

# Social Processes, Information Flow, and Anonymized Network Data

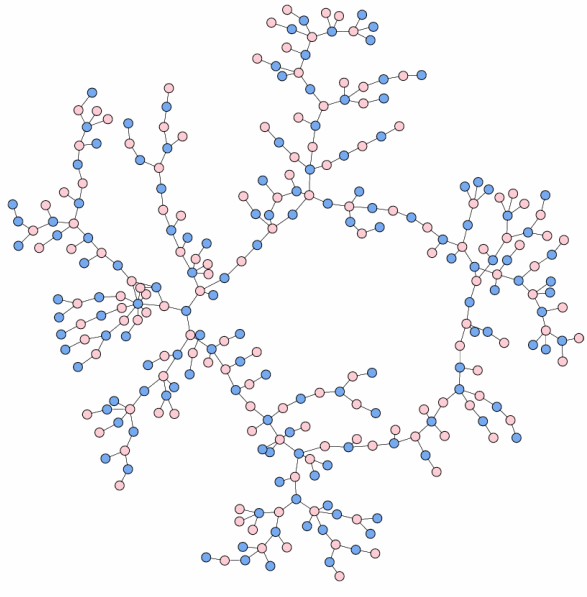
Jon Kleinberg

Cornell University

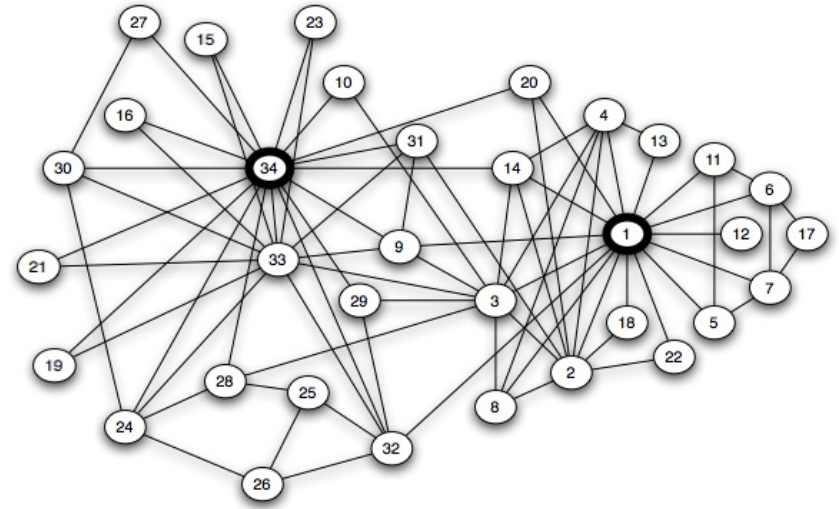


**Including joint work with Lars Backstrom, Cynthia Dwork,  
and David Liben-Nowell**

# Social Network Analysis



High-school dating (Bearman-Moody-Stovel 2004)

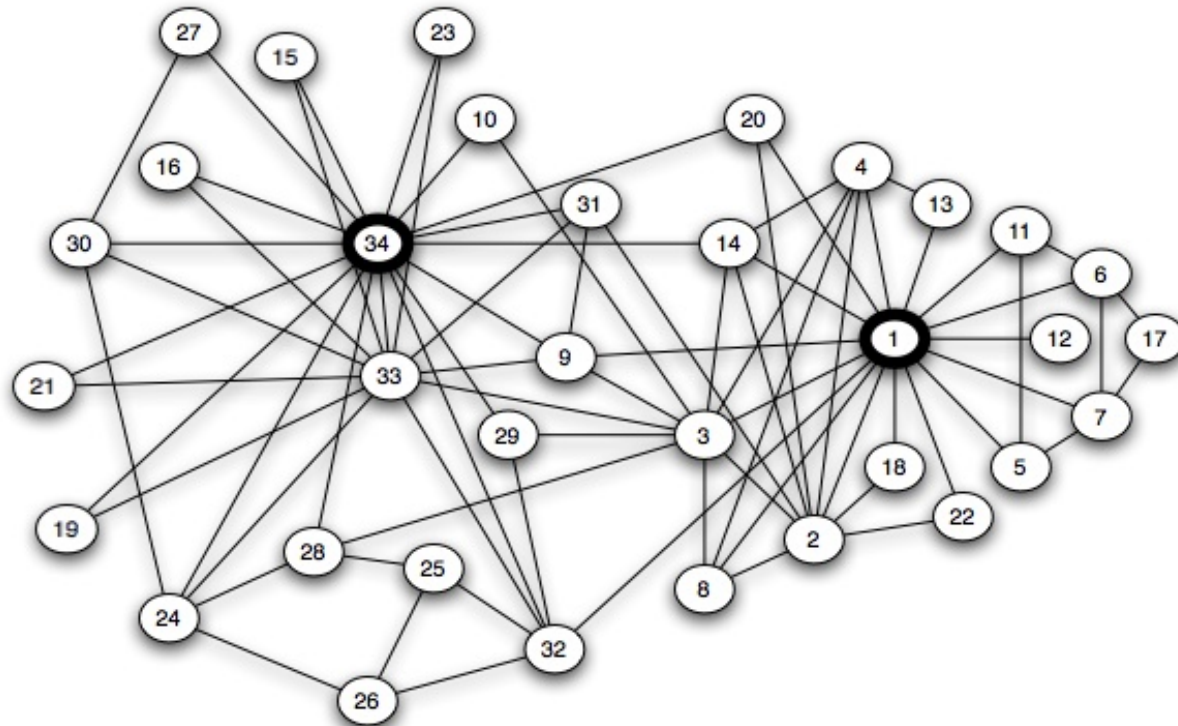


Karate club (Zachary 1977)

## Social network data

- Active research area in sociology, social psychology, anthropology for the past half-century.
- Today: Convergence of social and technological networks  
Computing and info. systems with intrinsic social structure.
- What can the different fields learn from each other?

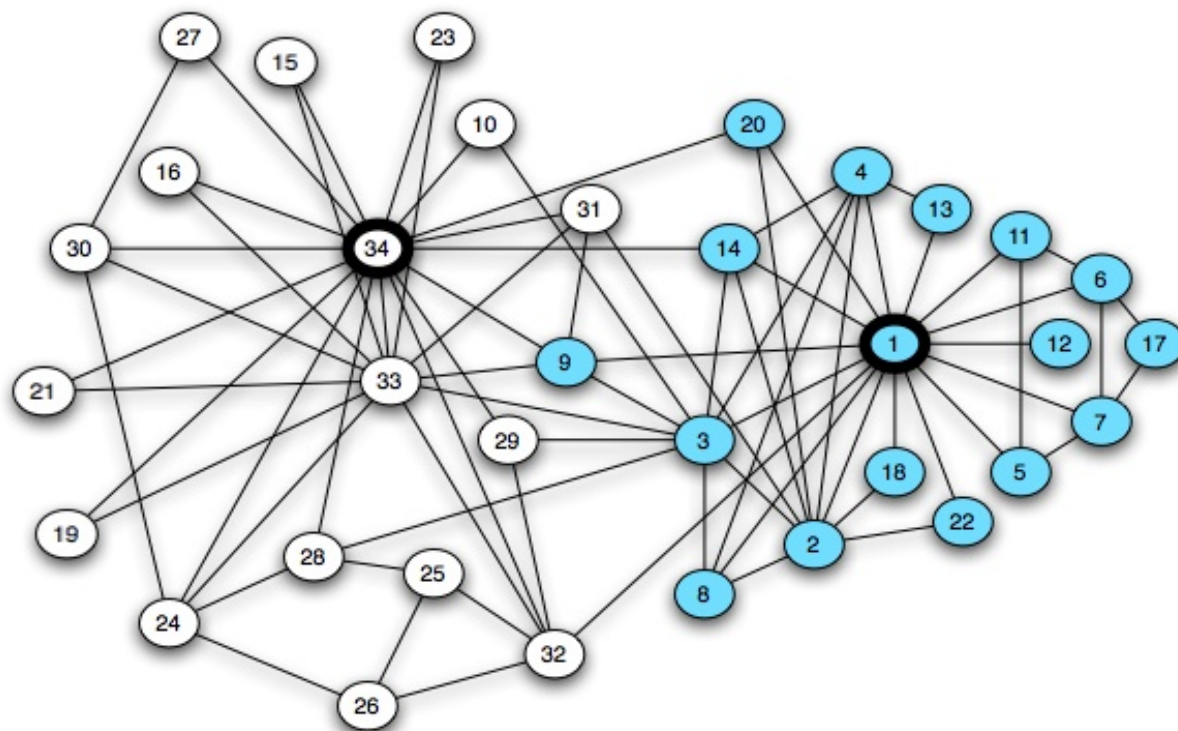
# Mining Social Network Data



Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.

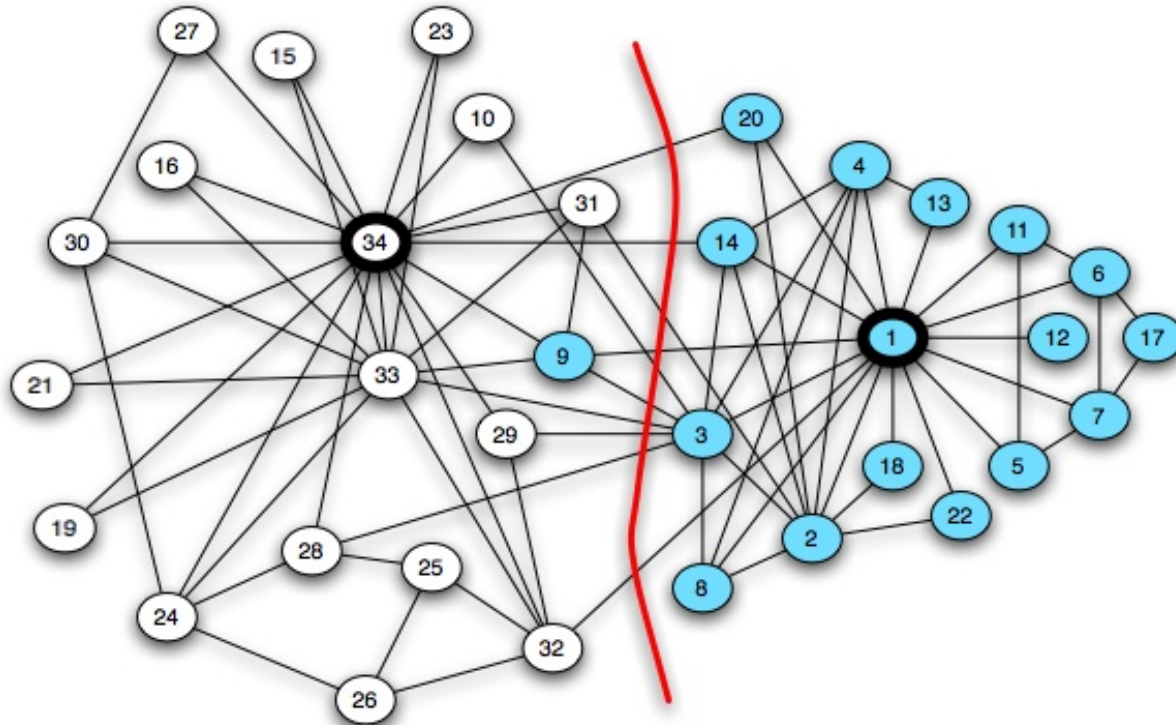
# Mining Social Network Data



Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.
- During his observation, conflicts intensified and group split.

# Mining Social Network Data



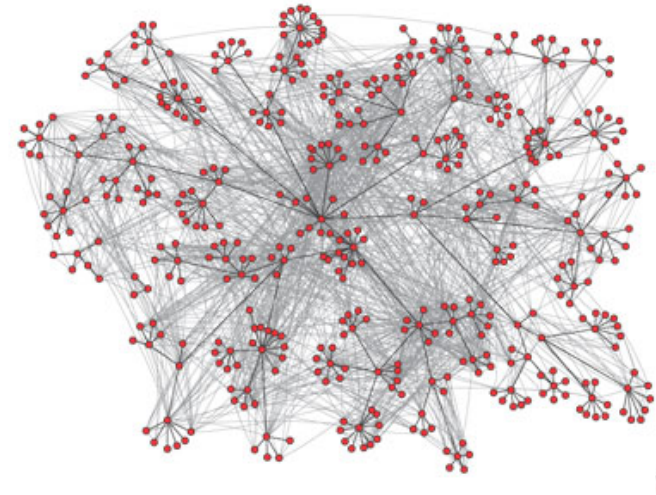
Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.
- During his observation, conflicts intensified and group split.
- Split could be explained by minimum cut in social network.

# A Matter of Scale

Social network data spans many orders of magnitude

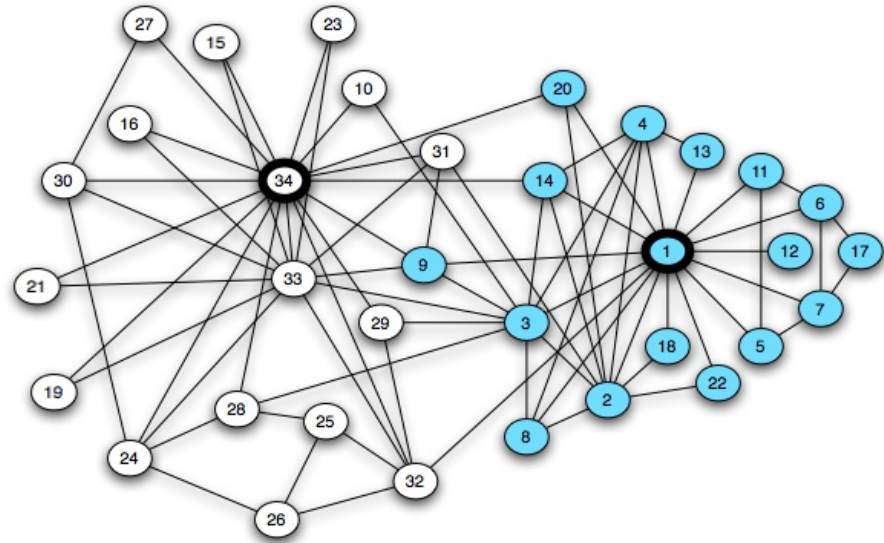
- 436-node network of e-mail exchange over 3 months at a corporate research lab (Adamic-Adar 2003)



- 43,553-node network of e-mail exchange over 2 years at a large university (Kossinets-Watts 2006)
- 4.4-million-node network of declared friendships on blogging community LiveJournal (Liben-Nowell et al. 2005, Backstrom et al. 2006)
- 240-million-node network of all IM communication over one month on Microsoft Instant Messenger (Leskovec-Horvitz'07)

# Not Just a Matter of Scale

- How does massive network data compare to small-scale studies?



Currently, massive network datasets give you both more and less:

- **More:** can observe global phenomena that are genuine, but literally invisible at smaller scales.
- **Less:** Don't really know what any one node or link means. Easy to measure things; hard to pose nuanced questions.
- **Goal:** Find the point where the lines of research converge.

# Outline

Several core computing ideas come into play:

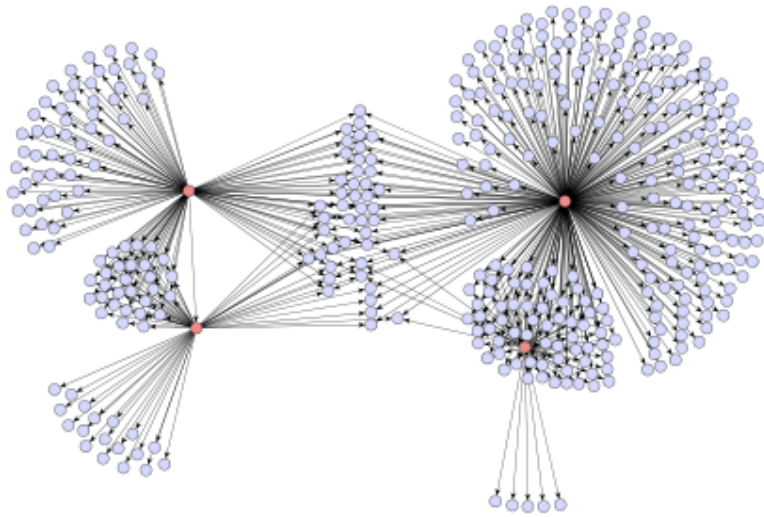
- Working with network data that is much messier than just nodes and edges.
- Algorithmic models as a basic vocabulary for expressing complex social-science questions on complex network data.
- Understanding social networks as datasets: privacy implications and other concerns.

Plan for the talk:

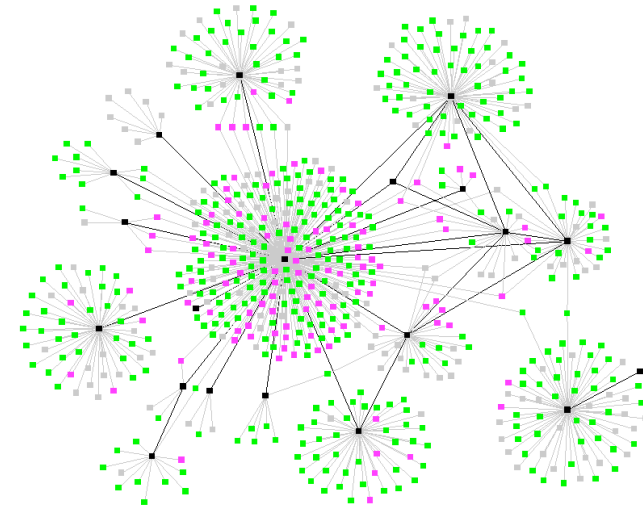
- Algorithmic models for cascading behavior in social networks: Formulating some fundamental unresolved questions.
- Evaluating anonymization as a standard approach for protecting privacy in social network data.



# Diffusion in Social Networks



Book recommendations (Leskovec et al 2006)



Contagion of TB (Andre et al. 2006)

Behaviors that cascade from node to node like an epidemic.

- News, opinions, beliefs, rumors, fads, ...
- Diffusion of innovations [Coleman-Katz-Menzel, Rogers]
- Viral marketing [Domingos-Richardson 2001]
- Localized collective action: riots, walkouts
- Modeling via
  - biological epidemics [Berger-Borgs-Chayes-Saberi 2005]
  - coordination games [Blume1993, Ellison1993, Jackson-Yariv2005]

# Chain-Letter Petitions

Chain-letter petitions as “tracers” through global social network  
[Liben-Nowell & Kleinberg 2008]

---

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

# Networks of Documents, Networks of People

Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them ... There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record.

