

The Geography of Social and Information Networks

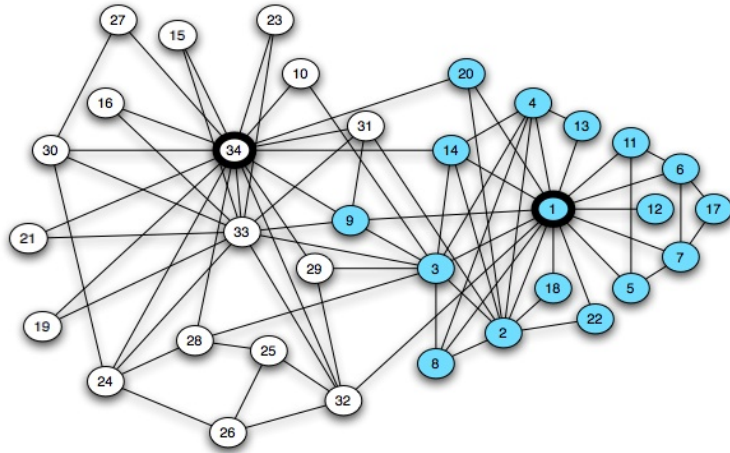
Jon Kleinberg

Cornell University

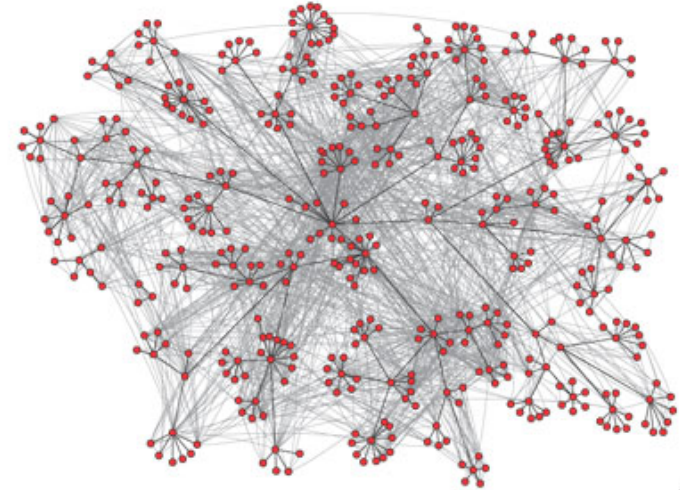


**Including joint work with Lars Backstrom, Cynthia Dwork,
and David Liben-Nowell**

Large-Scale Social Computing Applications



Karate club dissolving
(Zachary, 1977)



Corporate e-mail communication
(Adamic and Adar, 2005)

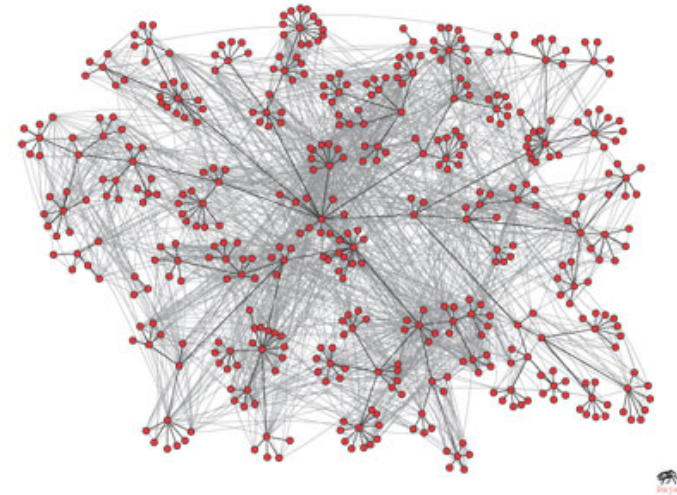
Science advances when we make the invisible become visible.

- Social interaction is leaving digital traces on-line.
- How can we use this to address long-standing social-science questions?
- How should it inform the design of new computing and information systems?
- What are the dangers and pitfalls?

A Matter of Scale

Social network data spans many orders of magnitude

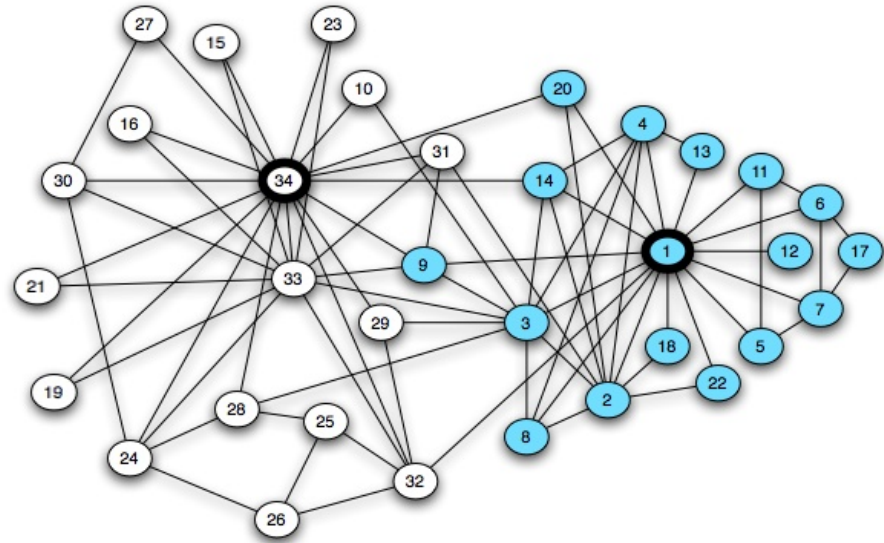
- 436-node network of e-mail exchange over 3 months at a corporate research lab (Adamic-Adar 2003)



