Meme-tracking, Diffusion, and the Flow of On-Line Information

Jon Kleinberg

Cornell University

Including joint work with Lars Backstrom, Jure Leskovec, and David Liben-Nowell

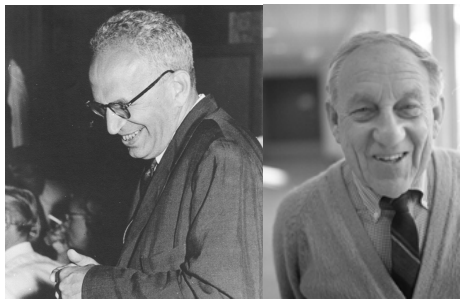
Information and the Media

Intersection of news media, technology, and the political process.

A rich history:

Lazarsfeld, Berelson, and Gaudet, *The People's Choice*, 1944, 1948.

Elihu Katz and Paul Lazarsfeld, *Personal Influence*, 1955.



From its early stages, a tension between global effects from the media and local effects carried by social structure.

- How does information transmitted broadly by the media interact with the personal influence arising from social networks?

Fragmentation and Acceleration

These issues increasingly involve considerations of the Internet, blogging, and social media.

- Social media means the dichotomy between global and local influence is evaporating, approaching more of a continuum.
- Speed of media reporting and discussion has intensified: very rapid progression of stories, with no pauses.

These forces contribute to the "24-hour news cycle":

- Difficult to define, but associated with technological acceleration and a challenge to healthy civic discourse.
  e.g. [Kovach-Rosenstiel 1999]

The News Cycle

Media theorists relate the exceeses of the news cycle to a "convergence of factors:"

"the news media's growing valorization of tabloidization and infotainment and audiences' consumption of it, and the importance and effects of speed on news practices, due to market competition; influence of the internet (especially blogs); the breakdown of traditional news gatekeeping; the fragmentation of mediated political discourse into soundbites on the news, and the tailoring of political messages to these constraints of the news format; and the 20th century drive for technocratization of mass democracies, accompanied by the infiltration of political communication by public relations, itself modeled on the strategies and successes of war-time propaganda."

— Jayson Harsin

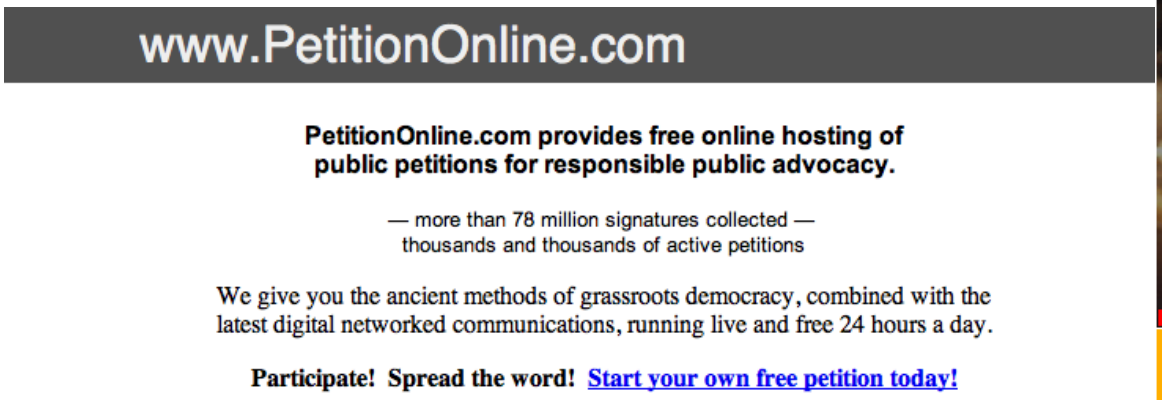
The Rumor that "John Kerry is French," i.e. Haughty, Foppish, Elitist, Socialist, Cowardly and Gay,

2007

On-Line Activism

On-line media are adding similar levels of complexity to how political information flows through social networks

- The spread of rumors and problems of credibility.
  [Fisher 1998, Harsin 2008]
- Traditional methods of political activism:
  petitions, boycotts, letter-writing campaigns
  [Ayres 1999, Earl and Schussman 2003]



Mechanisms of News and Information Flow

Plan for the talk:

Analyzing the underlying mechanisms in some of these processes.

- The dynamics of the global news cycle.
  [Leskovec-Backstrom-Kleinberg 2009]
- The spread of political messages through social networks.
  [LibenNowell-Kleinberg 2008]

Related to work on

- methods for on-line tracking, aggregation, visualization, and personalization of news
  [Allan et al 98, Gabrilovich et al 04, Gamon et al 08, Godbole et al 07, Havre et al 00, Yang et al 98]
- methods for analyzing cascading information on-line
  [Adar et al 04, Backstrom et al 06, Gruhl et al 04, Leskovec et al 06, 07]
- analysis of political blogging
  [Adamic-Glance 05, Mullen-Malouf 06, Singer 05, Tremayne et al 06]

Defining the News Cycle

Sept. 11, 2008 (New York Times): "Mr. McCain's increasingly aggressive campaign has sought to put Mr. Obama on the defensive in each news cycle, using any development at hand, like Mr. Obama's colloquial comment this week about 'putting lipstick on a pig.' "

Oct. 10, 2008 (New York Times): "Mr. McCain's traveling road show has veered from message to message ... . Each news cycle seems to bring another tactic as the campaign appears to be trying anything and everything to see what might work."

- Question: Is the "news cycle" simply a metaphorical construct describing our perception of the news, or is it visible in data? And if it's visible, can we measure some of its basic properties?
- To address this: data from Spinn3r on the three months leading up to the 2008 U.S. Presidential Election: about 1 millions news articles and blog posts per day, from 20,000 news sources, plus millions of blogs and forums.

Units of Analysis

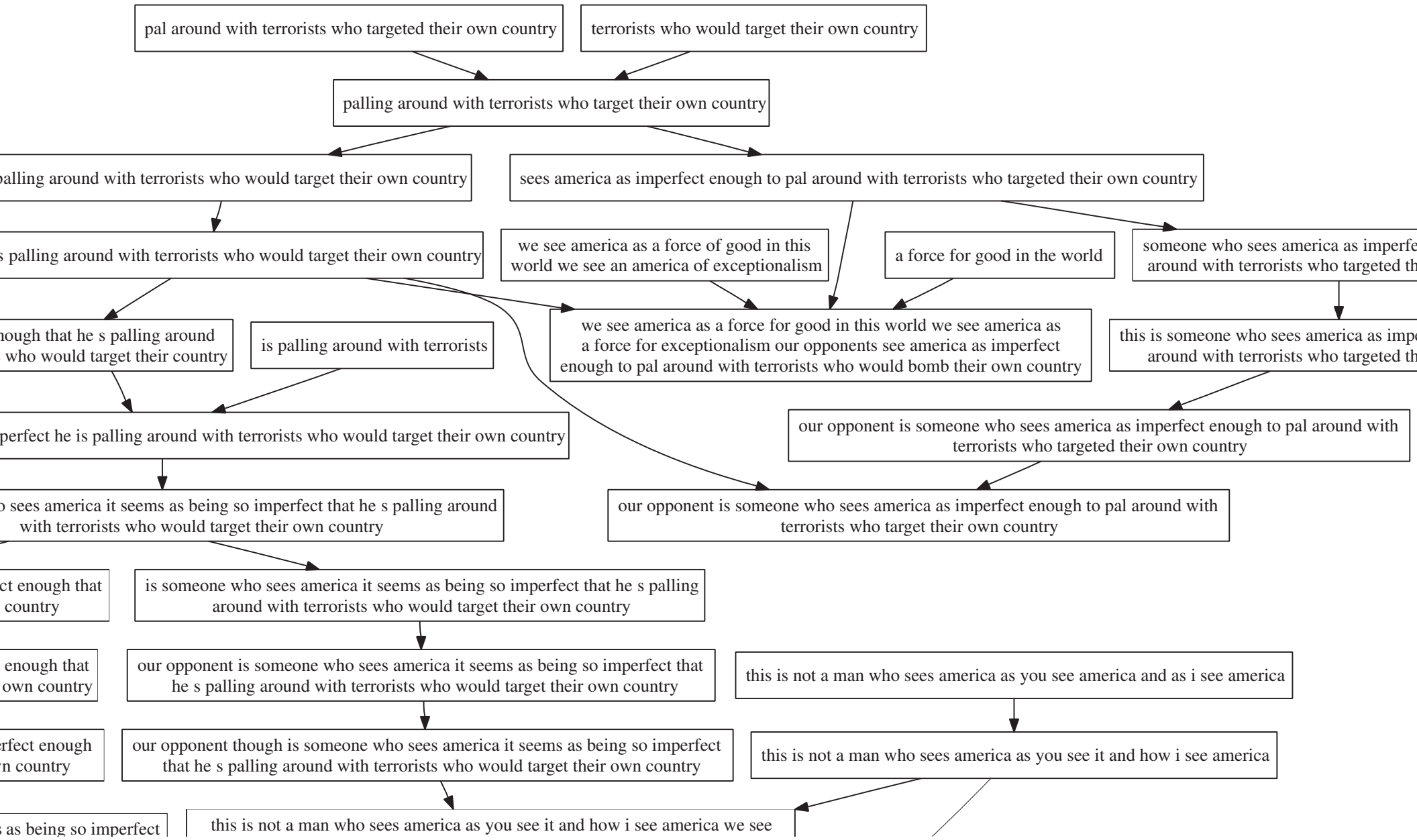
First question: what basic "units" make up the news cycle?

- Cascasding links to individual articles: too fine-grained.
  [Adar et al 04, Gruhl et al 04, Kumar et al 03, Leskovec et al 07]
- Topics as probabilistic term mixtures: too coarse-grained.
  [Blei-Lafferty 06, Wang-McCallum 06, Wang et al 07]

Need units of analysis that correspond to aggregates of articles, vary over the order of days, and can be handled at the half-terabyte scale of the data.

- Plan: identify text fragments, phrase, quotes, "memes" that travel relatively unchanged through many articles.
- Specifically: in Spinn3r data, look for phrases inside quotes. (About 1.25 quotes per document in the data.)