(Bush, 1945)

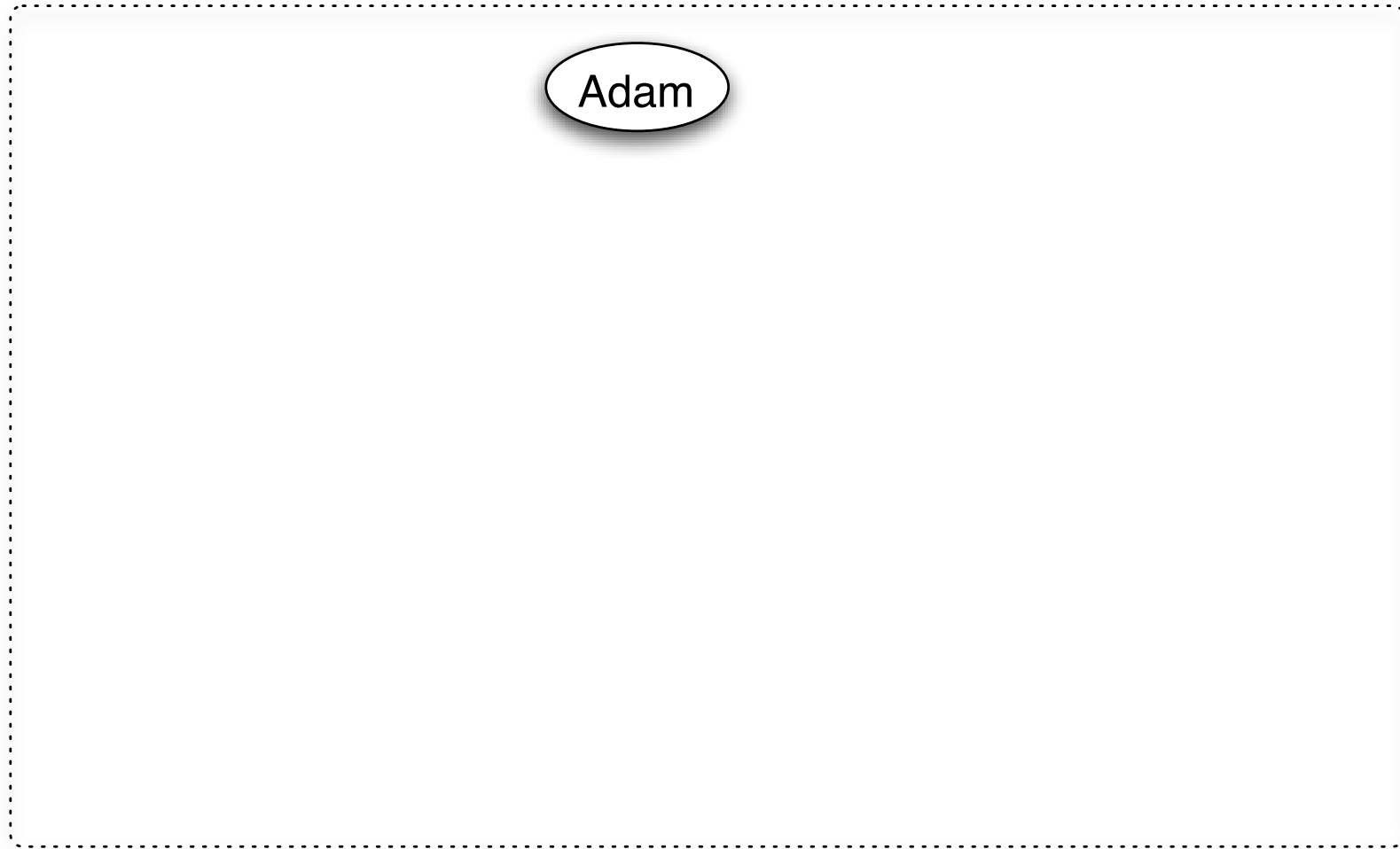
The chain-letter is a dual process:

A person blazing trails through a network of documents,

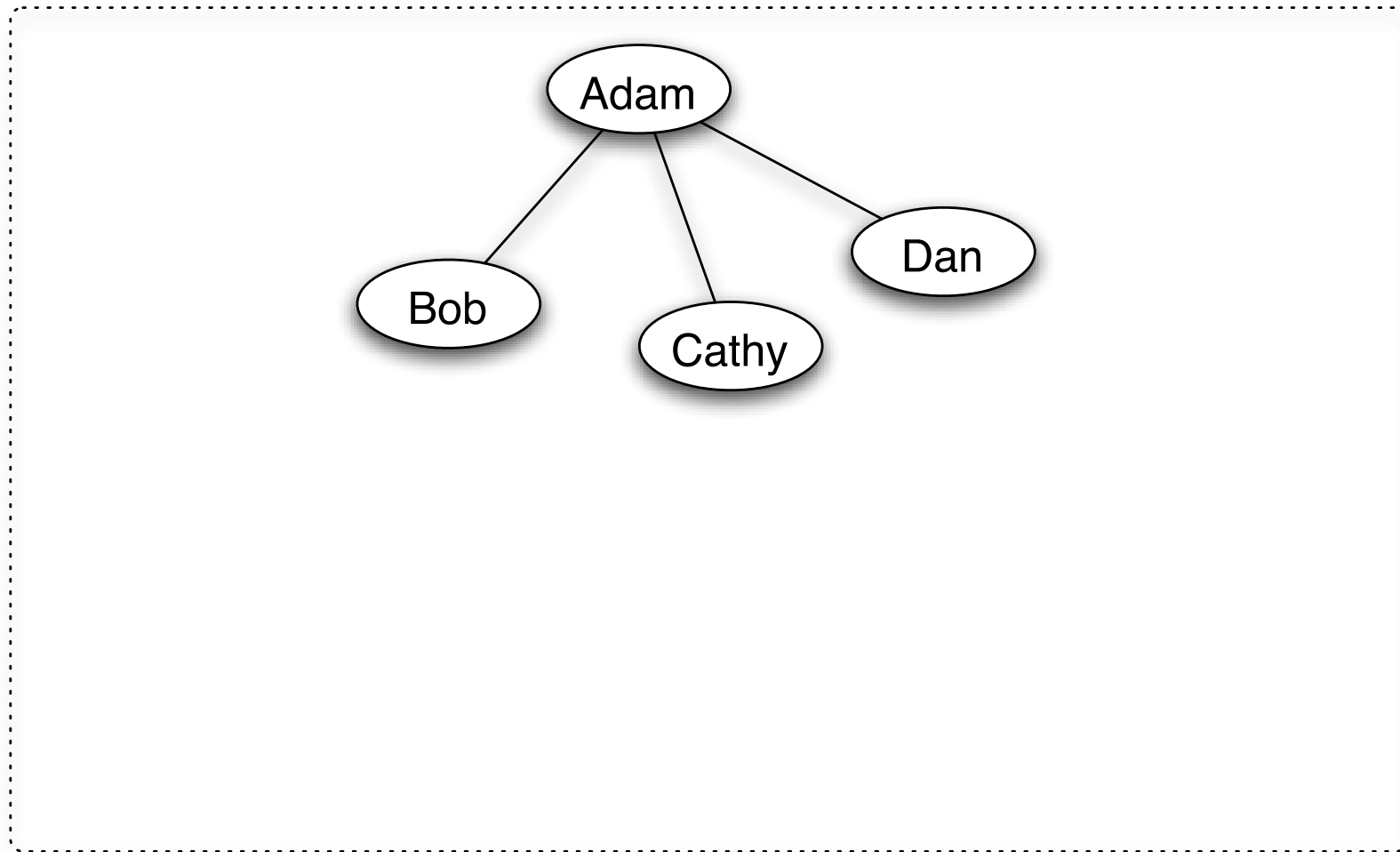
vs.

A document blazing trails through a network of people.

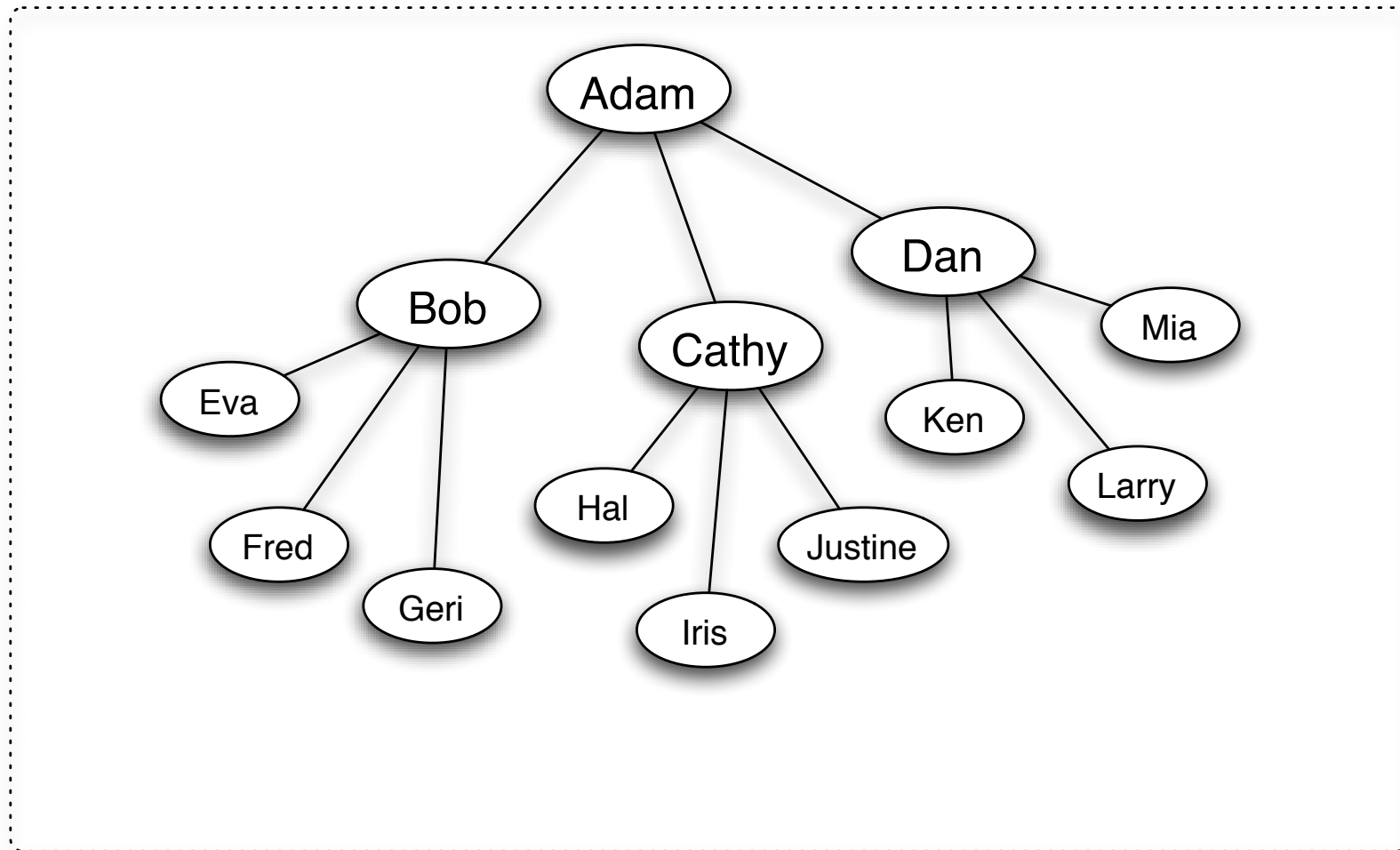
# How Information Spreads (Traditional Picture)



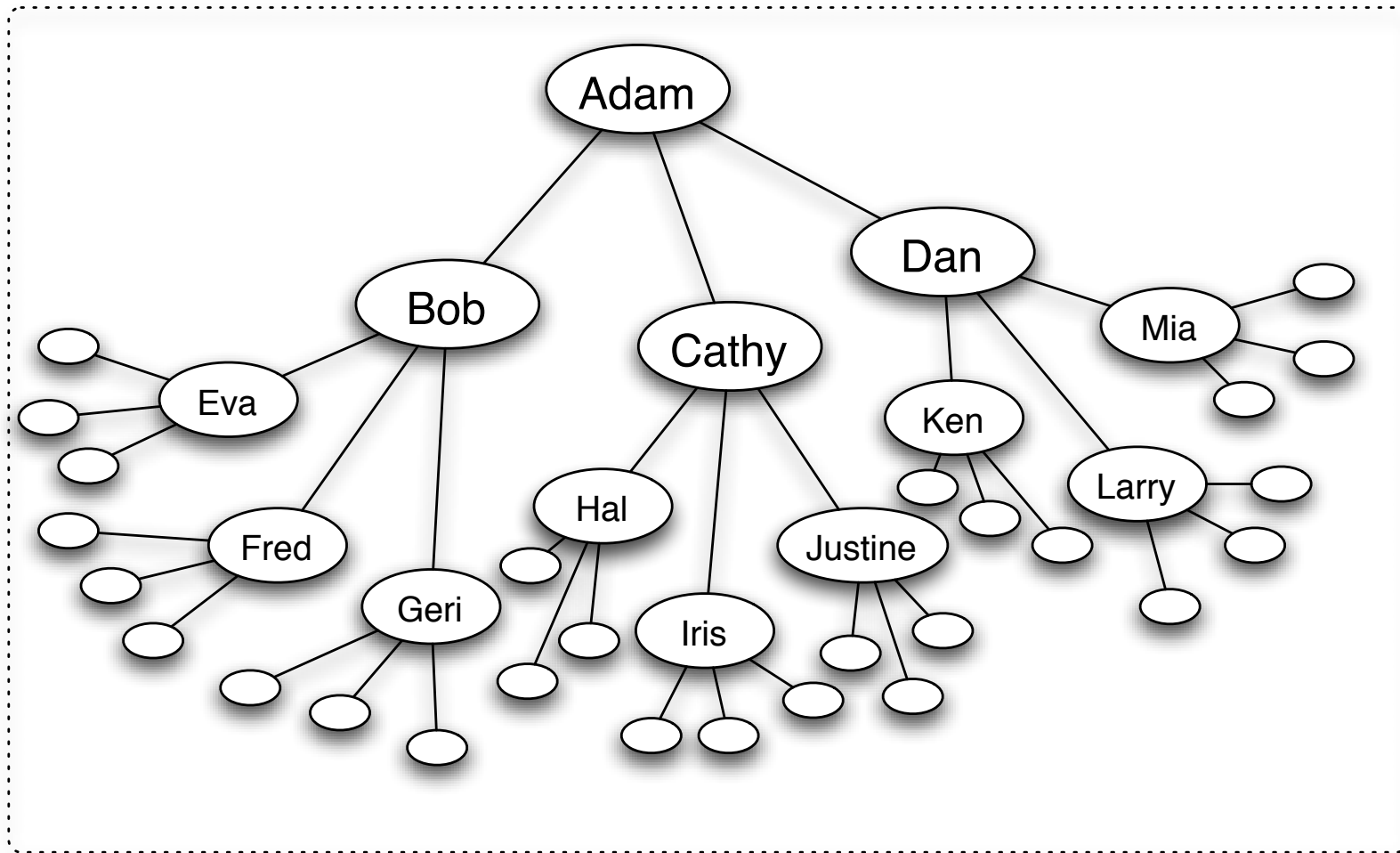
# How Information Spreads (Traditional Picture)



# How Information Spreads (Traditional Picture)

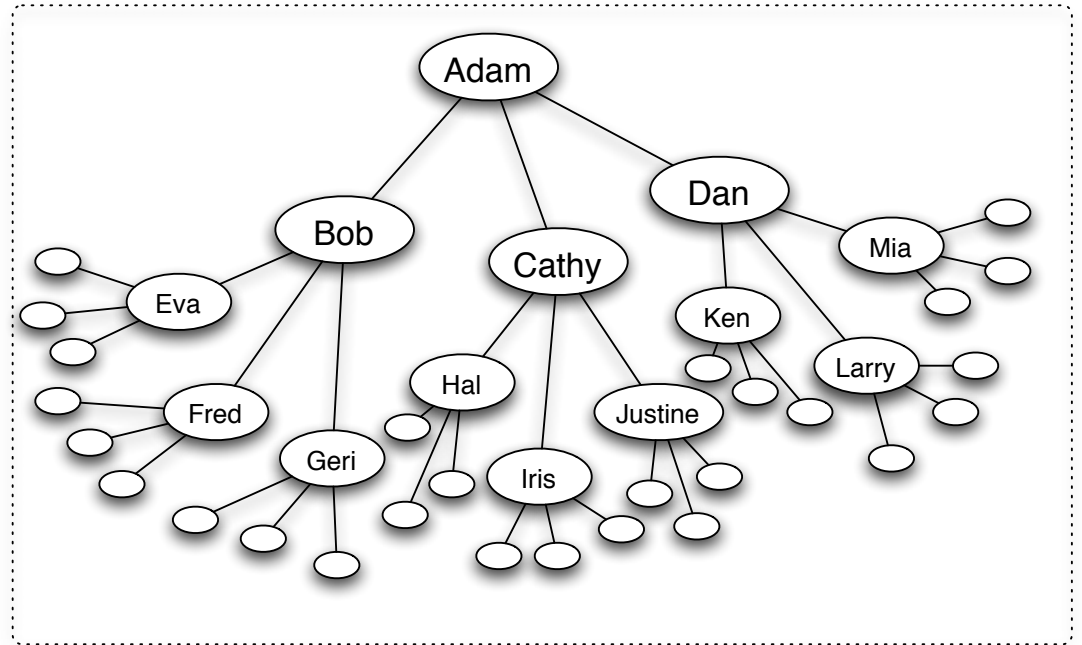


# How Information Spreads (Traditional Picture)



# Assembling a Chain-Letter Tree

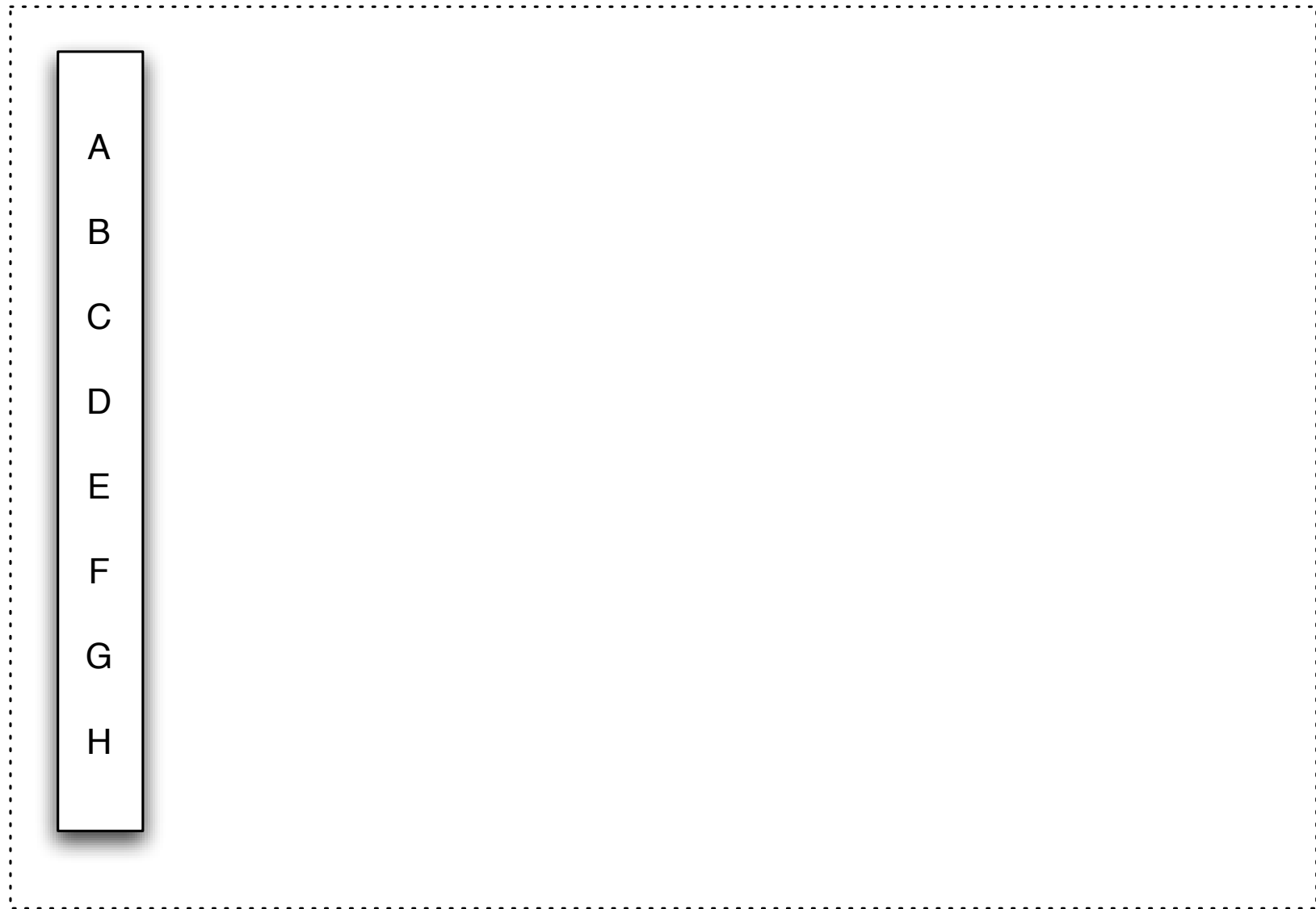
The full tree is unobservable.



