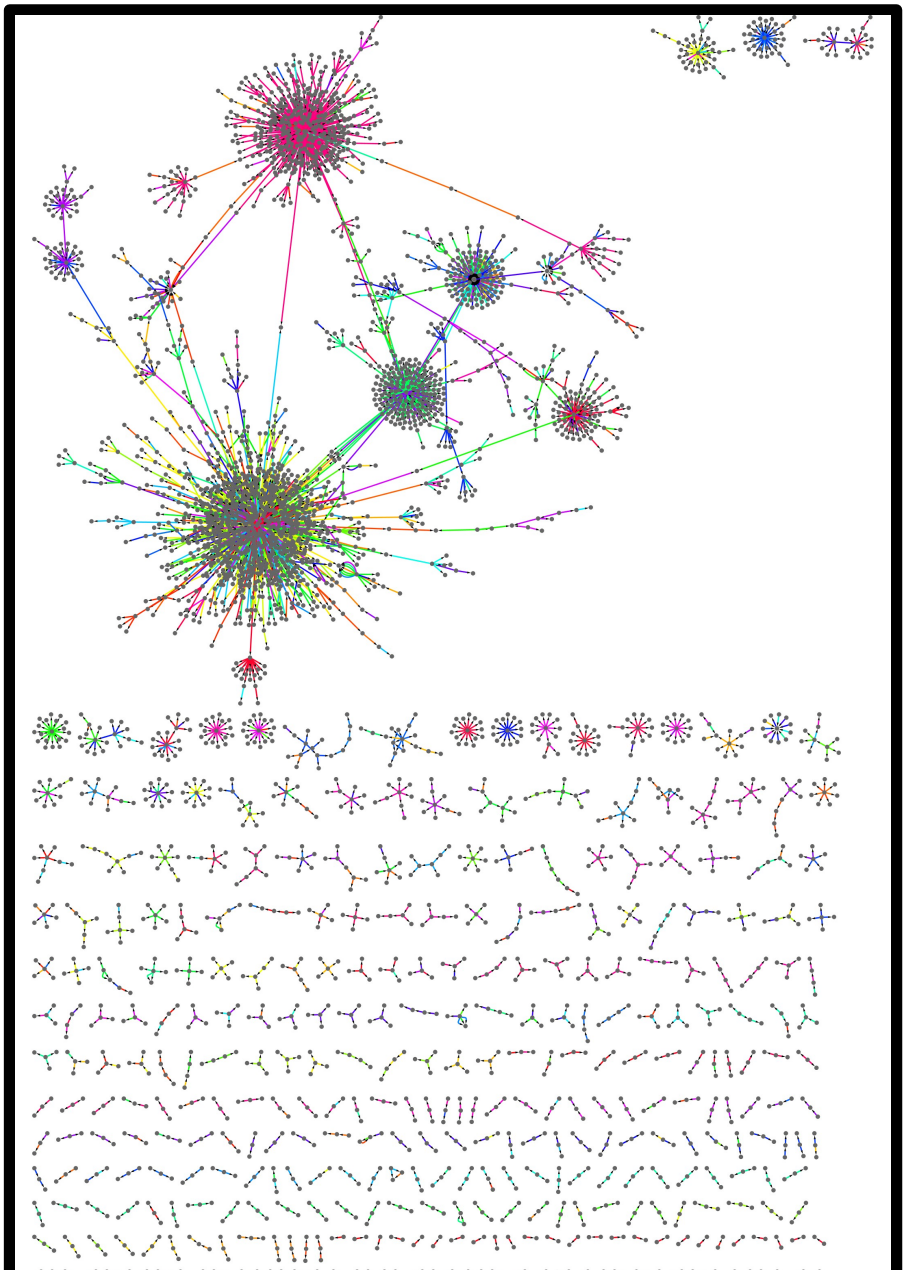
Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore
Sampling hidden populations
Inferred spread of topics among blogs/email/Twitter
Word of mouth/viral marketing
Digital traces from online social communities
- [Bakshy Kerrer Adamic 2009]: cloneable assets in Second Life
- [Cha Mislove Gummadi 2009]: Flickr favorites
- [Lerman Ghosh 2010] [Kwak Lee Park Moon 2010]: retweeting
- ...

Gestures in Second Life
[Bakshy Kerrer Adamic 2009]



Retweets in Twitter
[Kwak Lee Park Moon 2010]



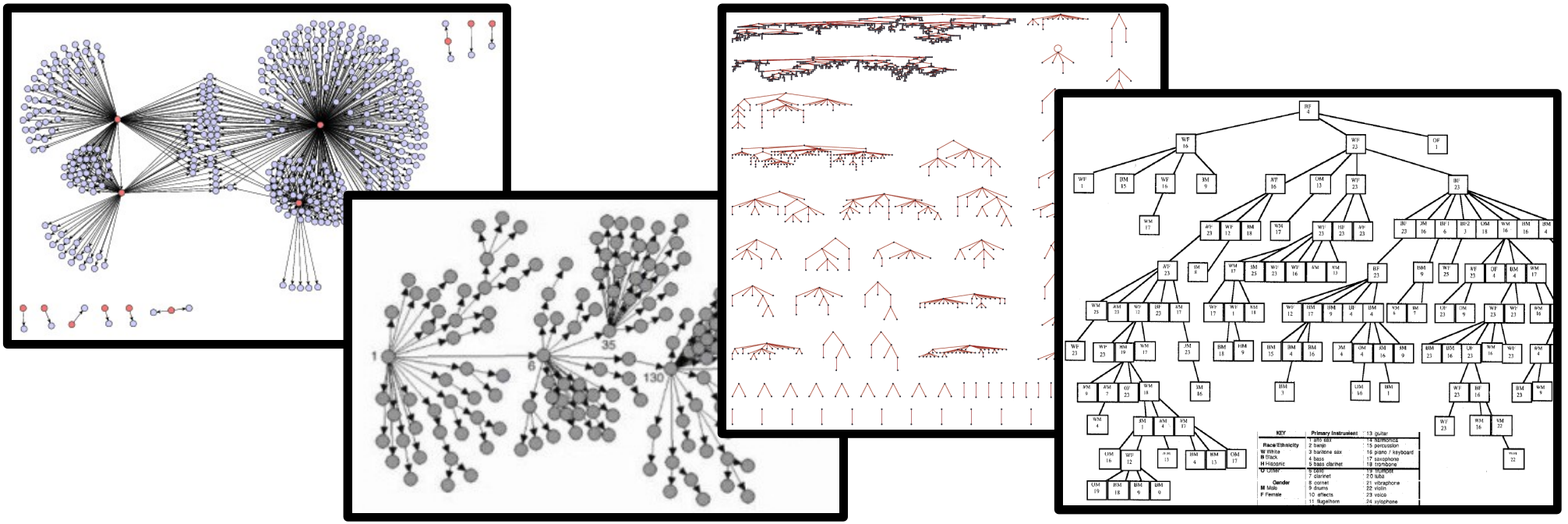
Generally hard to get genuine, large-scale data on a single entity's diffusion.
(But it's getting easier ...)

Diffusion of innovation
Contact tracing (epidemiology)
Folklore
Sampling hidden populations
Inferred spread of topics among blogs/email/Twitter
Word of mouth/viral marketing
Digital traces from online social communities

The rest of this talk:

Email-based chain-letter petitions as "tracers" of large-scale propagation through the social network.

[DLN Kleinberg, PNAS 2008]



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

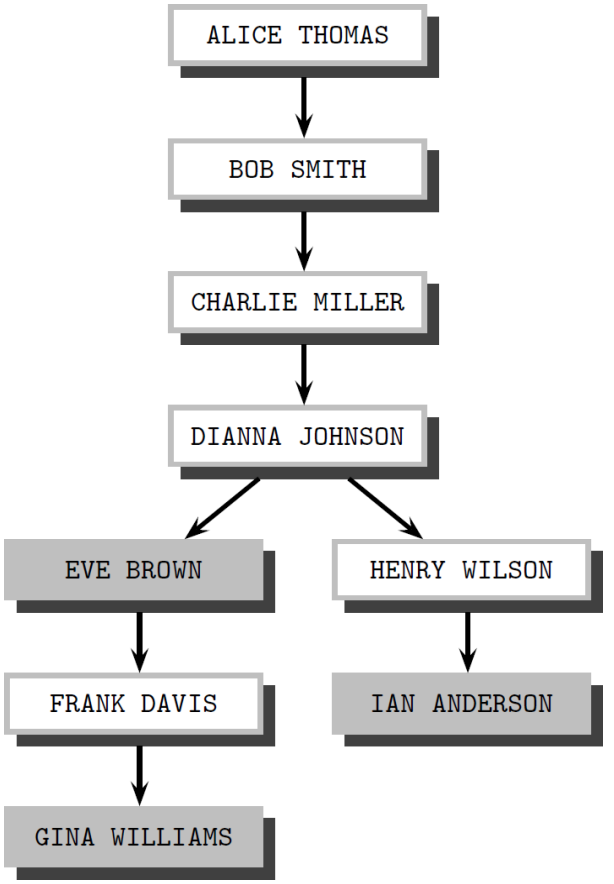
“Before we start”

(80% of the work)

1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

2) Build propagation tree from copies.
(x,y) edge = x immediately precedes y in some copy



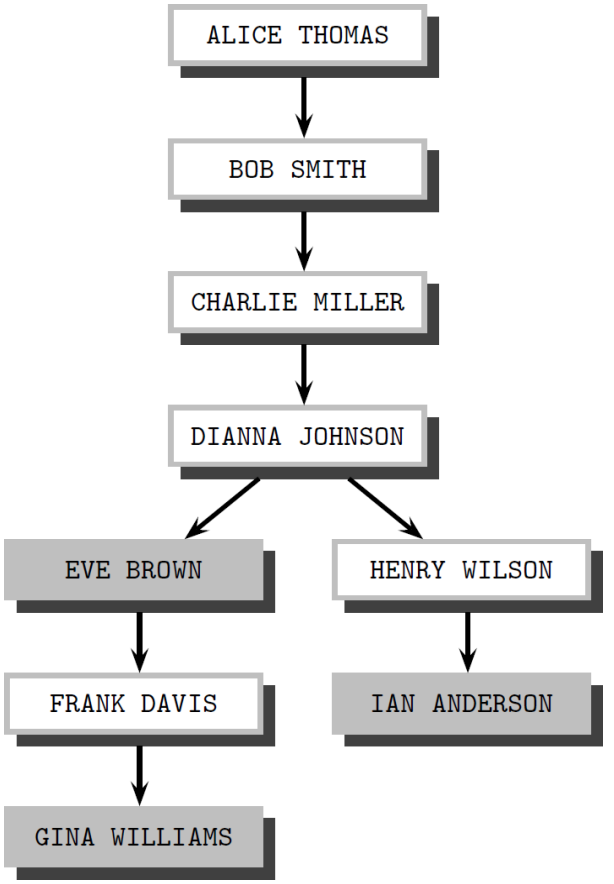
1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

2) Build propagation tree from copies.
(x,y) edge = x immediately precedes y in some copy

Some valuable and unusual features of the dataset:

Genuine large-scale trace of propagation through social network

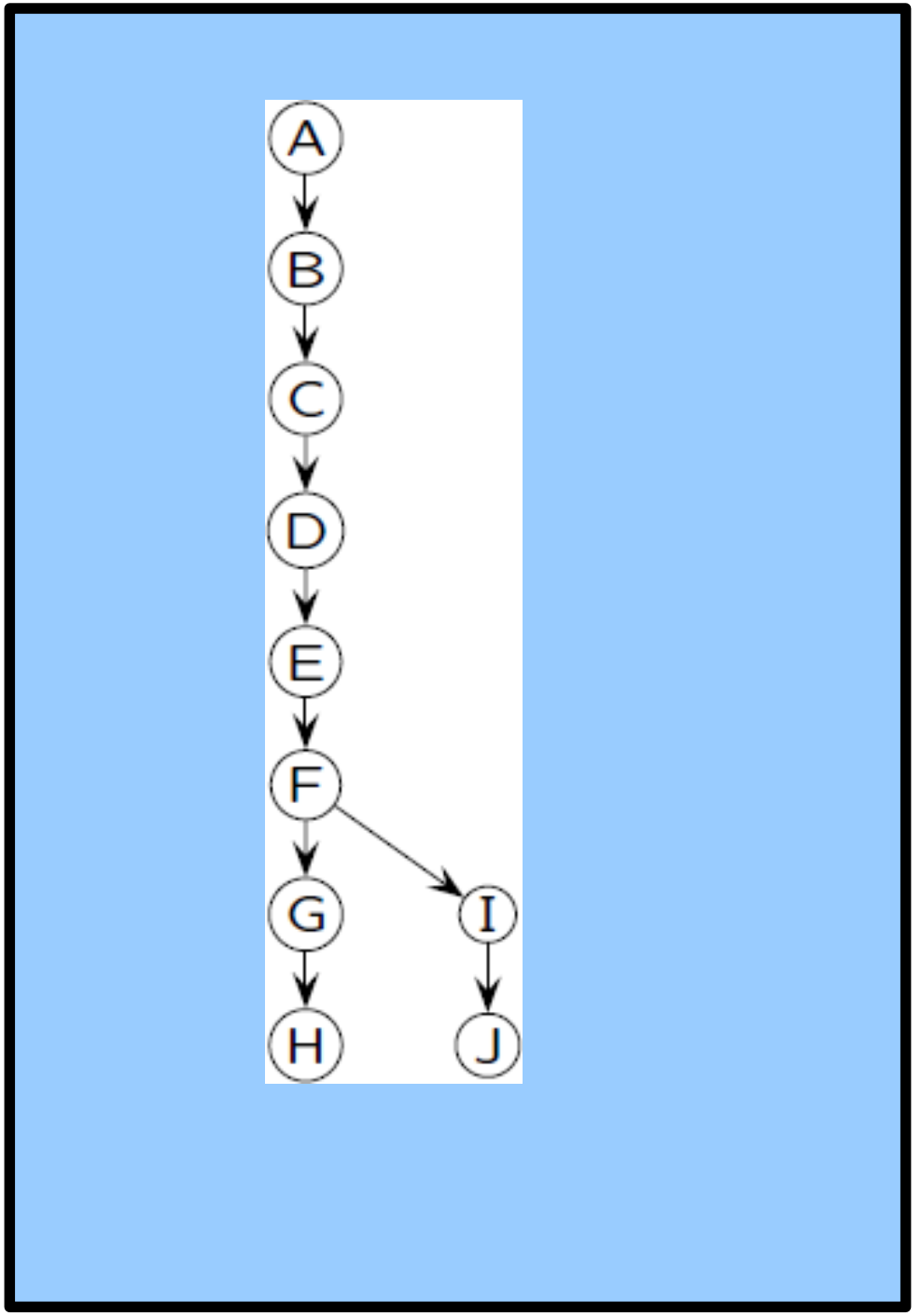
Each copy "lights up" a path to source (650 copies yield 20K people!)