Challenge: Quotes Mutate ... A Lot



Creating Clusters of Mutational Variants

B C D X Y Z

B C D

A B C D

A B C D E F G H

A B C

A B C E F G

A B C E F

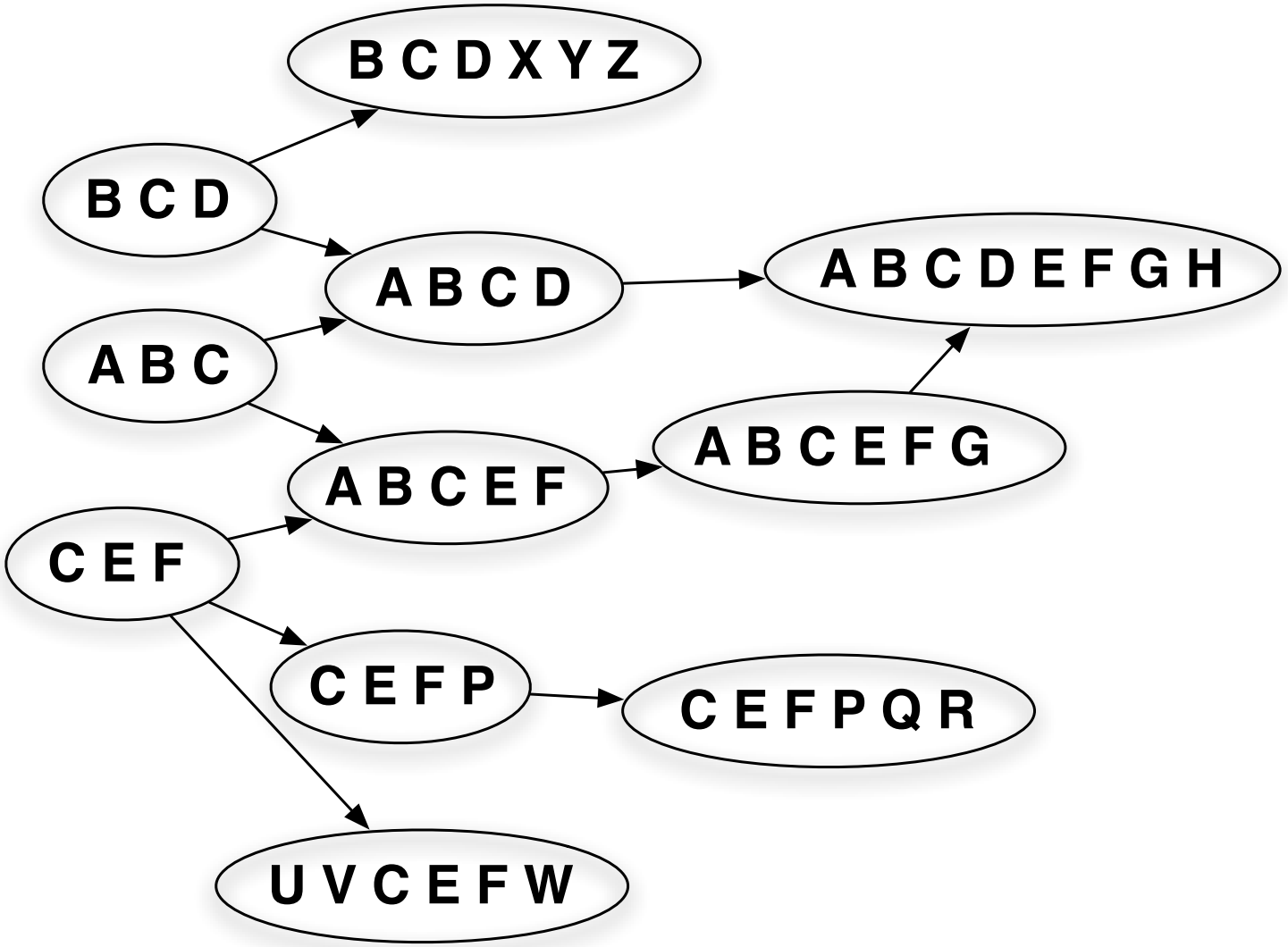
C E F

C E F P

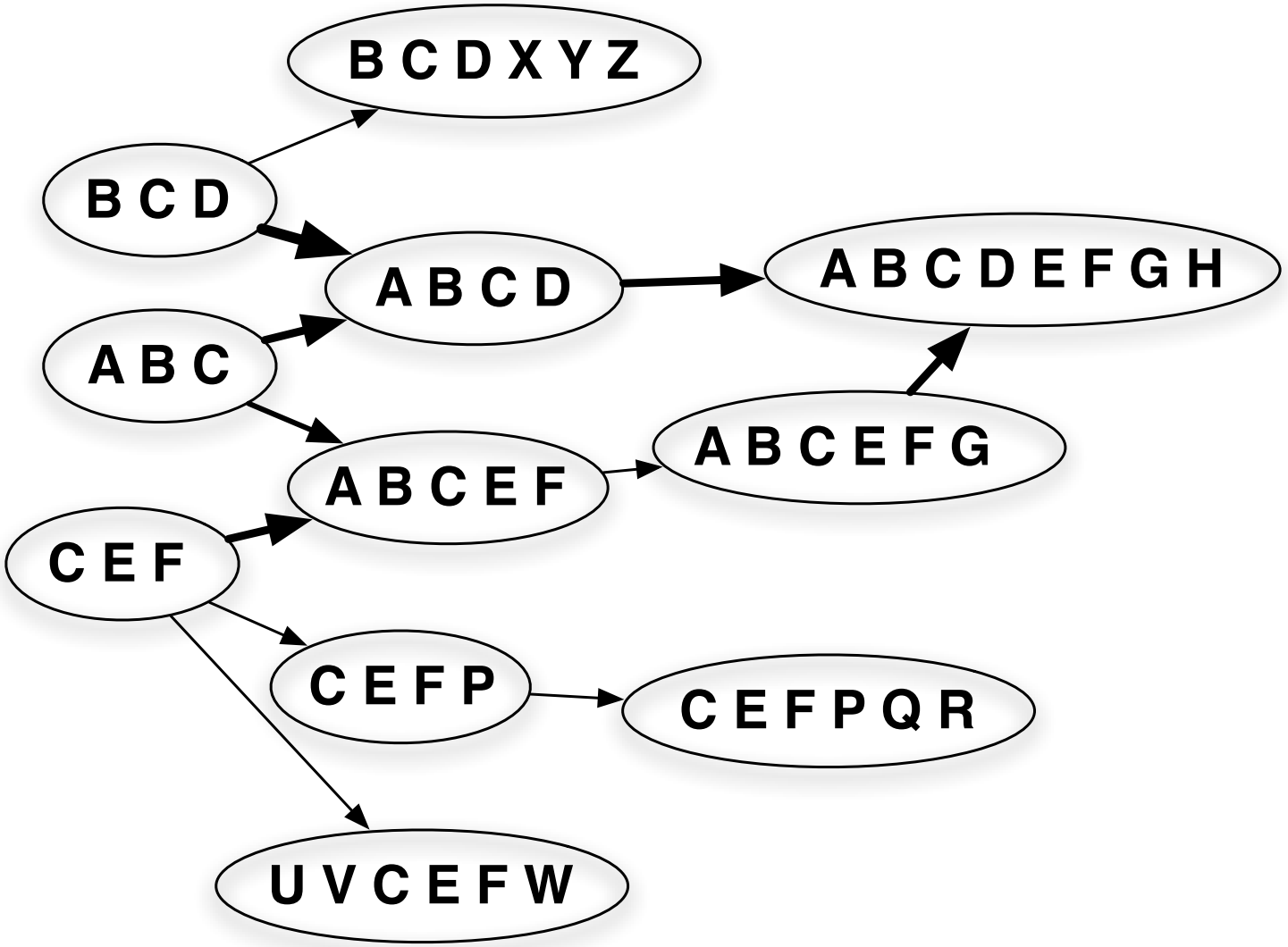
C E F P Q R

U V C E F W

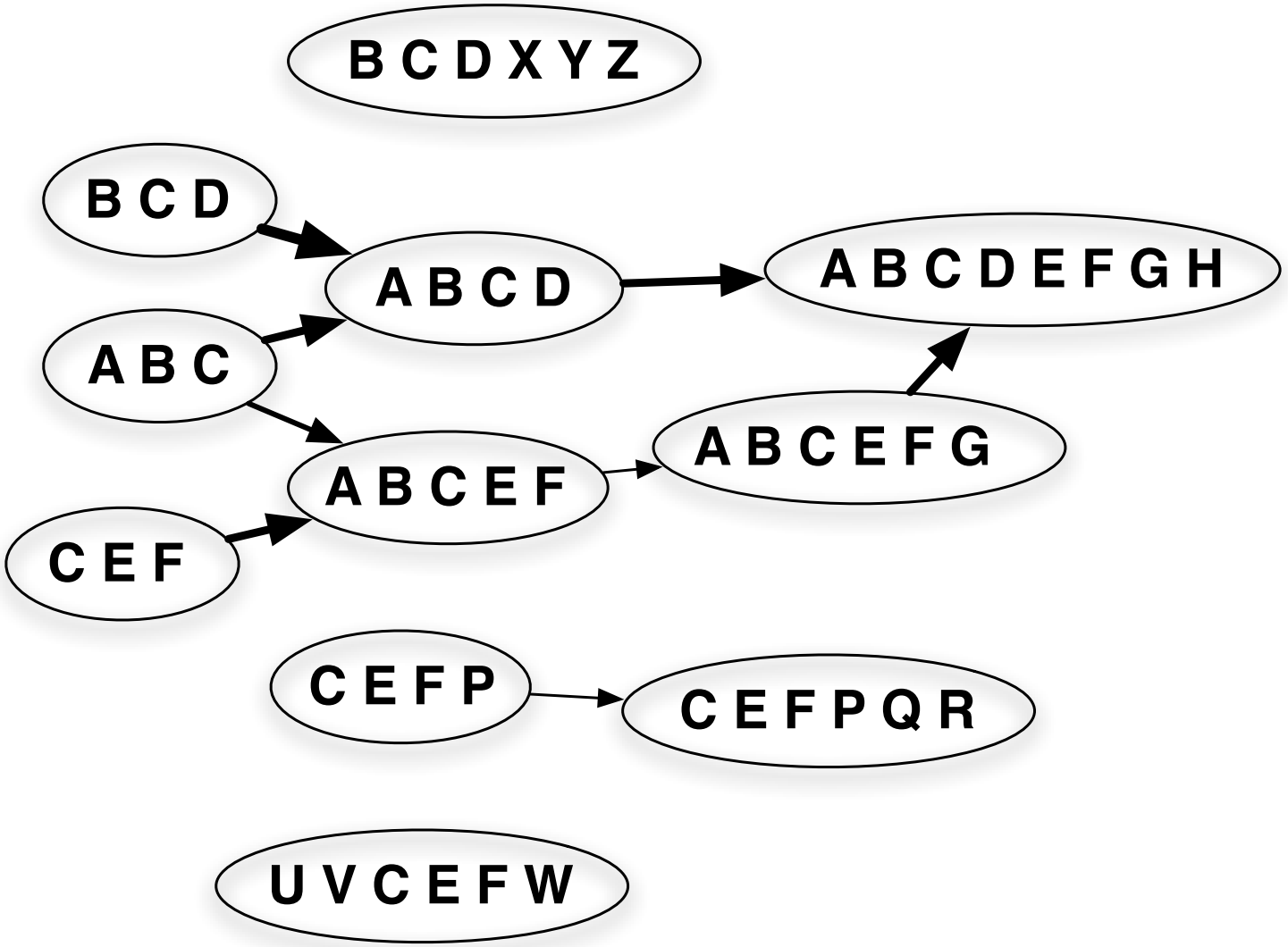
Creating Clusters of Mutational Variants



Creating Clusters of Mutational Variants

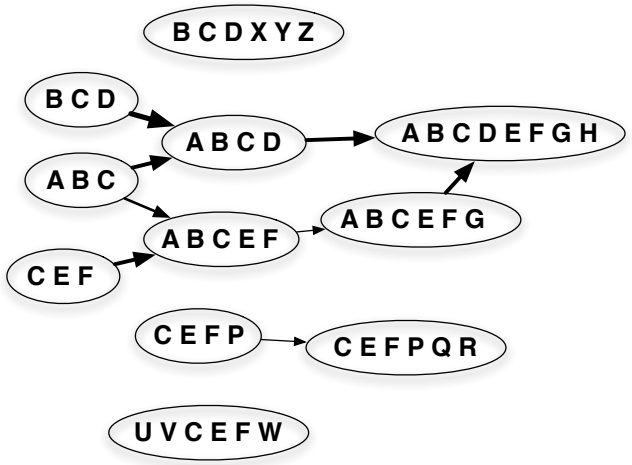


Creating Clusters of Mutational Variants



Quote Clusters

Objective: in directed acyclic graph (approximate quote inclusion), delete min. total edge weight s.t. each component has a single "sink" node.



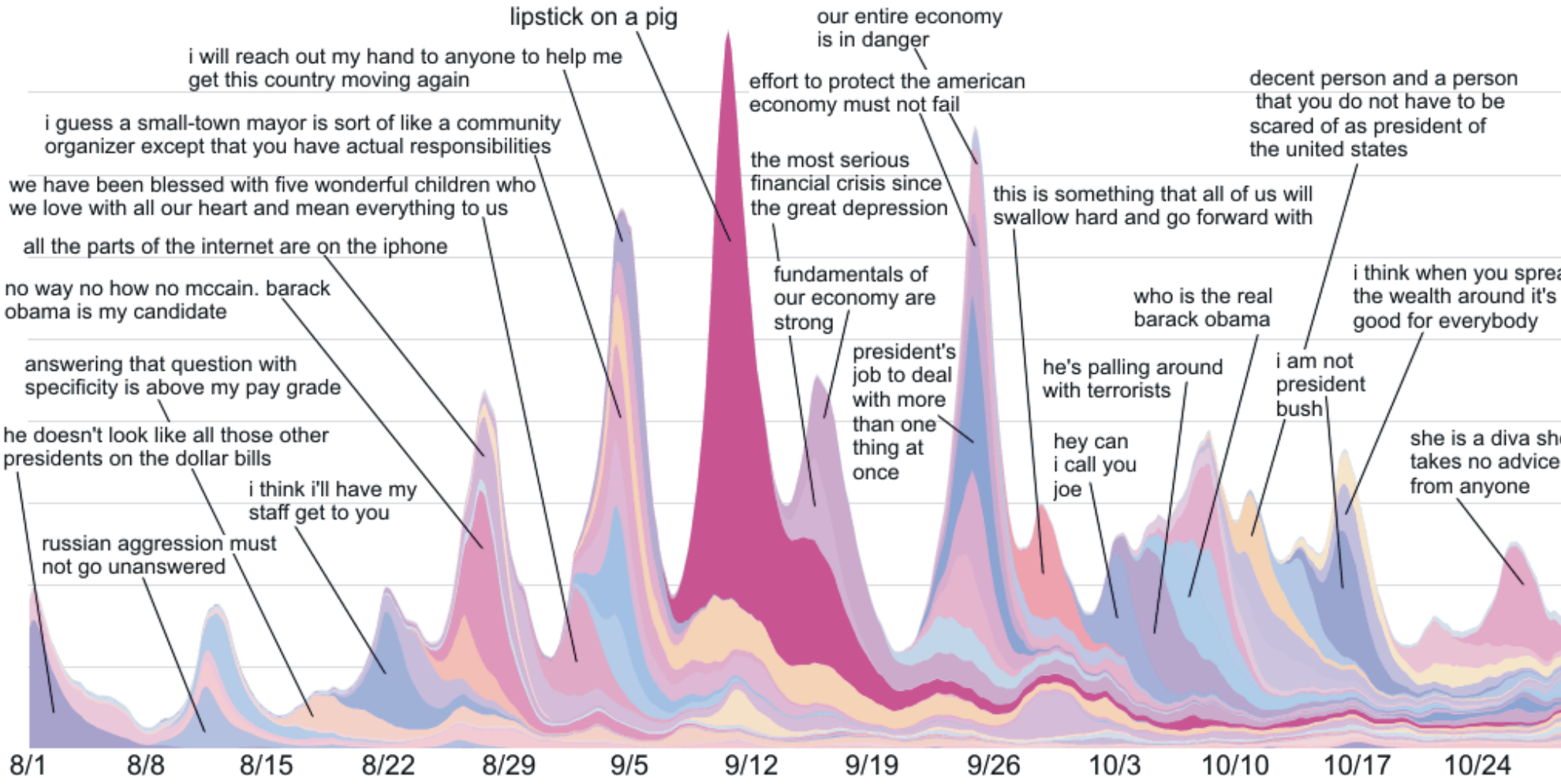
Our basic units will be clusters of quotes.

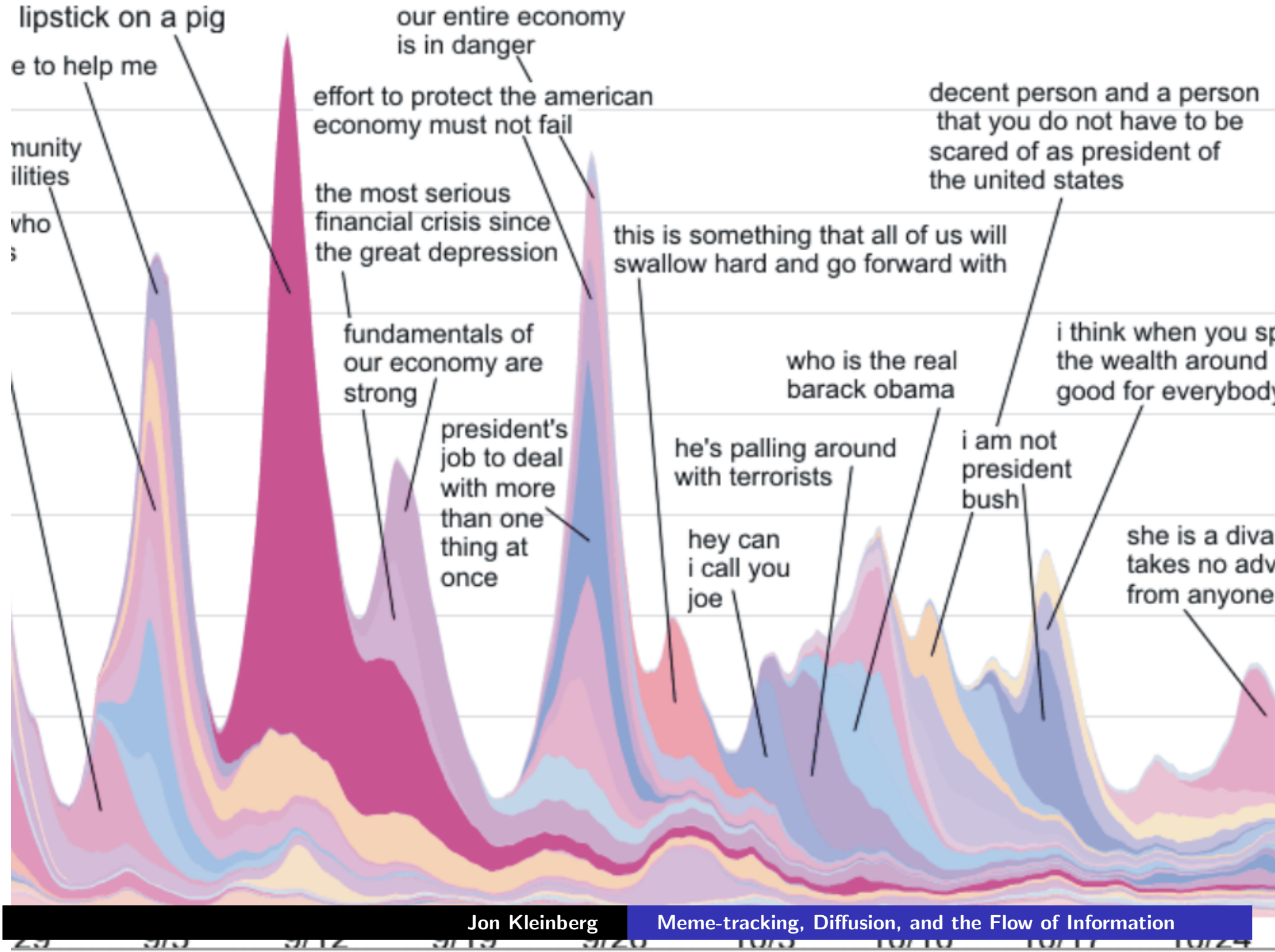
- Pre-processing: length ≥ 4, frequency ≥ 10, no more than $1/4$ of occurrences from a single domain.
- Produces 22 million distinct quotes.

Our graph partitioning formulation is NP-hard (related to multiway cut), but heuristics based on selecting a single edge out of each quote are effective.

- After graph partitioning, $\sim 35,000$ non-trivial components.