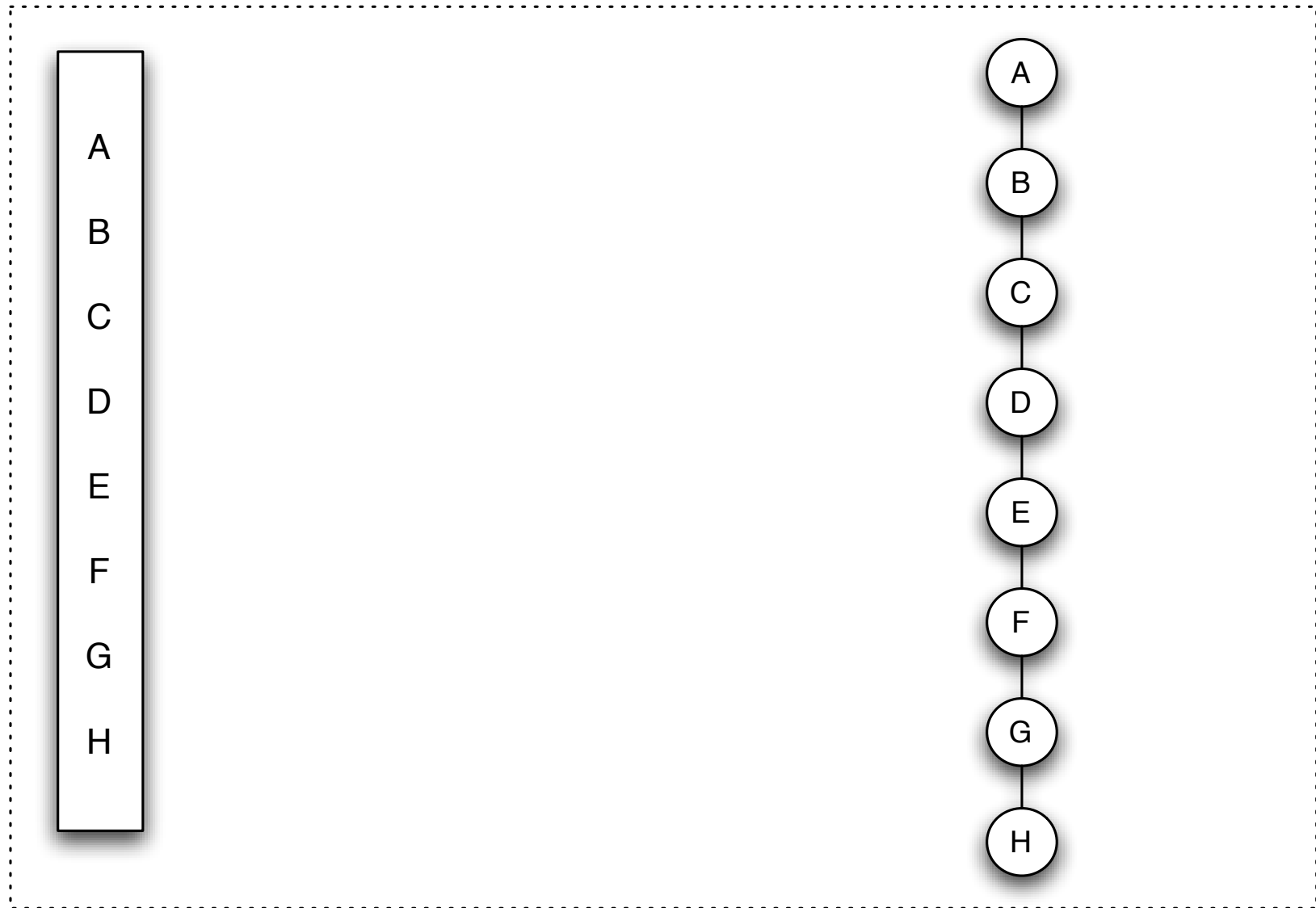
- But hundreds of copies with distinct recipient lists have been posted to mailing lists.
- We can obtain these by Web searches and then assemble a partial tree.