- 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

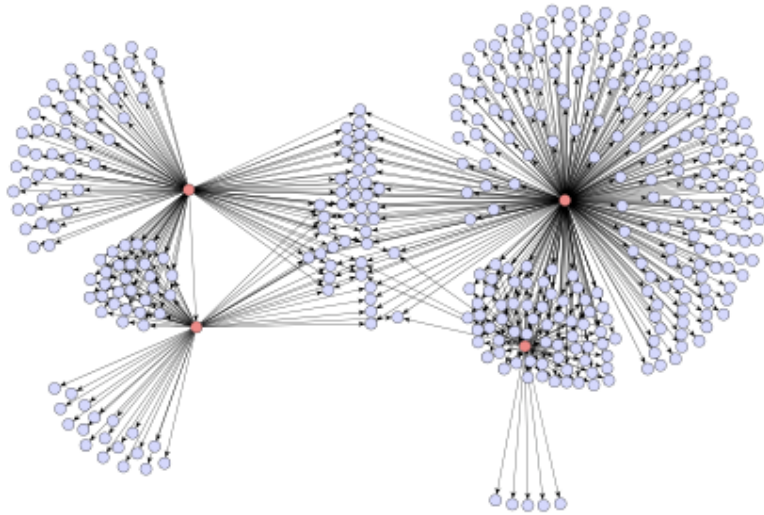
Fundamental mathematical and computational ideas come into play:

- 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.

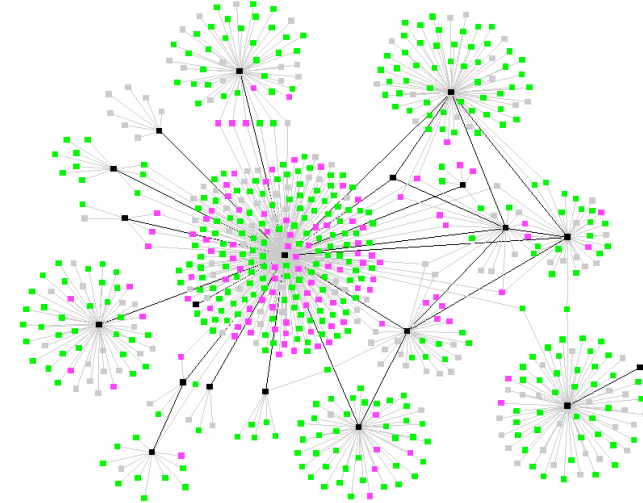
Plan for the talk:

- 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, fads, political mobilization
- Diffusion of innovations [Coleman-Katz-Menzel, Rogers]
- Viral marketing [Domingos-Richardson 2001]
- Public health (e.g. [Christakis-Fowler 2007])
- Localized collective action: riots, walkouts

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:

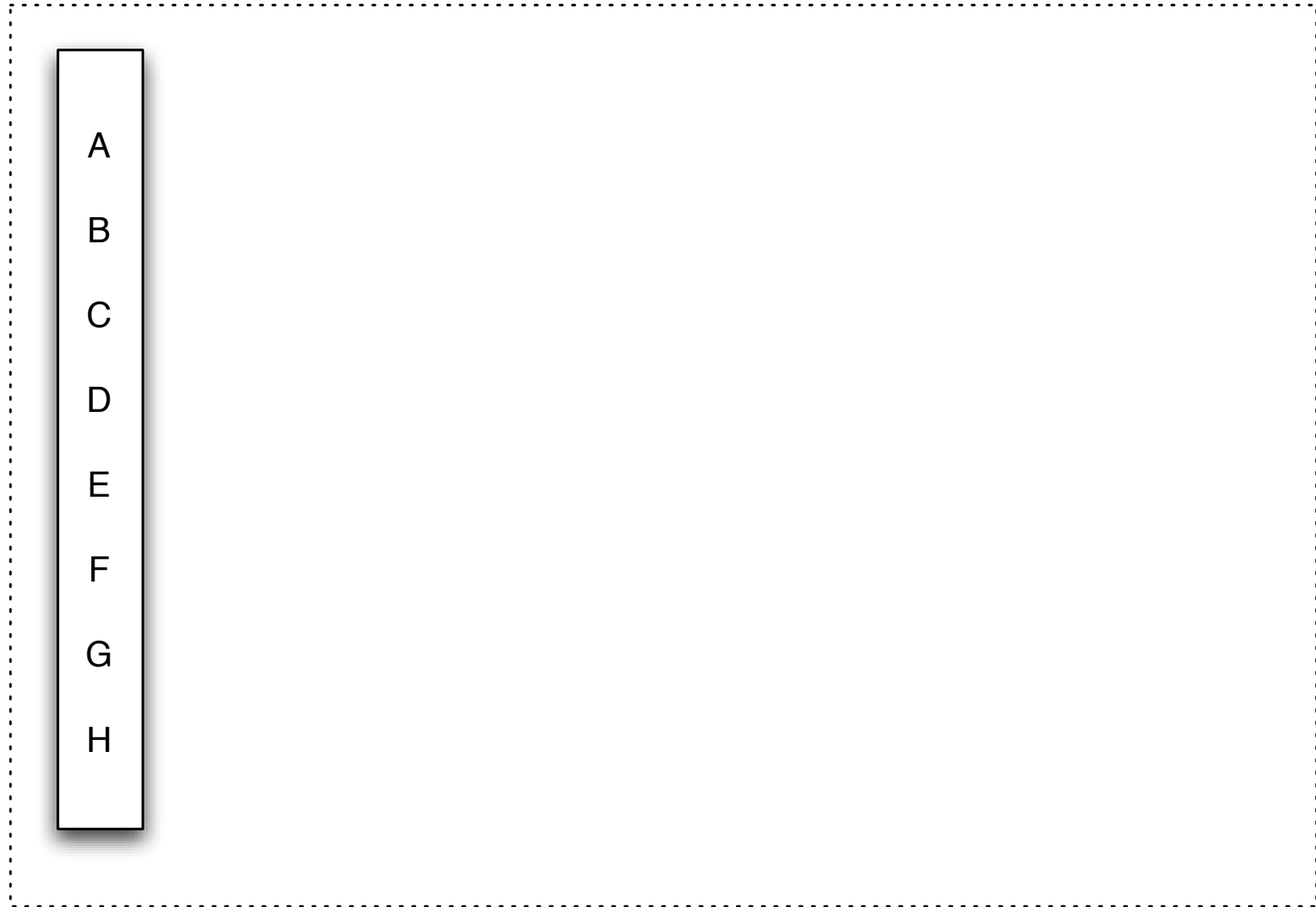
[...]