Cluster Volume over Time




lipstick on a pig
e to help me
nunity
ilities
who
our entire economy
is in danger
effort to protect the american
economy must not fail
the most serious
financial crisis since
the great depression
fundamentals of
our economy are
strong
president's
job to deal
with more
than one
thing at
once
this is something that all of us will
swallow hard and go forward with
he's palling around
with terrorists
hey can
i call you
joe
who is the real
barack obama
decent person and a person
that you do not have to be
scared of as president of
the united states
i am not
president
bush
i think when you sp
the wealth around
good for everybody
she is a diva
takes no adv
from anyone

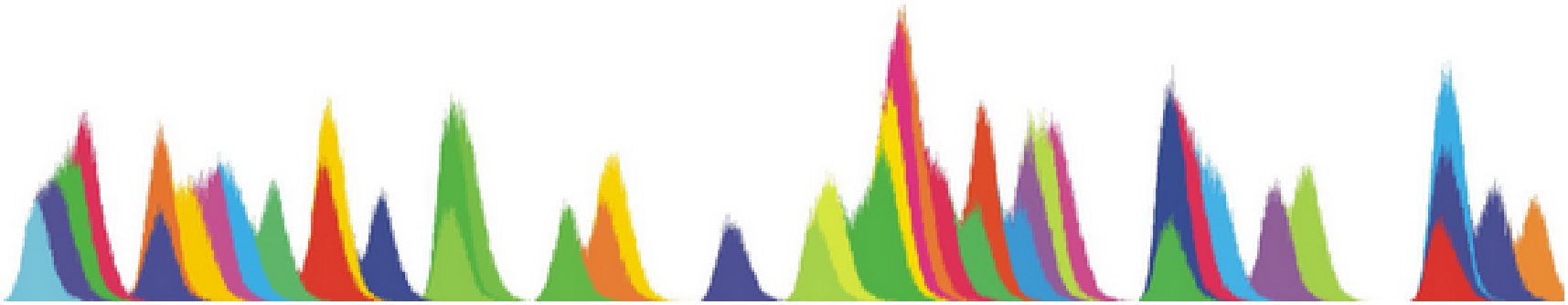
Modeling the Temporal Variation

Temporal variation has potential connections with natural processes

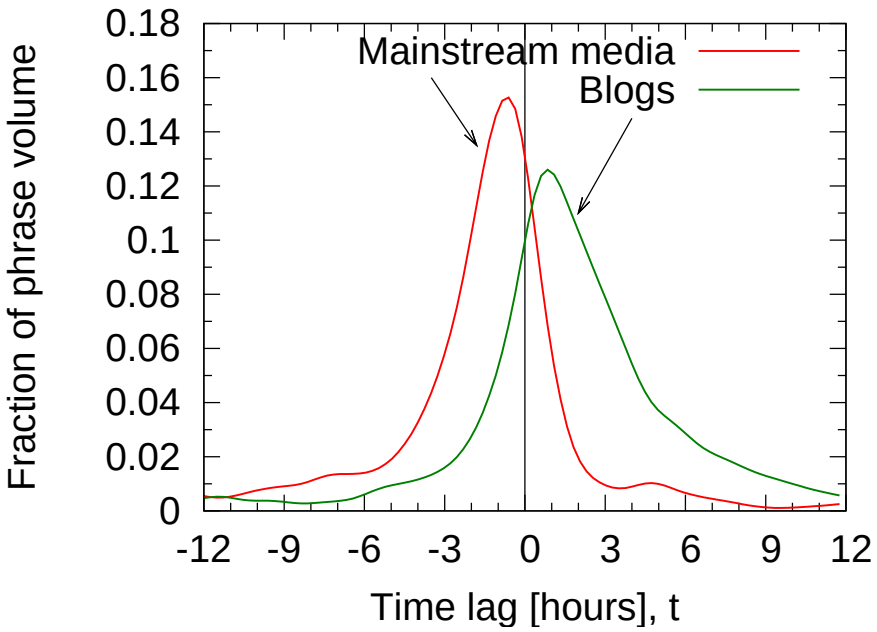
- Species competing for resources in an ecosystem.
- Biological systems that synchronize to favor a small number of individuals at any point in time [Lacker-Peskin 1981].

A model with N news sources, one new story per time step; each source's probabilistic choice of what to cover controlled by:

- Imitation: increasing in number of sources covering story.
- Recency: decreasing in time since story's appearance.



Interaction of News and Blogs



Rank	Lag [h]	Reported	Site
1	-26.5	42	hotair.com
2	-23	33	talkingpointsmemo.com
4	-19.5	56	politicalticker.blogs.cnn.com
5	-18	73	huffingtonpost.com
6	-17	49	digg.com
7	-16	89	breitbart.com
8	-15	31	thepoliticalcarnival.blogspot.com
9	-15	32	talkleft.com
10	-14.5	34	dailykos.com
16	-14	54	blogs.abcnews.com
30	-11	32	uk.reuters.com
34	-11	72	cnn.com
40	-10.5	78	washingtonpost.com
48	-10	53	online.wsj.com
49	-10	54	ap.org

Can study a coarse division of sources into news and blogs (using Google News classification).

- Peak intensity from blogs typically comes about 2.5 hours after peak intensity from news.
- Can classify individual sources by their typical timing relative to the peak aggregate intensity.

Stories Catalyzed by Blogs

Can use the data to formulate queries for different temporal "signatures": e.g., stories catalyzed by blogs.

- An $[x, y, t]$-query: between x and y fraction of the total quote volume occurred on blogs at least t days before the aggregate peak.

The highest-volume quotes matching the $[0.15, 0.50, 7]$-query:

- ...answering that question with specificity ... is above my pay grade.
- A changing environment will affect Alaska more than any other state ... . I'm not one though who would attribute it to being man-made.
- It was Ronald Reagan who said that freedom is always just one generation away from extinction.
- After trying to make experience the issue of this campaign, John McCain celebrated his 72nd birthday by appointing a former small-town mayor and brand-new governor as his vice-presidential nominee.
- Clarion fund recently financed the distribution of some 28 million DVDs containing the film "Obsession: Radical Islam's War Against the West" ...

Further Questions

A framework for tracking memes (quote clusters) through the news, to quantify the dynamics of the news cycle.

Many further questions:

- Which elements of the news cycle do we miss by using quote clusters?
- Can this analysis of memes help identify dynamics of polarization? (cf. [Adamic-Glance, 2005])
- How are these memes actually spreading among people?

This last question hints at the interface between news media and social networks.

- To study this further, we need a setting where the overt spread of memes through a social network can be tracked.

Tracing Specific Political Messages

Chain-letter petitions as "tracers" through global social network

[Liben-Nowell & Kleinberg 2008]

An example:

This is for anyone who thinks NPR/PBS is a worthwhile expenditure of $1.12/year of their taxes . . . A petition follows. ... PBS, NPR (National Public Radio), and the arts are facing major cutbacks in funding.

[...]

Please add your name to this list and forward it to friends if you believe in what we stand for. This list will be forwarded to the President of the United States, the Vice President of the United States, ... If you happen to be the 150th, 200th, 250th etc. signer of this petition, please forward a copy to wein2688@blue.univnorthco.edu.

Chain-Letter Petitions

Another example; possibly the largest one of all time:

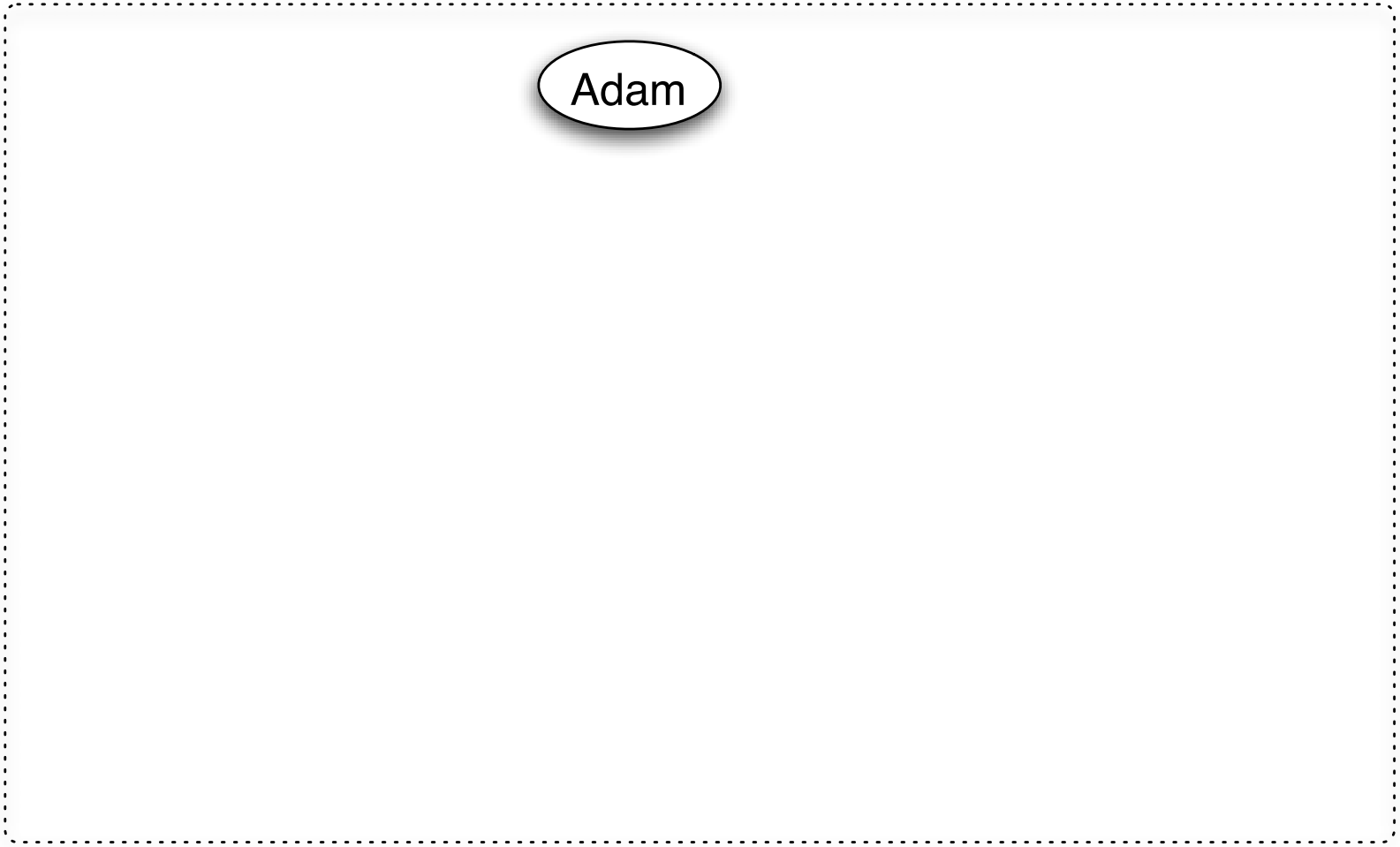
Dear All, The US Congress has authorised the President of the US to go to war against Iraq. Please consider this an urgent request. UN Petition for Peace:

[...]

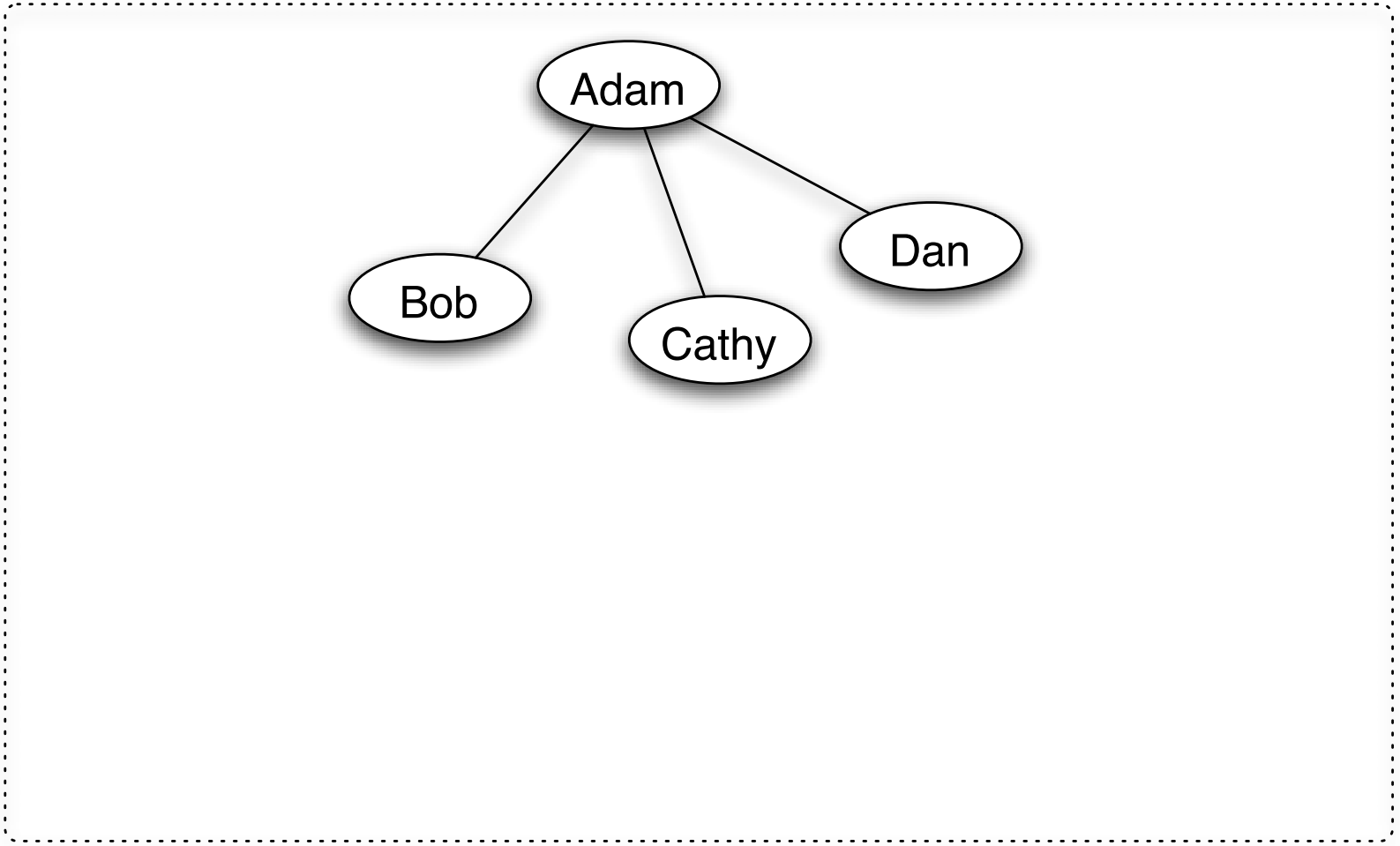
Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know. If you receive this list with more than 500 names signed, please send a copy of the message to:

usa@un.int
president@whitehouse.gov

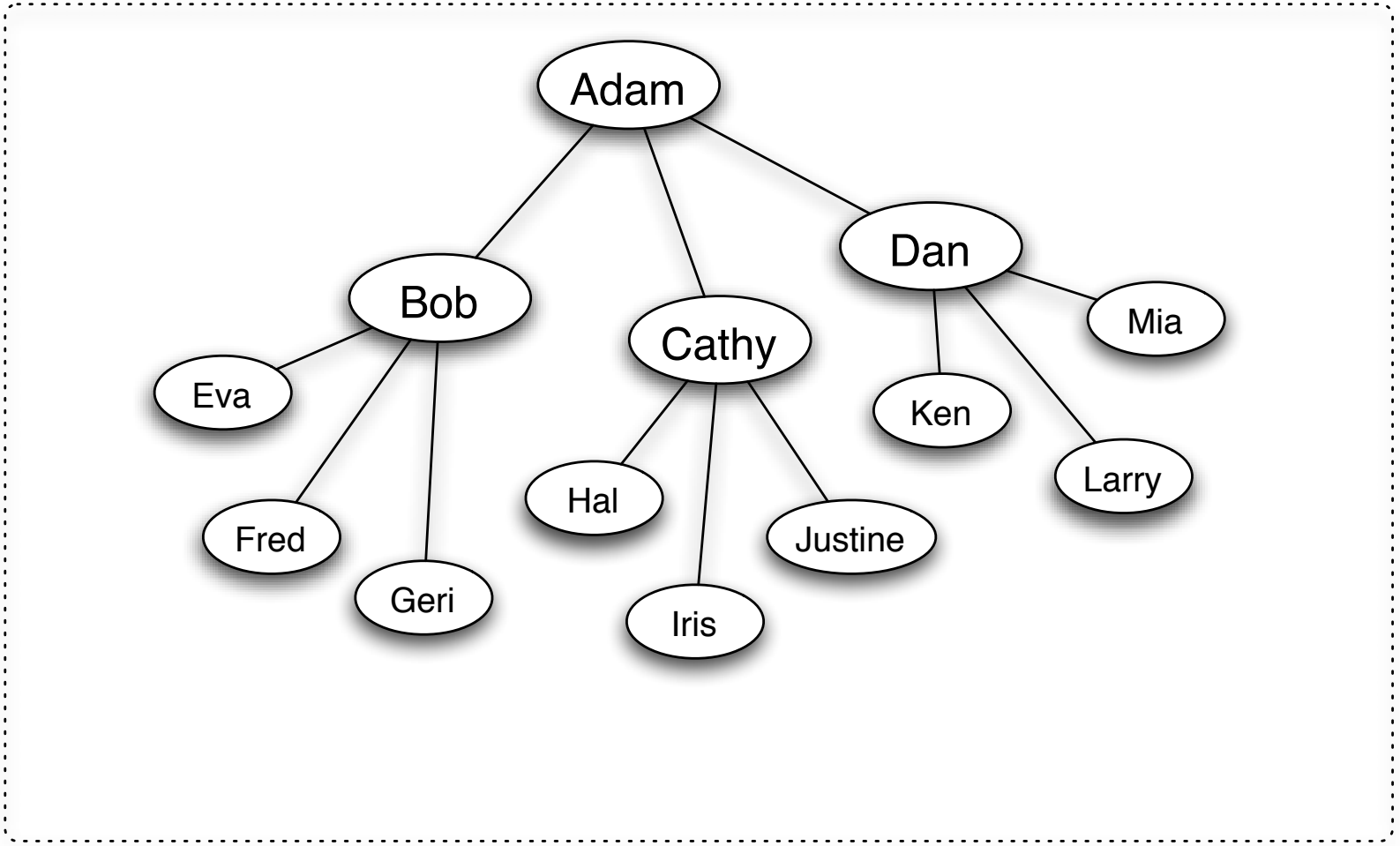
How Information Spreads (Traditional Picture)