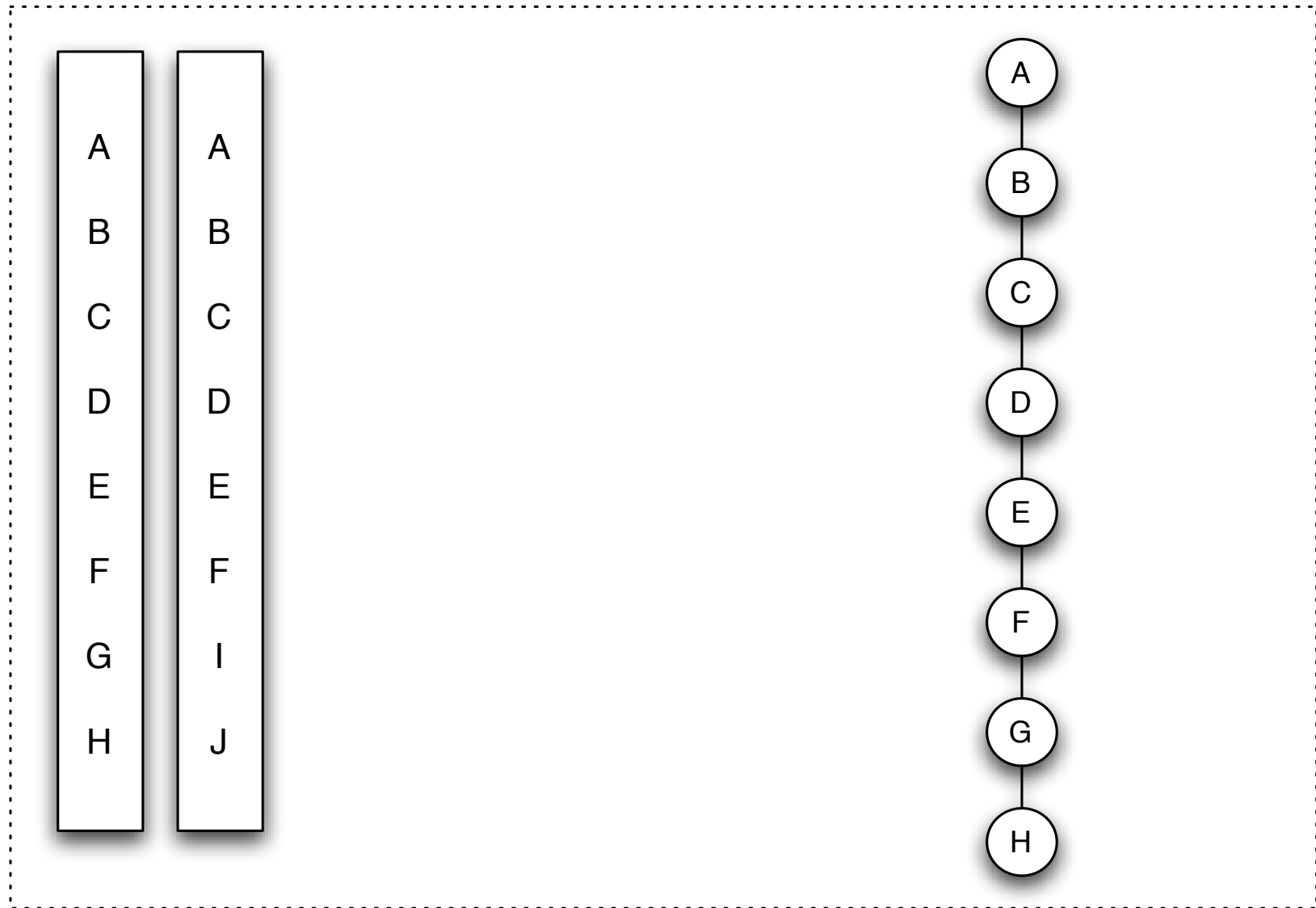
# Assembling a Chain-Letter Tree



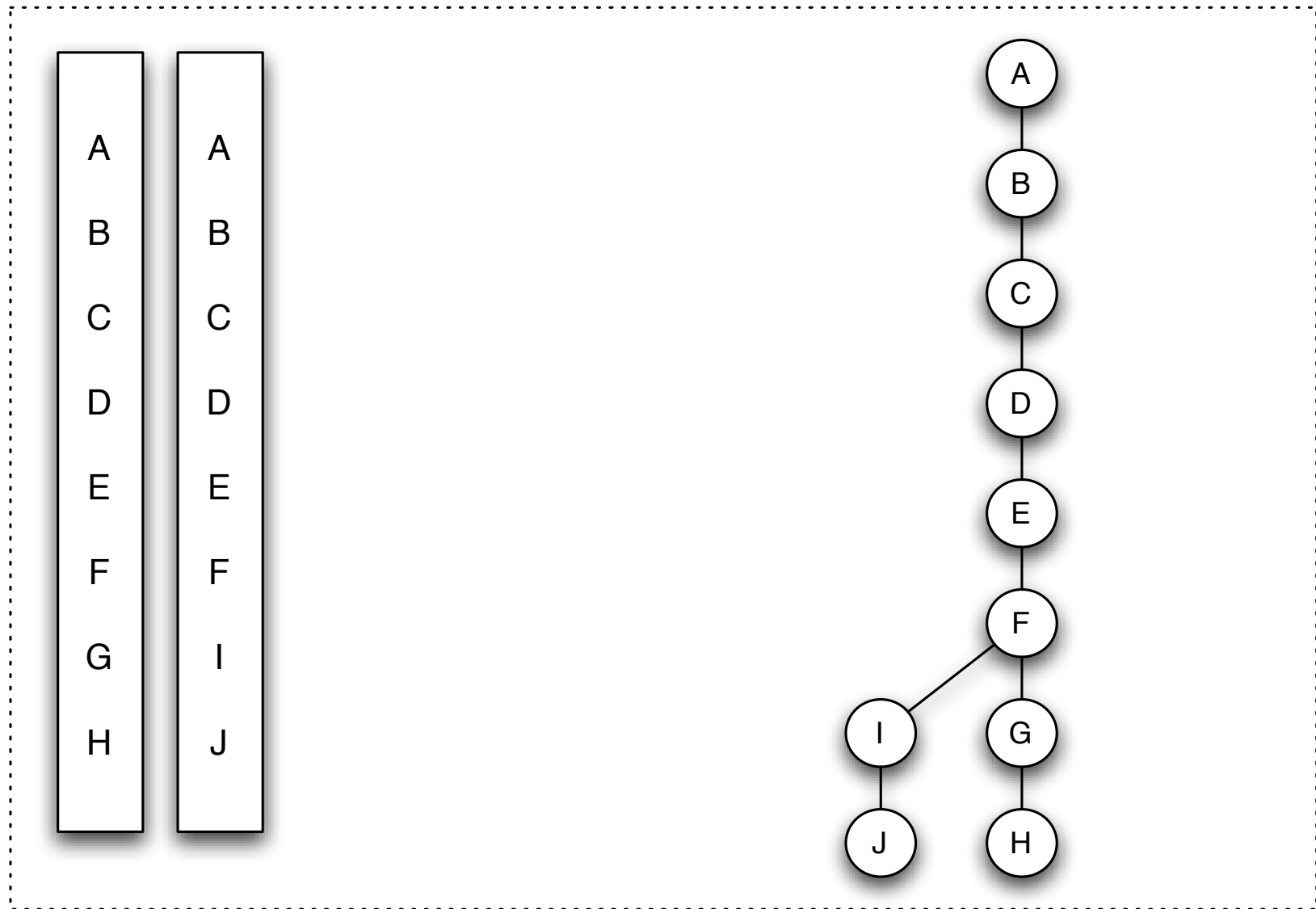
# Assembling a Chain-Letter Tree



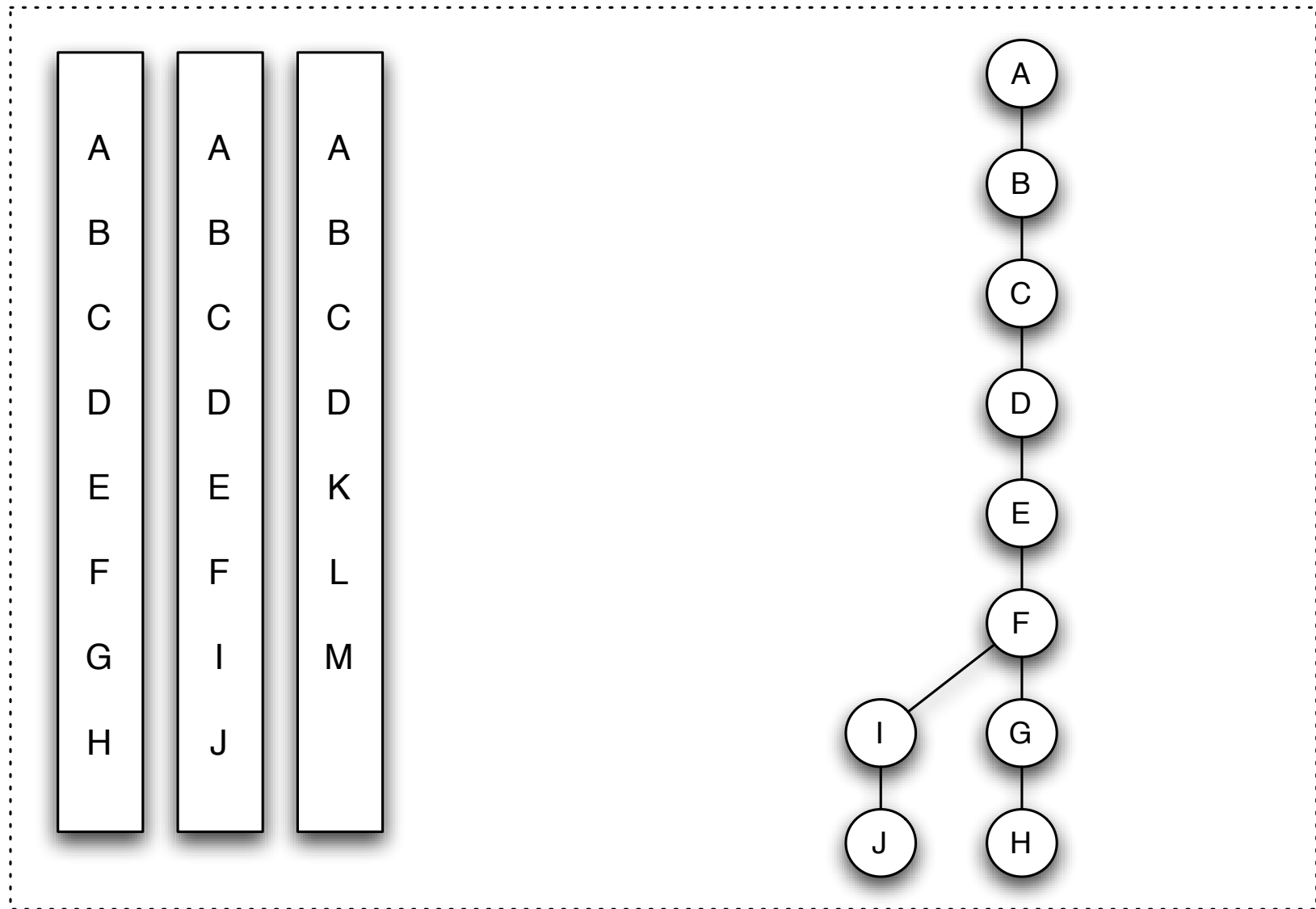
# Assembling a Chain-Letter Tree



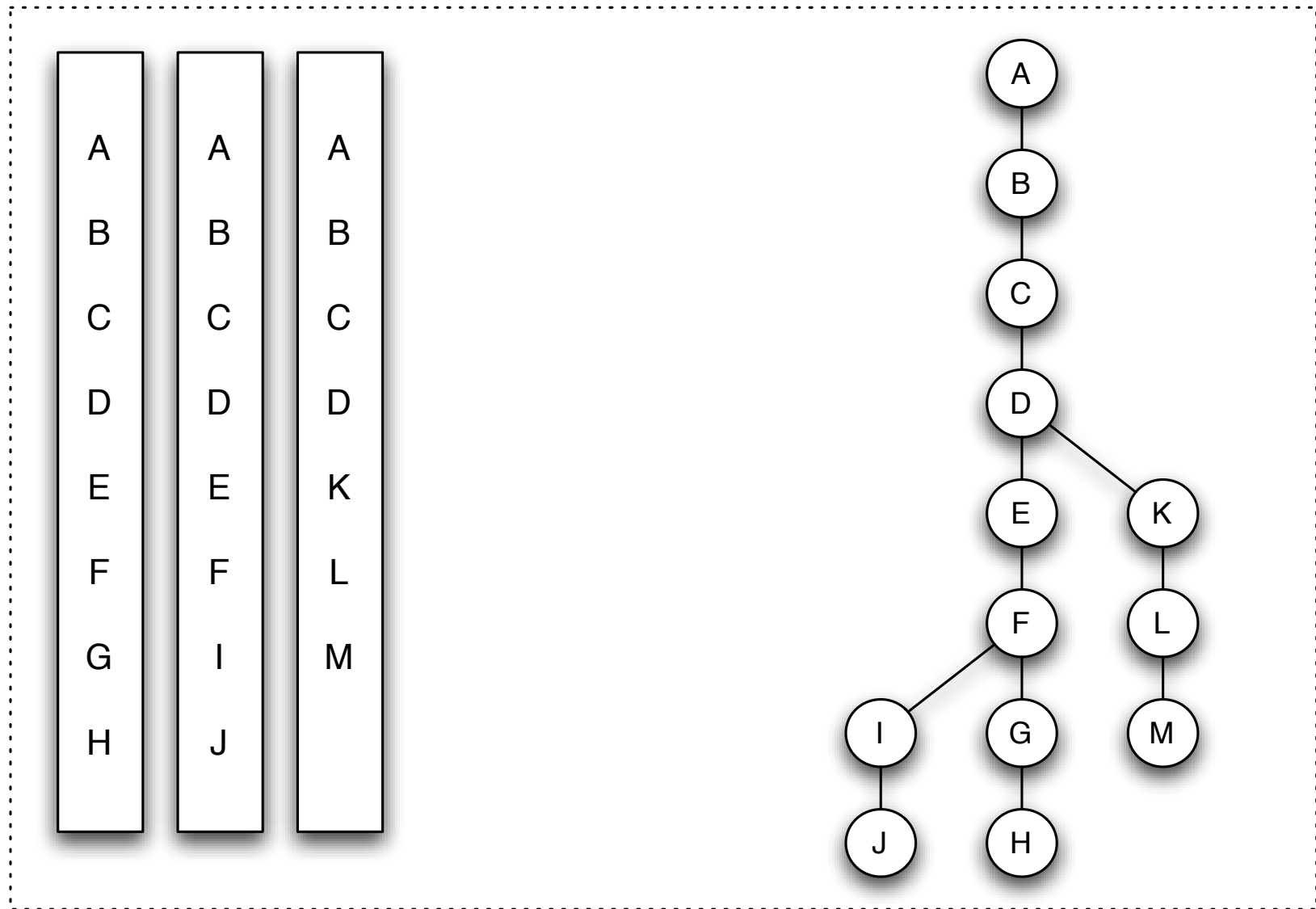
# Assembling a Chain-Letter Tree



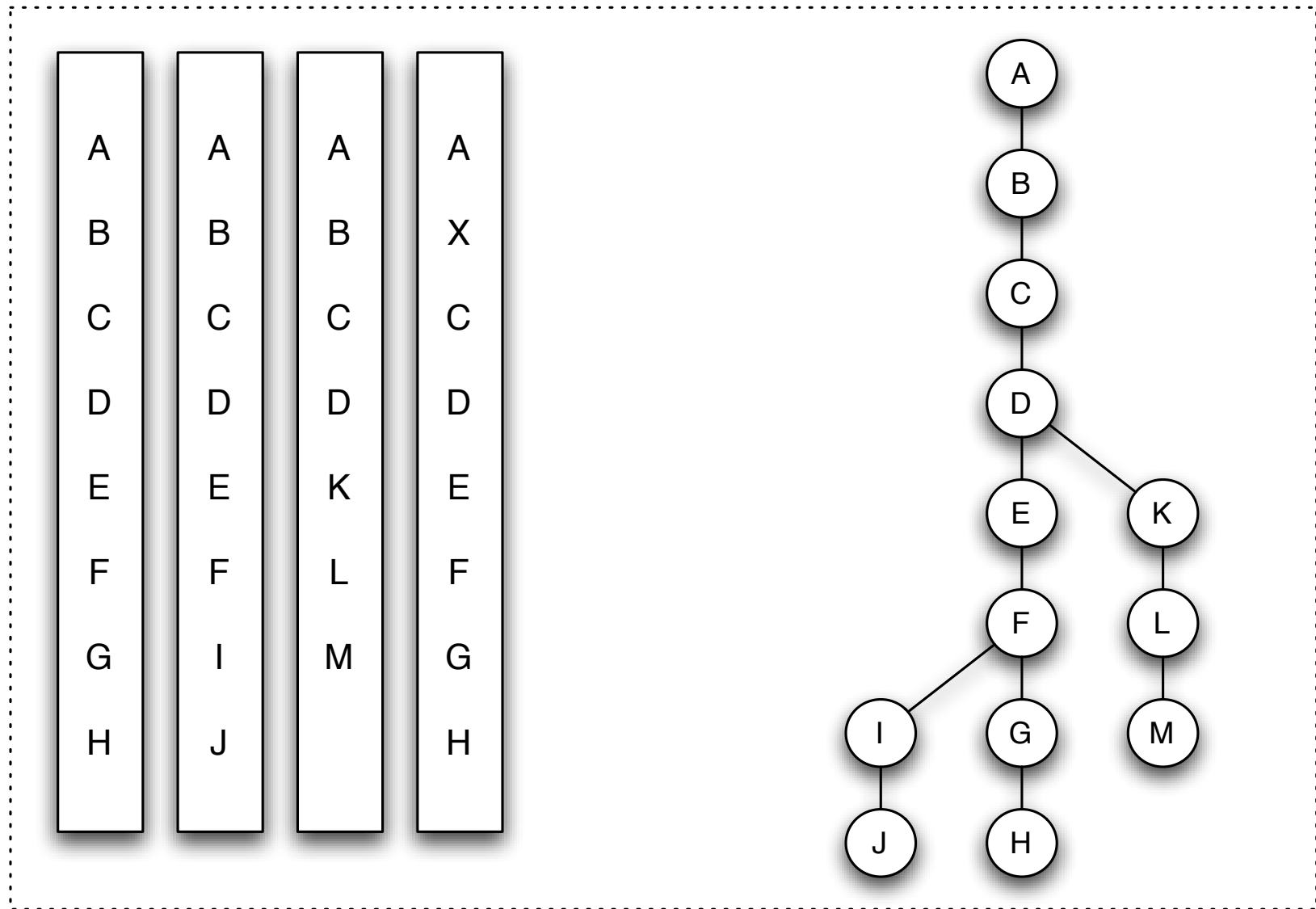
# Assembling a Chain-Letter Tree



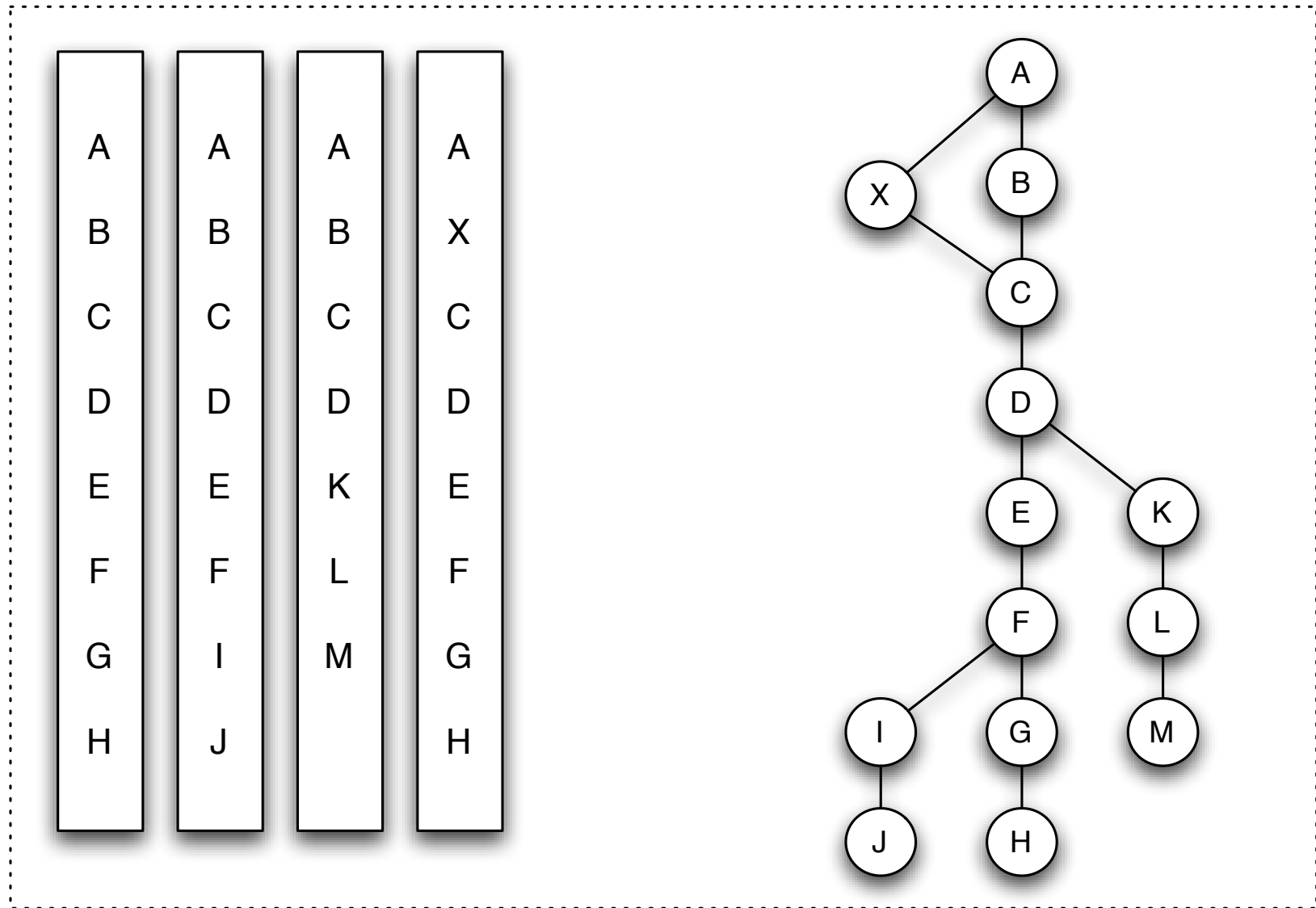
# Assembling a Chain-Letter Tree



# Assembling a Chain-Letter Tree

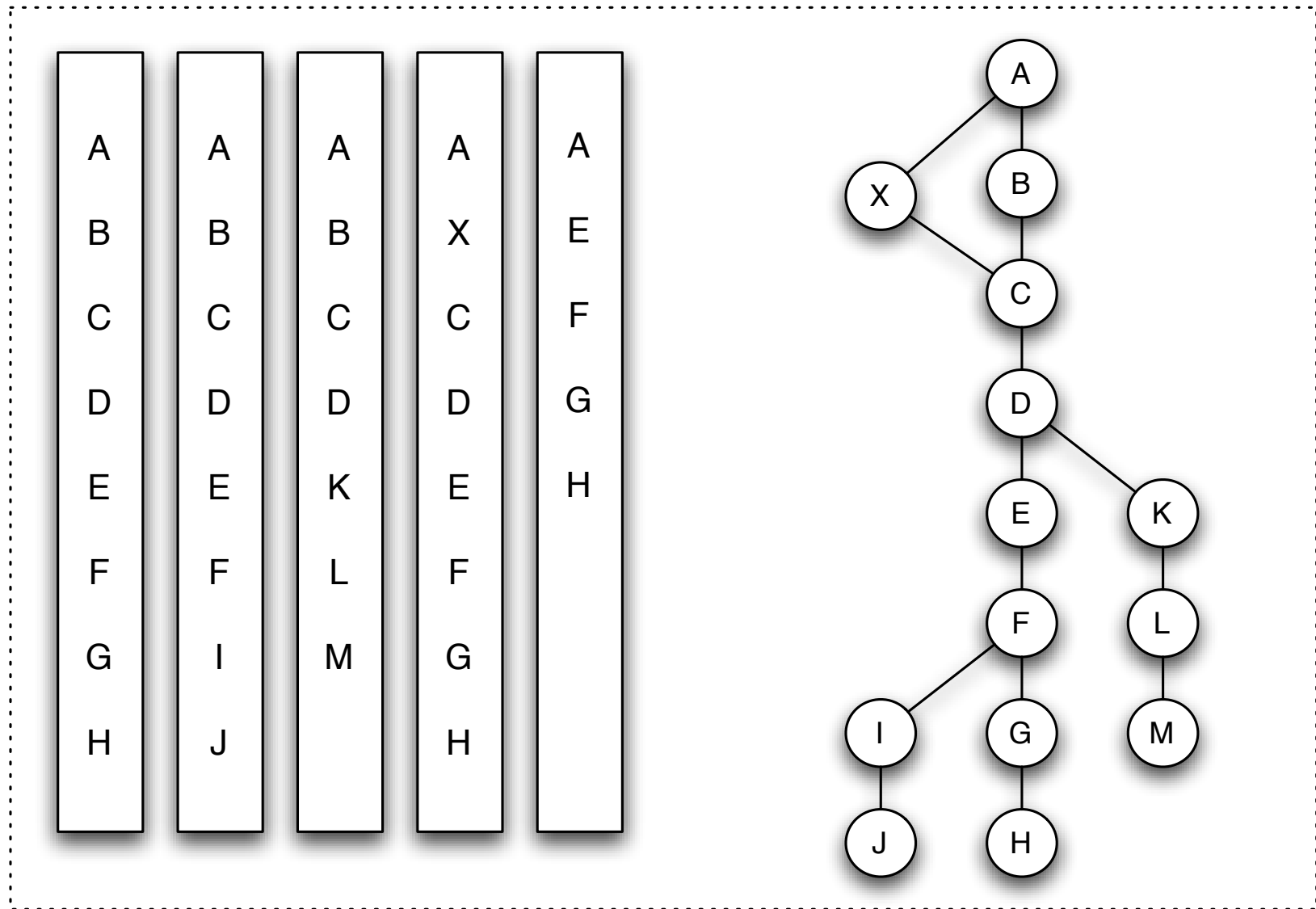


# Assembling a Chain-Letter Tree

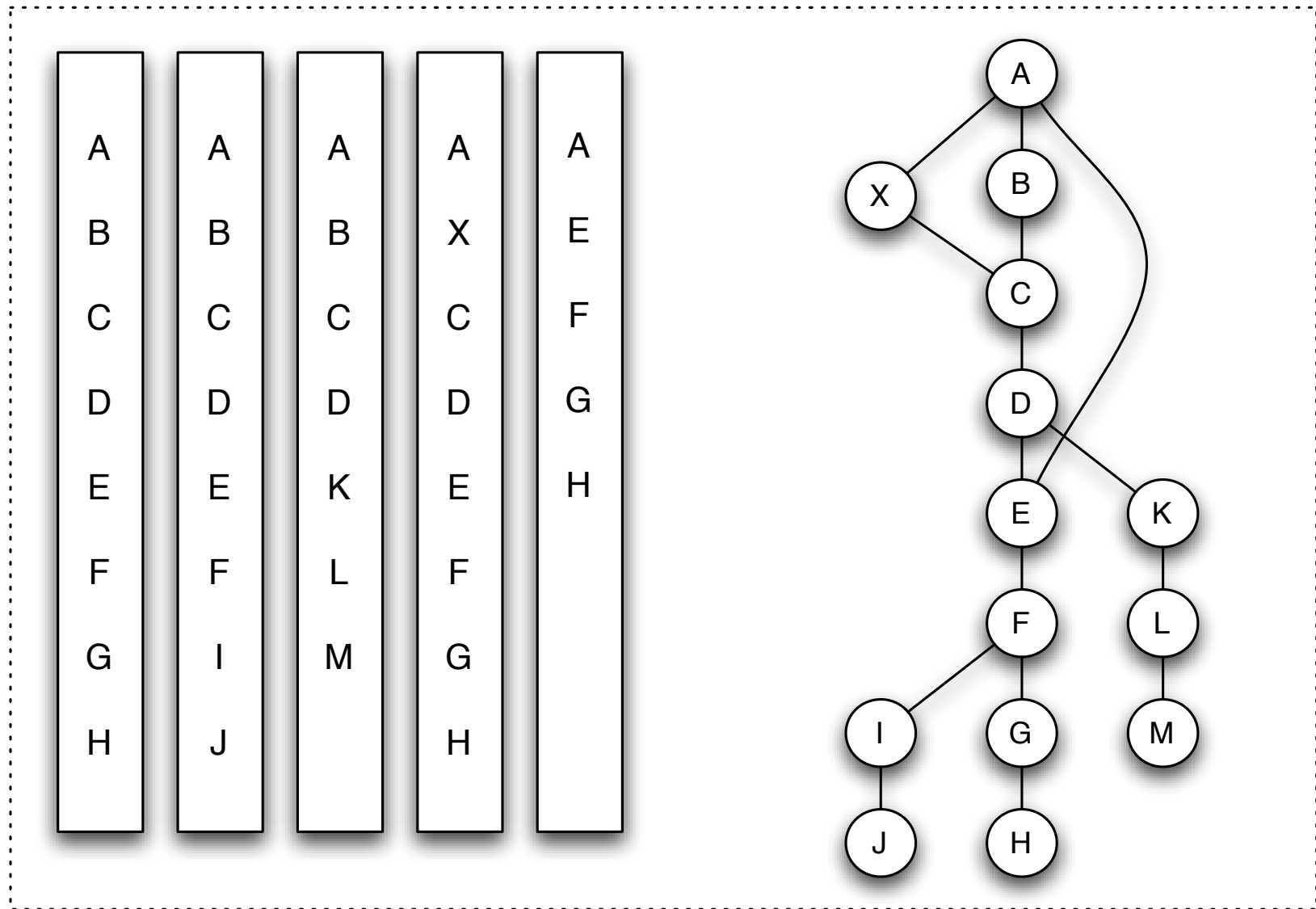




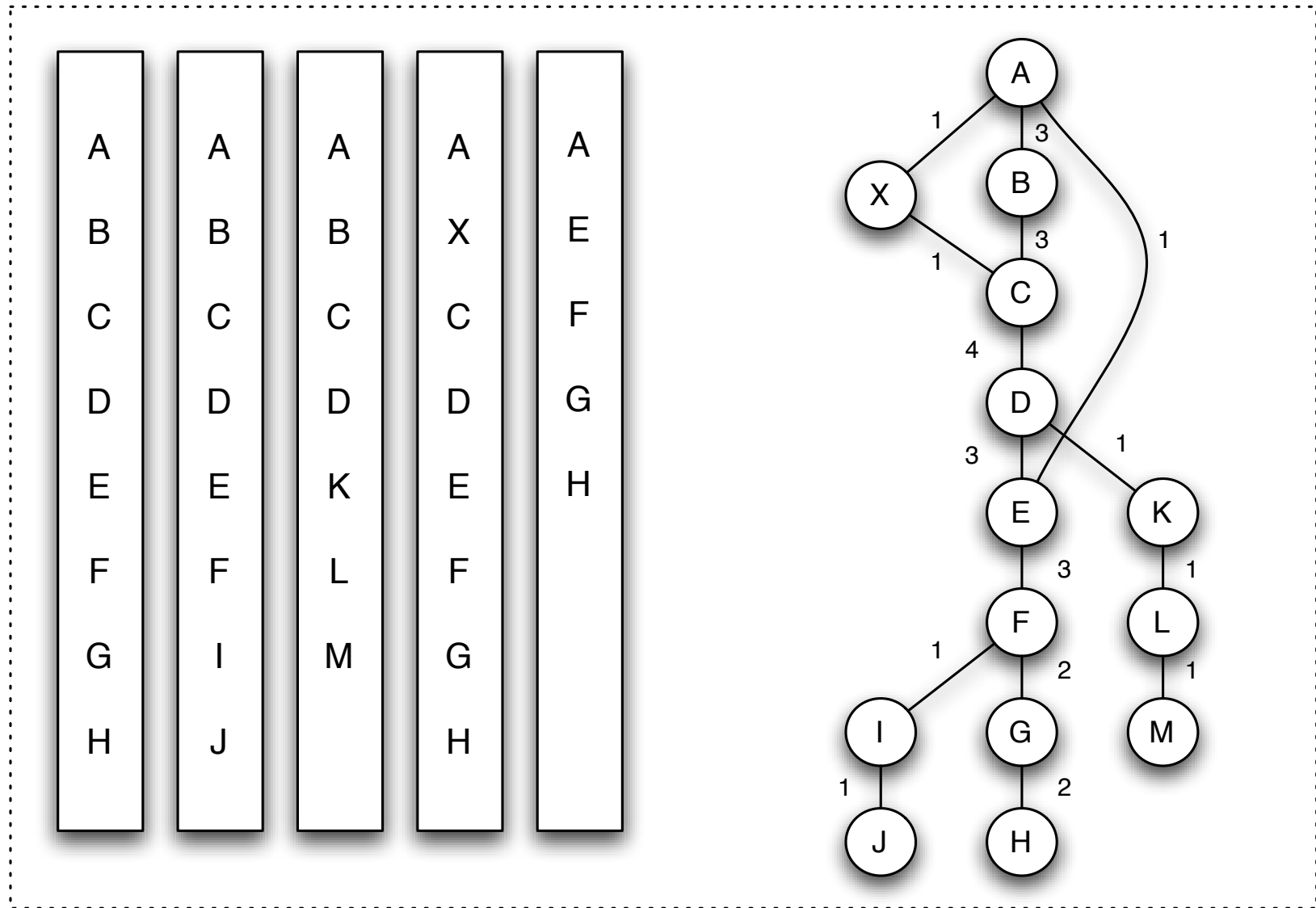
# Assembling a Chain-Letter Tree



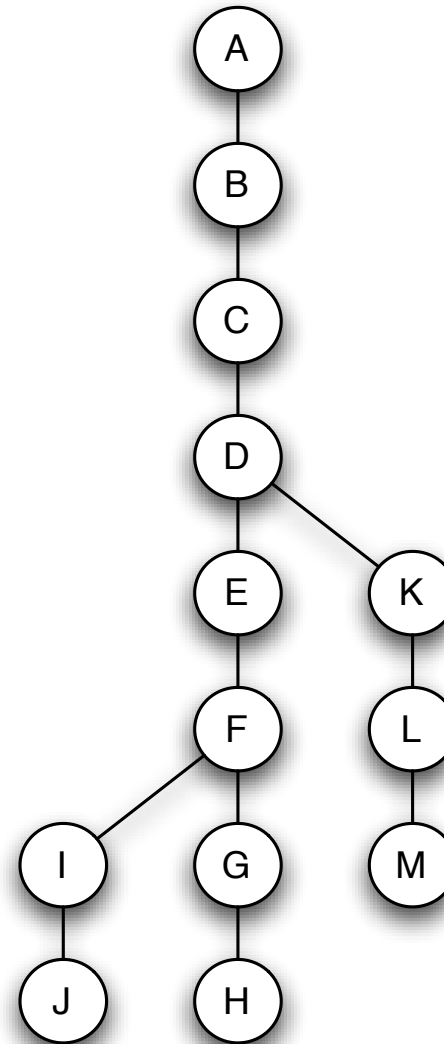
# Assembling a Chain-Letter Tree



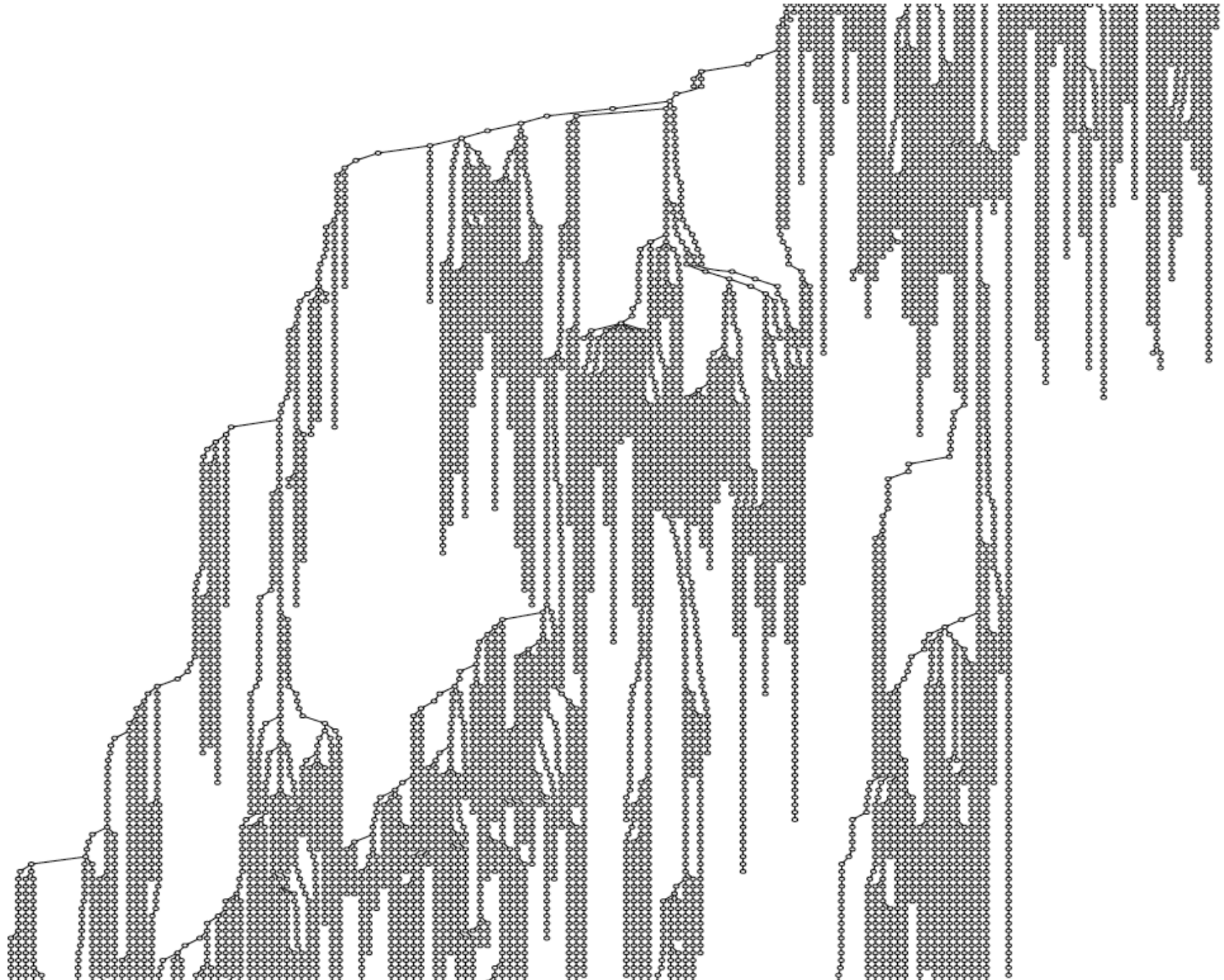