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:

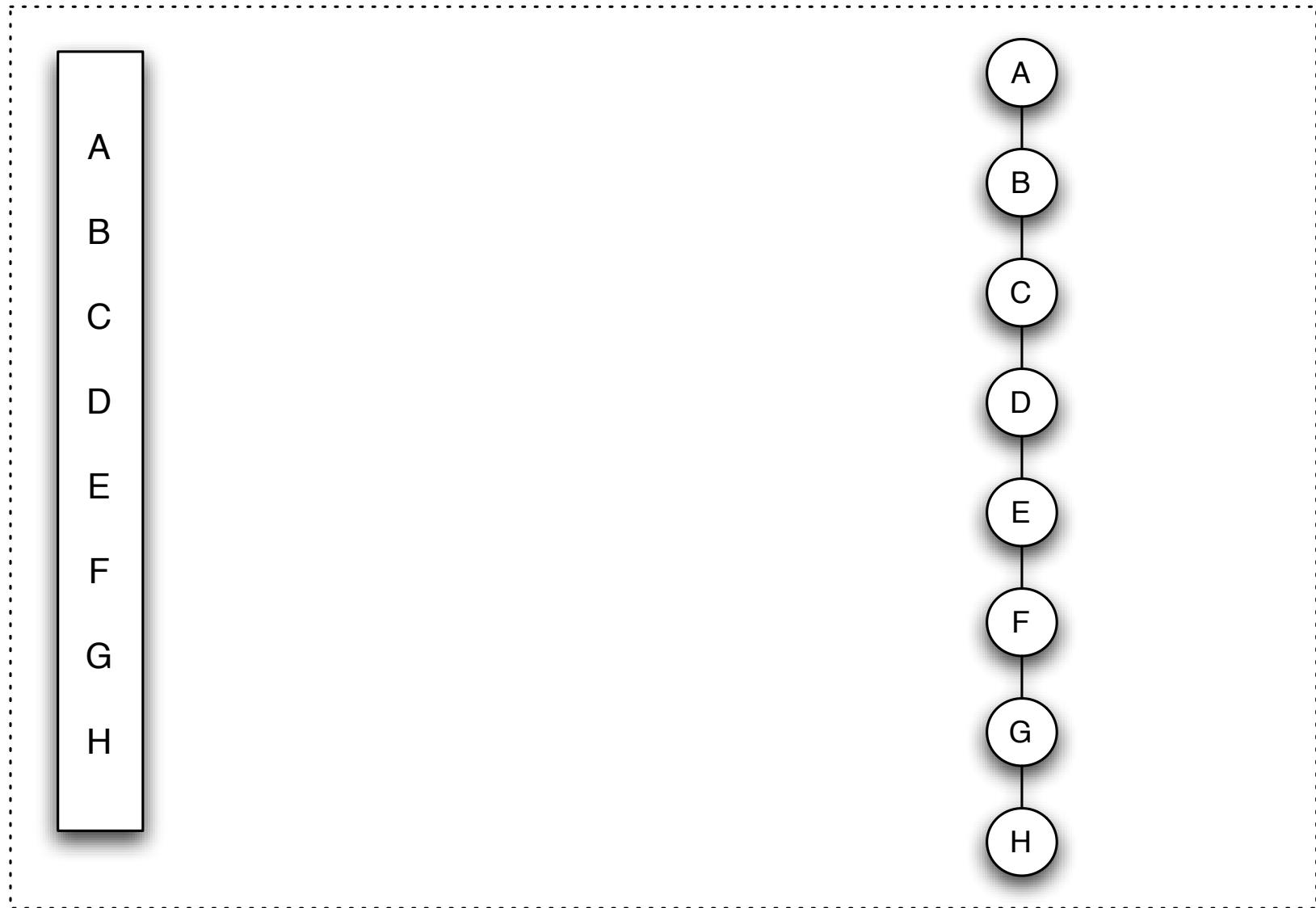
usa@un.int