A
B
C
D
E
F
G
H

A
B
C
D
E
F
G
H

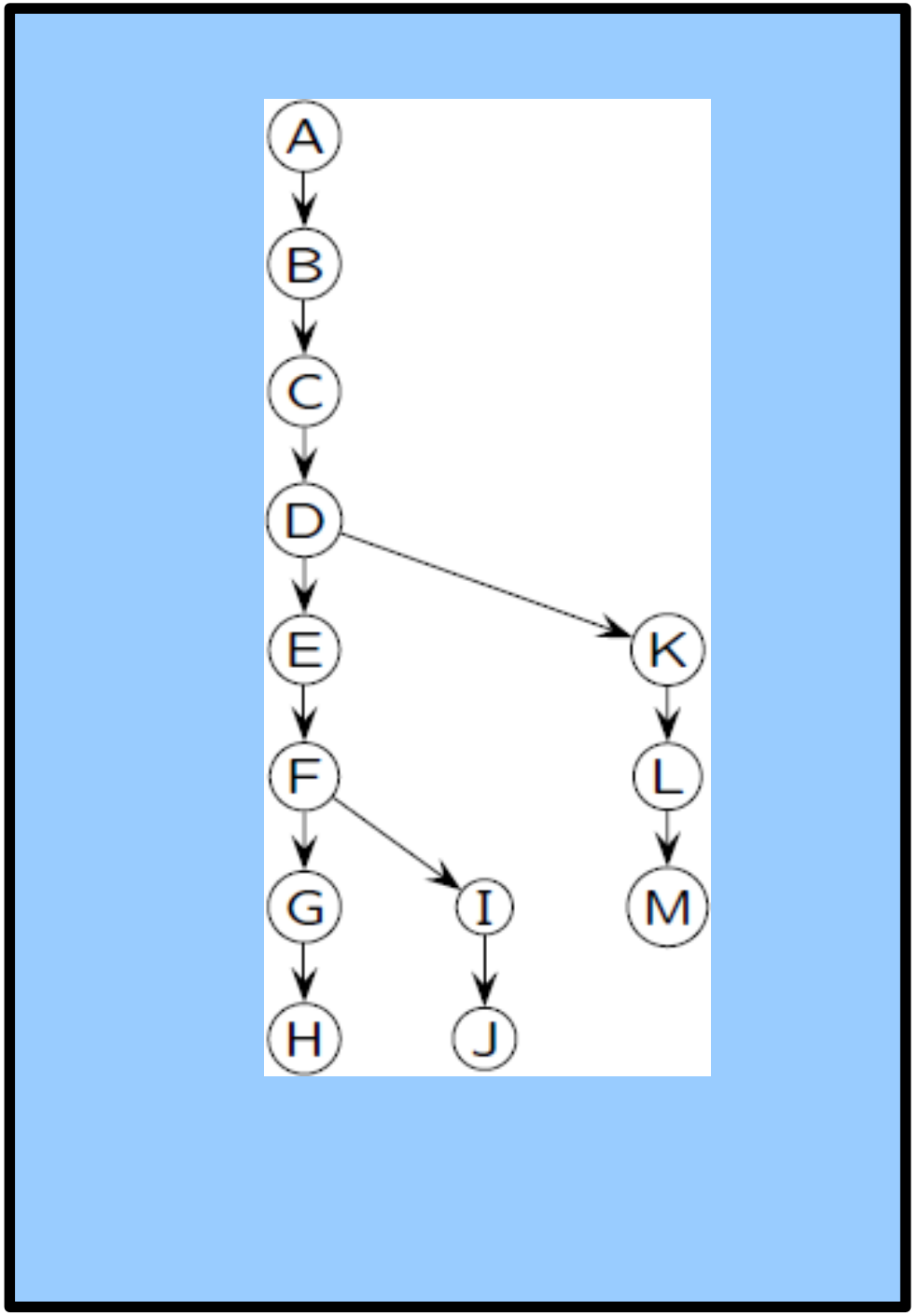
A
B
C
D
E
F
I
J



A B C D E F G H

A B C D E F I J

A B C D K L M



Is the propagation tree *really* a tree?

Is the propagation tree *really* a tree?

- No: some respond twice (= have 2 parents).
  (Rare.)

Is the propagation tree *really* a tree?

- No: some respond twice (= have 2 parents).
  (Rare.)
- Apparently not: many typographical changes.
  (Frequent.)

Change you can believe in:

[redacted] believes in health care but is also super excited that she got the job!!!!!!!!!

11 minutes ago · Comment · Like

[redacted] and am especially glad to have a job as my health benefits just ran out.

8 minutes ago

[redacted] Congrats!

8 minutes ago

[redacted] Congratulations! You go woman!

2 minutes ago

Write a comment...

[redacted] No one should die because they cannot afford health care. No one should go broke because they get sick. No one should be denied medical care by their own insurer. If you agree, post as your status for today.

27 minutes ago · Comment · Like

[redacted] believes that no one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.

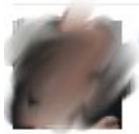
28 minutes ago · Comment · Like

[redacted] likes this.

Write a comment...

[redacted] No one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.

35 minutes ago · Comment · Like



[redacted] thinks that no one should die because they cannot afford health care, and no one should go broke because they get sick. If you agree, please post this as your status for the rest of the day.

42 minutes ago · Comment · Like

Change you can believe in:

No one should die because they cannot afford health care. No one ...

believes that no one should die because they cannot afford health care, *and* no one ...

No one should die because they cannot afford health care, *and* no one ...

thinks that no one should die because they cannot afford health care, *and* no one ...

#copies	Version of the name of the same signatory
520	137) Name Withheld, Barcelona, Catalonia
138	138) Name Withheld, Barcelona, Catalonia
17	137) Name WIthheld, Barcelona, Catalonia
6	137. Name Withheld, Barcelona, Catalonia
3	139) Name Withheld, Barcelona, Catalonia
3	137) Name Withheld, Barcelona, Catalonia, Spain
2	147) Name Withheld, Barcelona, Spain
2	147) Name Withheld, Barcelona, Catalonia
2	137) Name Withheld, Barcelona, España
2	137) Name Withheld, Barcelona, Catalonia &
2	137) Name Withheld, Barcelona, Ca talonia
1	138) Name Withheld, Barcelona, Spain
1	137) Name Withheld, Barcelona, Spain
1	137) Name Withheld, Barcelona, SPAIN
1	137) Name Withheld, Barcelona, CATALONIA

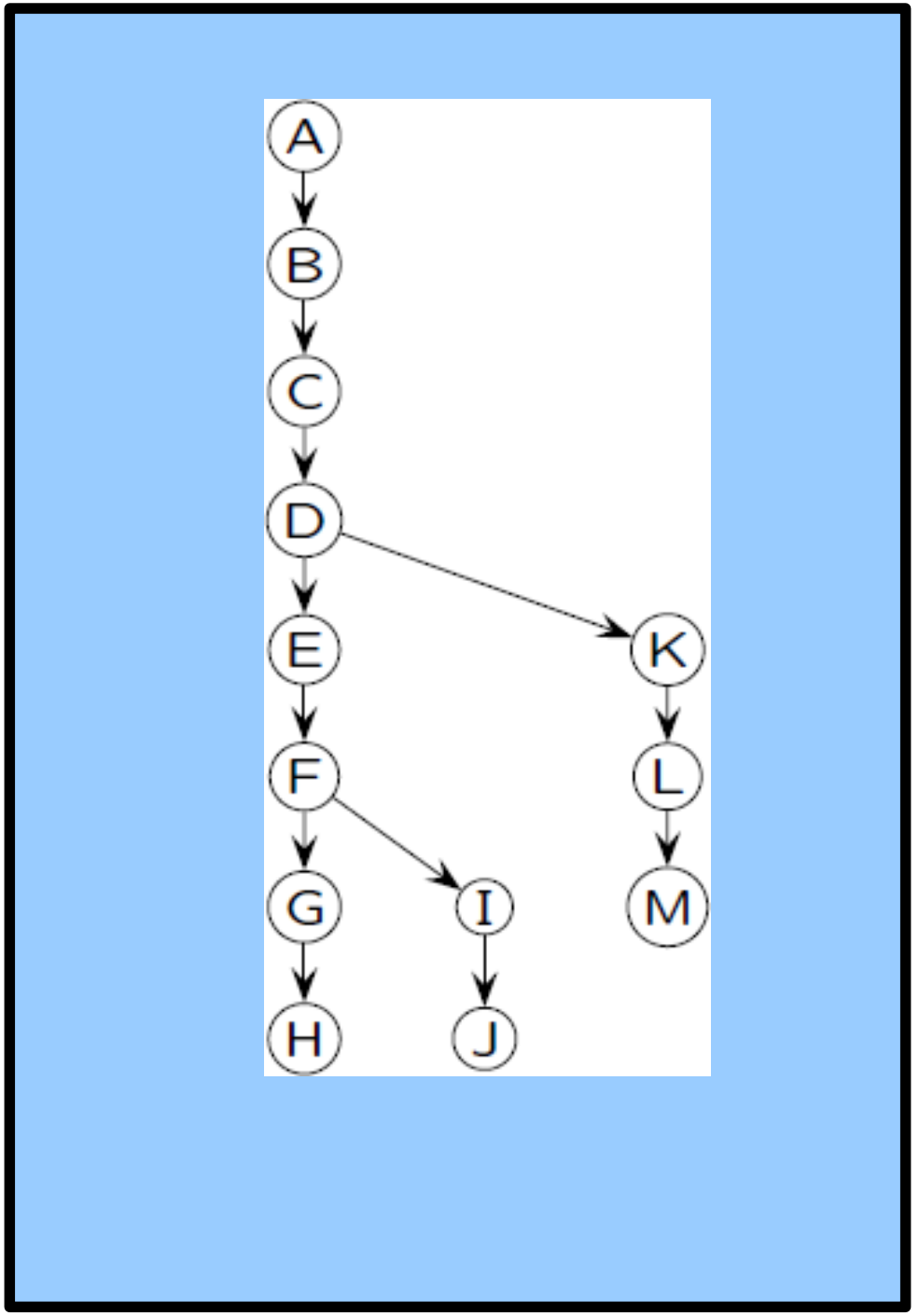
Is the propagation tree *really* a tree?

- No: some respond twice (= have 2 parents).
  (Rare.)
- Apparently not: many typographical changes.
  (Frequent.)
- Apparently not: many list rearrangements.
  (Often enough!)