# Assembling a Chain-Letter Tree

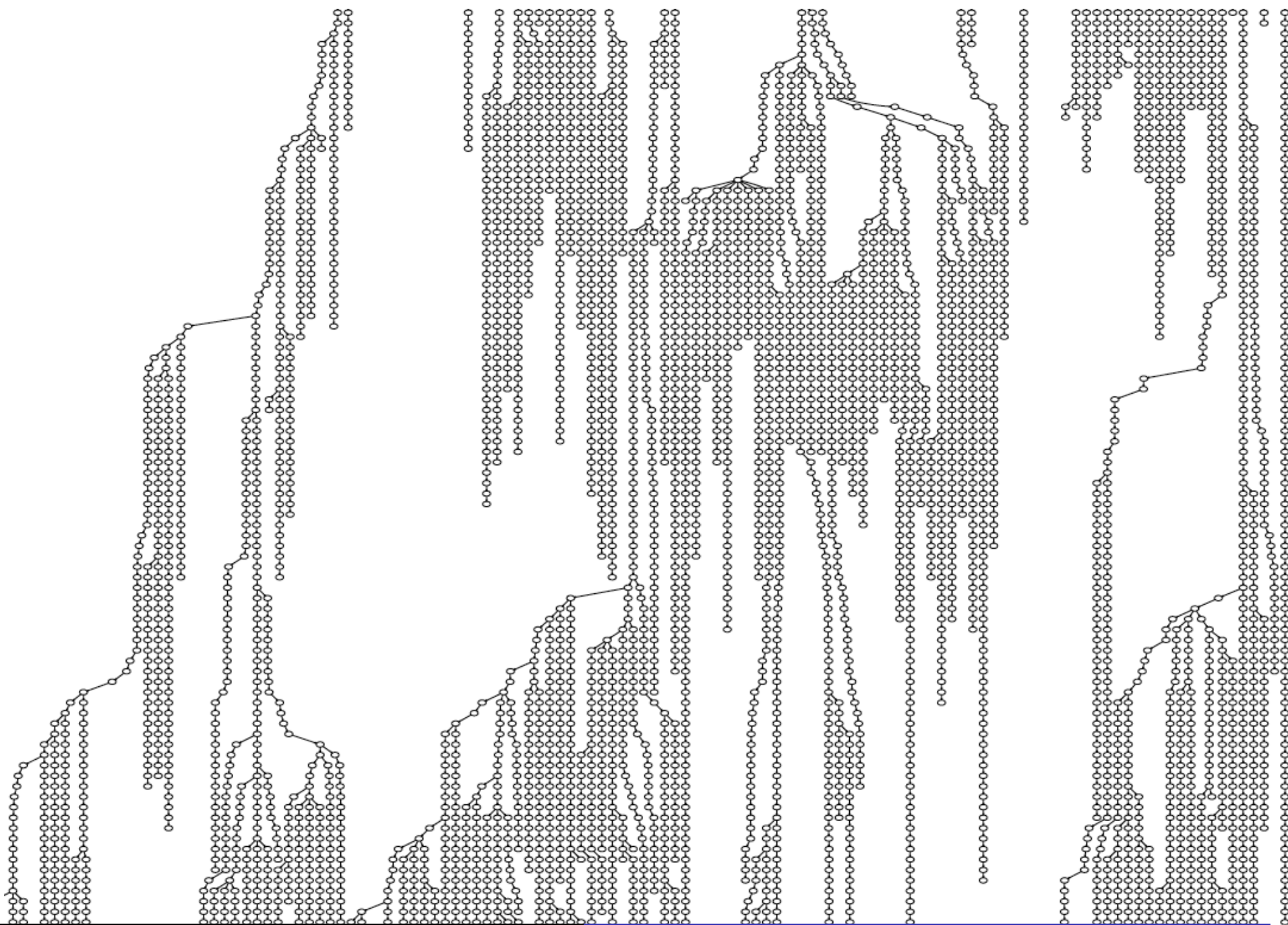


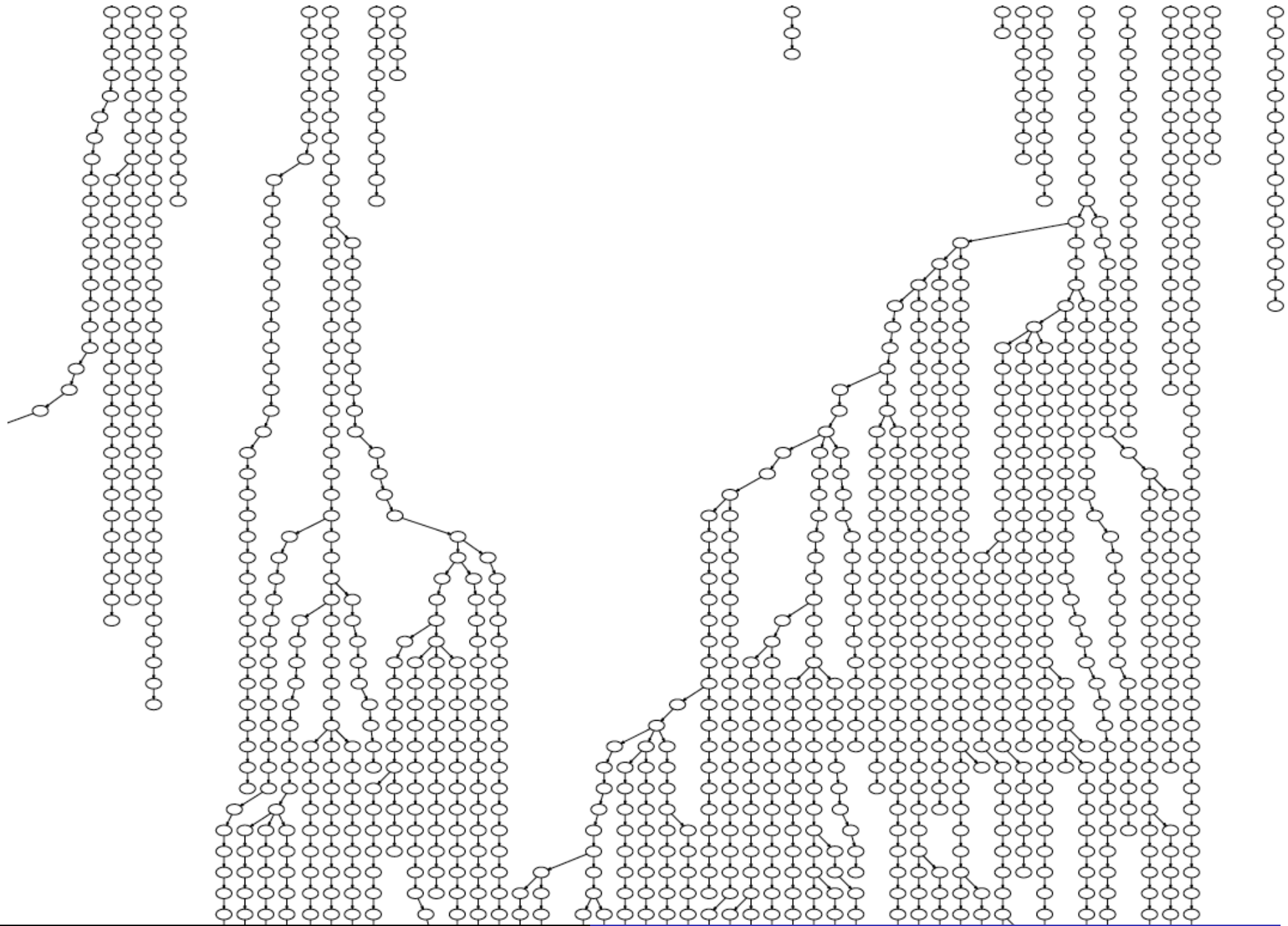
# Assembling a Chain-Letter Tree





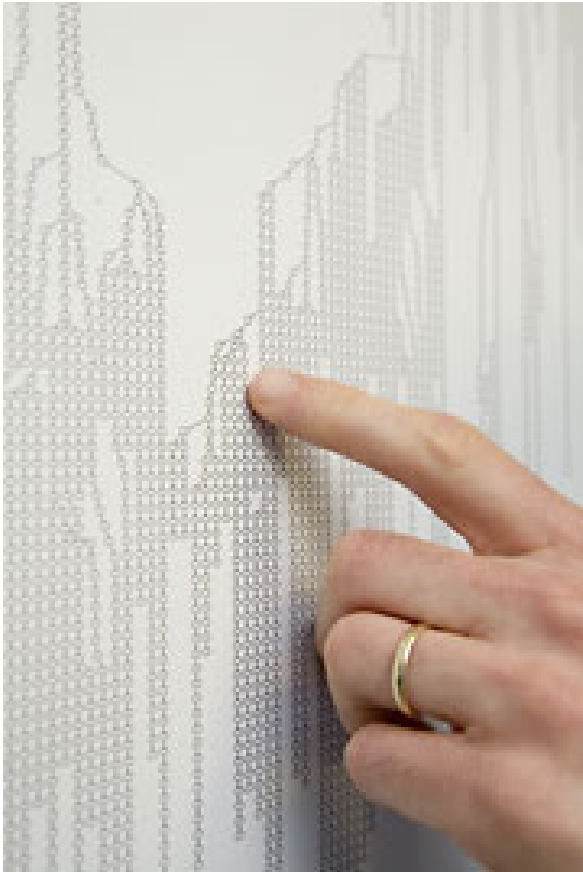








# Modeling the Structure of the Tree



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.

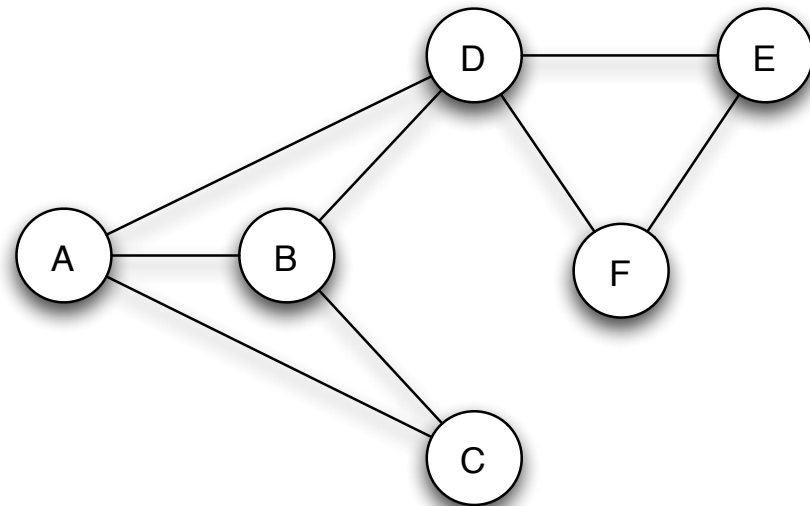
Some plausible models that can produce trees of this shape:

- (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?

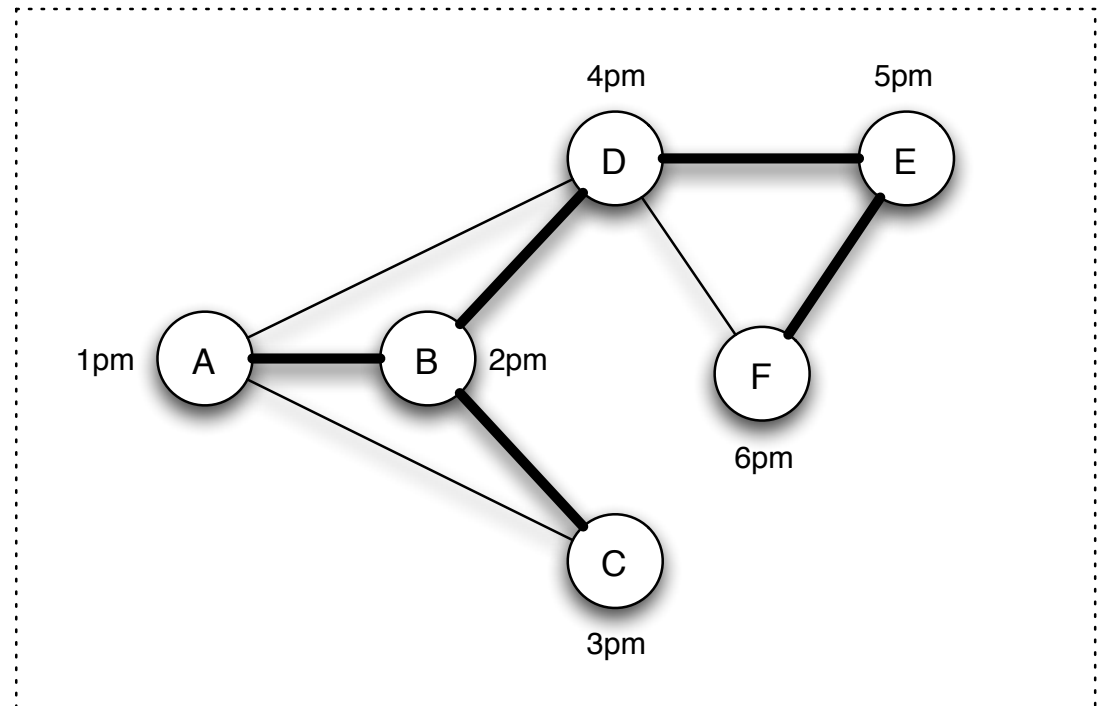
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



# Why is the Tree So Deep and Narrow?

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

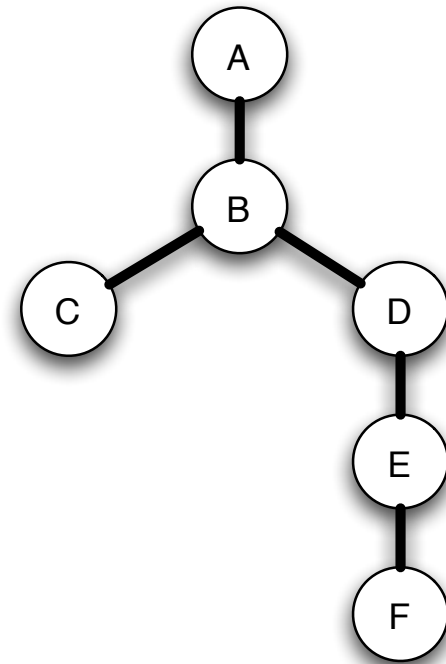
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