president@whitehouse.gov

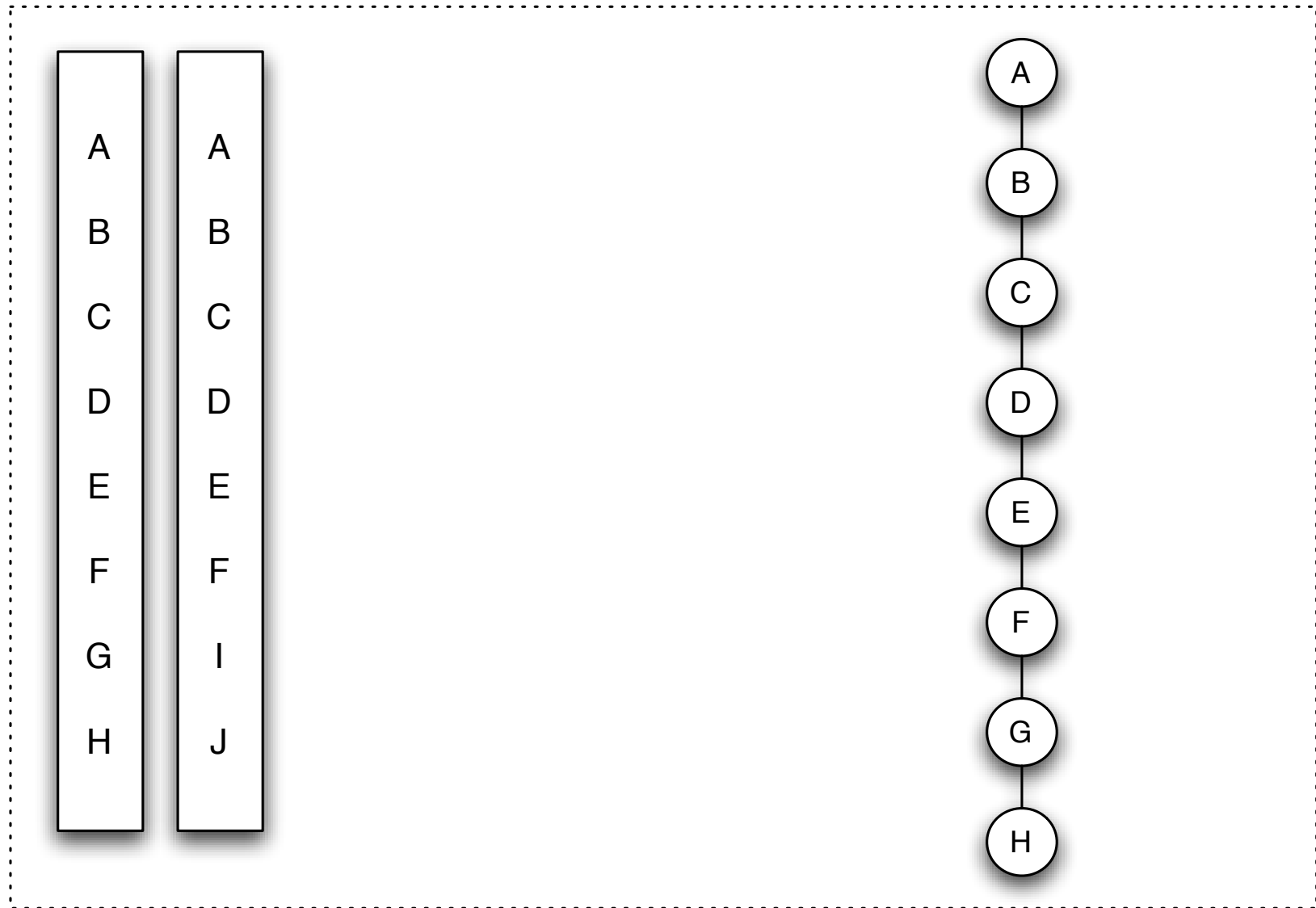
Assembling a Chain-Letter Tree