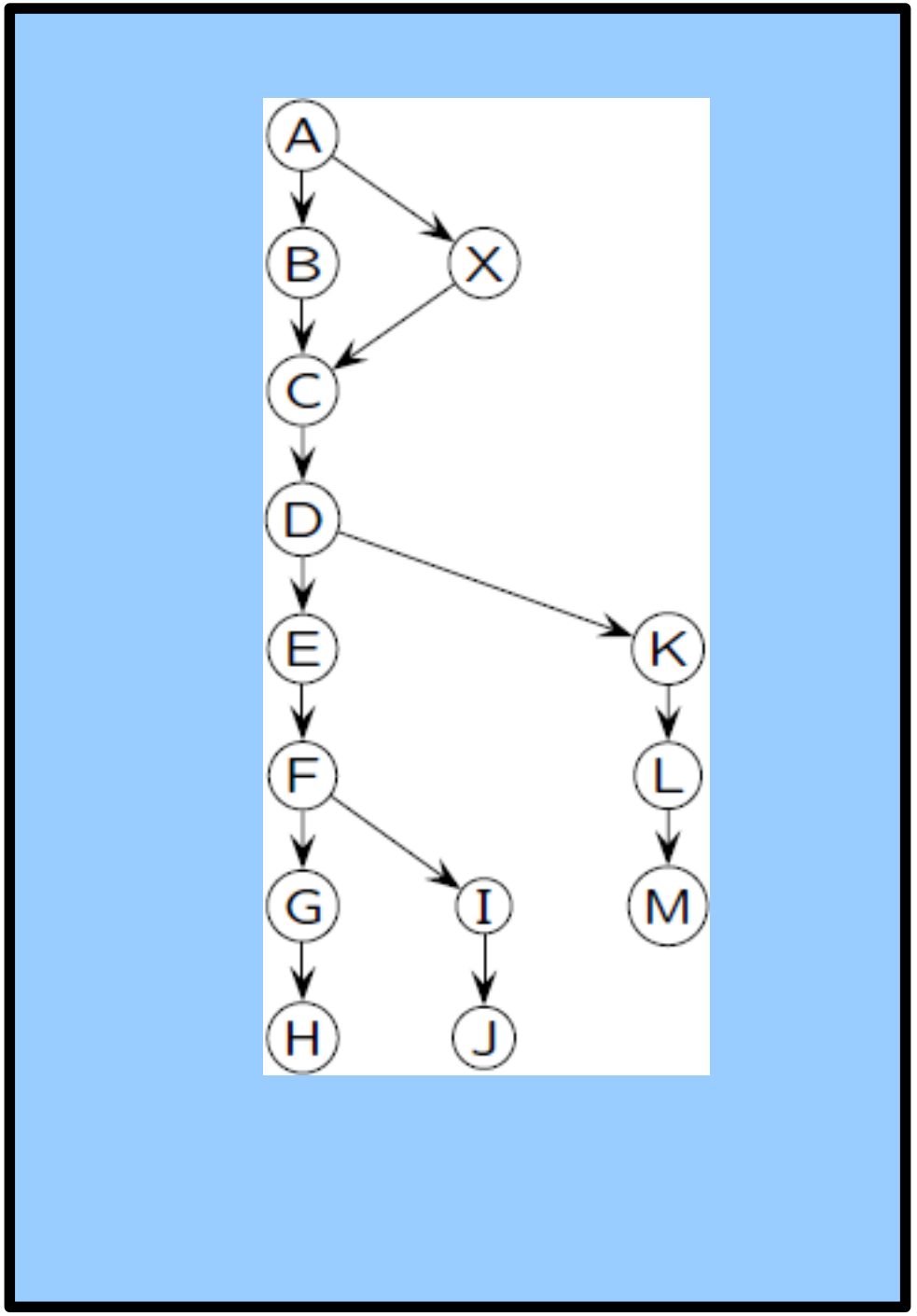
A B C D E F G H

A B C D E F I J

A B C D K L M

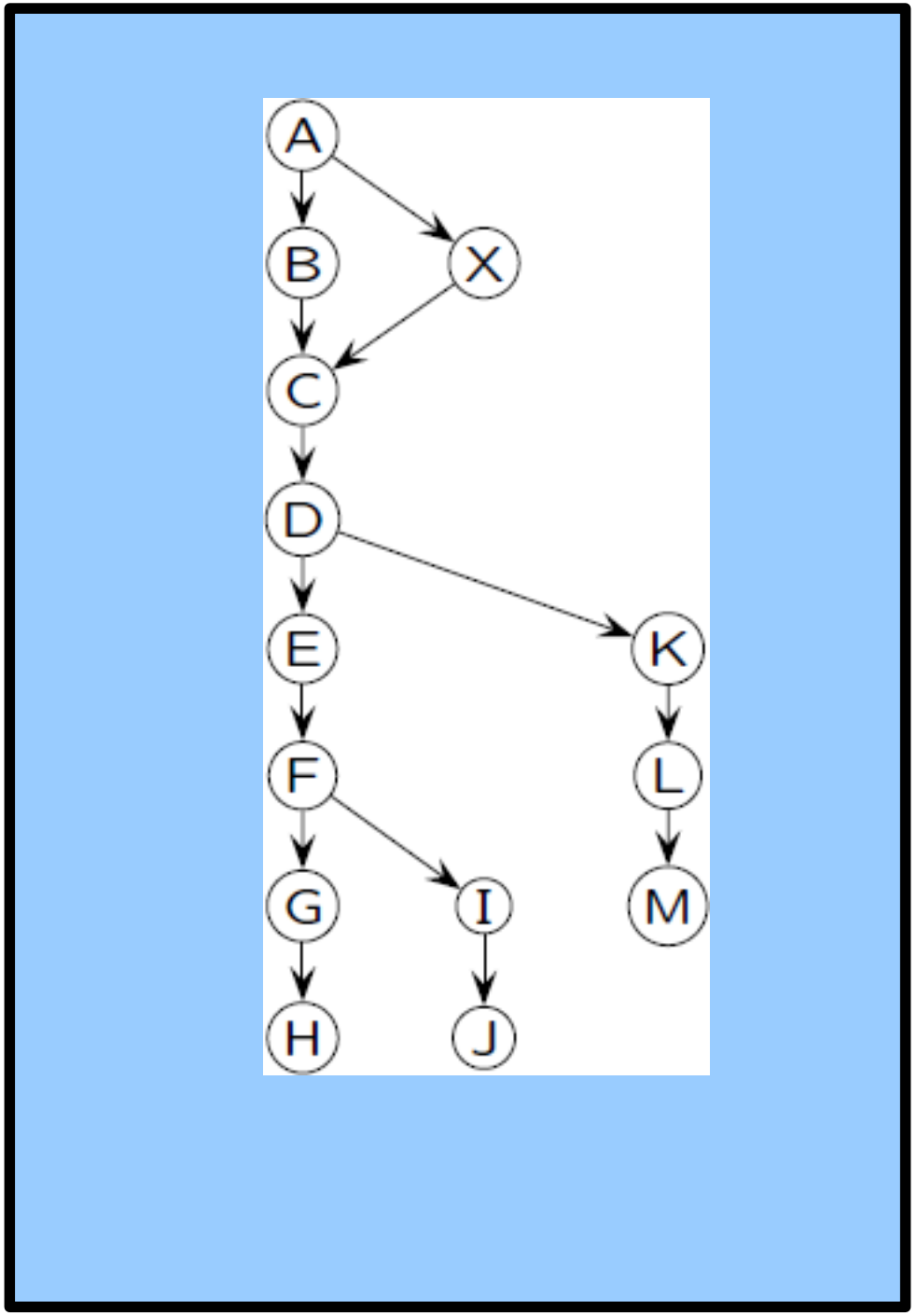


A	A	A	A
B	B	B	X
C	C	C	C
D	D	D	D
E	E	K	E
F	F	L	F
G	I	M	G
H	J		H

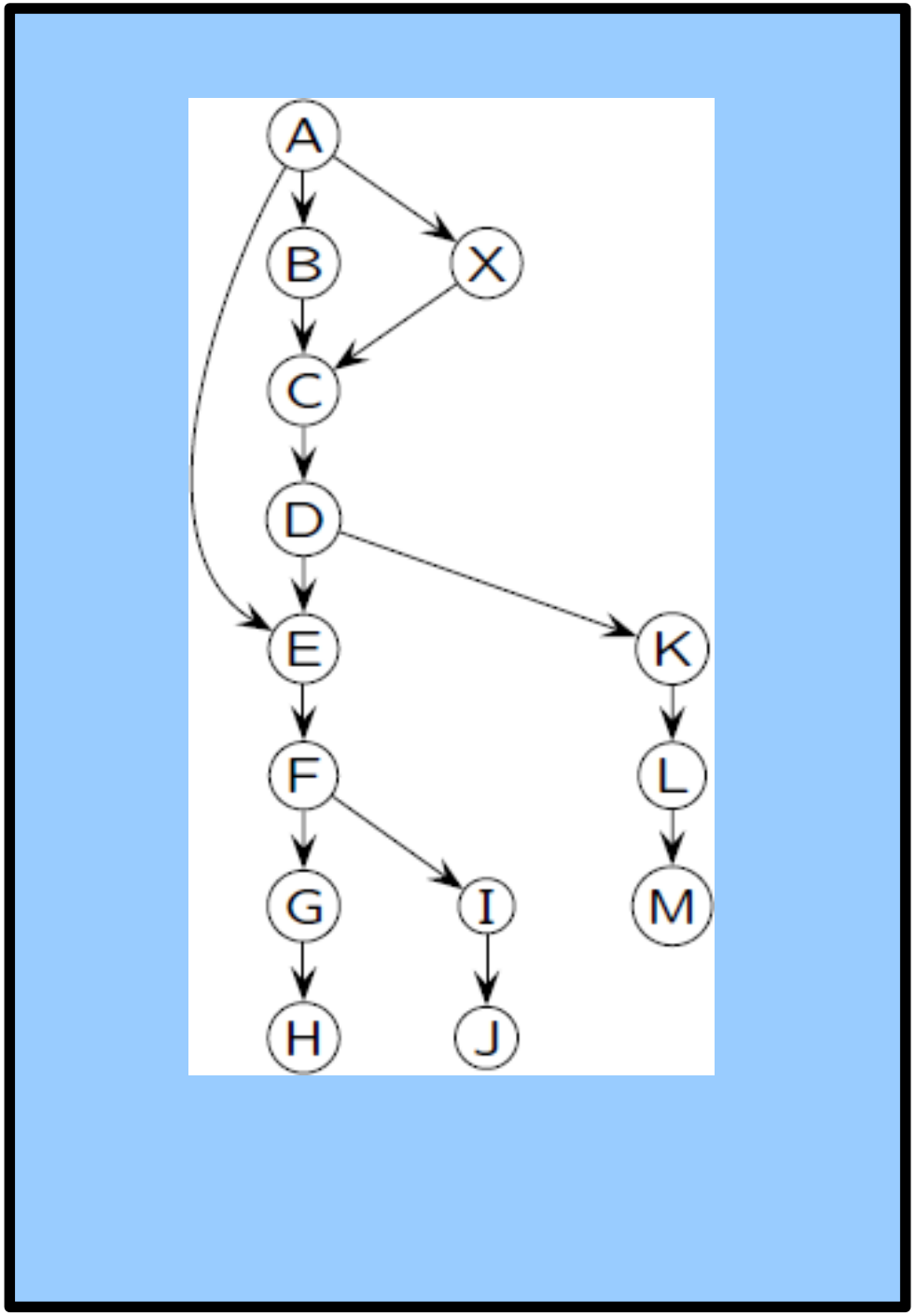


Point mutation: names replaced by political figures

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



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



Point mutation: names replaced by political figures

Insertion/deletion: blocks of ~5 names missing/included

Point mutation: names replaced by political figures

Insertion/deletion: blocks of ~5 names missing/included

Duplication: blocks of ~10 names sometimes repeated

Point mutation: names replaced by political figures

Insertion/deletion: blocks of ~5 names missing/included

Duplication: blocks of ~10 names sometimes repeated

Transposition: two blocks of ~5 names swapped

Point mutation: names replaced by political figures

Insertion/deletion: blocks of ~5 names missing/included

Duplication: blocks of ~10 names sometimes repeated

Transposition: two blocks of ~5 names swapped

Hybridization: two lists interleaved in third list(!)

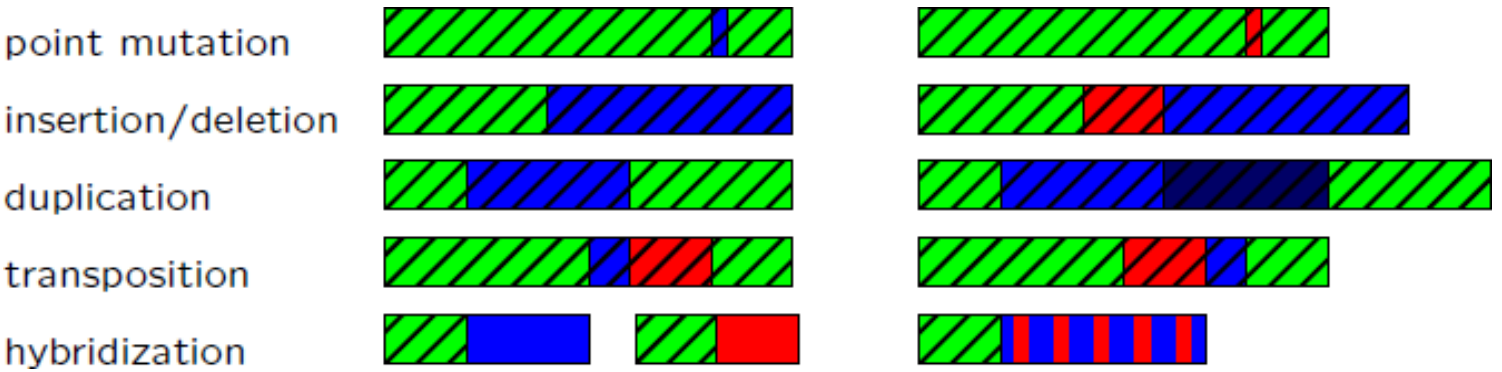
Is the propagation tree *really* a tree?

- No: some respond twice (= have 2 parents).
  (Rare.)

- Apparently not: many typographical changes.
  (Frequent.)

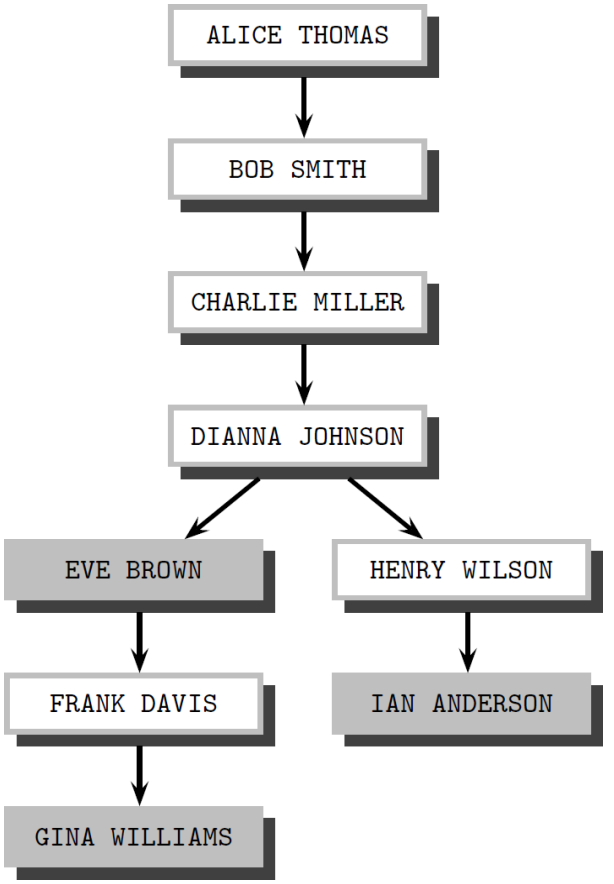
```
137) Name Withheld, Barcelona, Catalonia
147) Name Withheld, Barcelona, Spain
137) Name Withheld, Barcelona, España
```

- Apparently not: many list rearrangements.
  (Often enough!)



1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

2) Build propagation tree from copies.
(x,y) edge = x immediately precedes y in some copy



1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

2) Build propagation *graph G* from copies.
(x,y) edge = x immediately precedes y in some copy
treat names within small edit distance as "same"

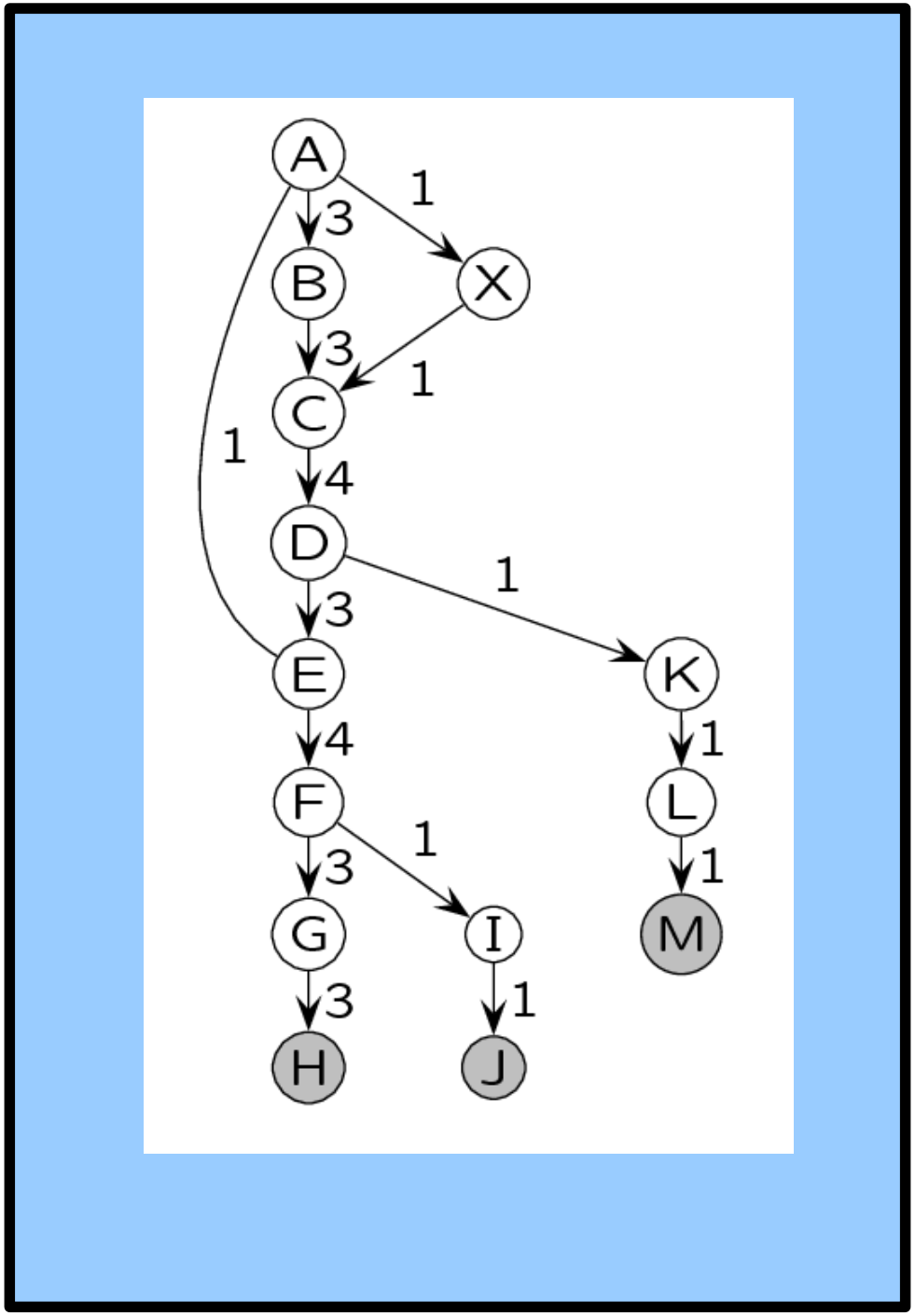
1) Query search engines to find copies of petitions. *(Got 650 distinct copies, ~20K names.)*