# Why is the Tree So Deep and Narrow?

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

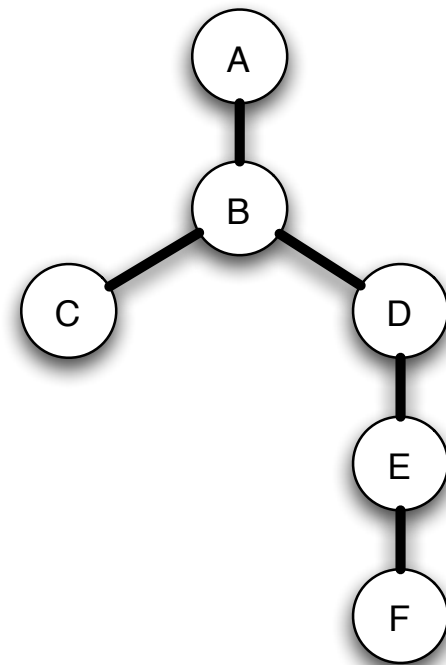
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



# Why is the Tree So Deep and Narrow?

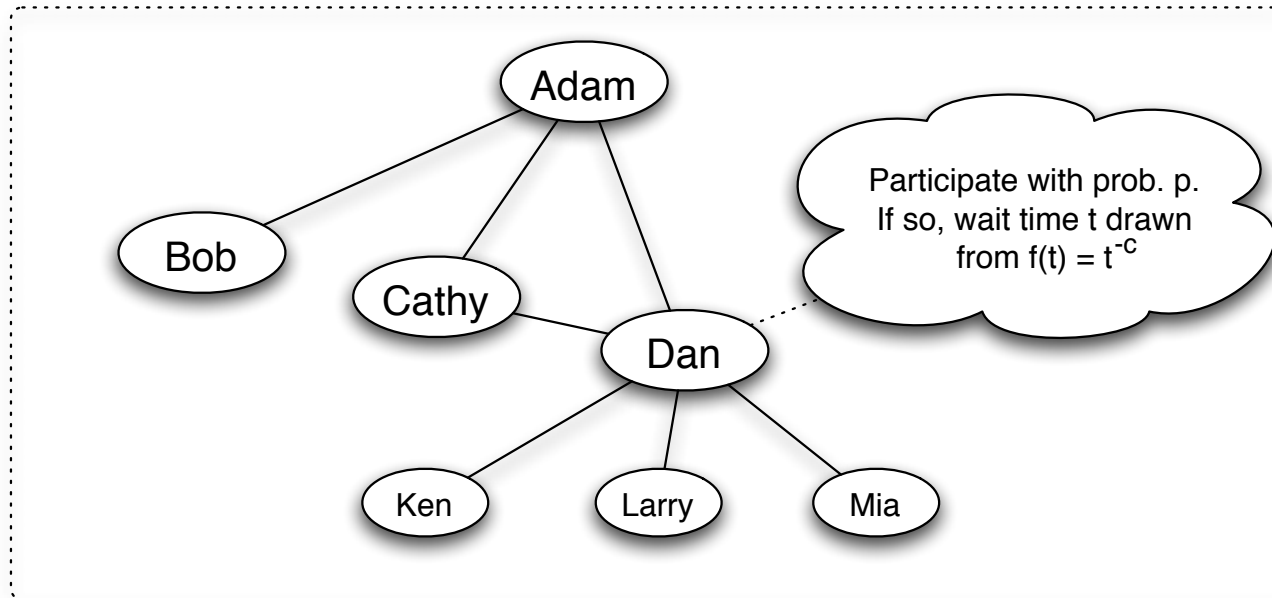
It looks like a depth-first search tree. But why?

- 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 Iraq-war chain letter.

# Timing-Based Models for Tree Structure



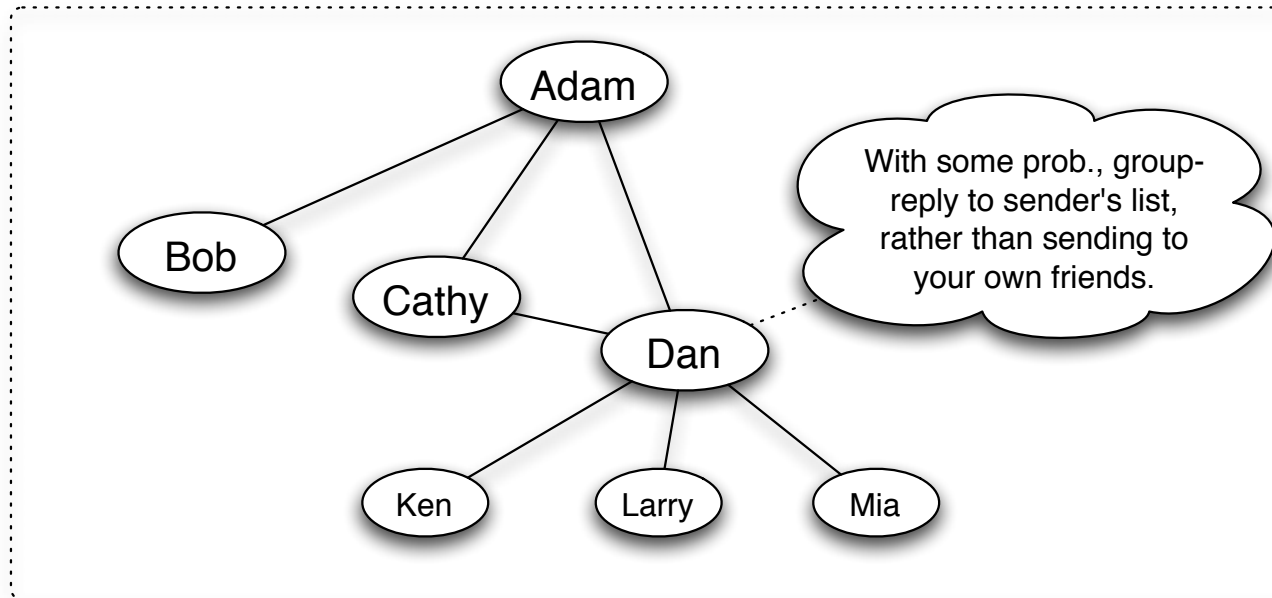
When a node  $v$  in the network first gets a copy of the message,

- $v$  participates in the chain-letter with prob.  $p$ .
- If so, waits time  $t$  before forwarding ( $t$  from  $f(t) = t^{-c}$ .)

Produces “elongated” trees when simulated in real networks.

- To get depth of real tree, need to let nodes “group-reply.”
- Open: Prove this yields asymptotically deeper trees in natural random-graph model.

# Timing-Based Models for Tree Structure



When a node  $v$  in the network first gets a copy of the message,

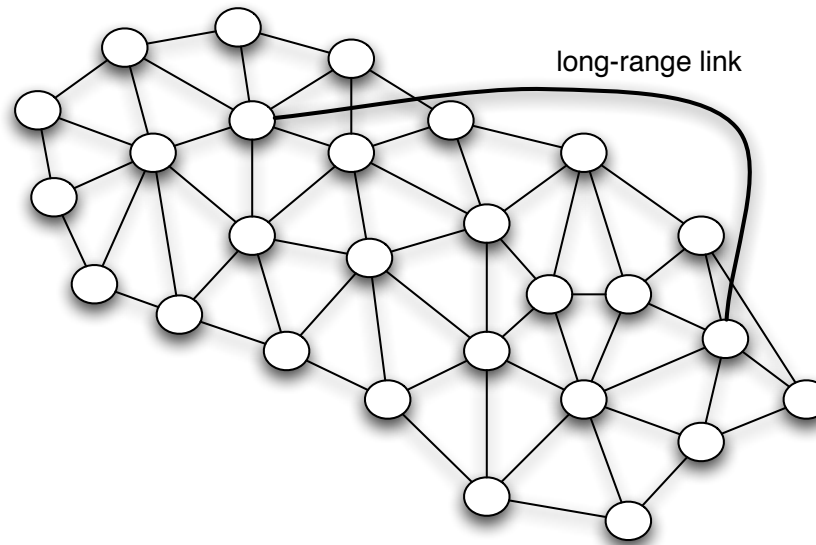
- $v$  participates in the chain-letter with prob.  $p$ .
- If so, waits time  $t$  before forwarding ( $t$  from  $f(t) = t^{-c}$ .)

Produces “elongated” trees when simulated in real networks.

- To get depth of real tree, need to let nodes “group-reply.”
- Open: Prove this yields asymptotically deeper trees in natural random-graph model.

# 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]

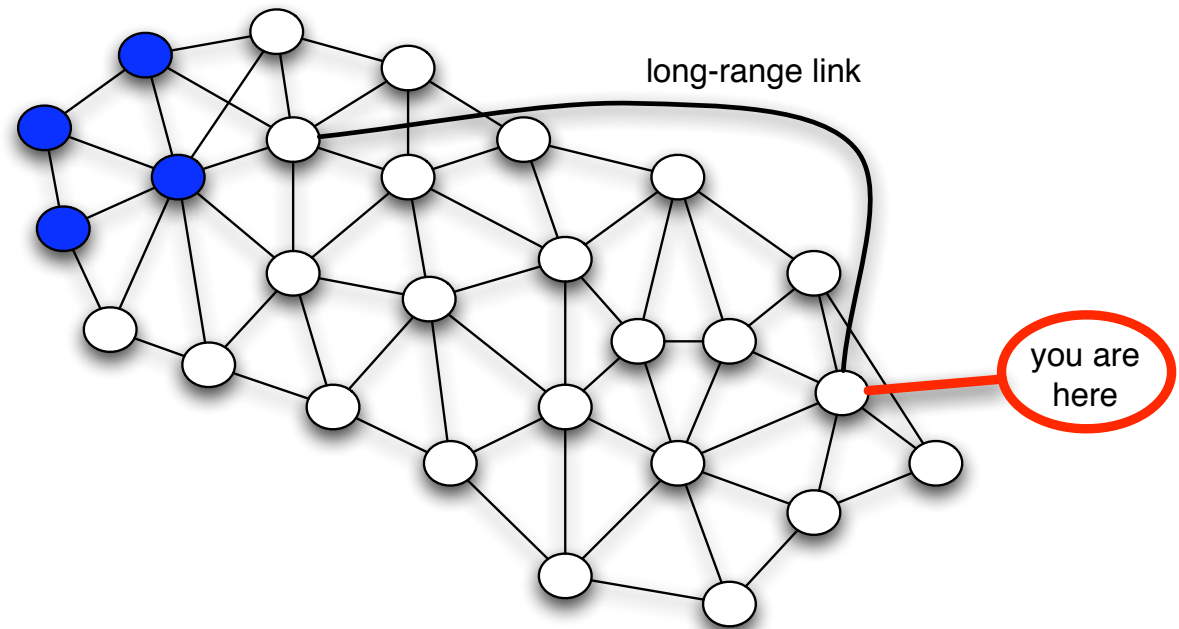


# Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

- Suppose people needed two stimuli to be willing to participate.

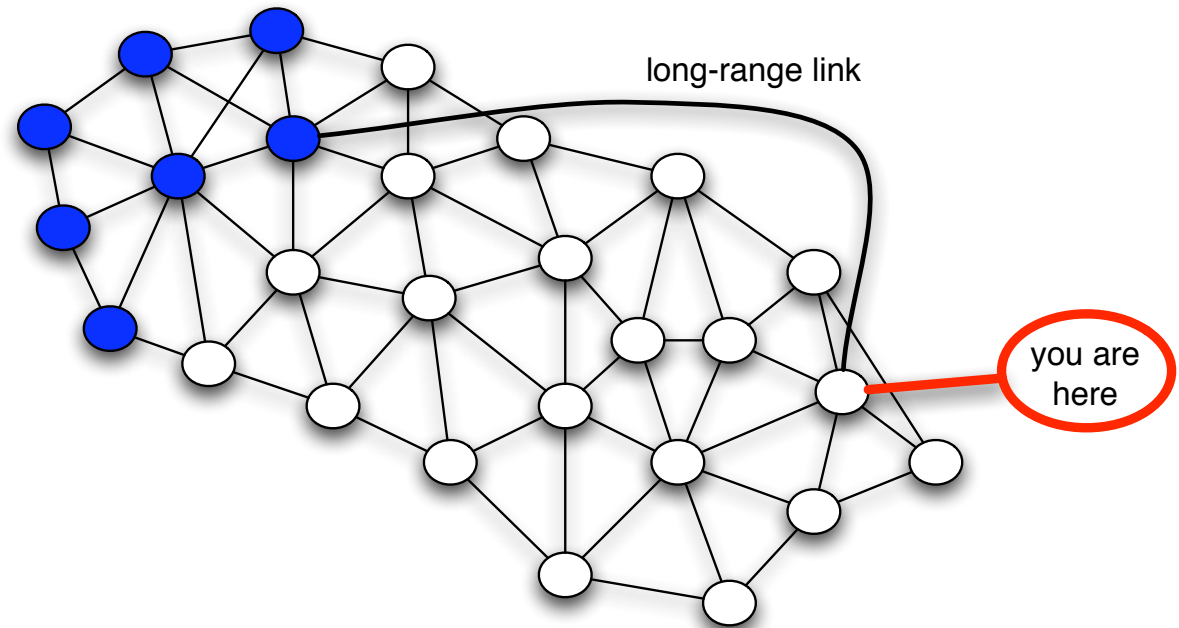


# Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

- Suppose people needed two stimuli to be willing to participate.

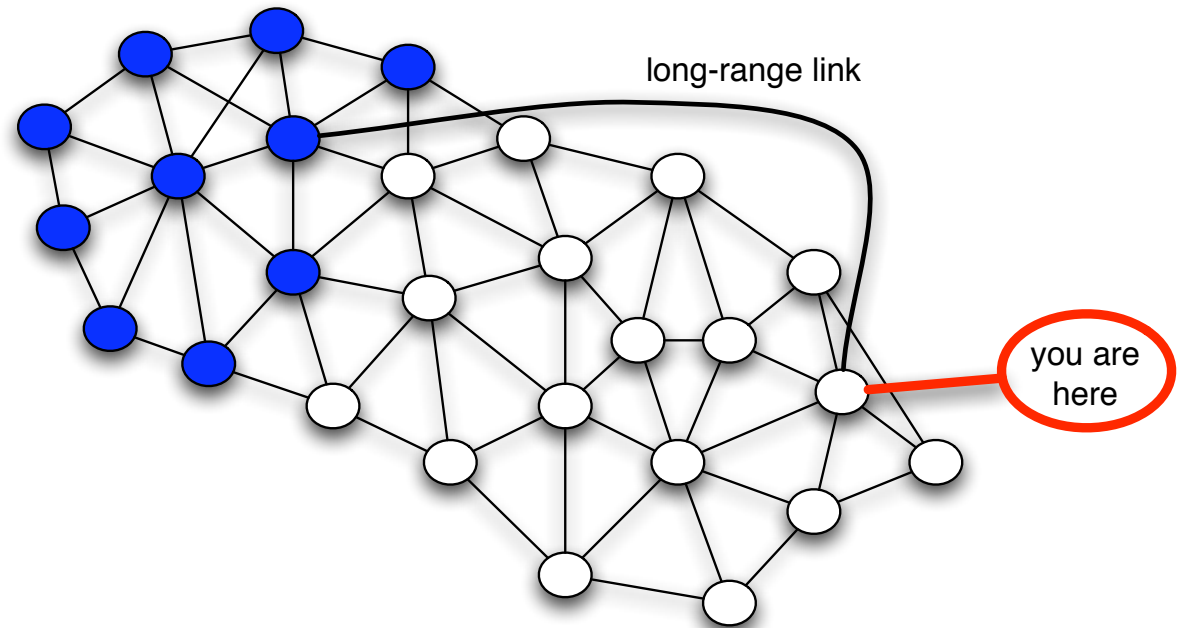


# Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

- Suppose people needed two stimuli to be willing to participate.

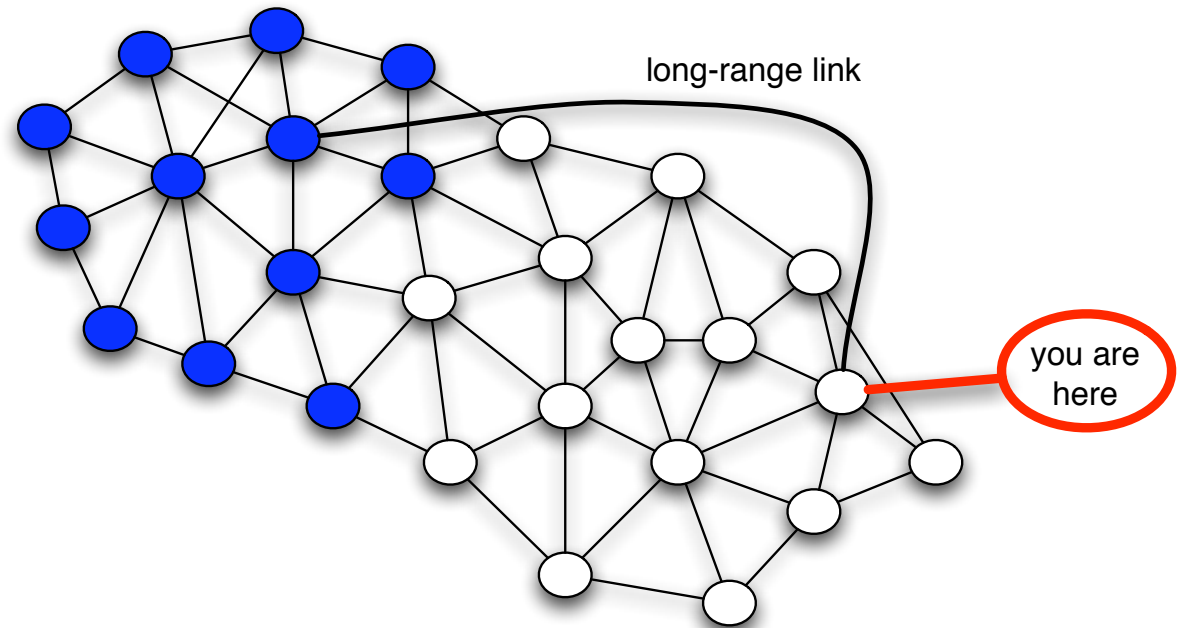


# Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

- Suppose people needed two stimuli to be willing to participate.

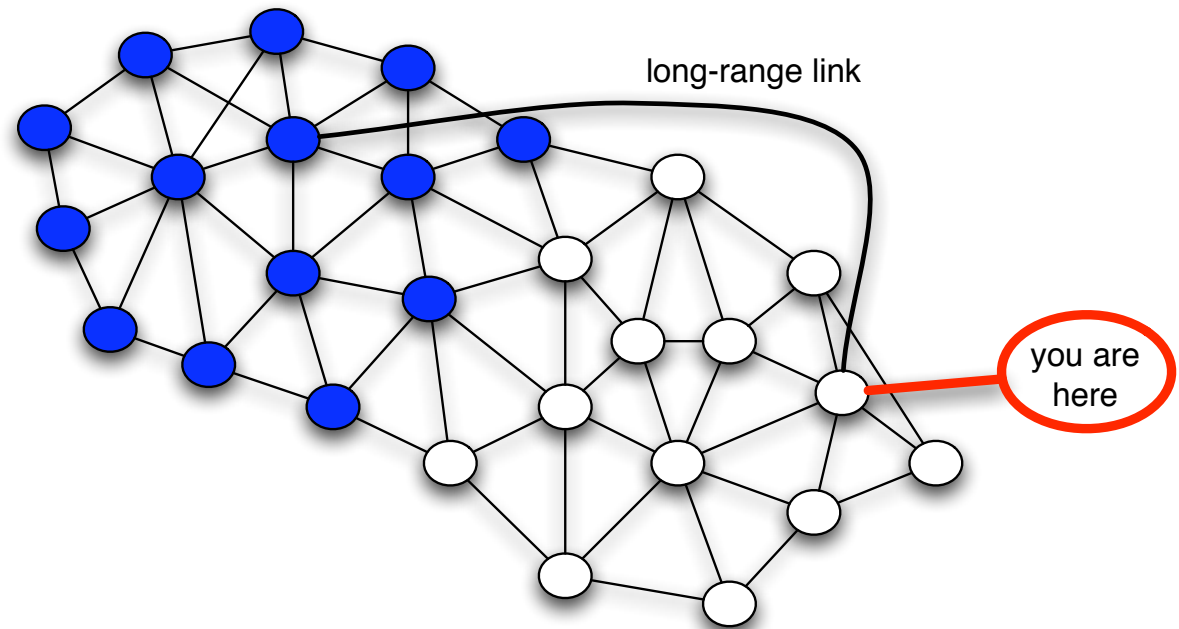


# Spatial Clustering and Thresholds

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.