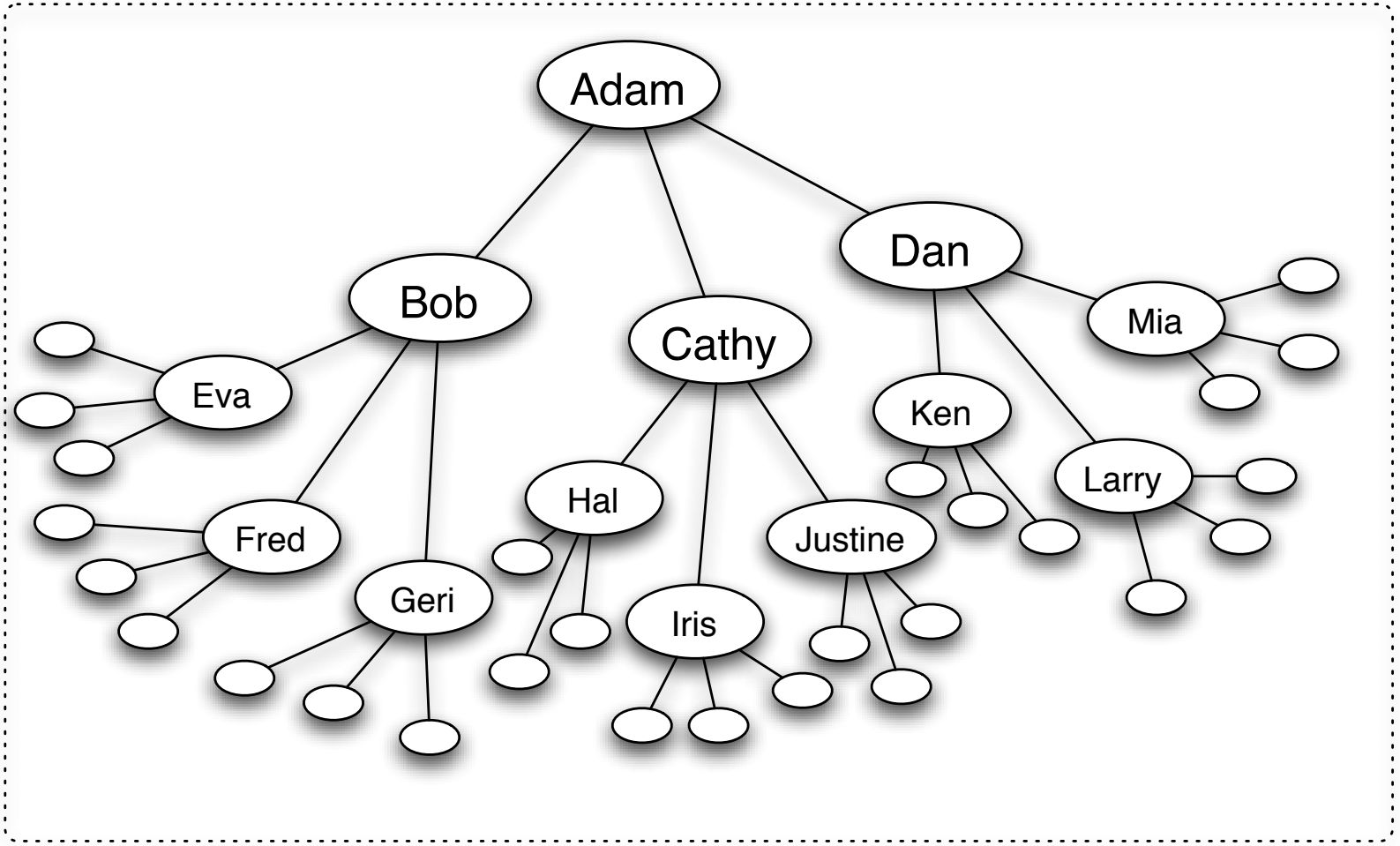
How Information Spreads (Traditional Picture)



How Information Spreads (Traditional Picture)

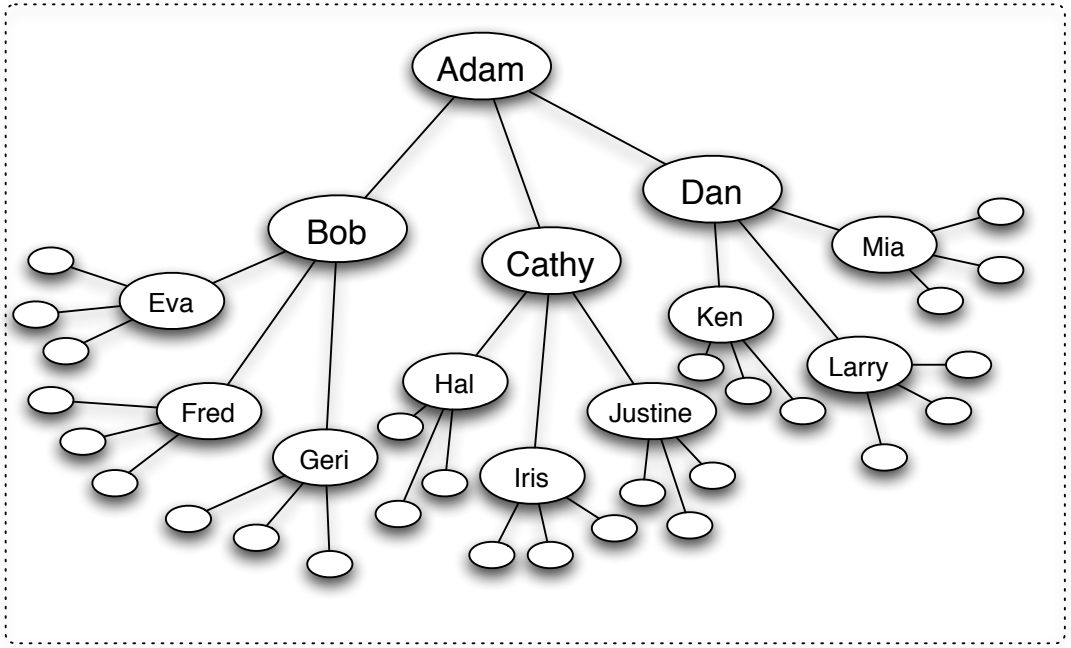


How Information Spreads (Traditional Picture)



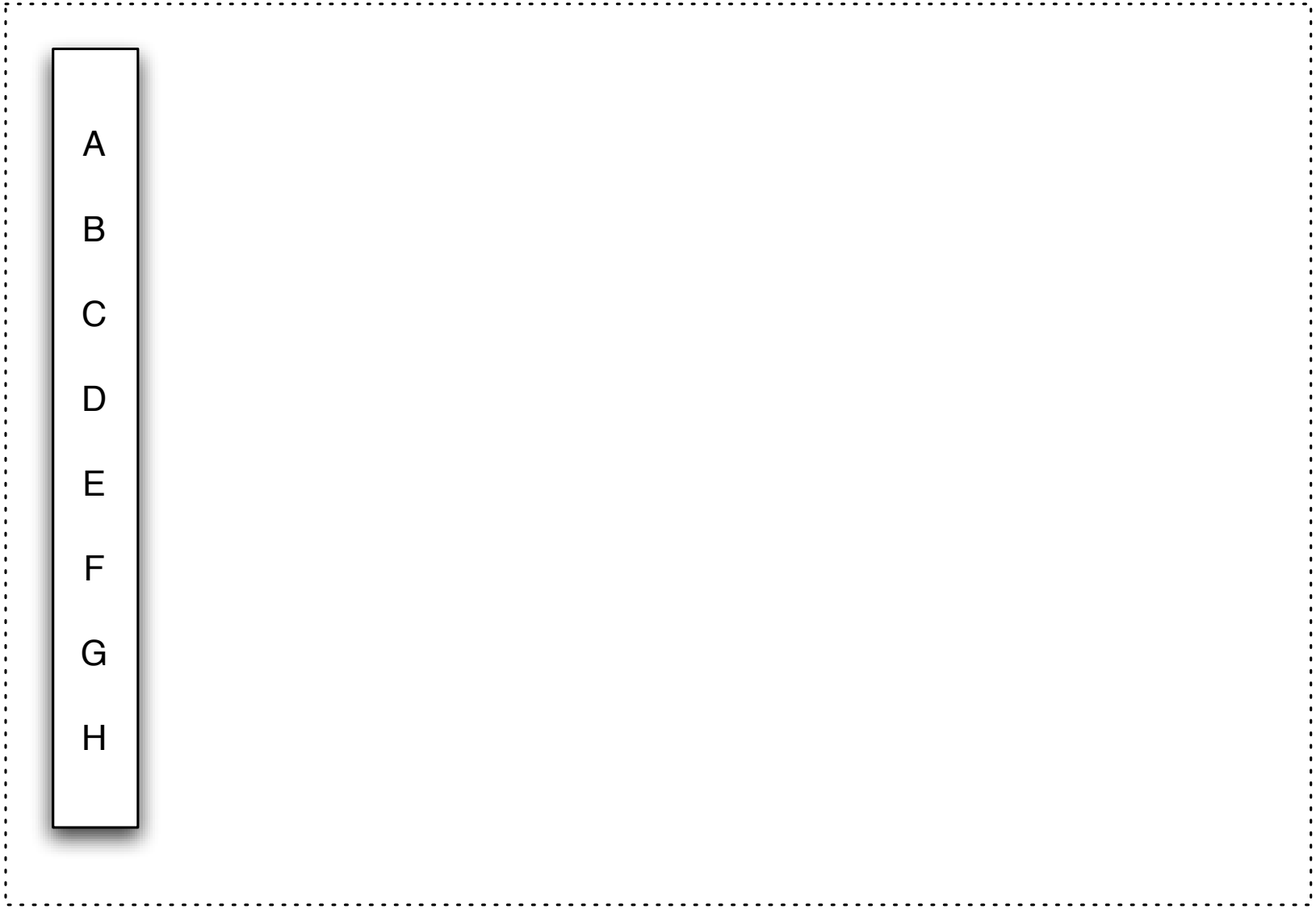
Assembling a Chain-Letter Tree

The full tree is unobservable.

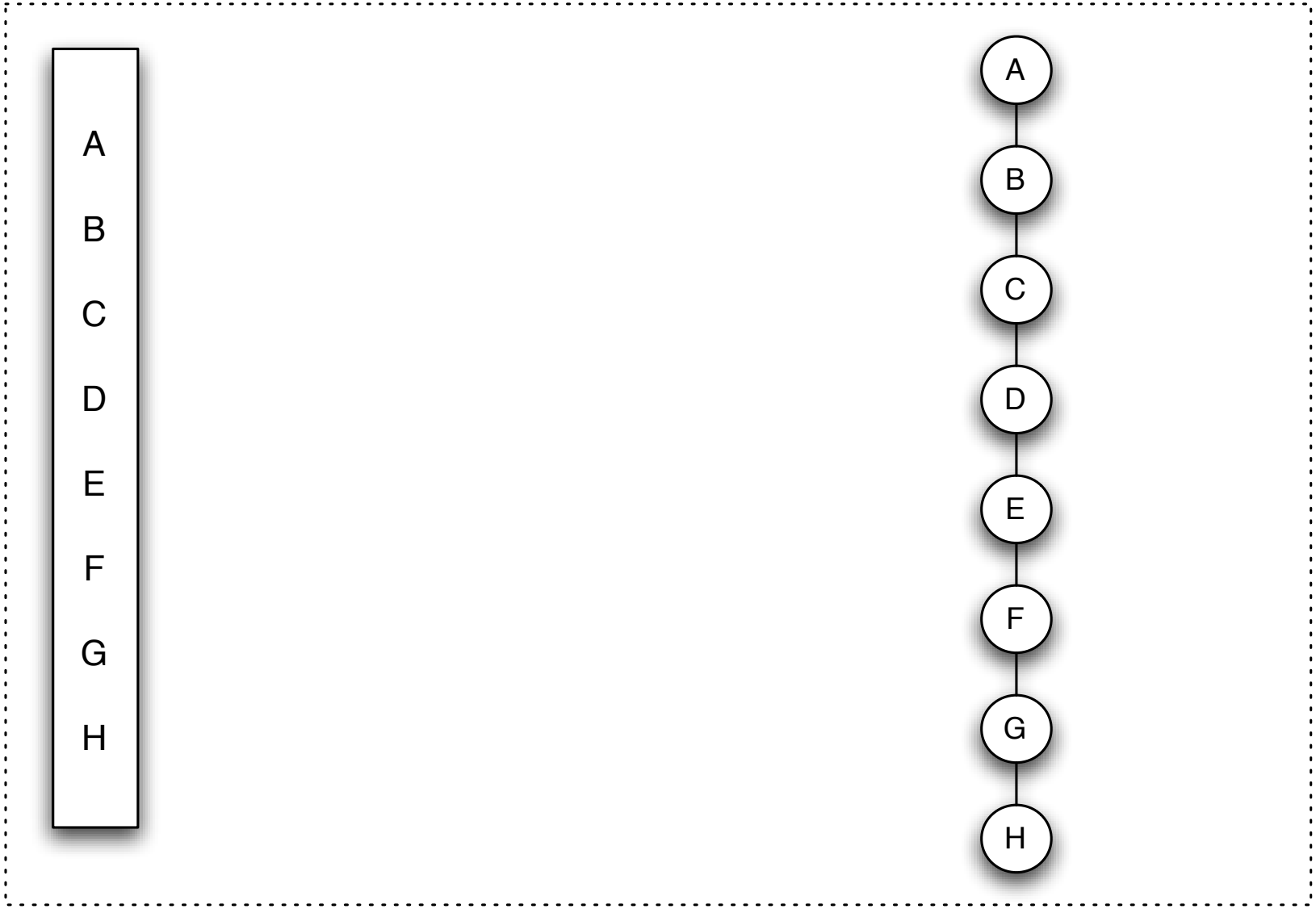


- But hundreds of copies with distinct recipient lists have been posted to mailing lists.
- We can obtain these by Web searches and then assemble a partial tree.