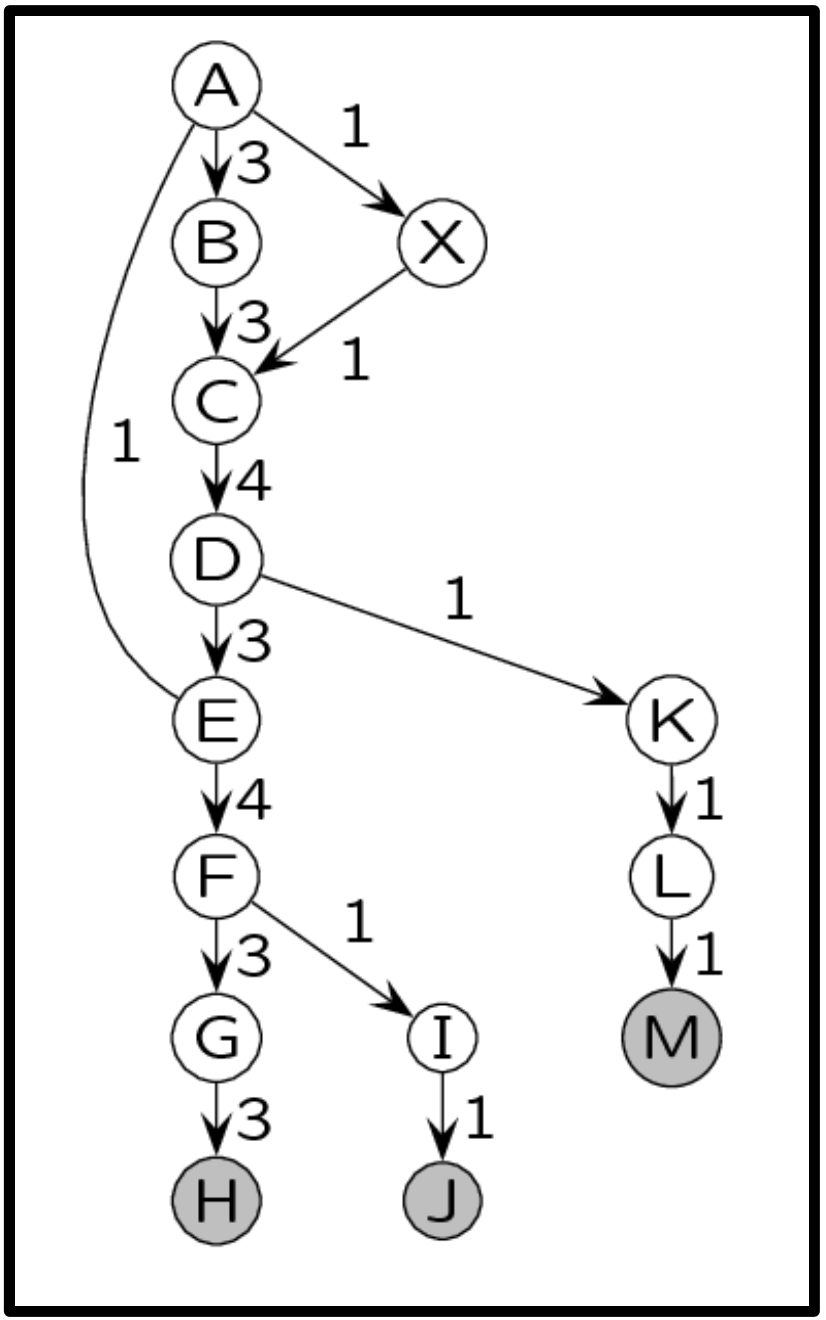
2) Build propagation *graph G* from copies.
(x,y) edge = x immediately precedes y in some copy
treat names within small edit distance as "same"

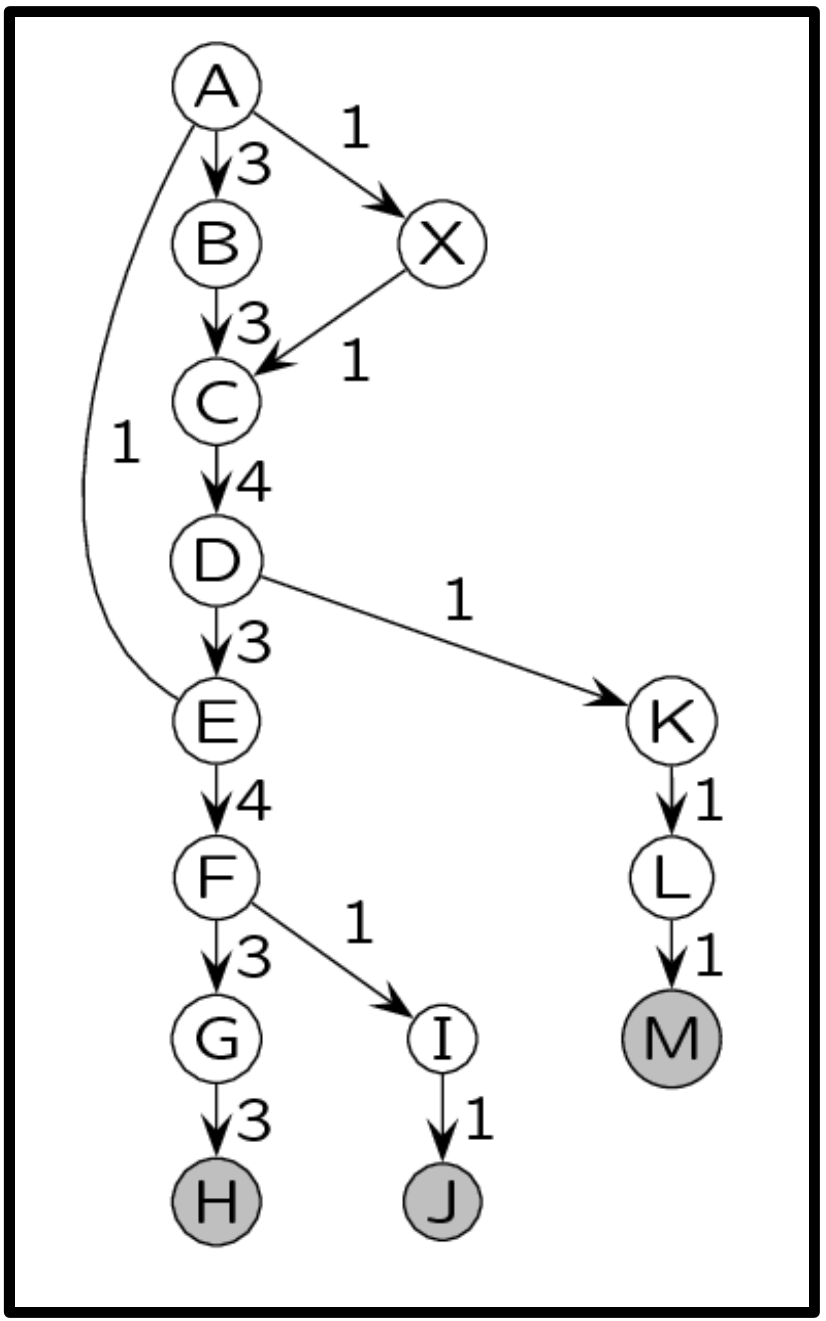
3) Delete the fewest edges possible from G to form a tree T; prune unsupported nodes.
[Edmonds 1964 "max weight spanning arborescence"]
really: delete "lightest" set of edges to form T, using number of copies containing (x,y) as edge weight.

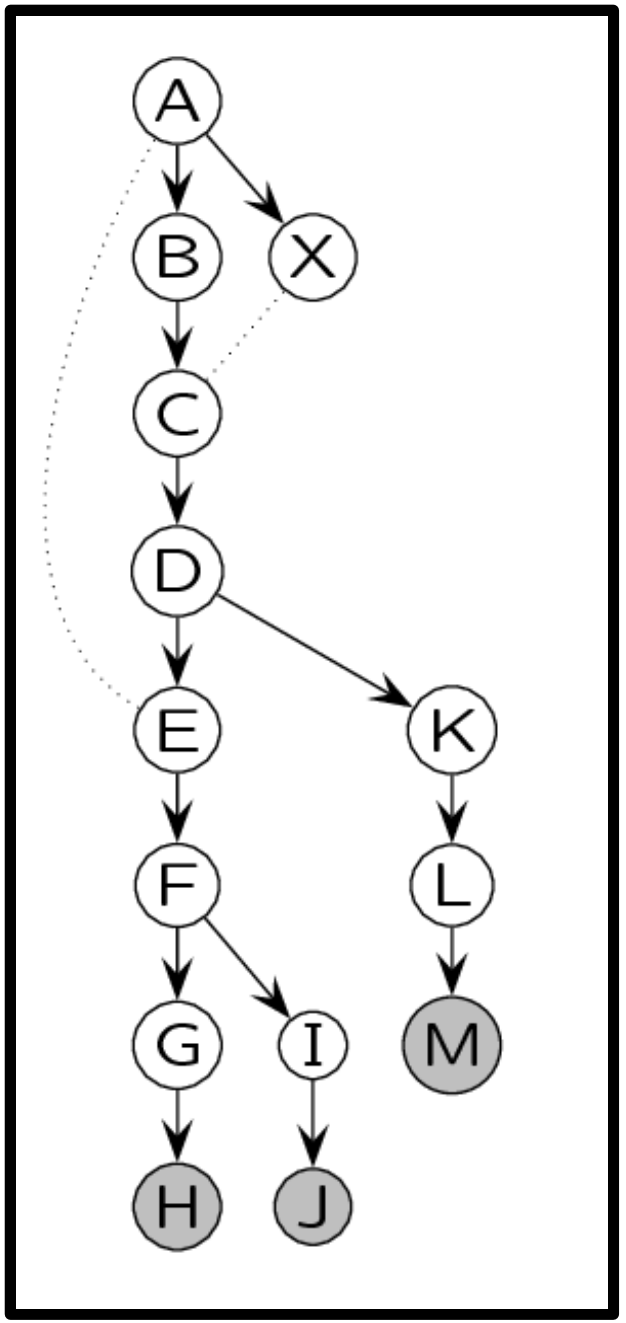
Impact is small; >97% edges remain intact.