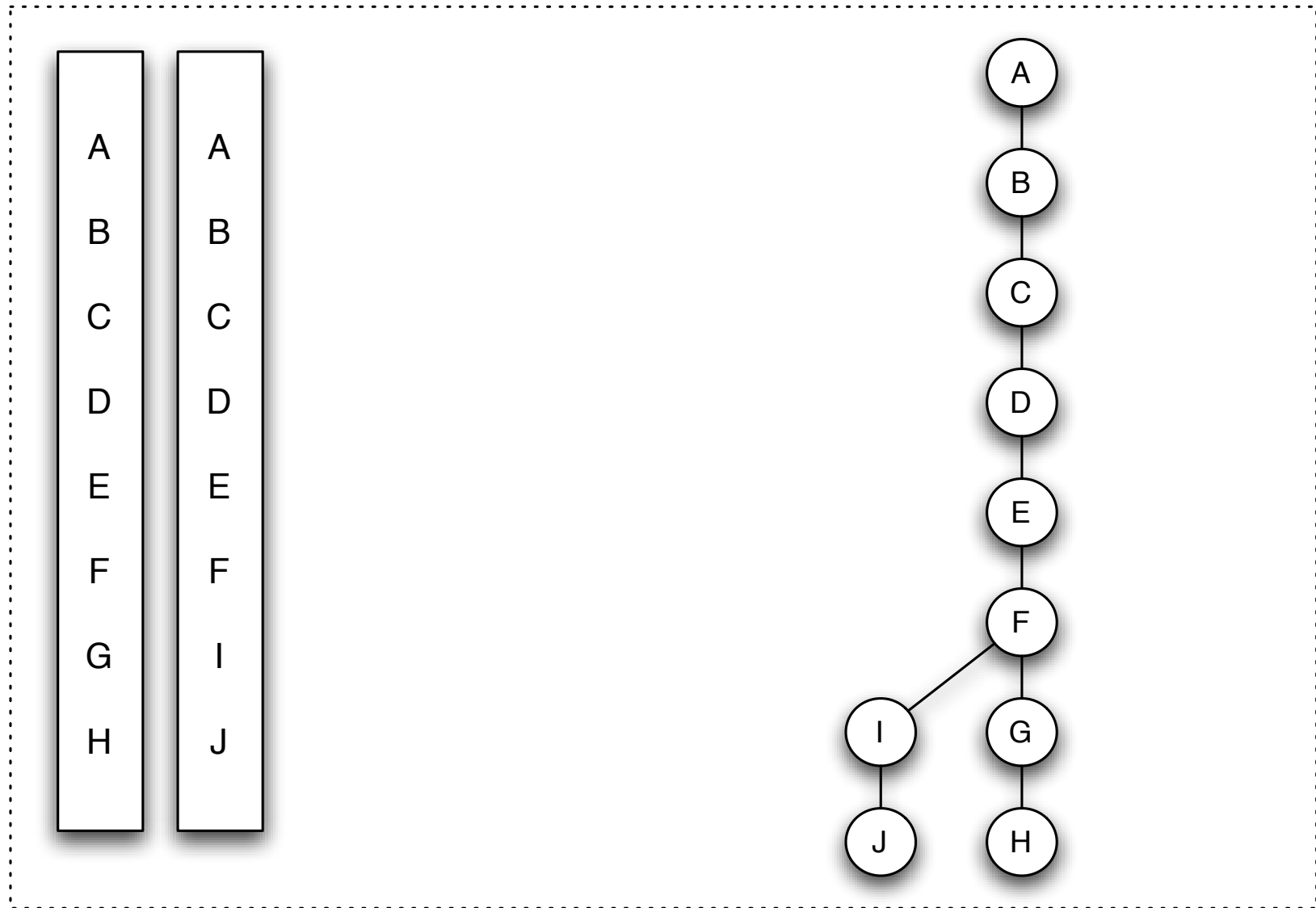
Assembling a Chain-Letter Tree



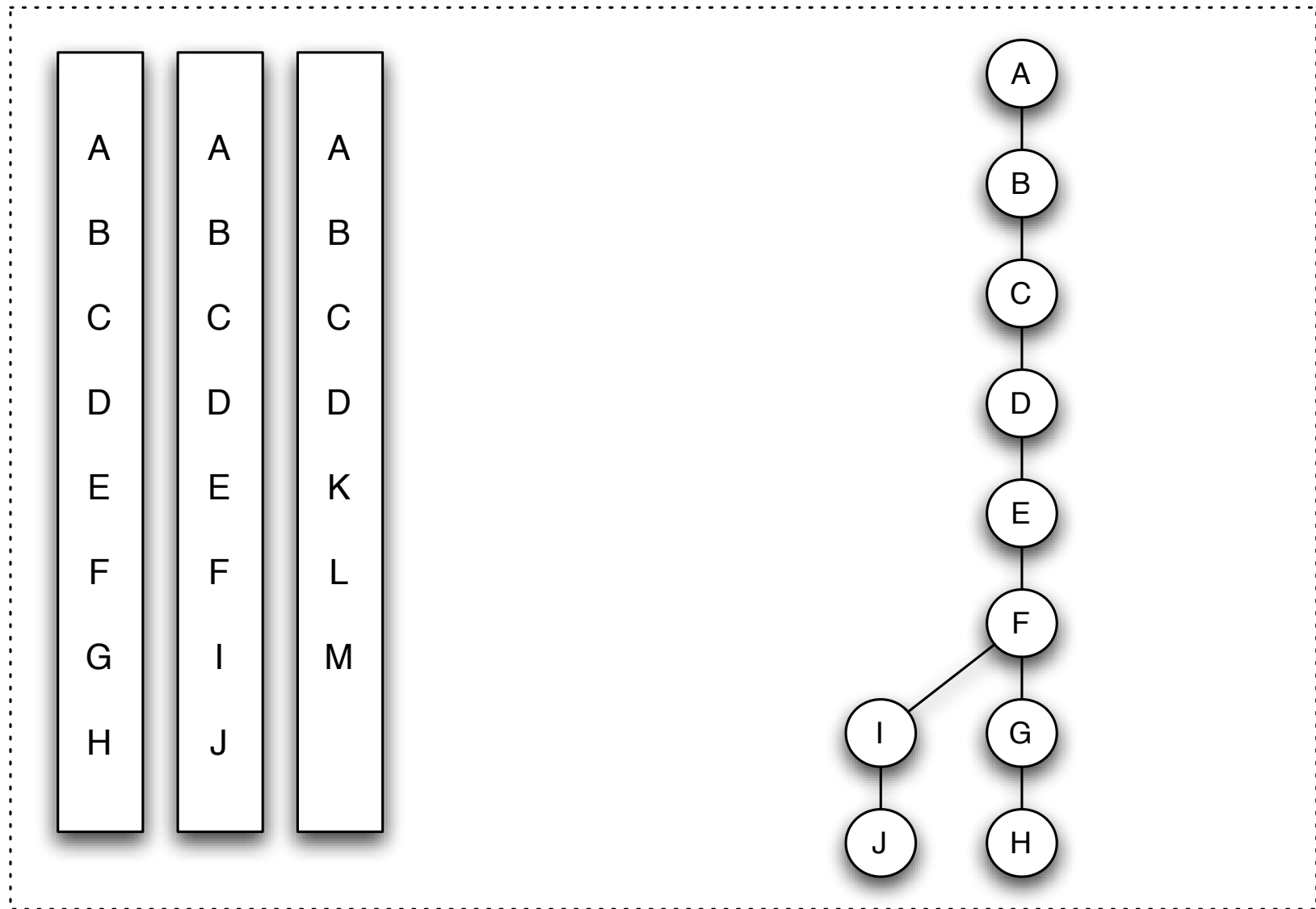