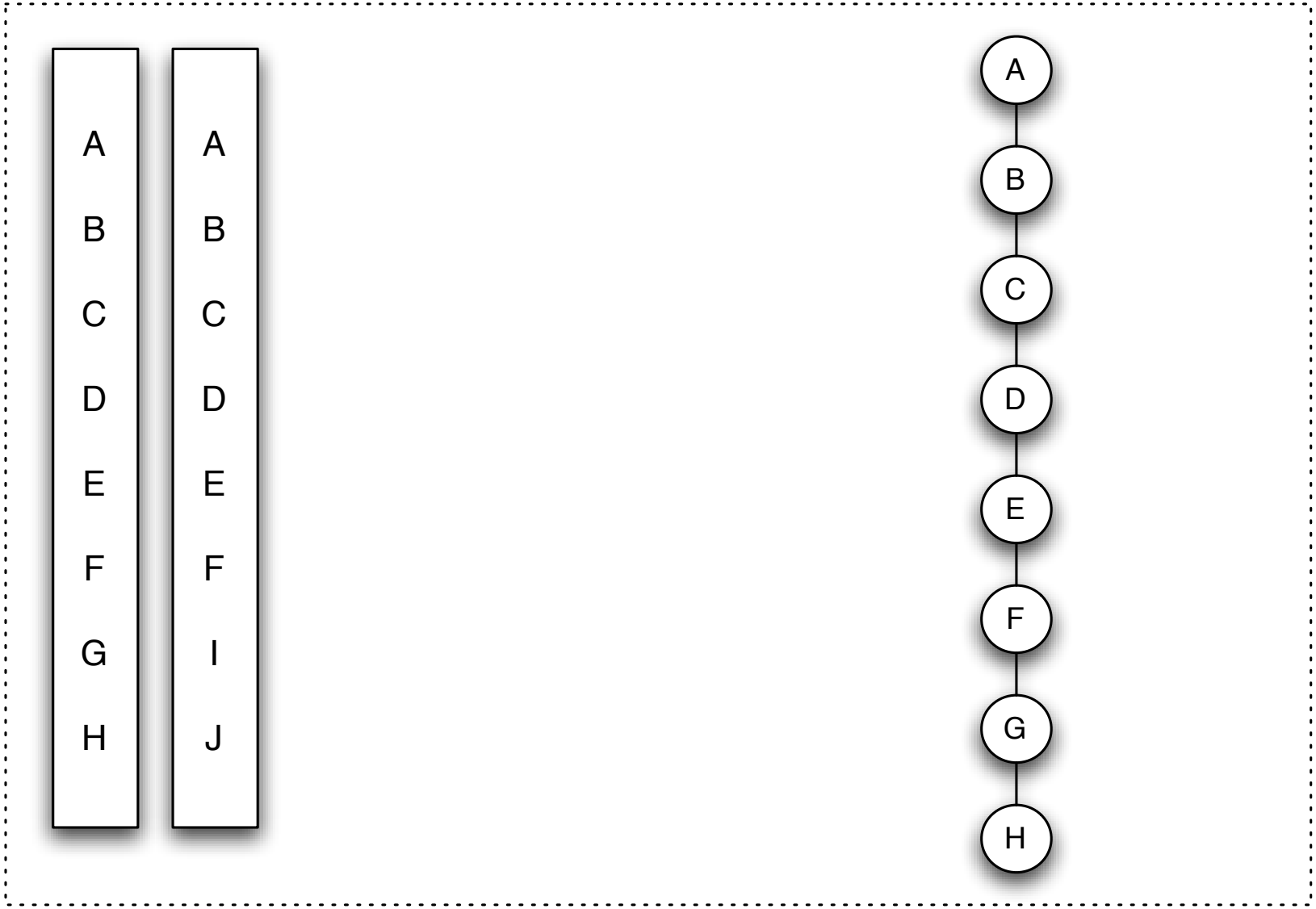
Assembling a Chain-Letter Tree



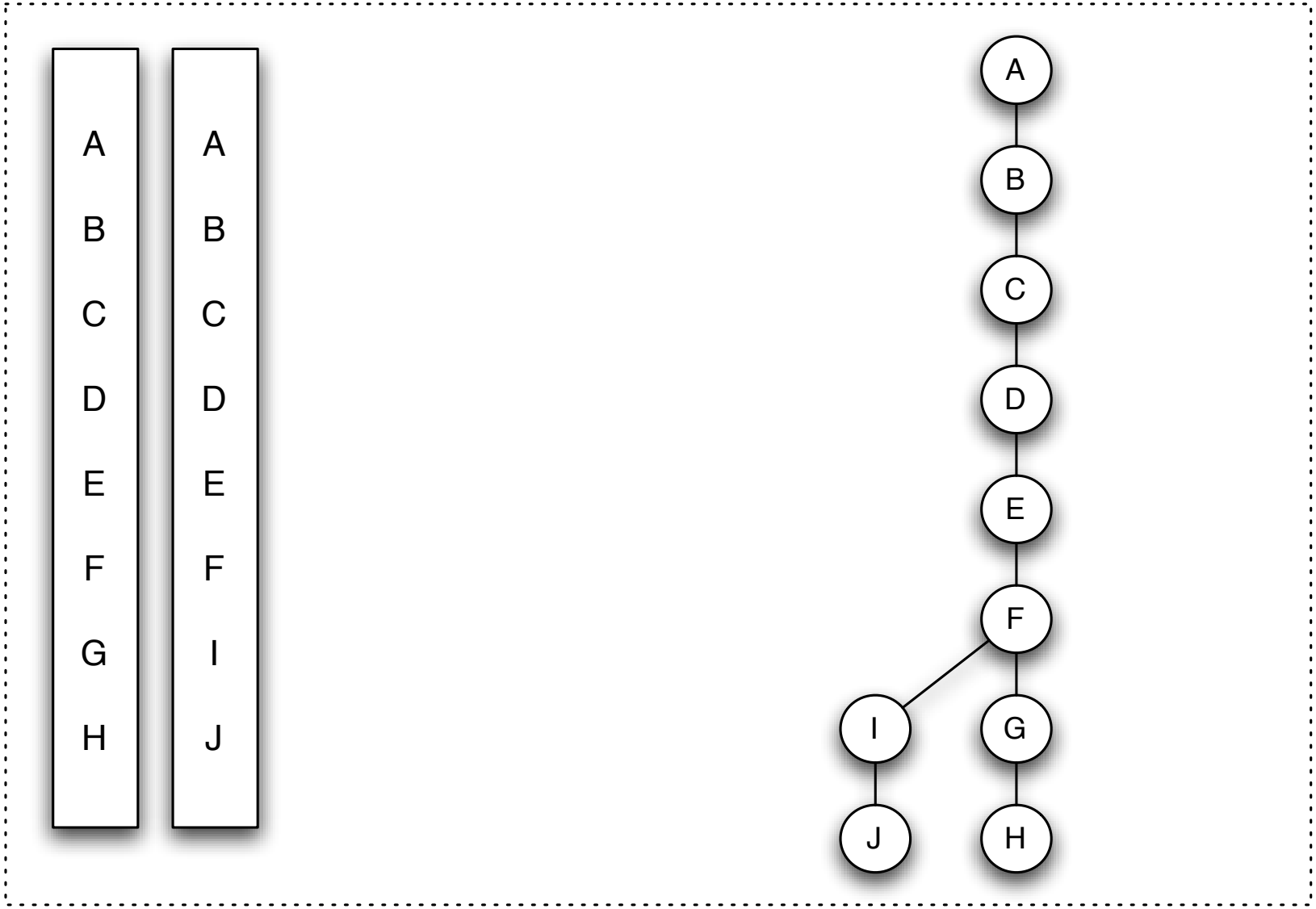
Assembling a Chain-Letter Tree



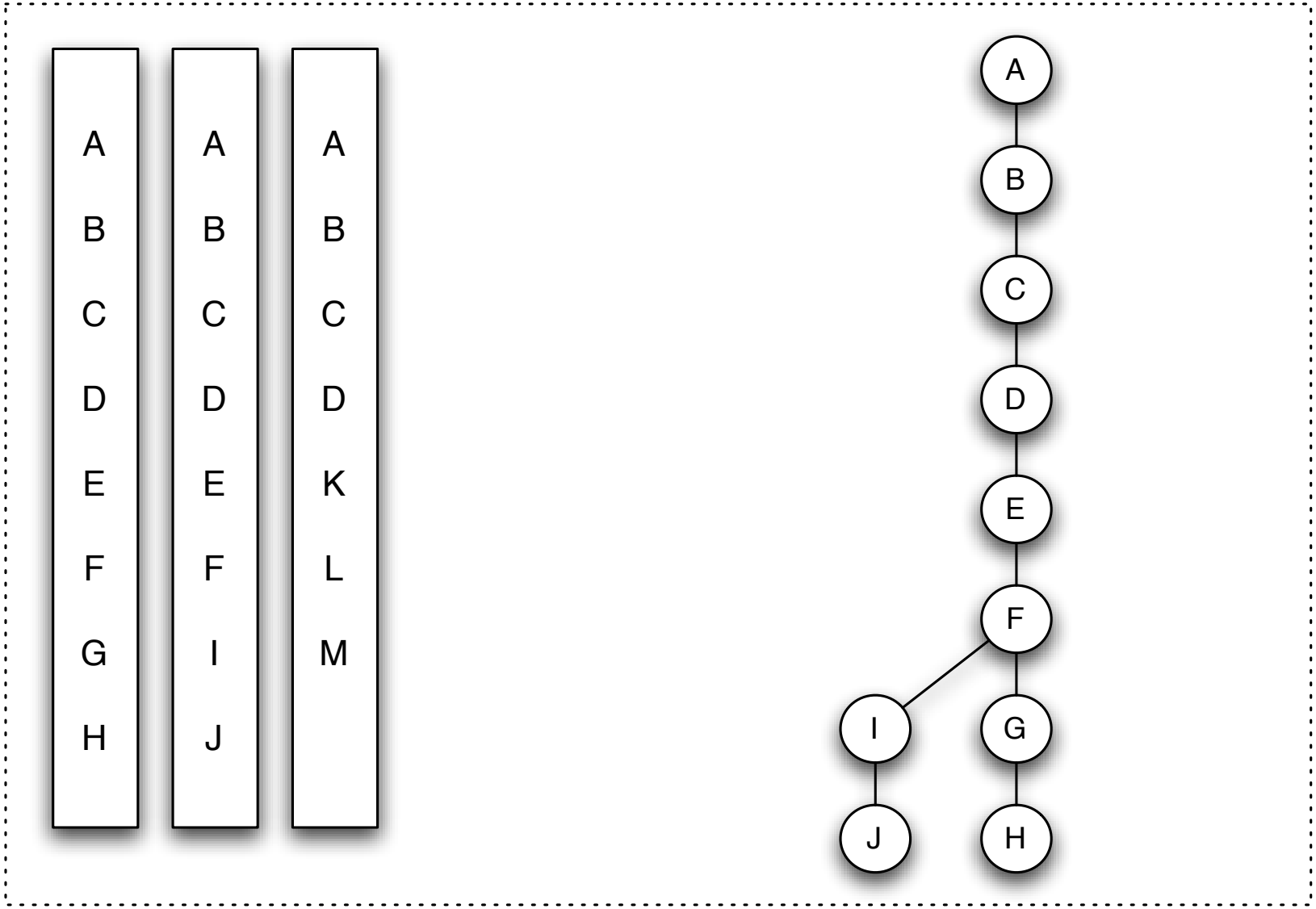
Assembling a Chain-Letter Tree



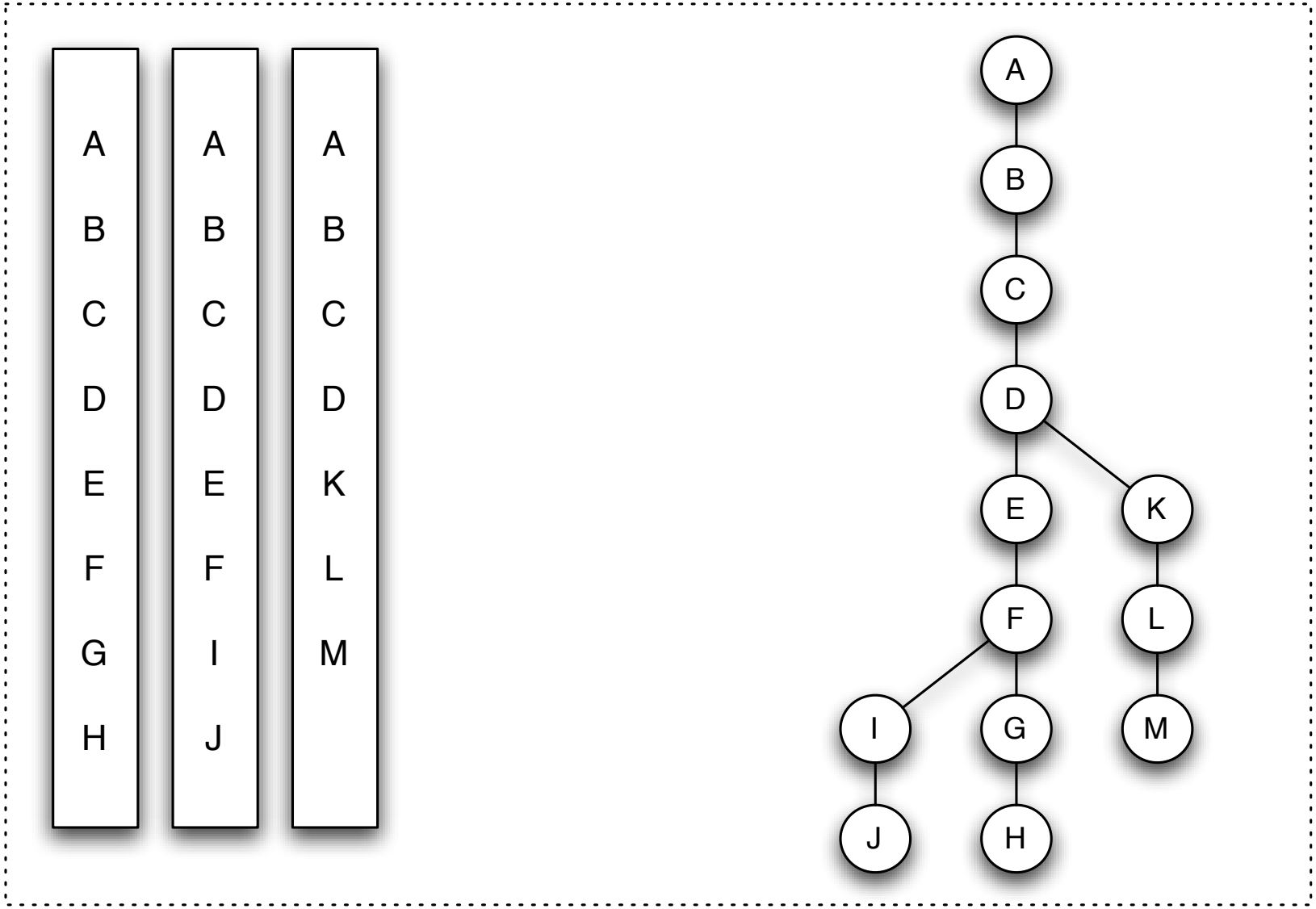
Assembling a Chain-Letter Tree



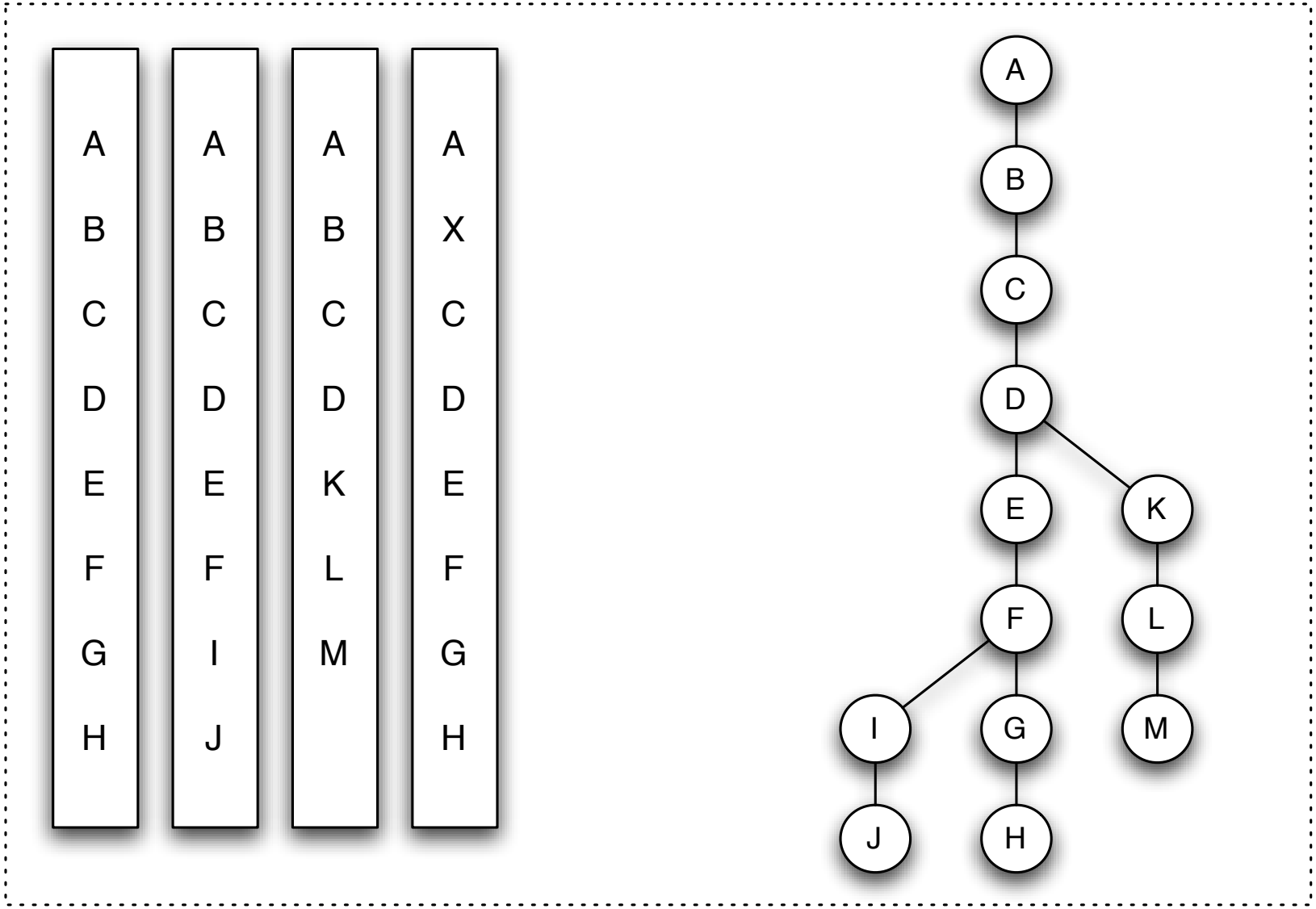
Assembling a Chain-Letter Tree



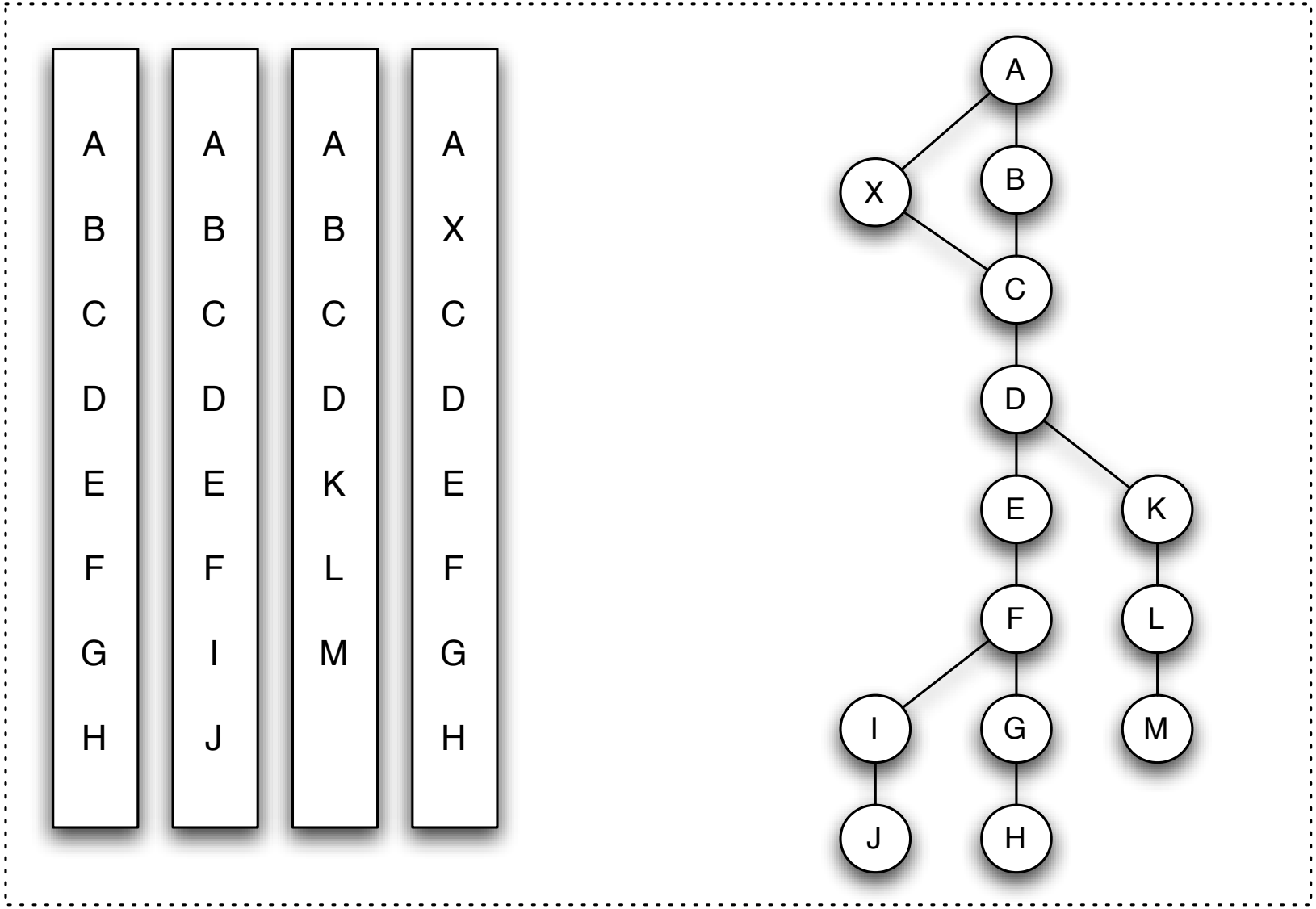
Assembling a Chain-Letter Tree



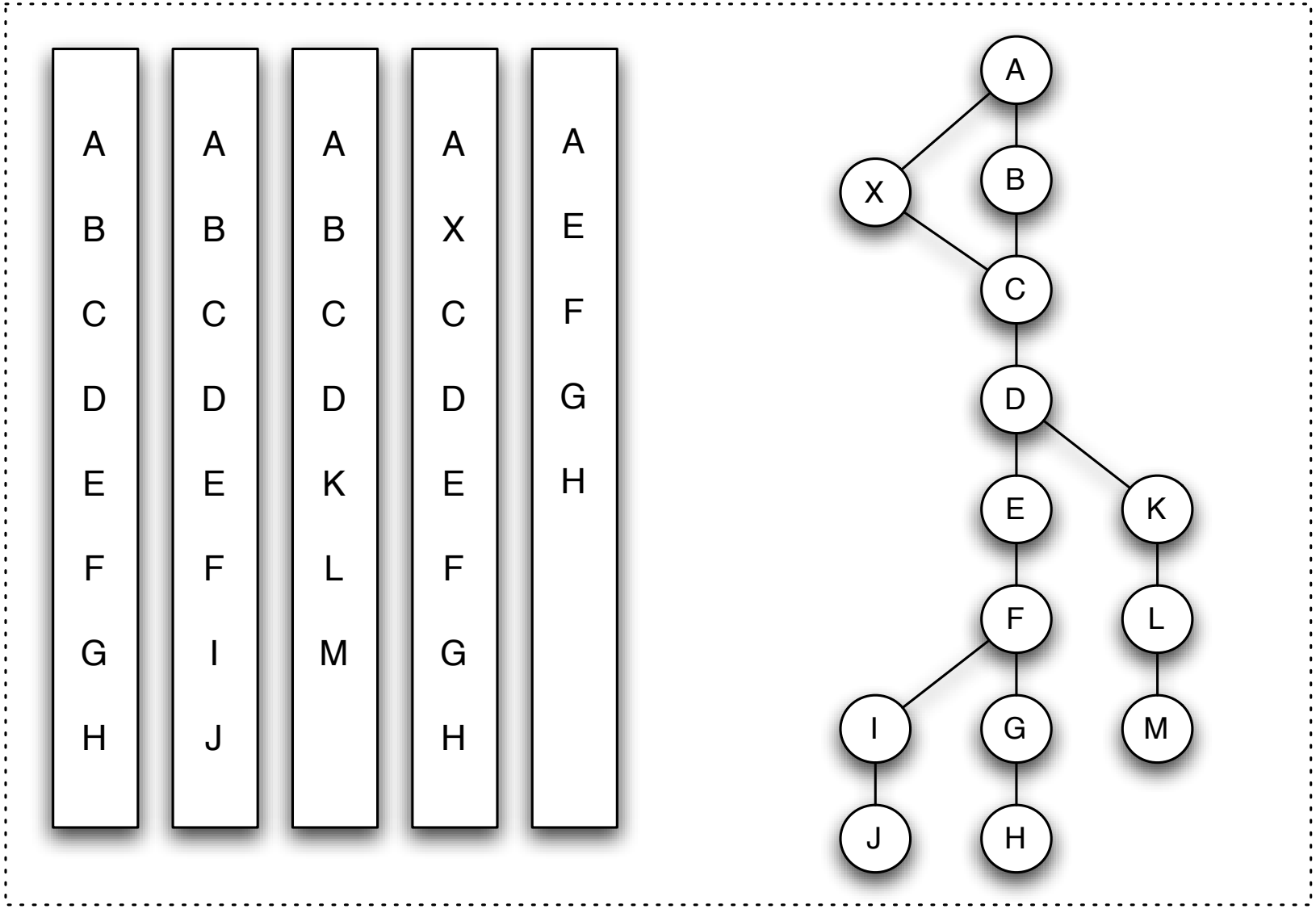