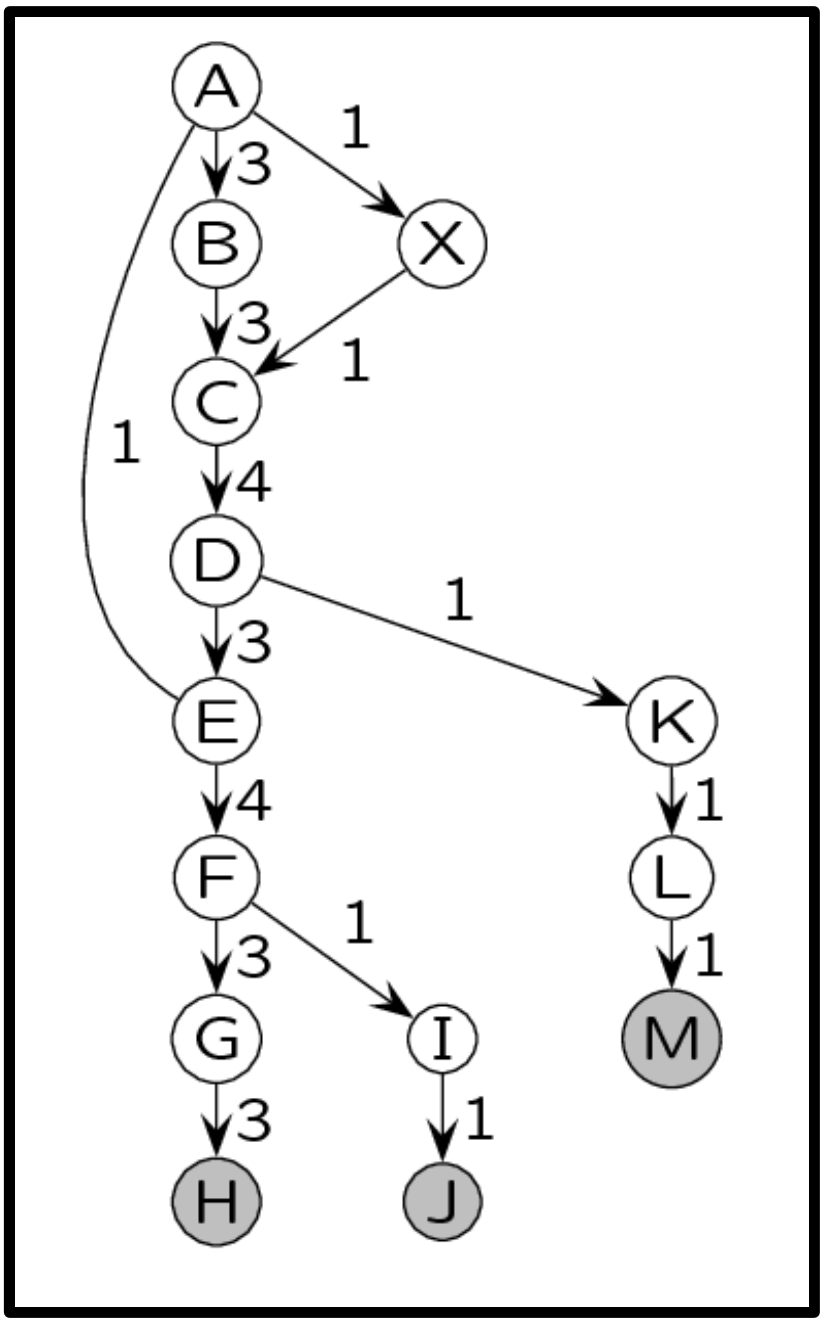
A	A	A	A	A
B	B	B	X	E
C	C	C	C	F
D	D	D	D	G
E	E	K	E	H
F	F	L	F	
G	I	M	G	
H	J		H	

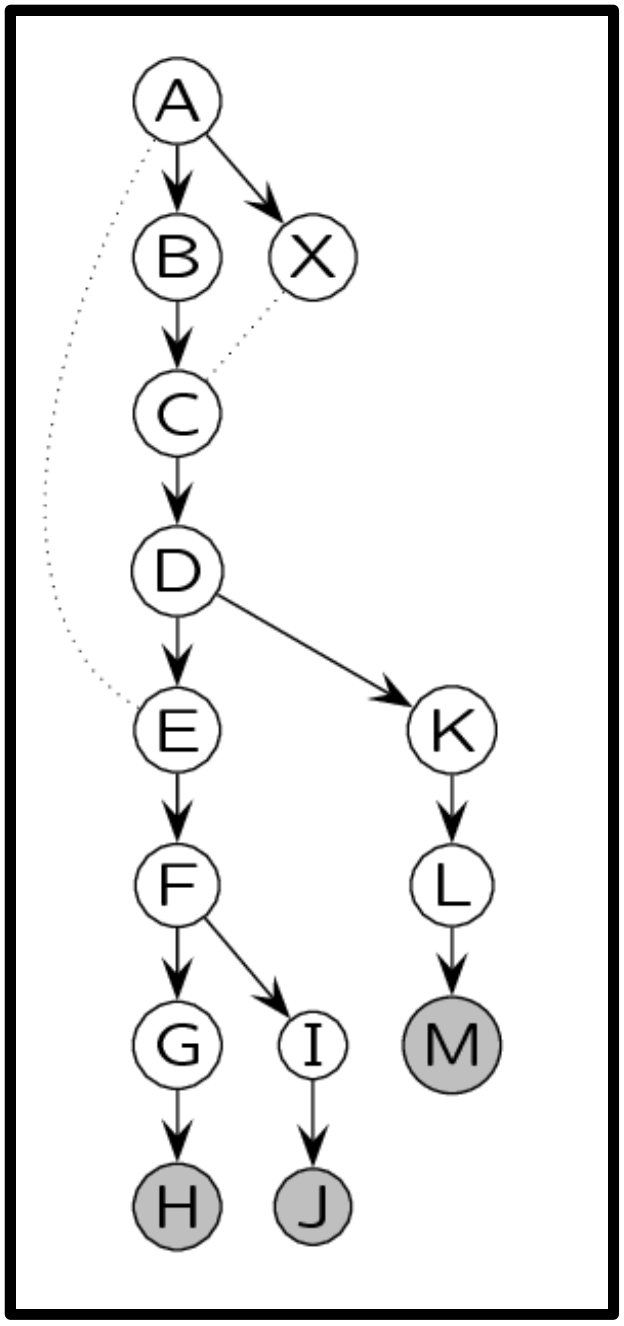


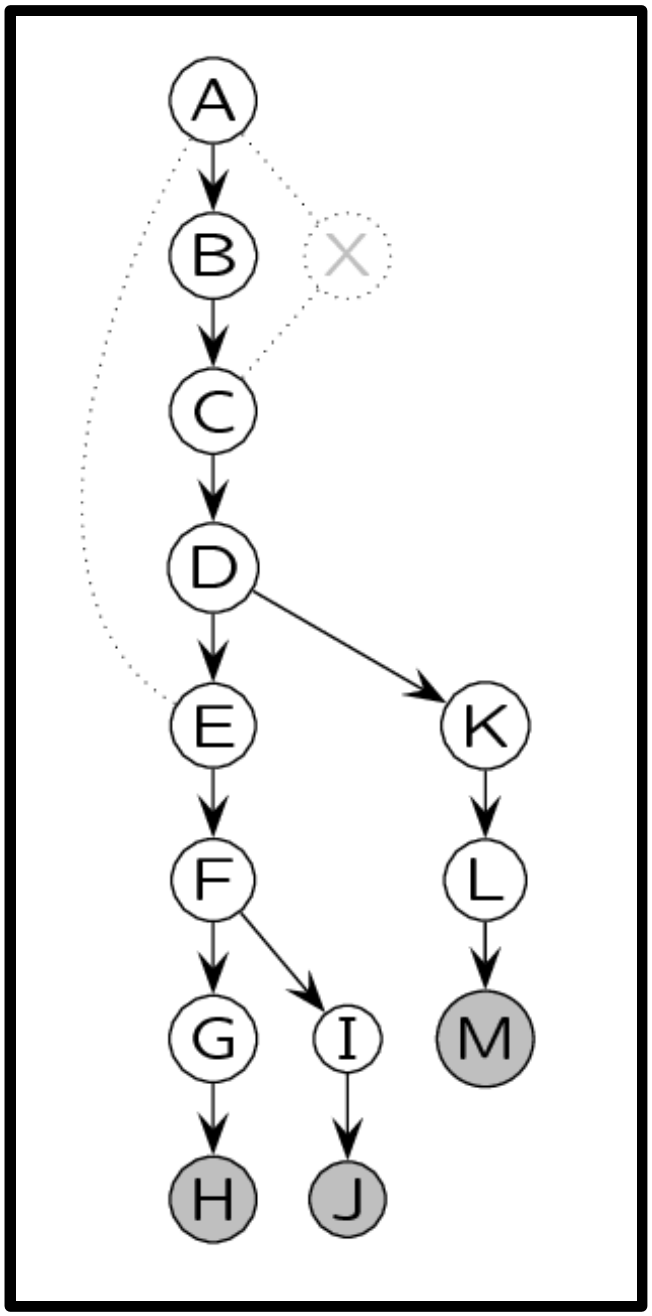

A
B
C
D
E
F
G
H
I
J
K
L
M
X
3
1
3
1
1
4
1
3
4
3
1
1
1


A
B
X
C
D
E
K
F
L
G
I
M
H
J
1
3
1
3
1
4
1
3
4
1
1
3
1
3
1


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


A
B
X
C
D
E
K
F
L
G
I
M
H
J
1
3
1
3
1
1
4
1
3
4
1
1
3
1
3
1
1


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


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

Central question:

How do ideas spread?

Intuition:

exponential growth ("going viral")

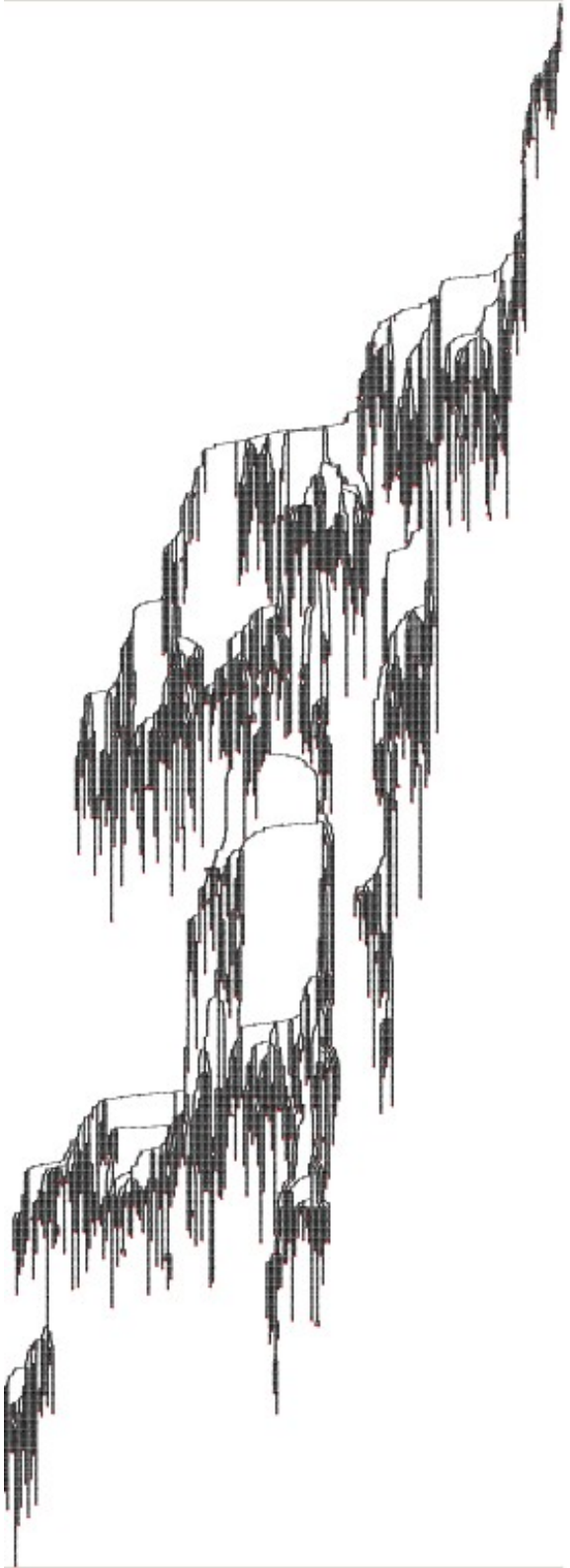
short chains ("small-world phenomenon")

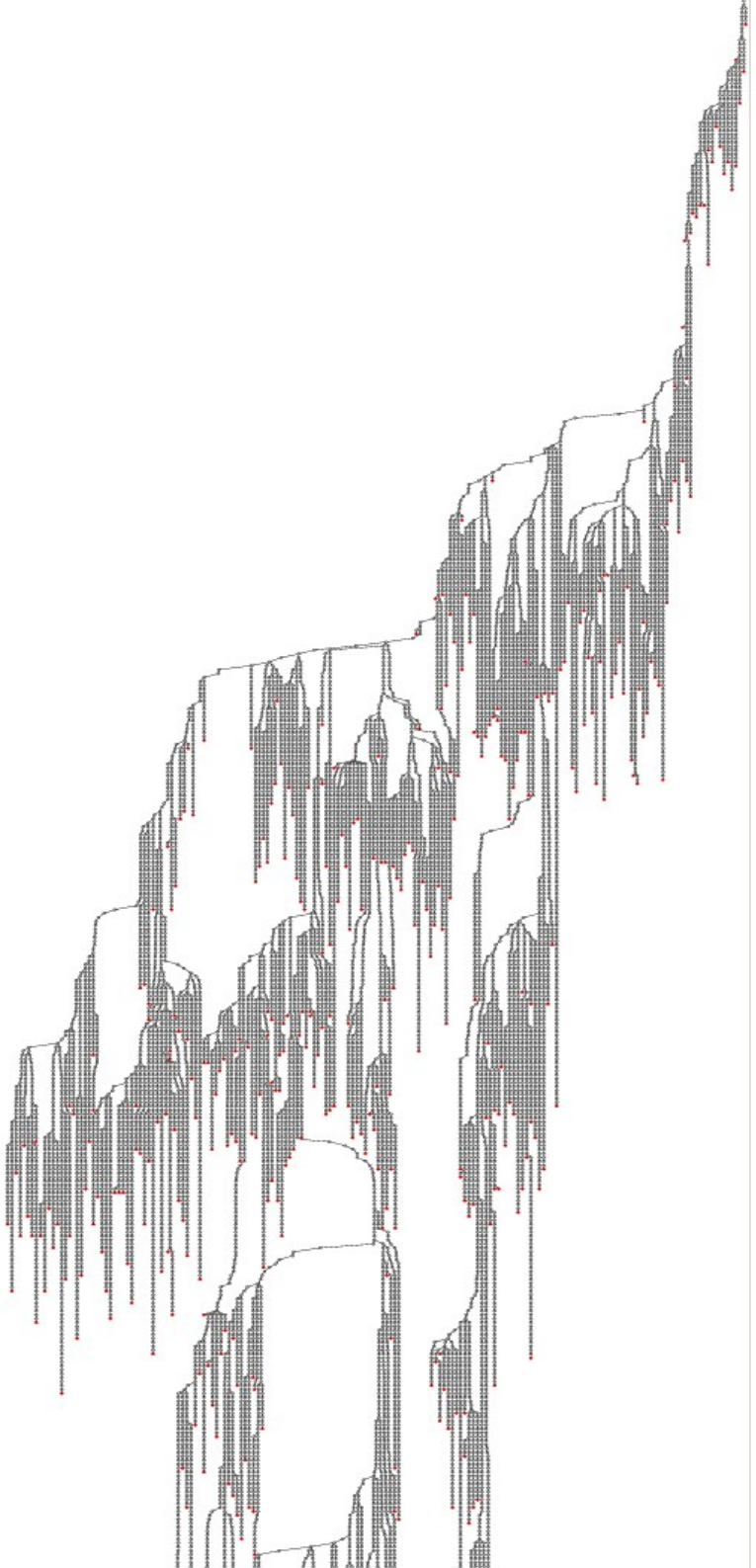
So we'd expect:

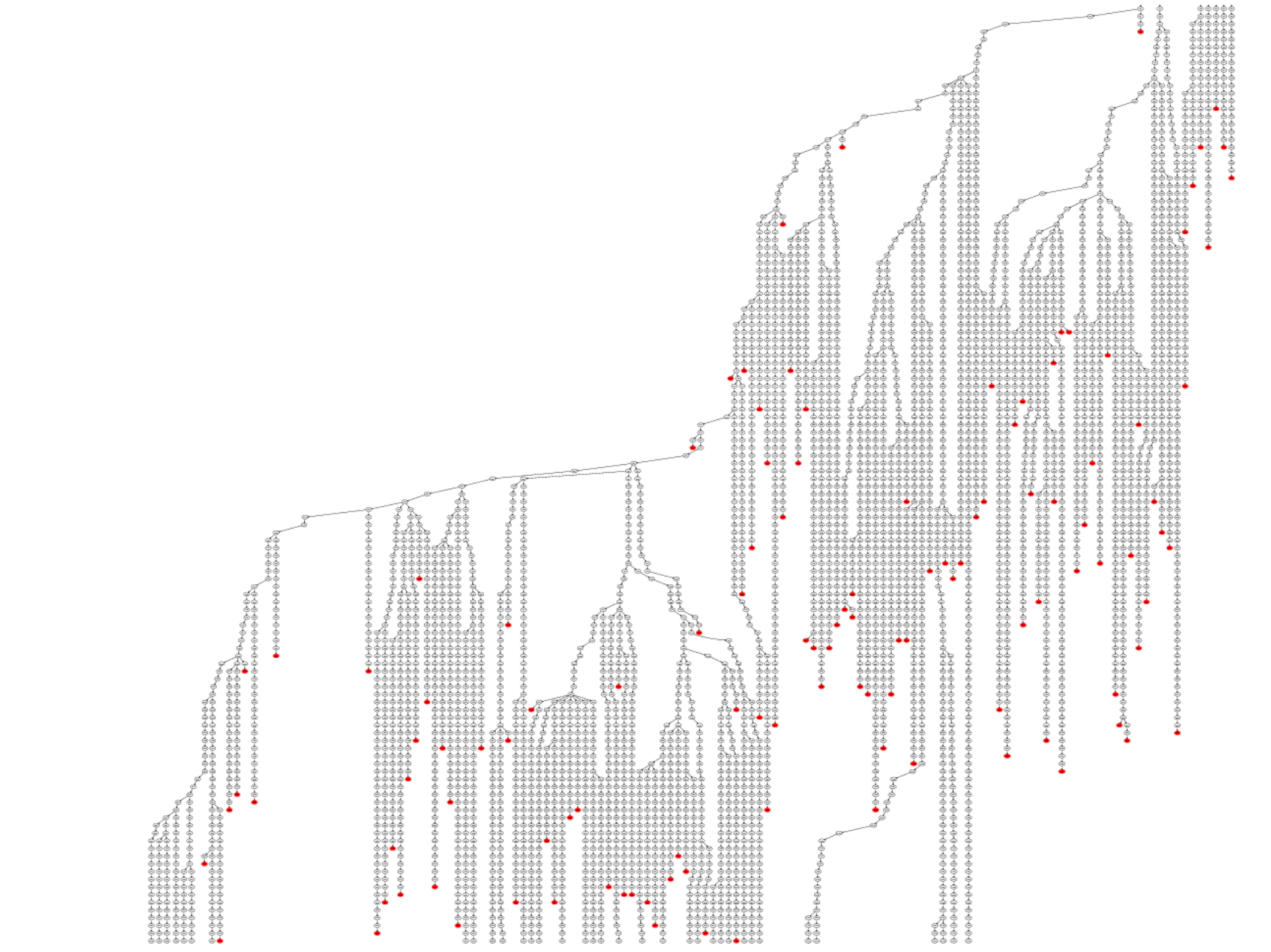
– small depth (small world)