Assembling a Chain-Letter Tree



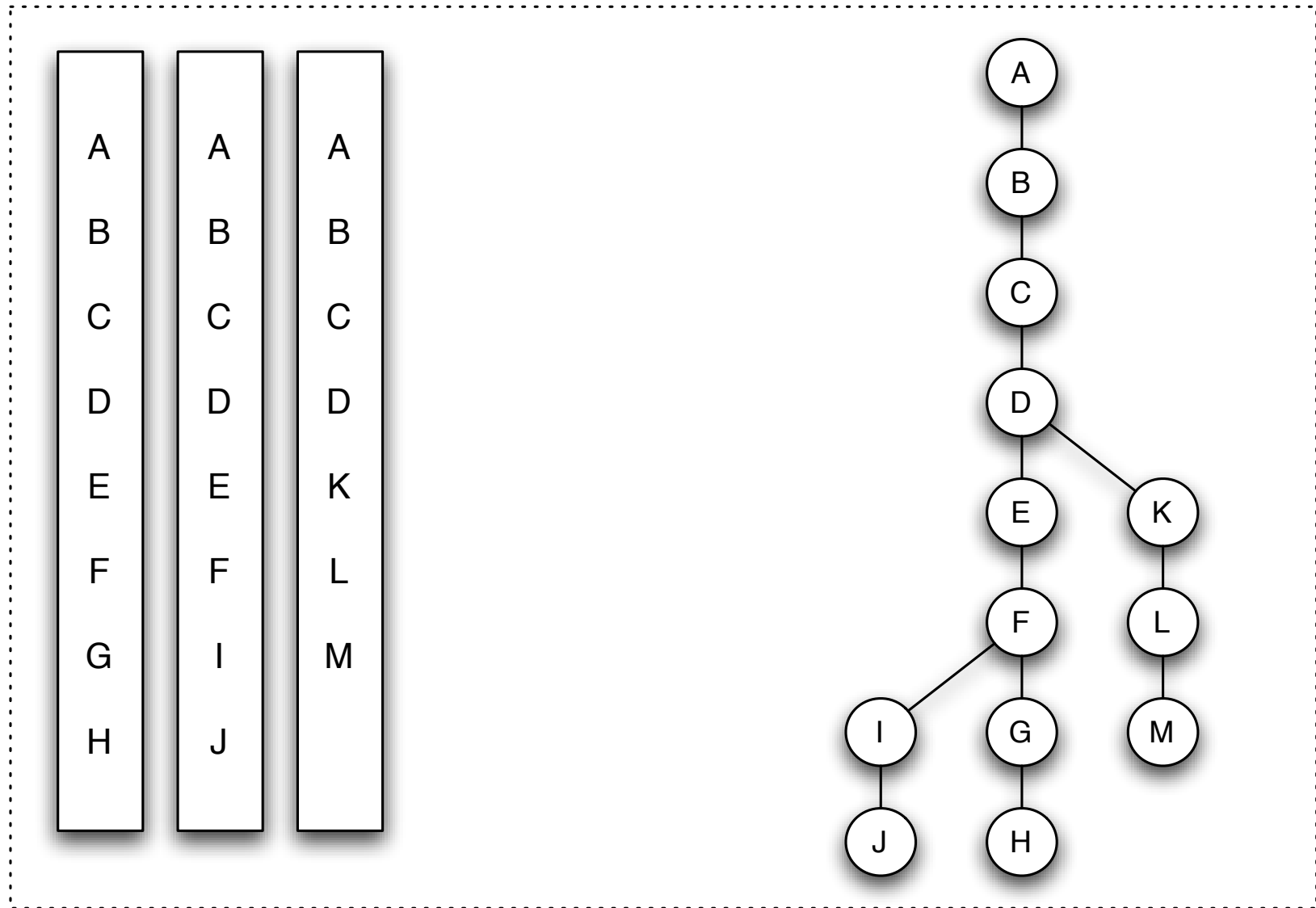
Assembling a Chain-Letter Tree