Assembling a Chain-Letter Tree



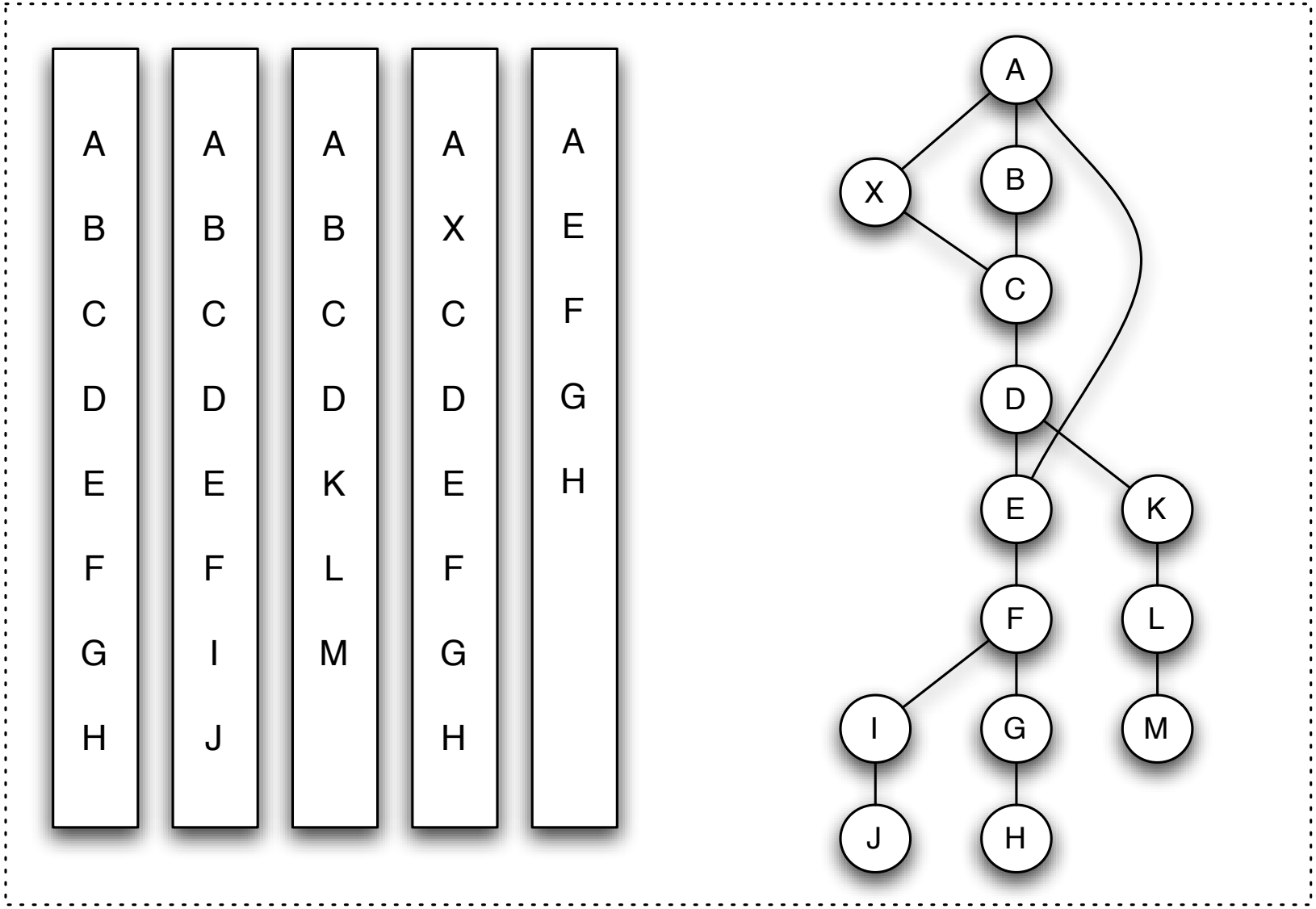
Assembling a Chain-Letter Tree



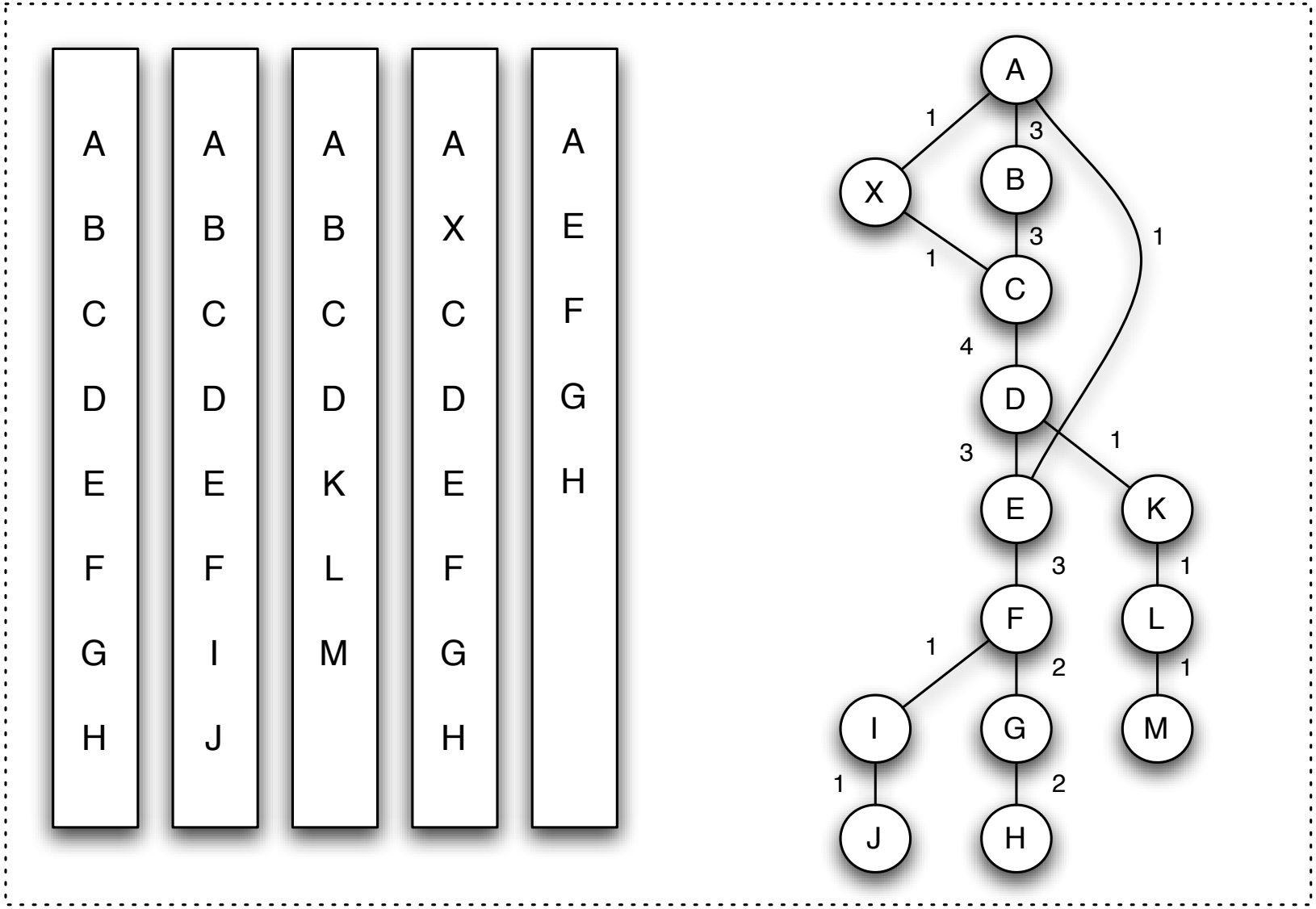
Assembling a Chain-Letter Tree



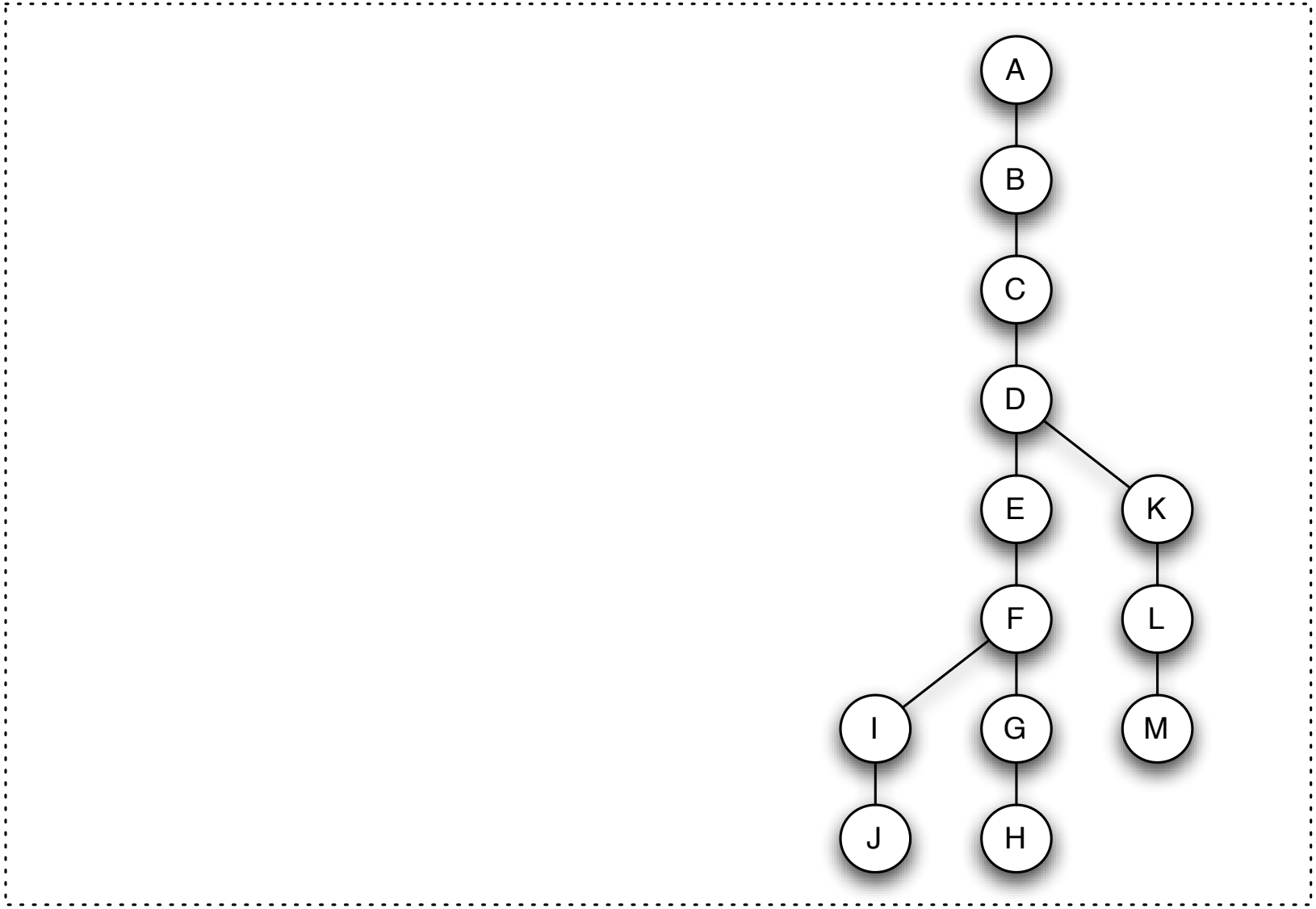
Assembling a Chain-Letter Tree

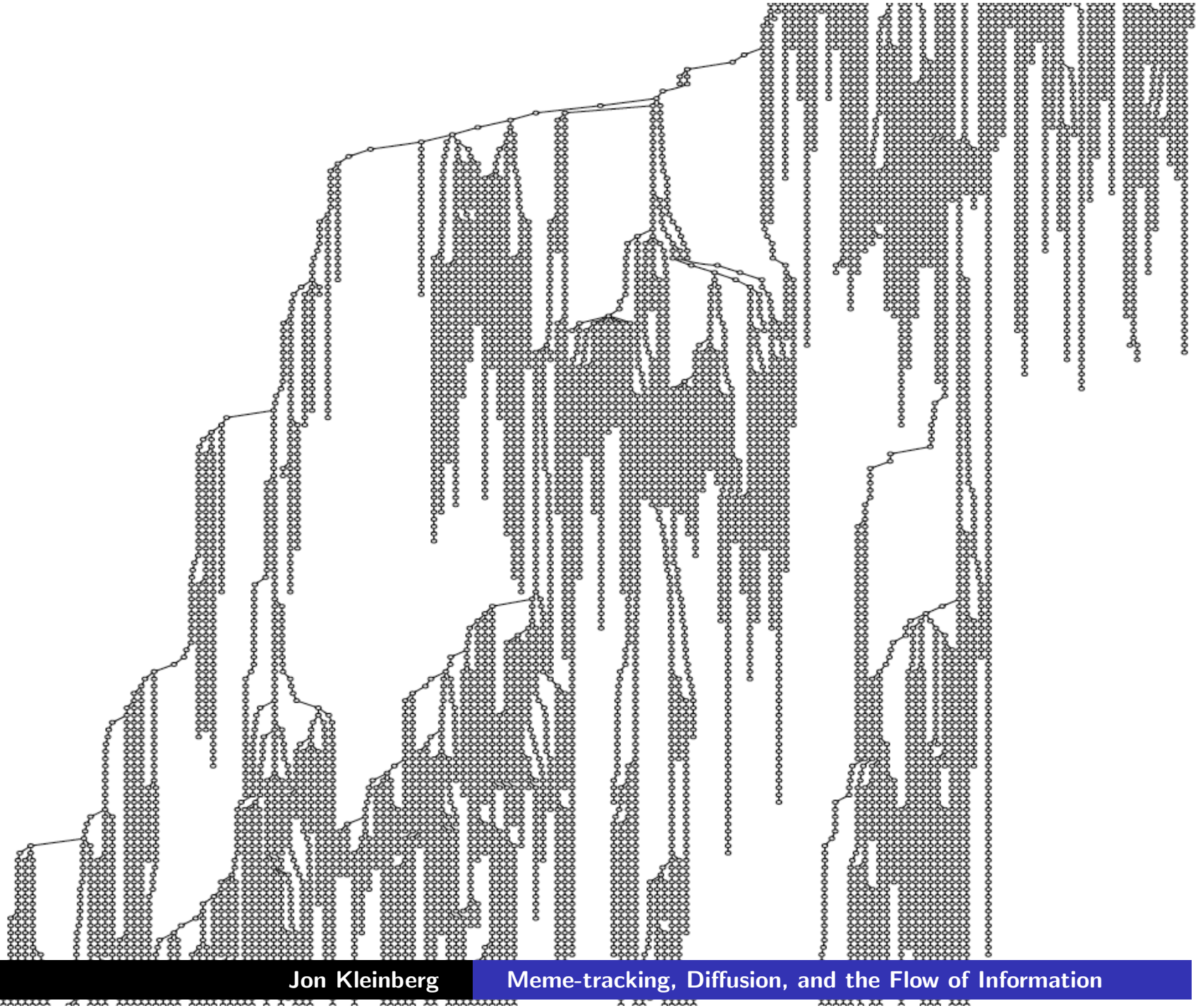


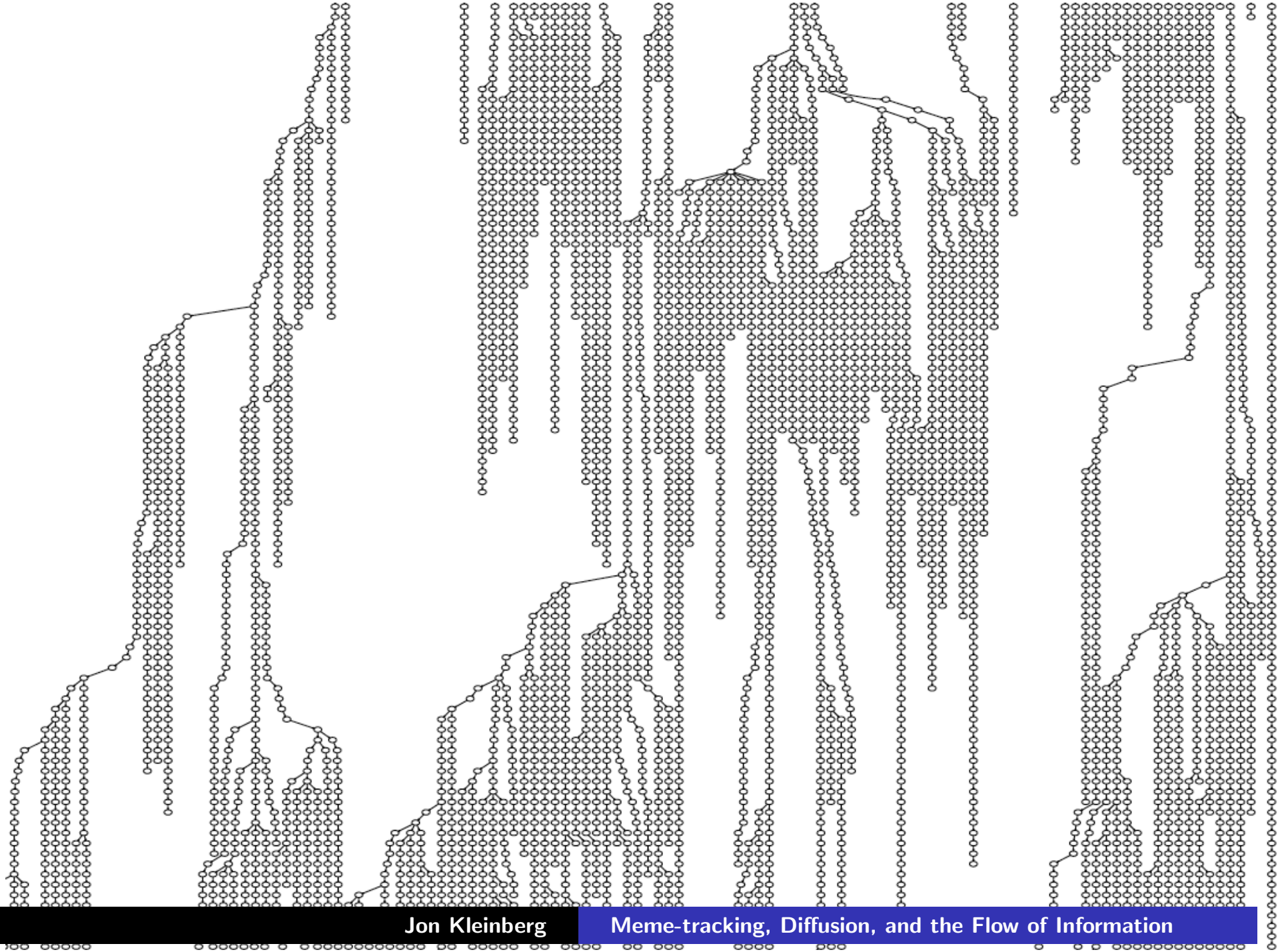
Assembling a Chain-Letter Tree



Assembling a Chain-Letter Tree







Modeling the Structure of the Tree

We're all a few steps apart in social network ("six degrees"), but the tree is very deep and narrow.