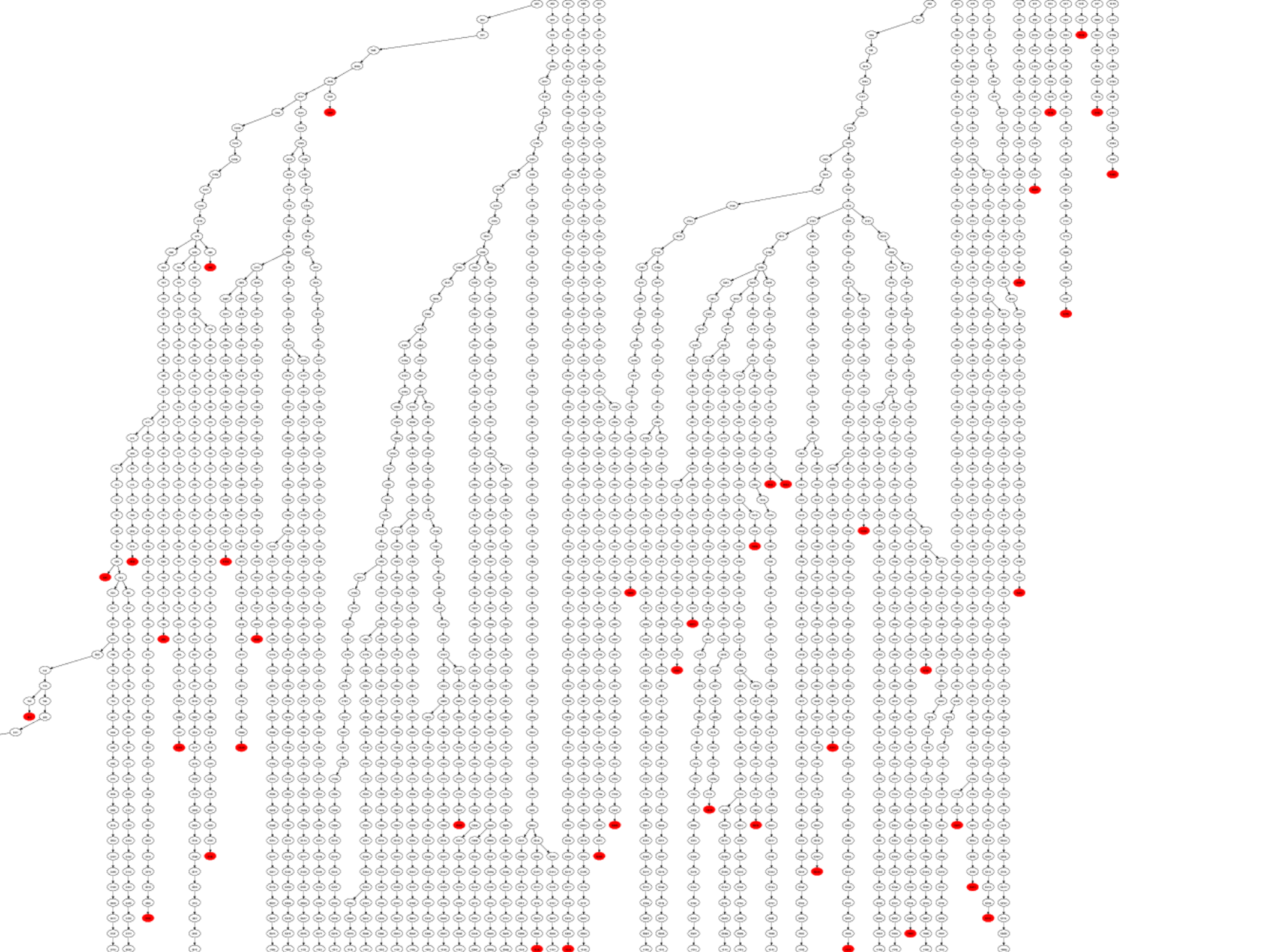
– high branching (10s to 100s of friends)