# Protecting Privacy in Social Network Data

Many large datasets based on communication (e-mail, IM, voice) where users have strong privacy expectations.

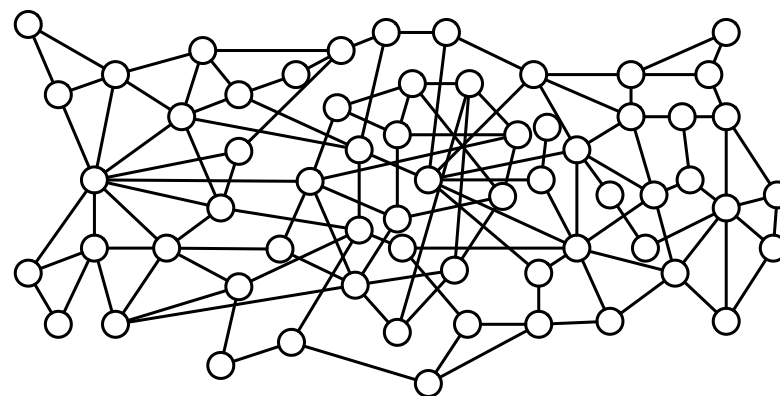
- Current safeguards based on anonymization: replace node names with random IDs.

With more detailed data, anonymization has run into trouble:

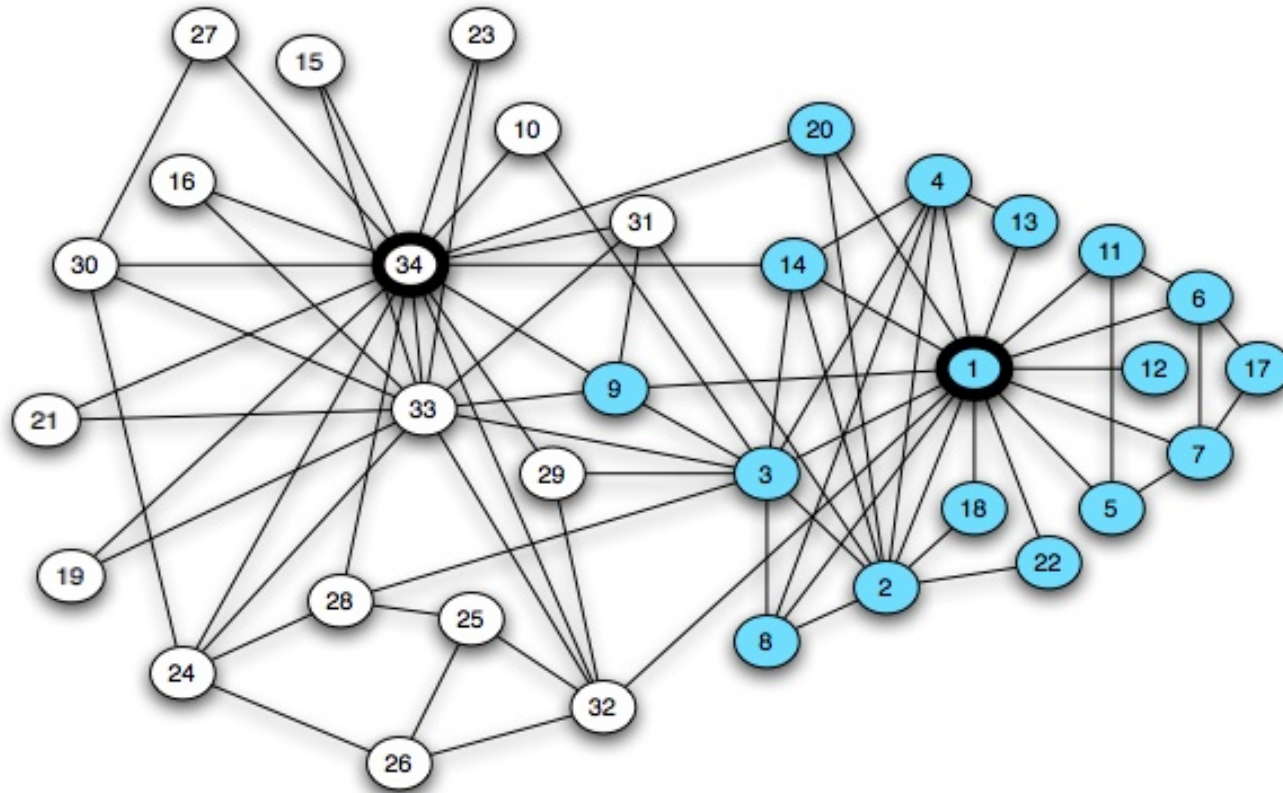
- Identifying on-line pseudonyms by textual analysis [Novak-Raghavan-Tomkins 2004]
- De-anonymizing Netflix ratings via time series [Narayanan-Shmatikov 2006]
- Search engine query logs: identifying users from their queries.

Our setting is much starker;  
does this make things safer?

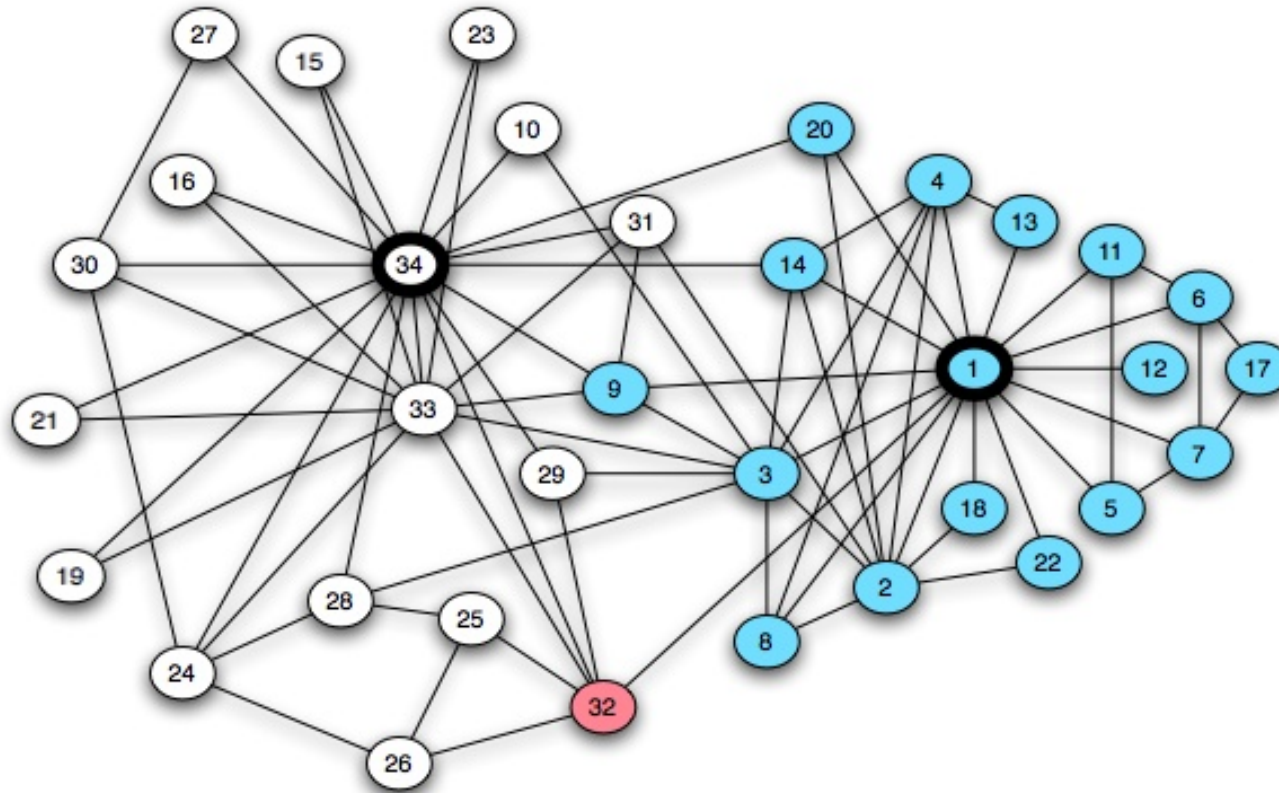
- E.g. no text, time-stamps, or node attributes
- Just a graph with nodes numbered  $1, 2, 3, \dots, n$ .



# An Example of What Can Go Wrong



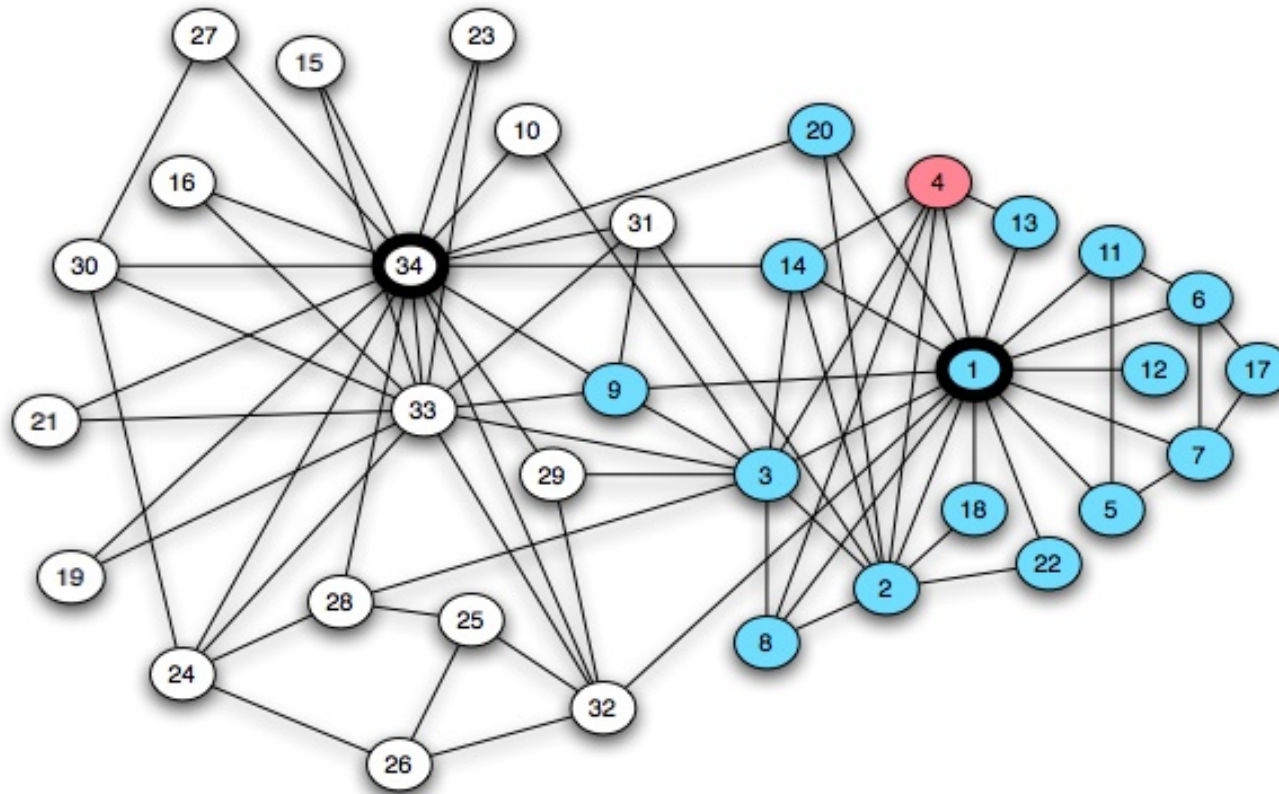
# An Example of What Can Go Wrong



- Node 32 can find himself: only node of degree 6 connected to both leaders.



# An Example of What Can Go Wrong



- Node 32 can find himself: only node of degree 6 connected to both leaders.
- Node 4 can find herself: only node of degree 6 connected to defecting leader but not original leader.

# Attacking an Anonymized Network

What we learn from this:

- Attacker may have extra power if they are part of the system.
- In large e-mail/IM network, can easily add yourself to system.
- But “finding yourself” when there are 100 million nodes is going to be more subtle than when there are 34 nodes.

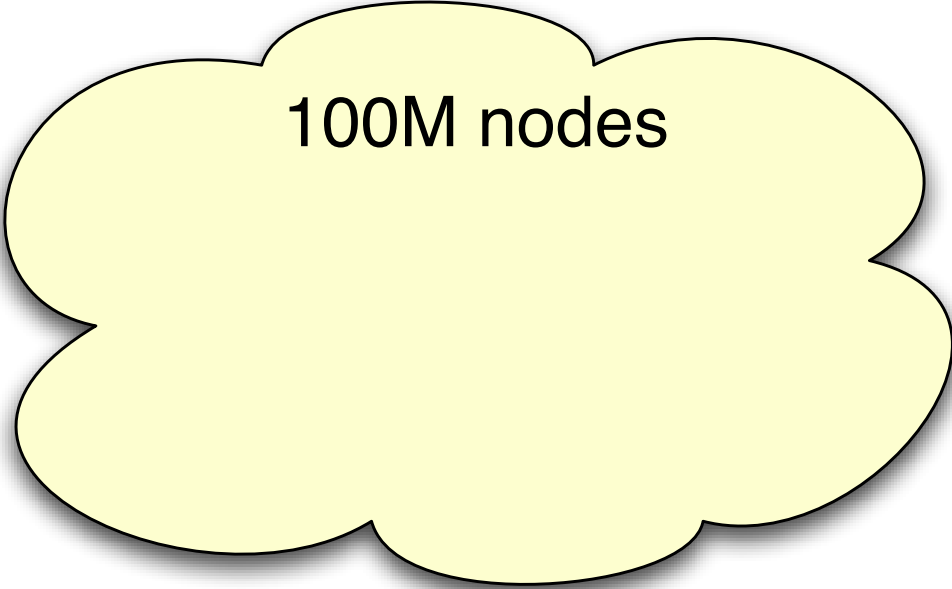
Template for an active attack on an anonymized network  
[Backstrom-Dwork-Kleinberg 2007]

- Attacker can create (before the data is released)
  - nodes (e.g. by registering an e-mail account)
  - edges incident to these nodes (by sending mail)
- Privacy breach: learning whether there is an edge between two existing nodes in the network.
- Note: attacker’s actions are completely “innocuous.”

Main result: active attacks can easily compromise privacy.

Idea is to exploit the incredible richness of link structures.

# An Attack

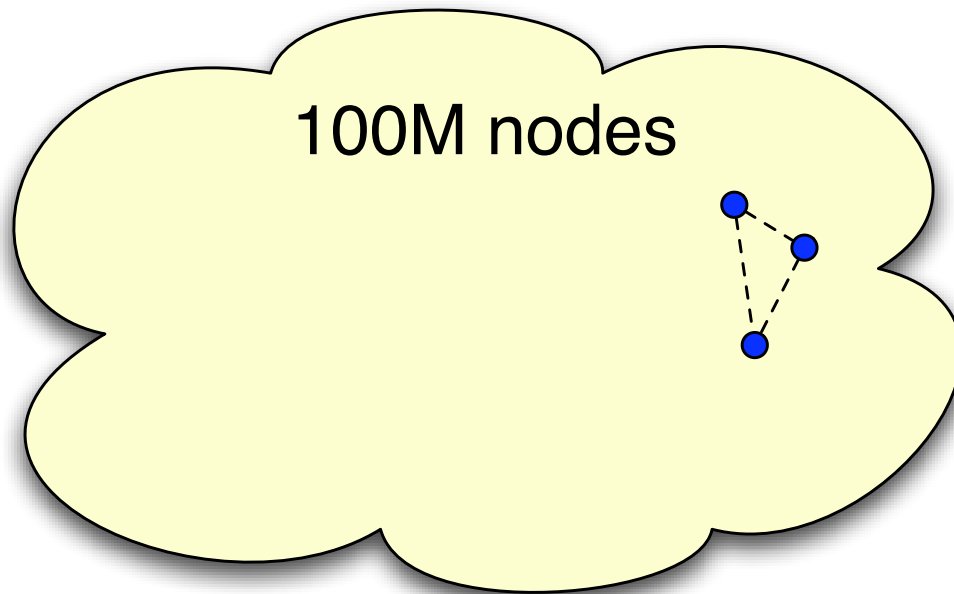


100M nodes

Scenario:

Suppose an organization were going to release an anonymized communication graph on 100 million users.

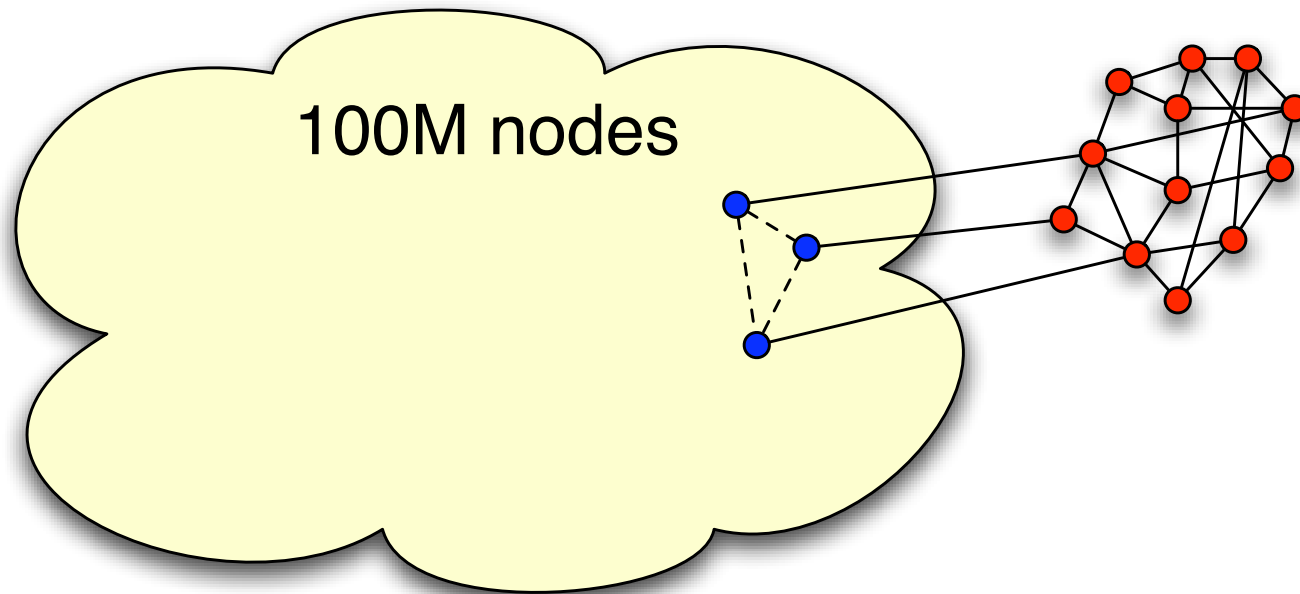
# An Attack



An attacker chooses a small set of user accounts to “target”:

Goal is to learn edge relations among them.