- Trees for other chain letters have very similar structure.
- Modeling non-participation and missing data doesn't account for this.

Some plausible models that can produce trees of this shape:

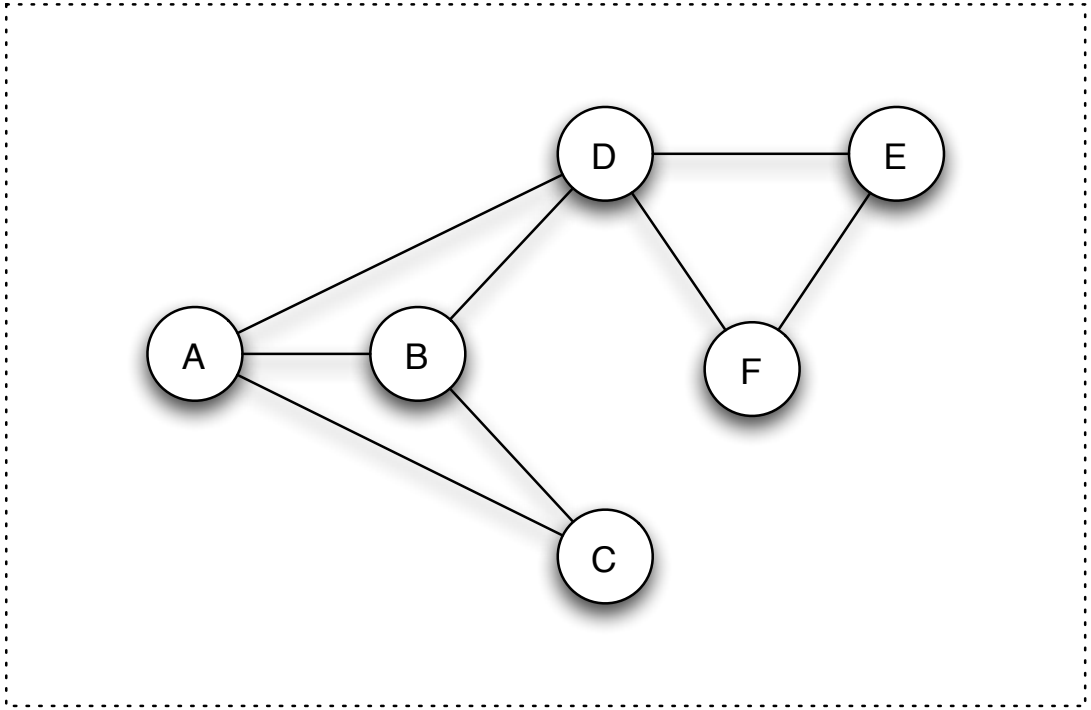
(1) Based on temporal ideas: people act on messages at very different speeds.

(2) Based on spatial ideas: social networks are geographically clustered.

Why is the Tree So Deep and Narrow?

It looks like a depth-first search tree. But why?