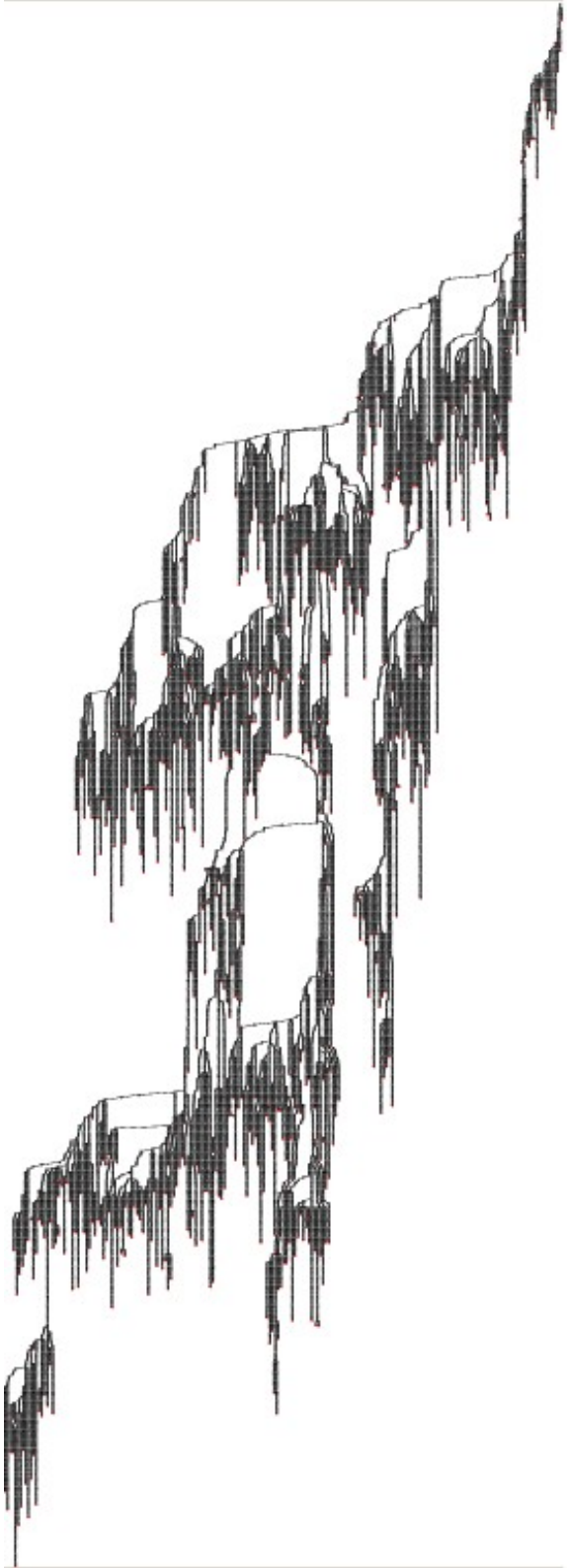
– *shallow & wide propagation tree*

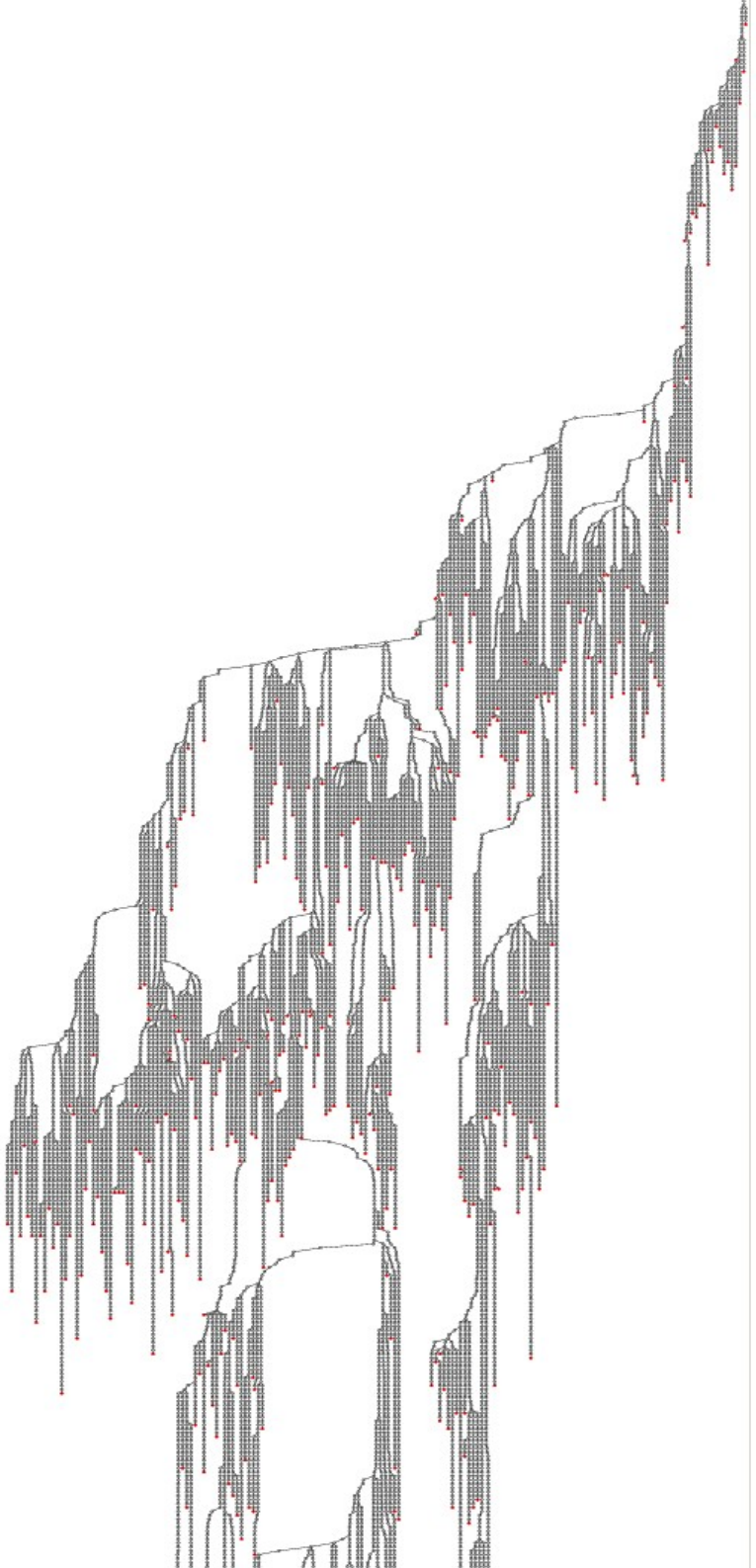


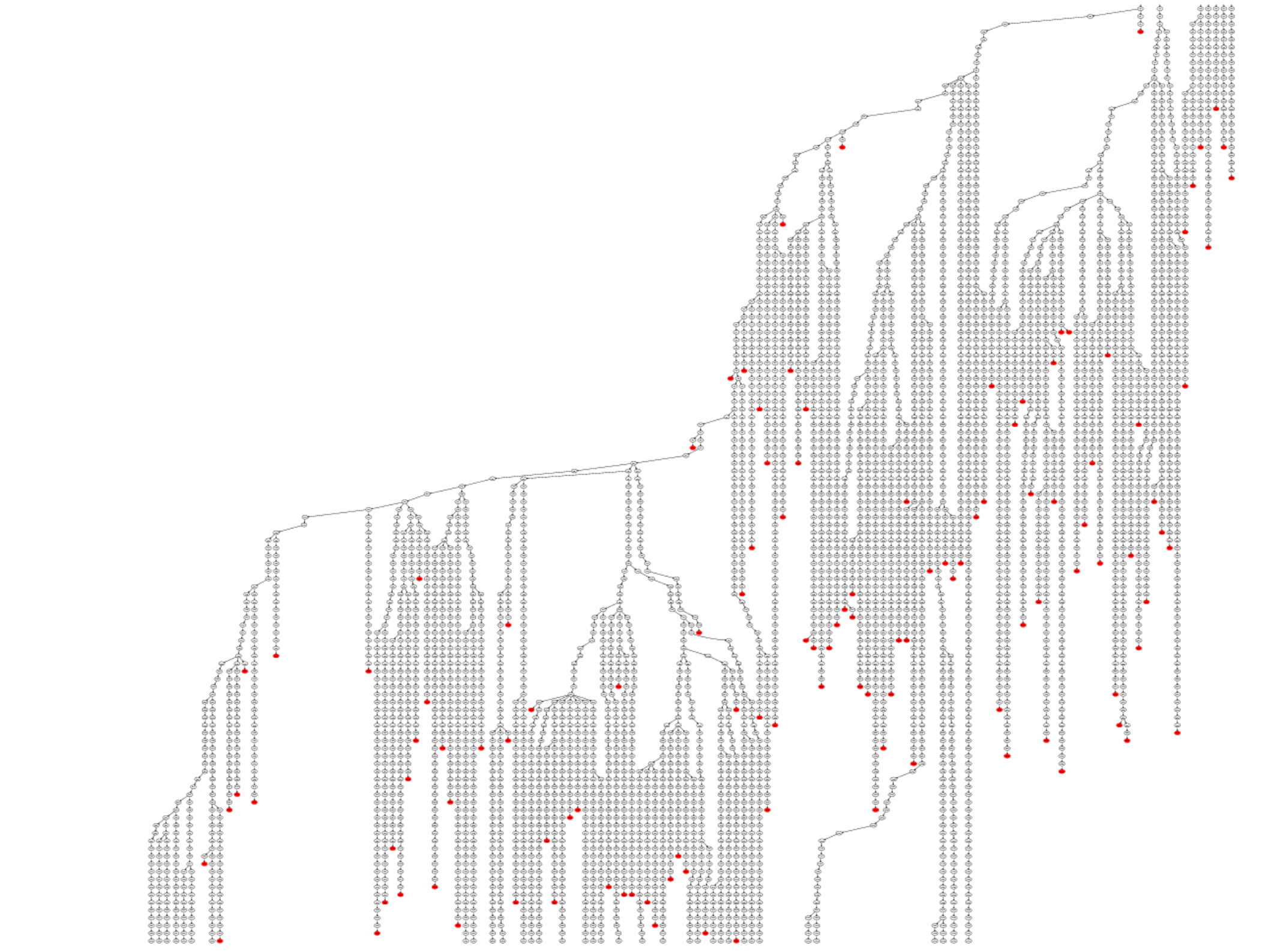


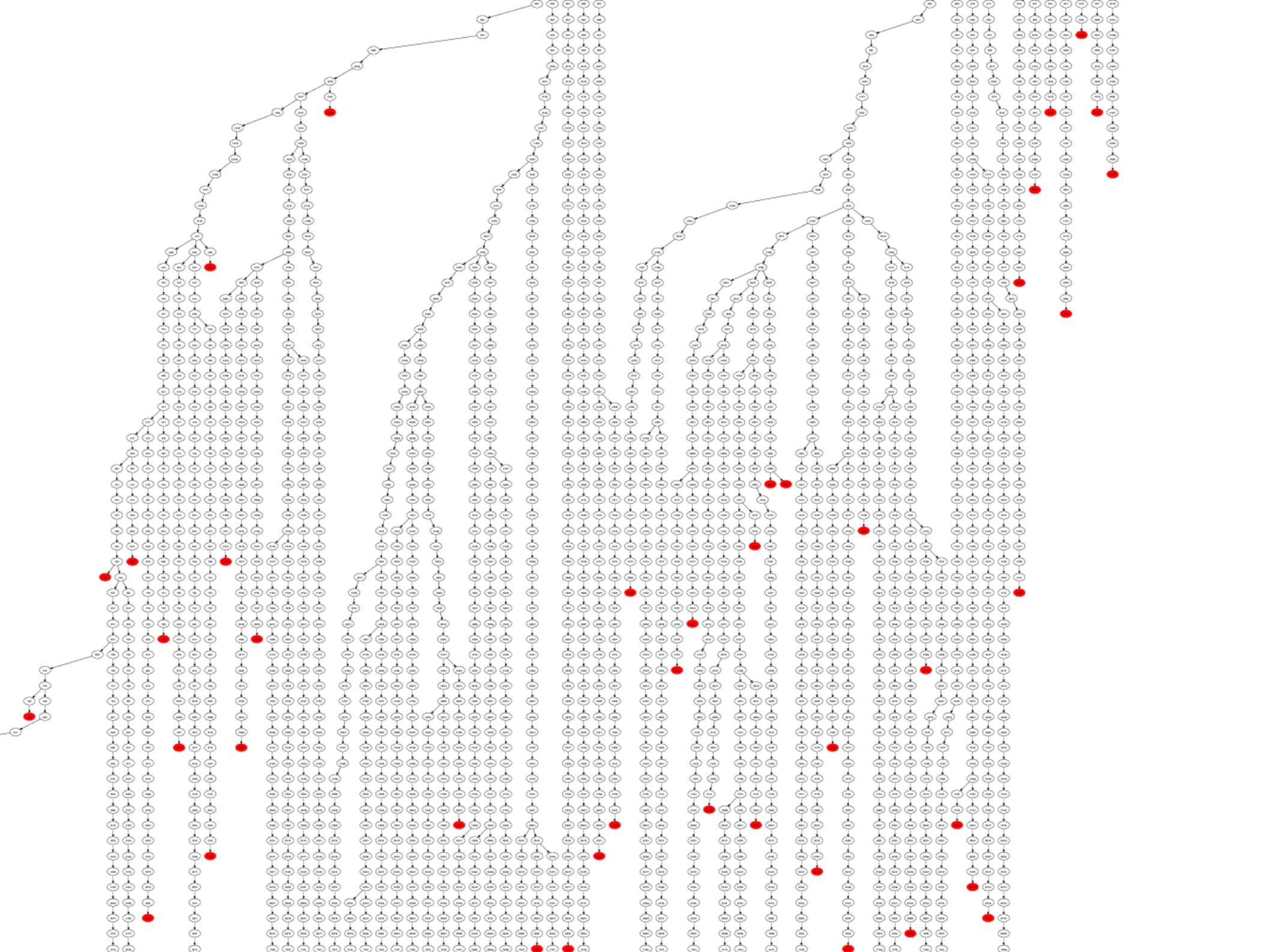








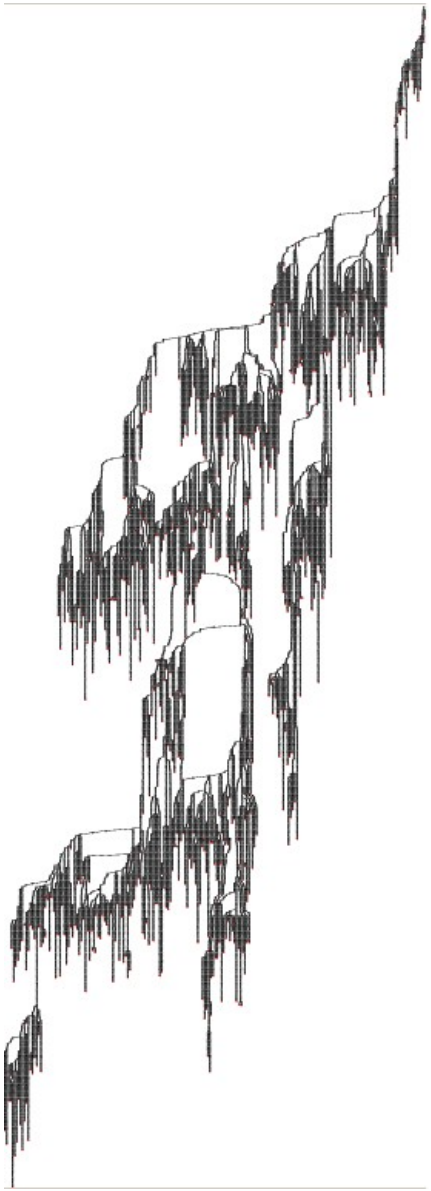




Expectations:

shallow/wide tree with high branching factor

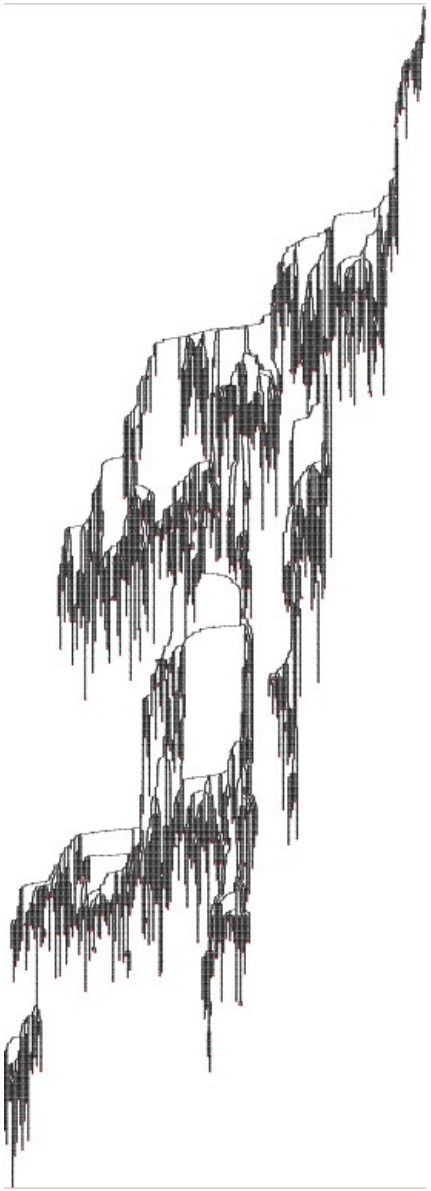
(unless it dies out quickly)



Expectations:

shallow/wide tree with high branching factor

(unless it dies out quickly)



Reality:

- the process doesn't die out quickly

  20K nodes in posted copies

- the tree is very deep.

  median node depth ~288

- *the tree is very narrow.*

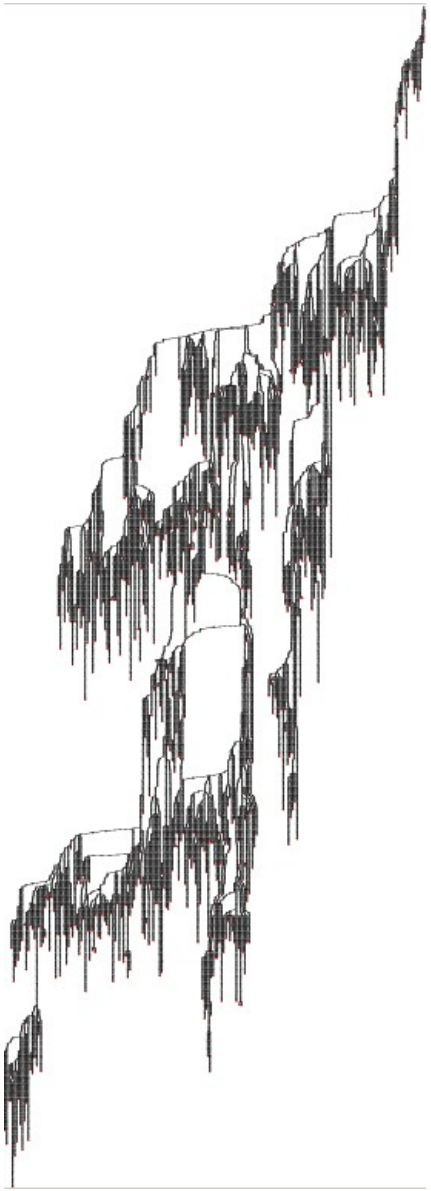
  >94% of nodes have exactly 1 child

Modeling and Implications

(20% of the work)

Modeling goals:

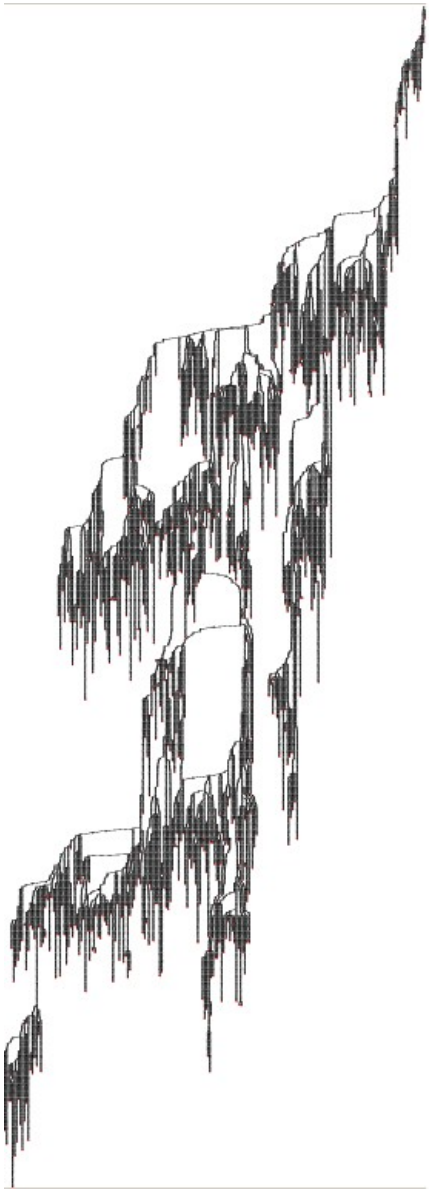
"good" trees: large median depth, small width, high single-child fraction



Goal:

simple, plausible generative model that reproduces the observed features.

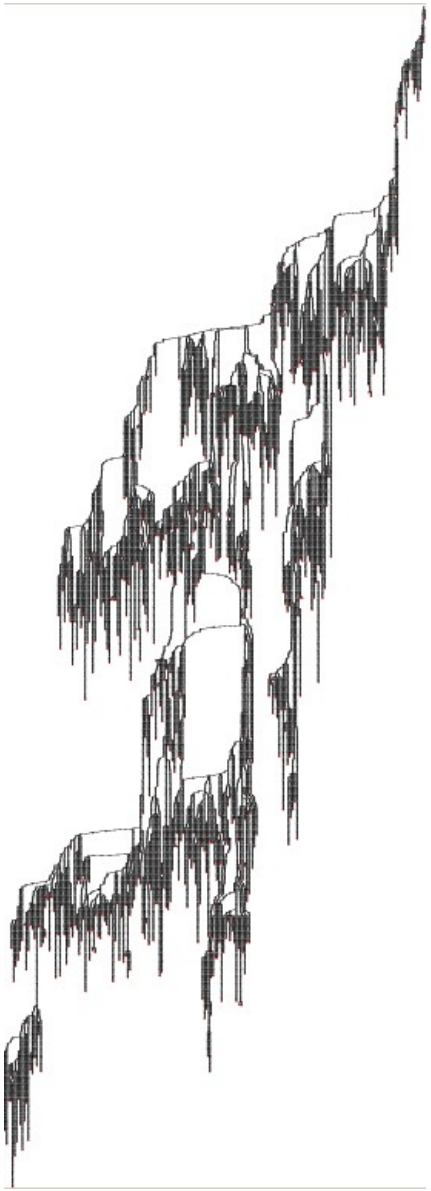
Ingredients for a model:



– *start from random seed (person in the social network).*

– *recipients can discard or forward.*
discard rate $= 0.65$ [Dodds Mohamad Watts 2003]

Ingredients for a model:

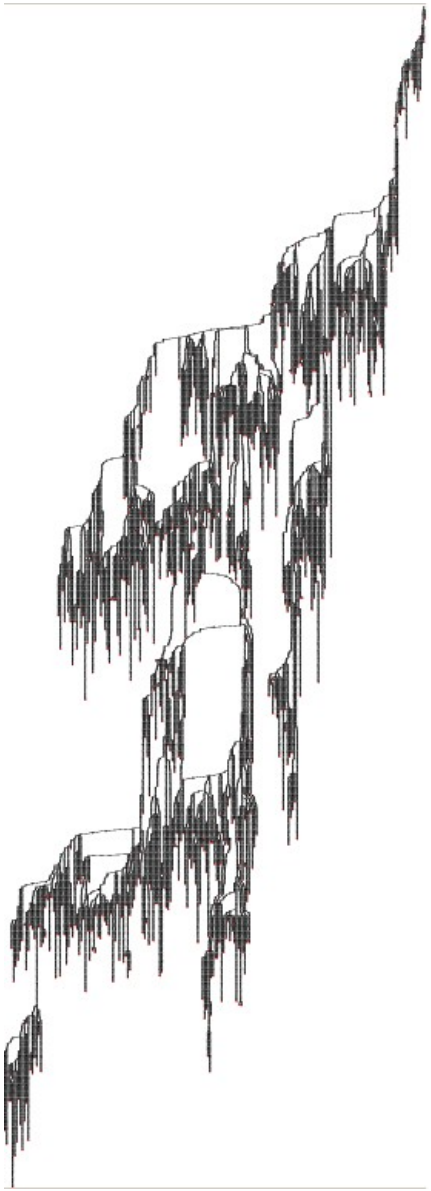


– *start from random seed (person in the social network).*

– *recipients can discard or forward.*
discard rate $= 0.65$ [Dodds Mohamad Watts 2003]

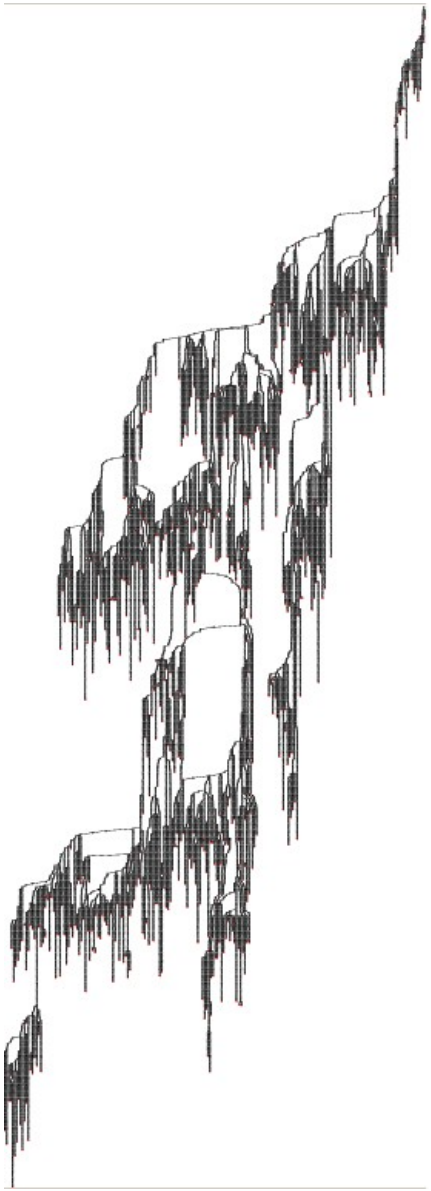
– *forwarders can "post" with some probability; path to root is lit up.*

Ingredients for a model:



- *start from random seed (person in the social network).*
- *recipients can discard or forward.*
  discard rate $= 0.65$ [Dodds Mohamad Watts 2003]
- *forwarders can "post" with some probability; path to root is lit up.*
- *run until fizzle or tree size = Iraq.*

Ingredients for a model:



– *start from random seed (person in the social network).*

– *recipients can discard or forward.*
discard rate = 0.65 [Dodds Mohamad Watts 2003]

– *forwarders can "post" with some probability; path to root is lit up.*

– *run until fizzle or tree size = Iraq.*

Then compare (non-fizzling) *observable tree* statistics to real data.

Model #1: *mechanisms/real networks.*

[DLN Kleinberg 2008]

Test models using real social network data (4.4M LiveJournal nodes).

The epidemic model:
Every non-discarding node forwards to all LJ friends, and posts with some probability p.

Terrible!

Model #1: *mechanisms/real networks.*

[DLN Kleinberg 2008]

What goes wrong with epidemics?

Really about single-child fraction:

95% of nodes in Iraq tree have one child.