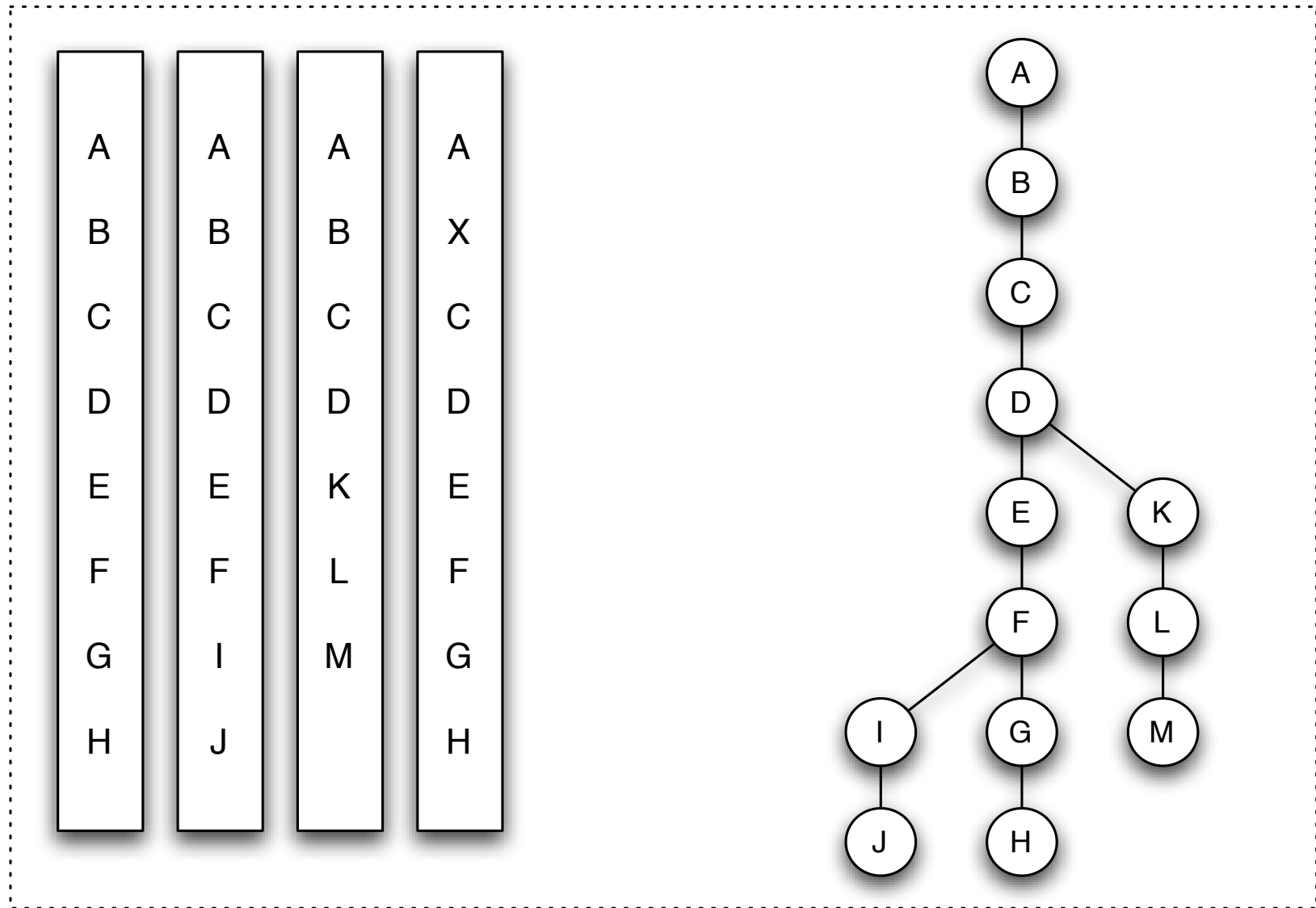
Assembling a Chain-Letter Tree



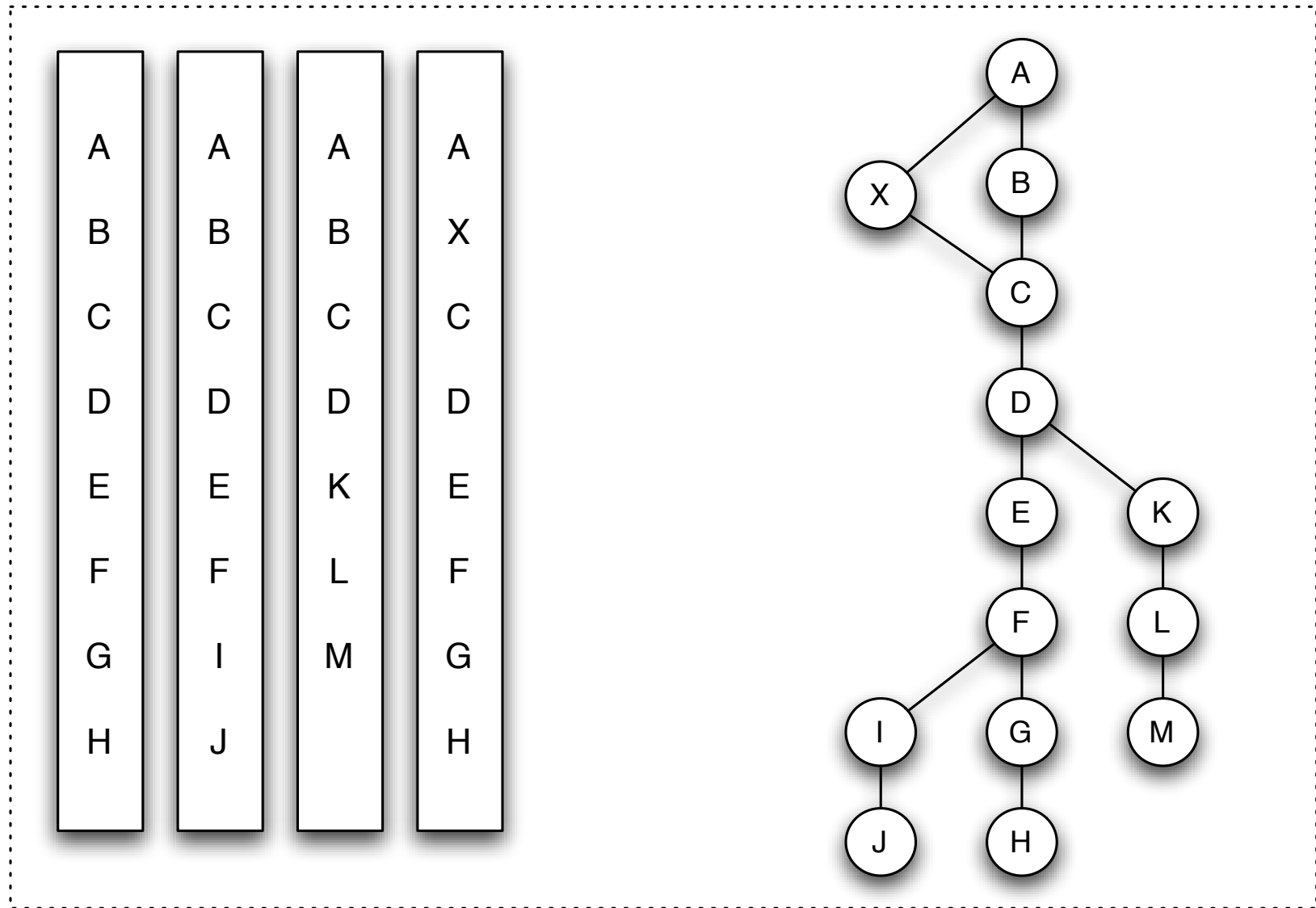