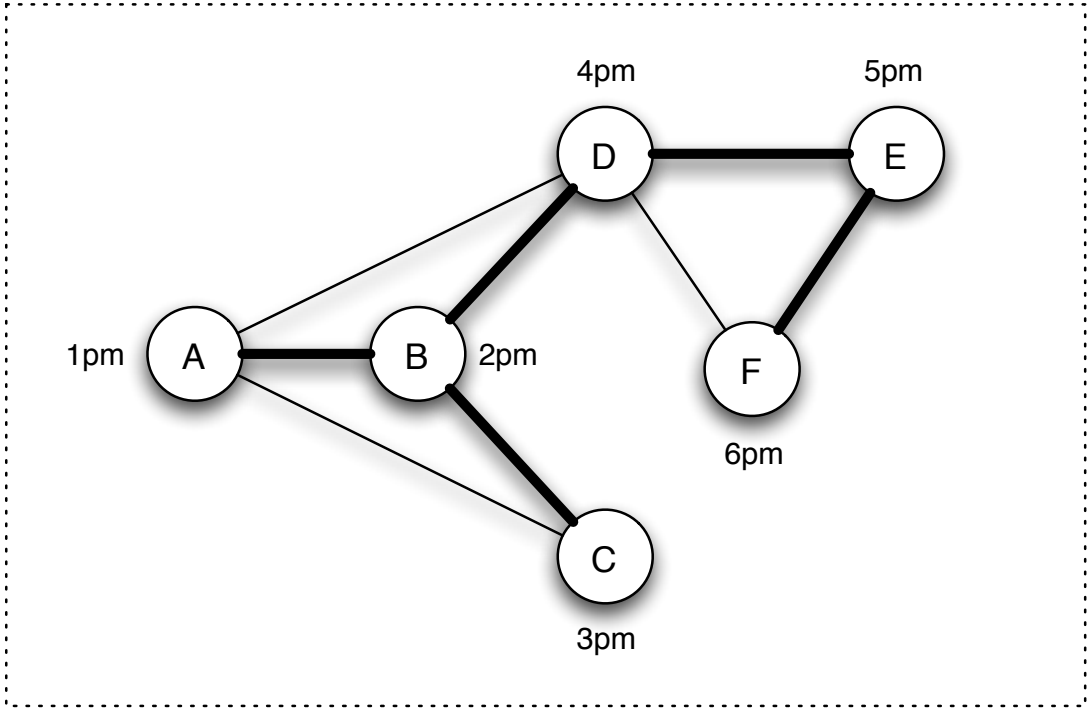
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

It looks like a depth-first search tree. But why?

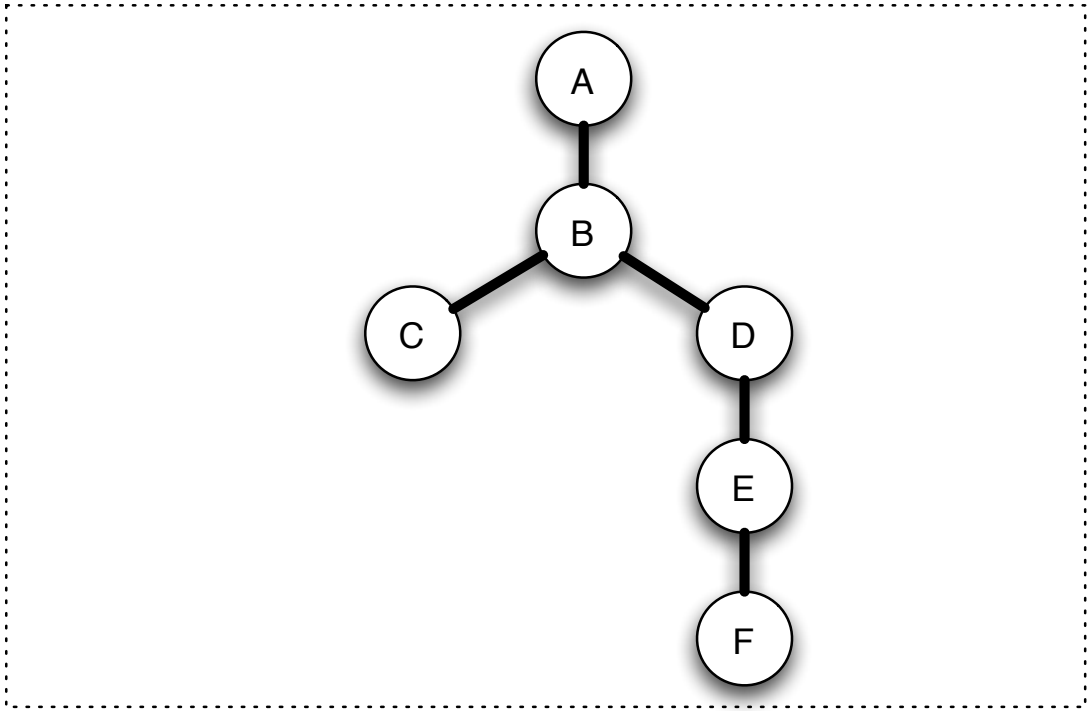
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

It looks like a depth-first search tree. But why?

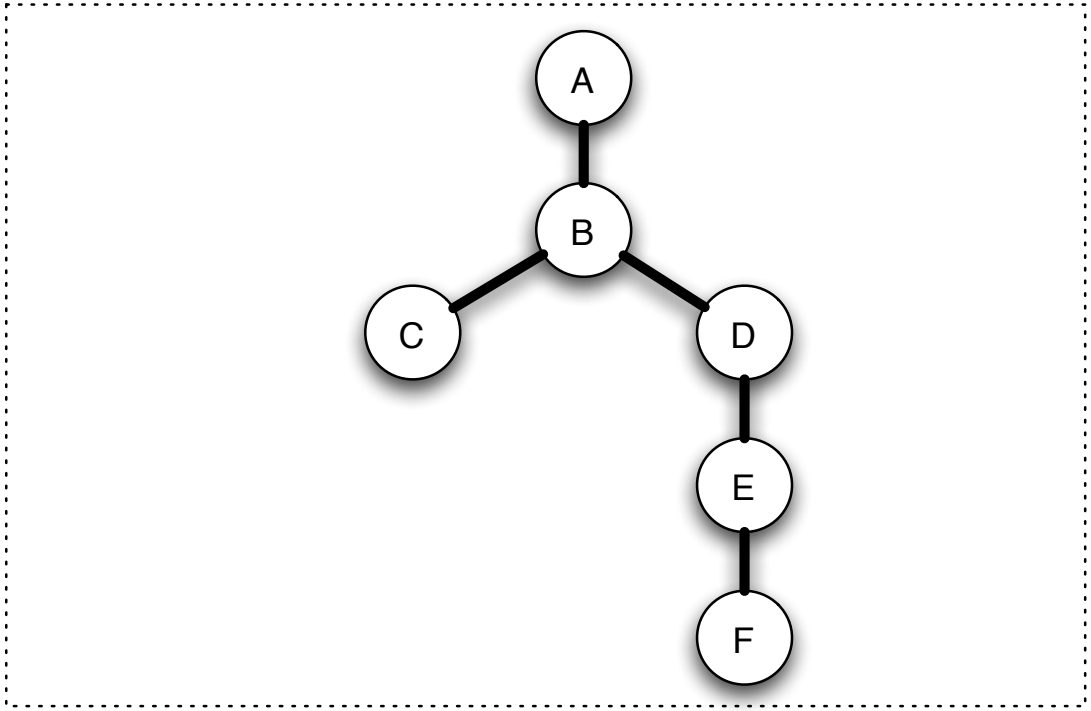
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

It looks like a depth-first search tree. But why?

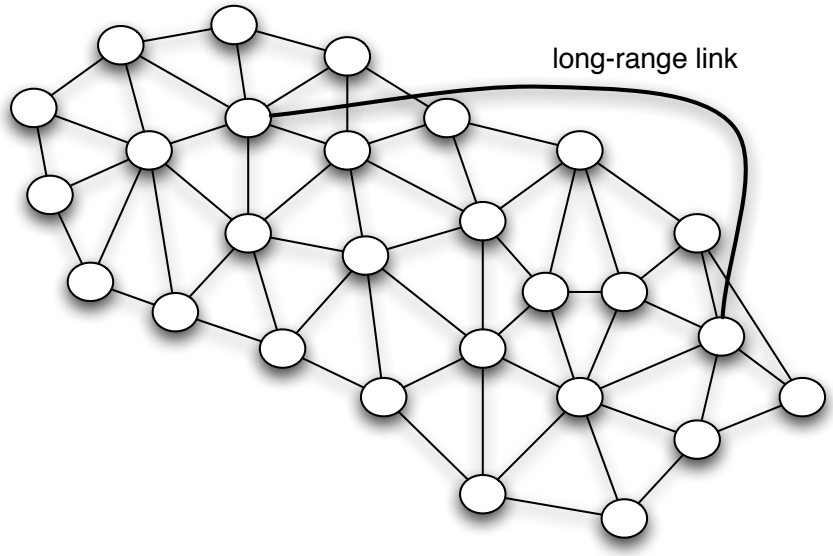
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



In simulations on 4.4-million-node LiveJournal friendship network, a generalization produces trees with height, depth, and width close to the actual chain letter trees.

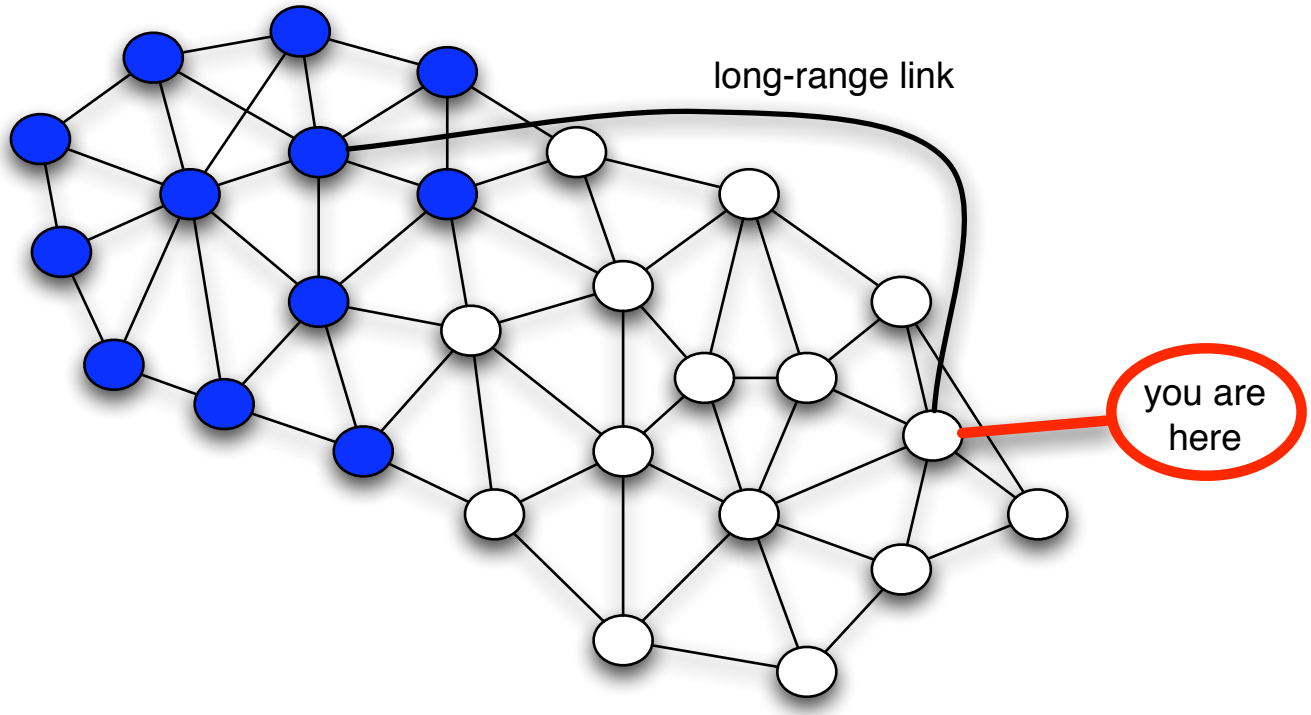
Spatial Clustering and Thresholds

A second class of theories based on spatial ideas.



- Even in on-line social networks, most friends are geographically (and demographically) similar to you [McPherson et al. 2001, Liben-Nowell et al. 2005]
- Decision rules for acting may involve thresholds: e.g., you may need to see multiple friends advocating a cause before signing on [Granovetter 1978, Schelling 1978]

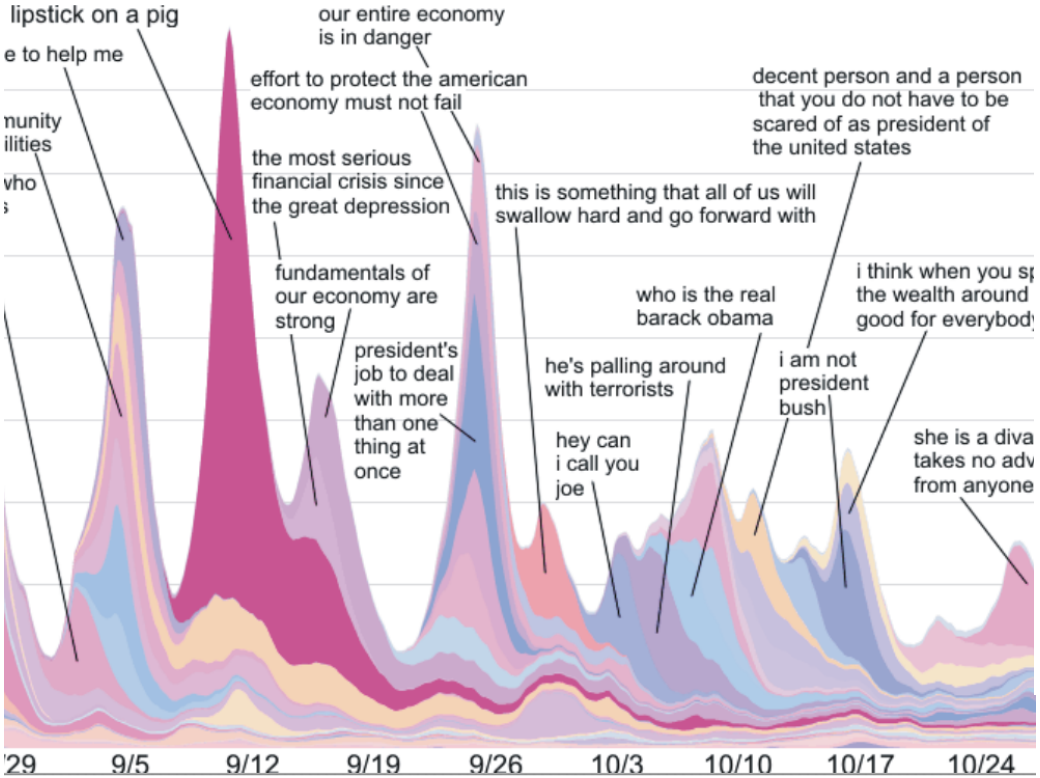
Spatial Clustering and Thresholds

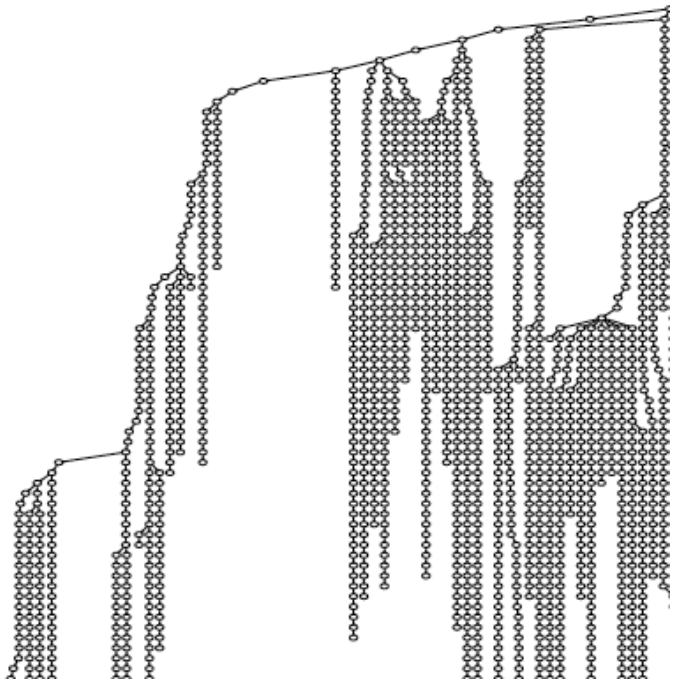


Interaction of local structure and thresholds [Centola-Macy 2007]

- Suppose people needed two stimuli to be willing to participate.
- Non-trivial thresholds make it hard to use long-range links.

Final Reflections: Temporal and Network Dynamics





- New levels of resolution on longstanding political phenomena.
- The temporal dynamics of the news and the network dynamics of diffusion are tied together.
- Social media amplifies this tie.
- Ideas from computing and on-line media will be crucial to the next steps.