How? *(Looks like a DFS tree, not a BFS tree!)*

Model #1: *mechanisms/real networks.*

[DLN Kleinberg 2008]

What goes wrong with epidemics?

Really about single-child fraction:

95% of nodes in Iraq tree have one child.

How? *(Looks like a DFS tree, not a BFS tree!)*

Most people send to $k>1$ friends?

Almost everyone has exactly k-1 discarding friends?

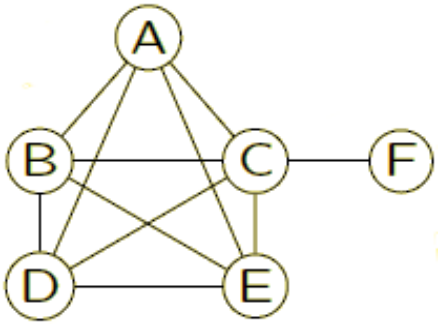
More discarders → fizzle.

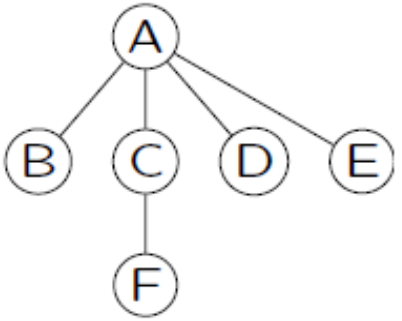
Fewer discarders → not enough single-child nodes.

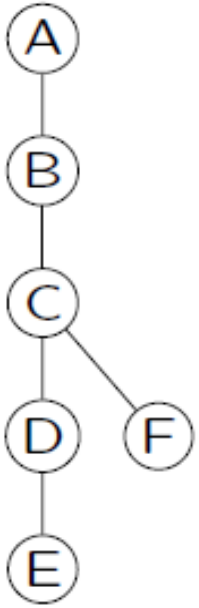
Model #1: *mechanisms/real networks.*

[DLN Kleinberg 2008]

- Serialize cliquey communities using broad distribution of delay in response times.



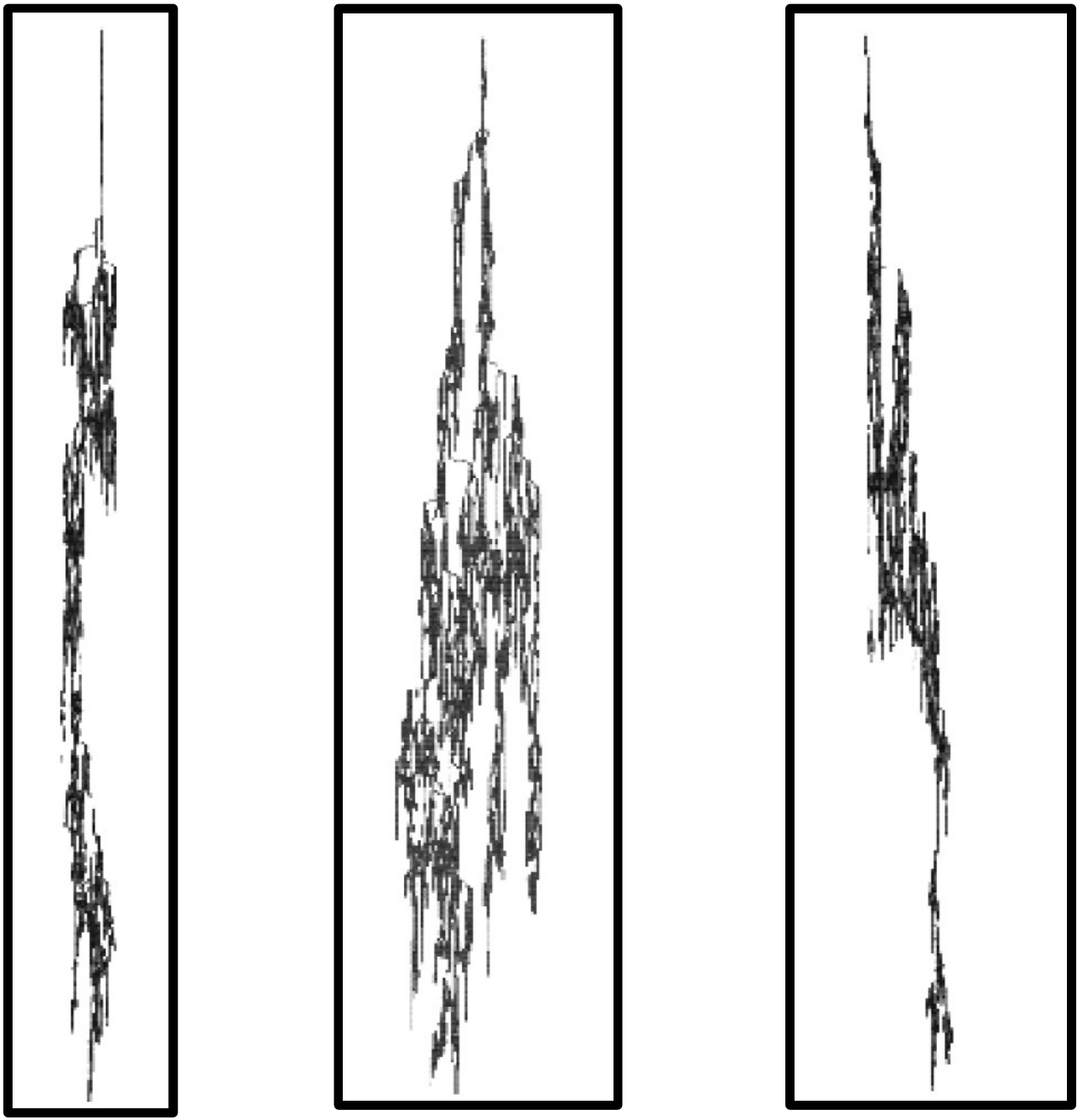




- Allow "reply all": a forwarder can
  - send to all friends; *or*
  - *reply to all corecipients.*

Two parameters: post rate p; reply-all rate b.

Model #1: *mechanisms/real networks.*



Simulations:
$b = 0.950$
$p = 0.22$

"All models are wrong, but some are useful."

– *George Box*

Model #2: *branching processes.*

[Golub Jackson 2010]

$D =$ degree distribution from real Iraq tree.
Define branching process using D.

Model #2: *branching processes.*

[Golub Jackson 2010]

D = degree distribution from real Iraq tree.
Define branching process using D.

E[degree] = (n-1)/n, so BP is barely subcritical.

Most generated trees are too small.

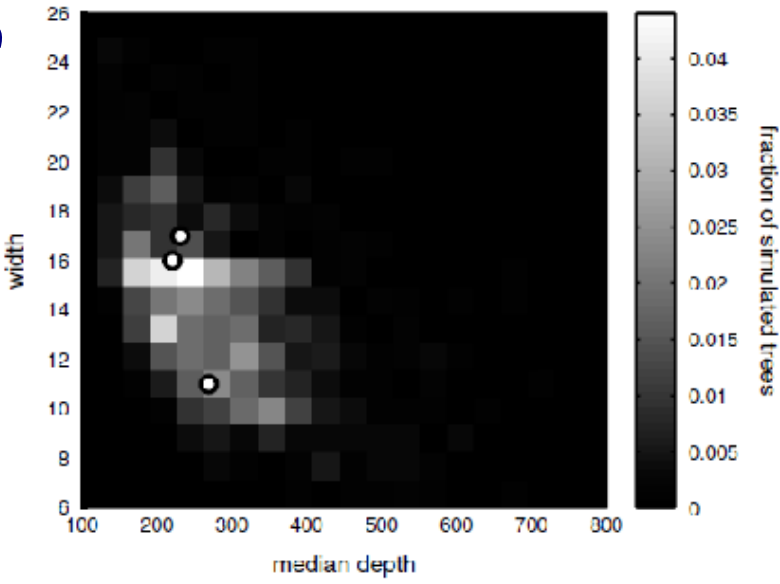
Model #2: *branching processes.*

[Golub Jackson 2010]

D = degree distribution from real Iraq tree.
Define branching process using D.

E[degree] = (n-1)/n, so BP is barely subcritical.

Most generated trees are too small.
But *conditioned on observable tree reaching Iraq size*, simulation shows depth/width of real Iraq tree is typical of trees generated by BP.



Model #2: *branching processes.*

[Golub Jackson 2010]

Golub/Jackson Branching Process Model:
Define branching process using real observed degree distribution D; condition on reaching Iraq size.

Leaves one central question: *how did observable degree distribution come to be D?*

Model #2: *branching processes.*

[Golub Jackson 2010]

Golub/Jackson Branching Process Model:
Define branching process using real observed degree distribution D; condition on reaching Iraq size.

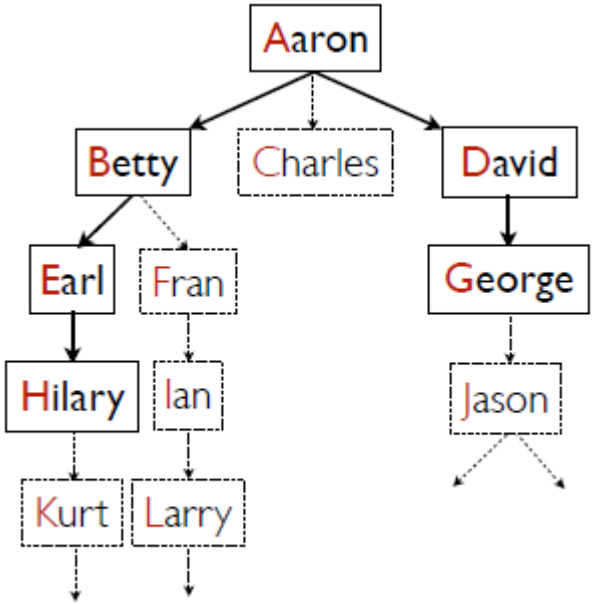
Leaves one central question: *how did observable degree distribution come to be D?*

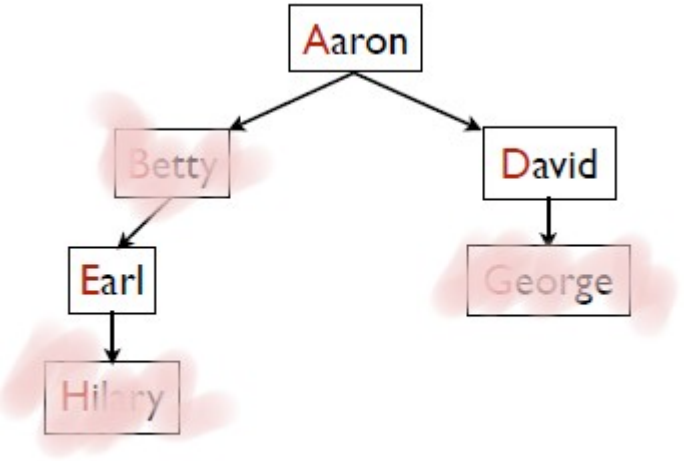
One way: [Golub Jackson 2010]
Forward to fixed number of friends with appropriate discard/posting rate.
Roughly, epidemic model without an underlying network.

Model #3: *it's all about observation.*

[Chierichetti Kleinberg DLN 2011]

Consider an arbitrary underlying tree T.
Let each node *expose* itself independently with probability p, revealing its path to root.
Let $T[p]$ denote result: observed tree (random).





Model #3: *it's all about observation.*

[Chierichetti Kleinberg DLN 2011]

Theorem:
If T's max degree is bounded and p is small enough, $T[p]$'s single-child fraction is $1-o(1)$.

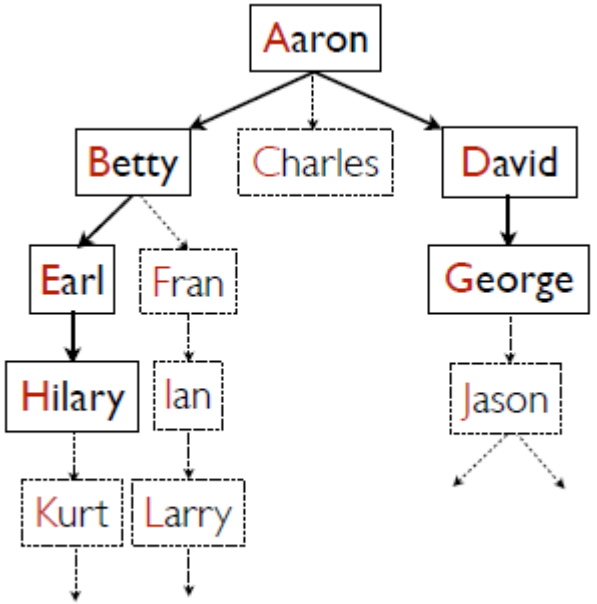
(In other words, the tree necessarily looks this way because of the way we *observe* it.)

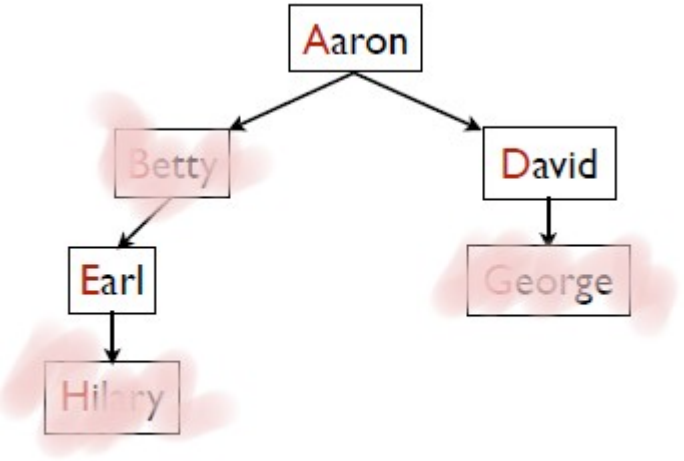
Proof idea: with high probability, the exposure process reveals many more internal nodes than leaves (or branch points).

Model #3: *it's all about observation.*

[Chierichetti Kleinberg DLN 2011]

We have a strange window (the observed tree) into true spreading process. But everything interesting is about what's invisible (the unobserved tree). *What can we say?*





Model #3: *it's all about observation.*

[Chierichetti Kleinberg DLN 2011]

Theorem 2:
If T's max degree is bounded, we can accurately estimate $|T|$ from $|T[p]|$, even without knowing p.

Estimate p as fraction of internal nodes of T[p] *that are exposed. Estimate* |T| *as (#exposers/p).*

For Iraq: 18K nodes in T[p] → *173K nodes in* T.

1) finding data on spread of a single piece of information is hard.

2) propagation tree has unexpected structure.

3) we can explain that structure with a model.

1) finding data on spread of a single piece of information is hard.

(chain letters!)

2) propagation tree has unexpected structure.

(narrow, deep, and stringy!)

3) we can explain that structure with a model.

(with a model that is wrong!)

Even the sampling process is oversimplified; posting decisions don't seem independent.

1) finding data on spread of a single piece of information is hard.

(chain letters!)

2) propagation tree has unexpected structure.

(narrow, deep, and stringy!)

3) we can explain that structure with a model.

(with a model that is wrong!)

Even the sampling process is oversimplified; posting decisions don't seem independent.

4) what else can we learn?

(size of underlying propagation? And??)

Patterns of Information Diffusion

David Liben-Nowell

dlibenno@carleton.edu

Carleton College, Department of Computer Science

Thank you!