# An Attack

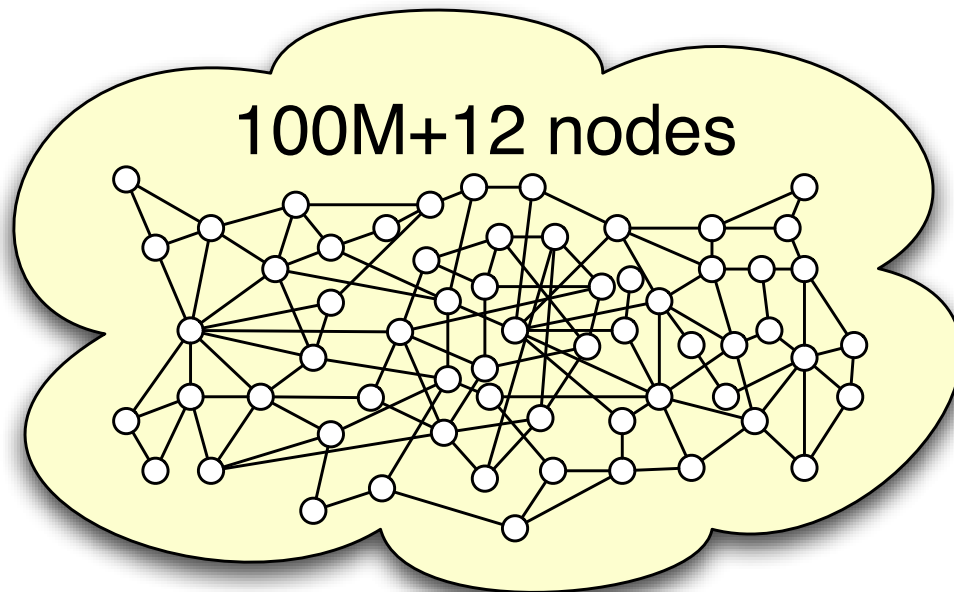


Before dataset is released:

Create a small set of new accounts, with links among them, forming a subgraph  $H$ .

Attach this new subgraph  $H$  to targeted accounts.

# An Attack

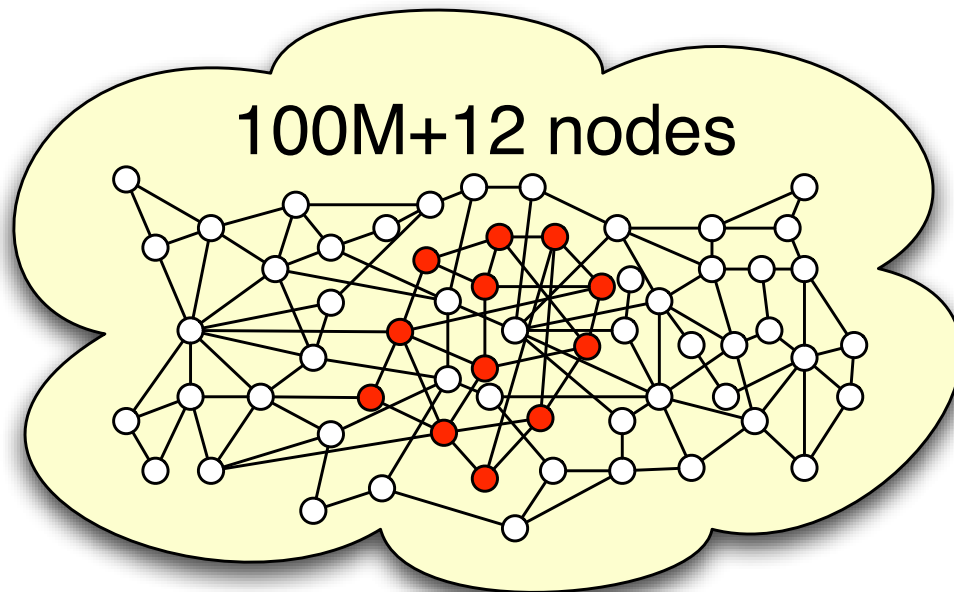


When anonymized dataset is released, need to find  $H$ .

Why couldn't there be many copies of  $H$  in the dataset?  
(We don't even know what the network will look like ... )

Why wouldn't it be computationally hard to find  $H$ ?

# An Attack

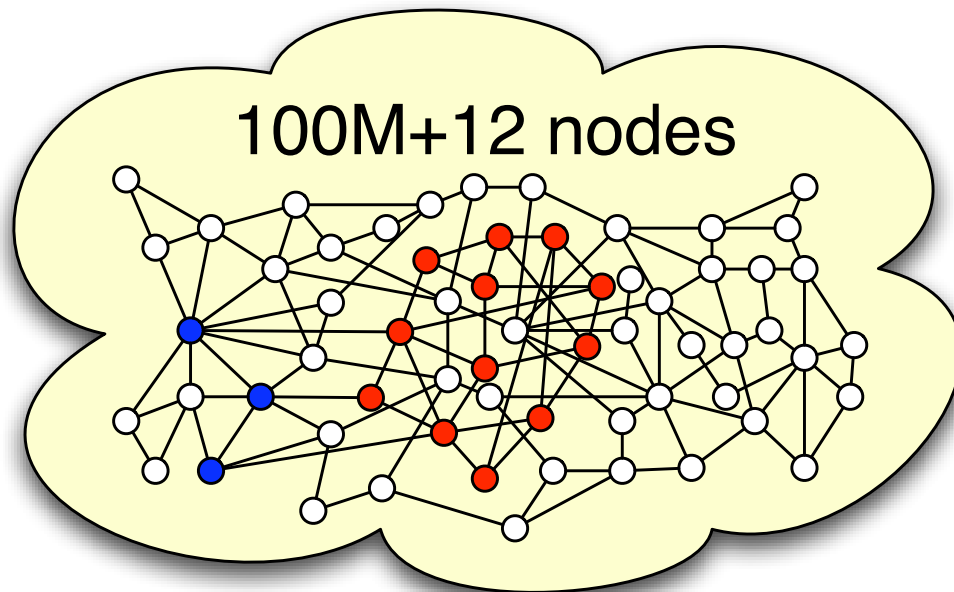


In fact,

Theorem: small random graphs  $H$  will likely be unique and efficiently findable.

Random graph: each edge present with prob.  $1/2$ .

# An Attack



Once  $H$  is found:

Can easily find the targeted nodes by following edges from  $H$ .



# Specifics of the Attack

First version of the attack:

- Create random  $H$  on  $(2 + \varepsilon) \log n$  nodes.
- In experiments on 4.4 million-node LiveJournal graph, 7-node graph  $H$  can compromise 70 targeted nodes (and hence  $\sim 2400$  edge relations).

Second version of the attack:

- Logarithmic size is not optimal.
- Can begin breaching privacy with  $H$  of size  $\sim \sqrt{\log n}$

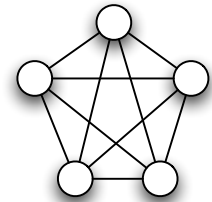
Passive attacks:

- In LiveJournal graph: with reasonable probability, you and 6 of your friends chosen at random can carry out the first attack, compromising about 10 users.

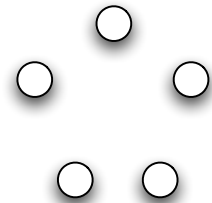
# Why is $H$ Unique? A Ramsey-Theoretic Calculation

Basic calculation at the foundation of

- Theorem (Erdős, 1947): There exists an  $n$ -node graph with no clique and no independent set of size  $> 2 \log n$ .
- Quantitative bound for Ramsey's Theorem; one of the earliest uses of random graphs.



clique



independent set

The calculation:

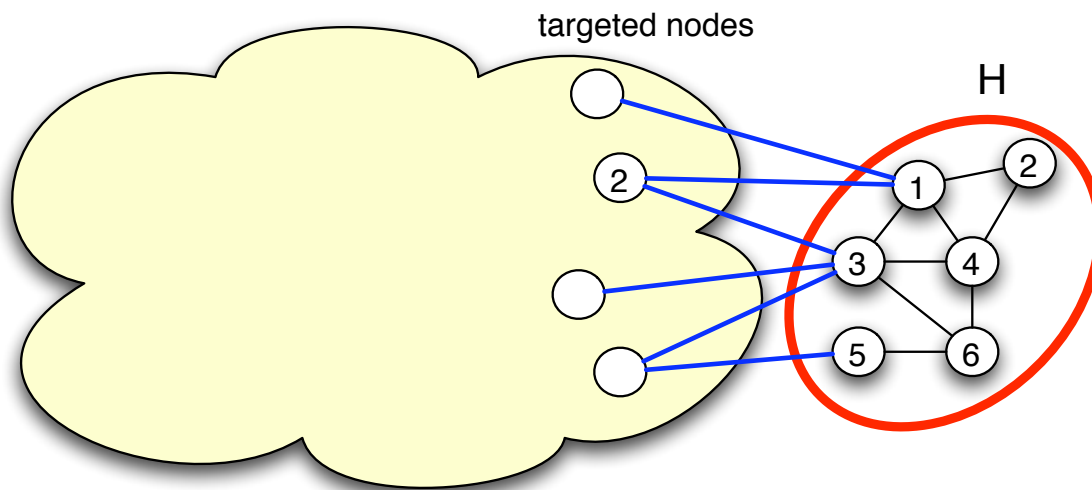
- Build random  $n$ -node graph, include each edge with prob.  $\frac{1}{2}$ .
- There are  $< n^k$  sets of  $k$  nodes; each is a clique or independent set with probability  $2^{-\binom{k}{2}} \approx 2^{-k^2/2}$ .
- Product  $n^k \cdot 2^{-k^2/2}$  upper-bounds probability of any clique or indep. set; it drops below 1 once  $k$  exceeds  $\approx 2 \log n$ .

# Why is $H$ Unique? A Ramsey-Theoretic Calculation

Erdős: Graph is random, subgraph is non-random.

Our case: Subgraph ( $H$ ) is random, graph is non-random.

But main calculation starts from same premise.

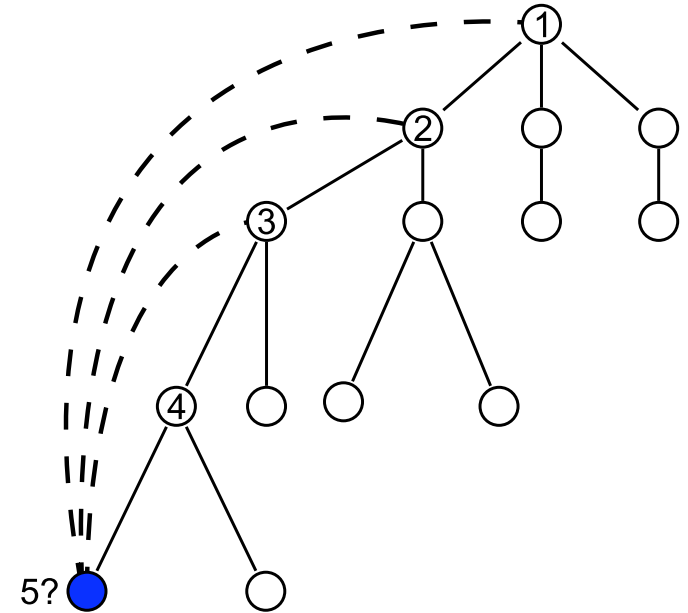


- Analysis is greatly complicated because  $H$  is plugged into full graph.
- New copies of  $H$  could partly overlap original copy of  $H$ .

# Finding the subgraph $H$

To find  $H$ :

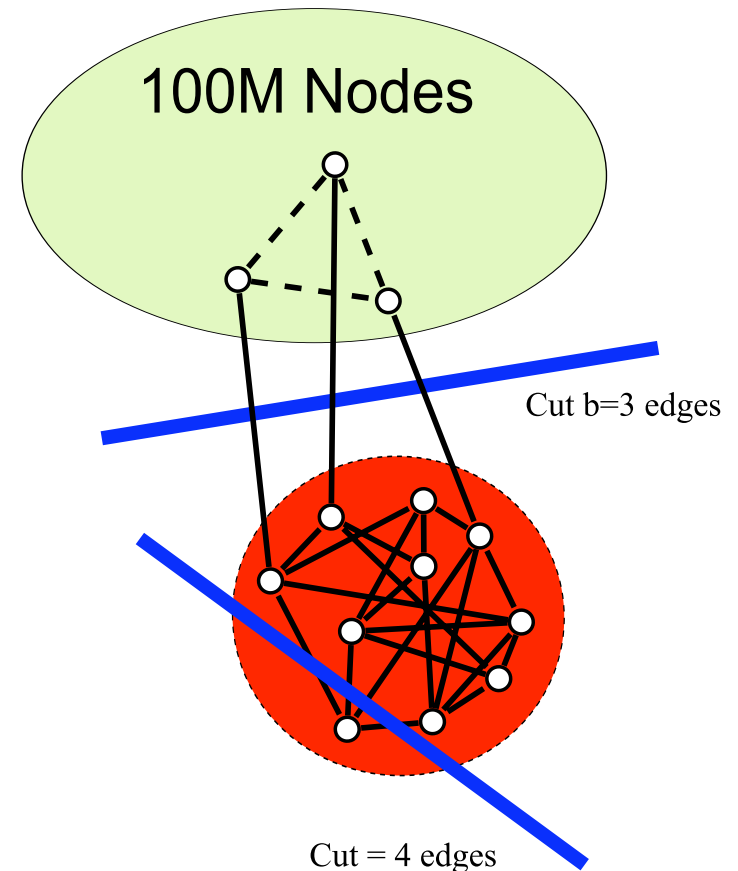
- Can assume there is a path through nodes  $1, 2, \dots, k$ .
- Start search at all possible nodes in  $G$ .
- Prune search path at depth  $j$  if edges back from node  $j$  don't match, or if degree of  $j$  doesn't match.



- Probability of a spurious path surviving to depth  $j$  is  $\approx 2^{-j^2/2}$  (modulo overlap worries).
- Overall size of search tree slightly more than linear in  $n$ .

# Stronger Theoretical Bound

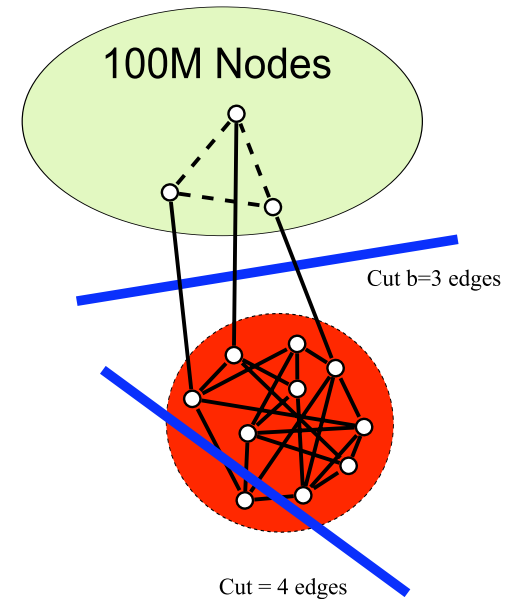
- Variant on construction breaches privacy with  $H$  of size  $\sim \sqrt{\log n}$ .
- Construct  $H$  as before on  $k$  nodes, but connect to  $b = \frac{k}{3}$  targeted nodes.
- With high prob., min. internal cut in  $H$  exceeds  $b =$  cut to rest of graph.



# Stronger Theoretical Bound

Recovery:

- Break graph up along cuts of size  $\leq b$ . Uses Gomory-Hu tree computation (e.g. Flake et al. 2004)
- Can prove that  $H$  will be one of the components after this decomposition.

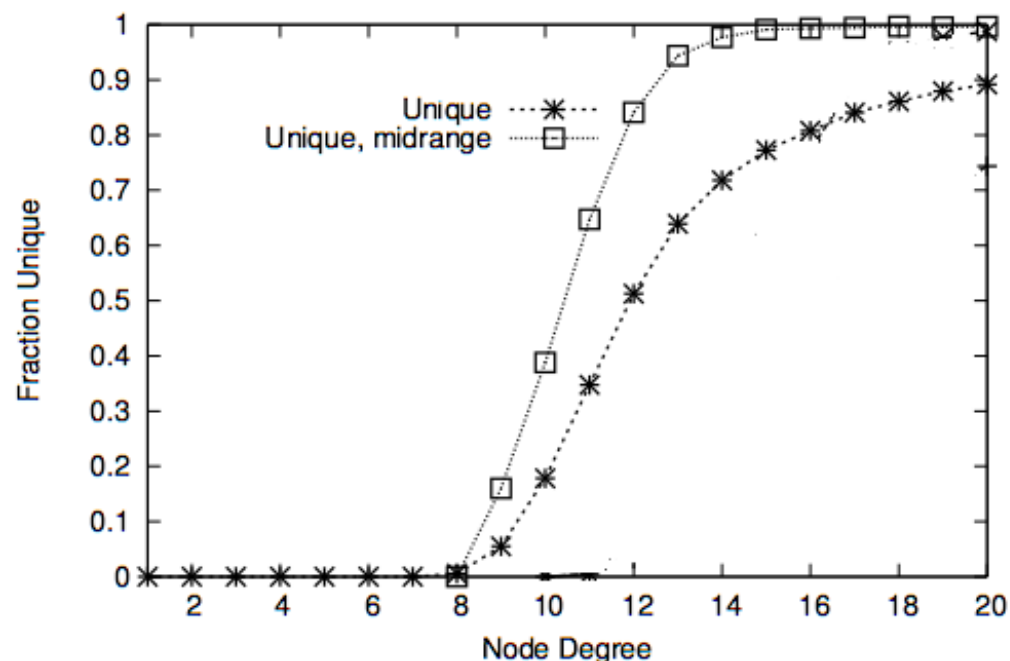


Uniqueness of  $H$ :

- After breaking apart the graph, there are  $\leq \frac{n}{k}$  size- $k$  components other than  $H$ .
- Each is isomorphic to  $H$  with probability  $\approx 2^{-k^2/2}$ .
- Now  $2^{-k^2/2}$  only has to cancel  $\frac{n}{k}$ , not  $n^k$ , so  $k \approx \sqrt{\log n}$  is enough.

# Passive Attacks and the Richness of Local Subgraphs

In 4.4-million-node LiveJournal network, once you have 10 neighbors, the subgraph on these neighbors is likely to be unique.



Friendship structures act like unique signatures.

- Passive attacks feasible with even smaller sets, using numbers of external neighbors in addition to internal network structure.
- Most of us have laid the groundwork for a privacy-breaching attack without realizing it.

# The Perils of Anonymized Data

- General release of an anonymized social network?  
Many potential dangers.
  - Note: earlier datasets additionally protected by legal/contractual/IRB/employment safeguards.
- Fundamental question: privacy-preserving mechanisms for making social network data accessible.
  - Interesting connections to issues in Sofya's talk:  
May be difficult to obfuscate network effectively;  
Interactive mechanisms for network data may be possible.  
(See also [Dinur-Nissim 2003, Dwork-McSherry-Talwar 2007])
- Recent proposals specifically aimed at
  - framework for safe public release [Blum-Ligget-Roth '08]
  - social networks [Hay et al '07, Zheleva et al '07, Korolova et al '08]
- Further issues
  - Even without overt attacks, increasingly refined pictures of individuals begin to emerge.



# Final Reflections: Glimpses into Massive Networks

Simultaneous opportunities and challenges.

How do we build deeper models of the processes at work inside large-scale social networks?

- A stronger vocabulary for analyzing operational models consistent with observed data.
- Understanding how much can be predicted.
- Social computing applications produce a new set of design constraints.

How do we make data available without compromising privacy?

- A need for guarantees: the dangers can be unexpected.

Algorithmic and mathematical models will be crucial to understanding all these developments.