Assembling a Chain-Letter Tree



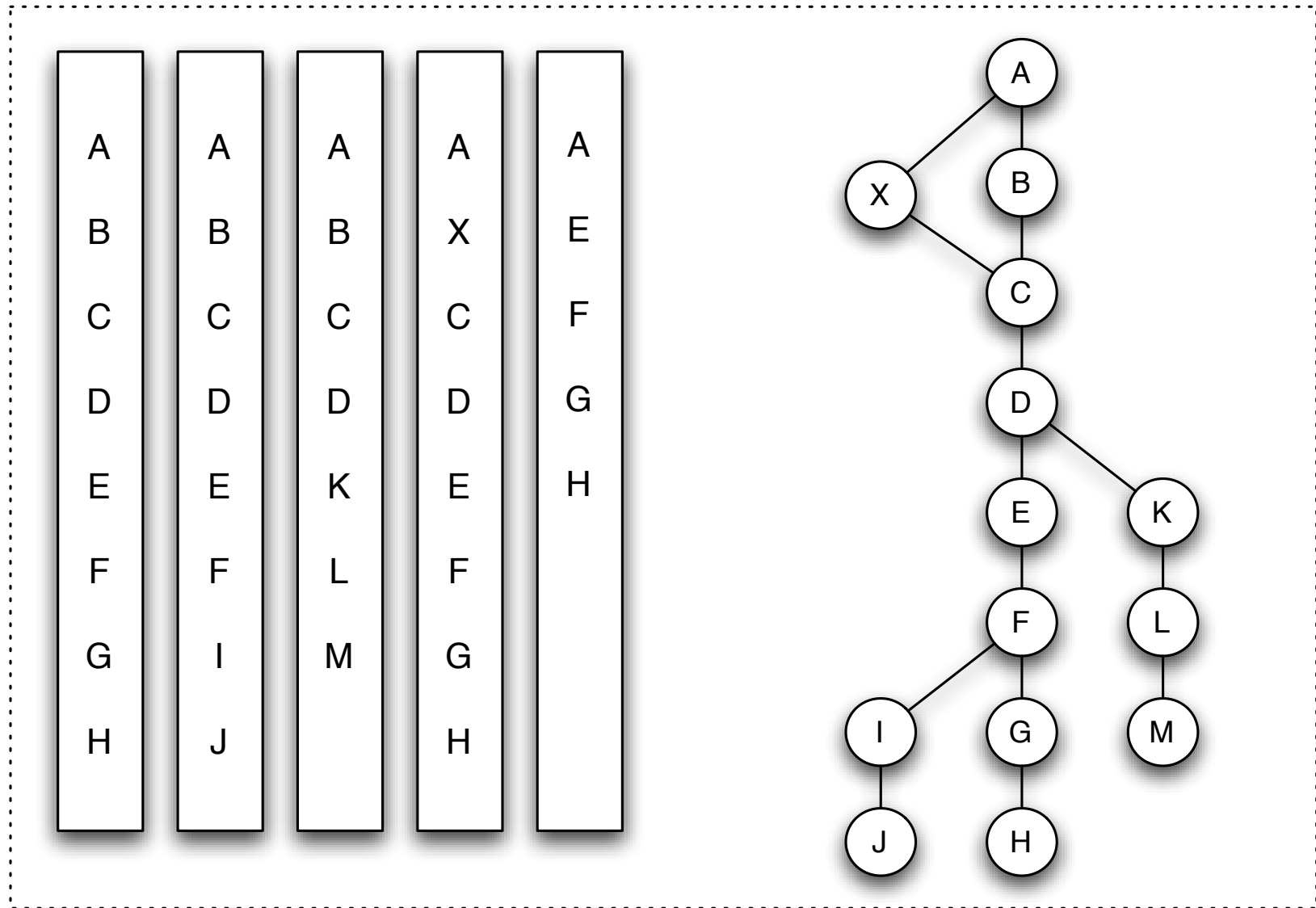
Assembling a Chain-Letter Tree



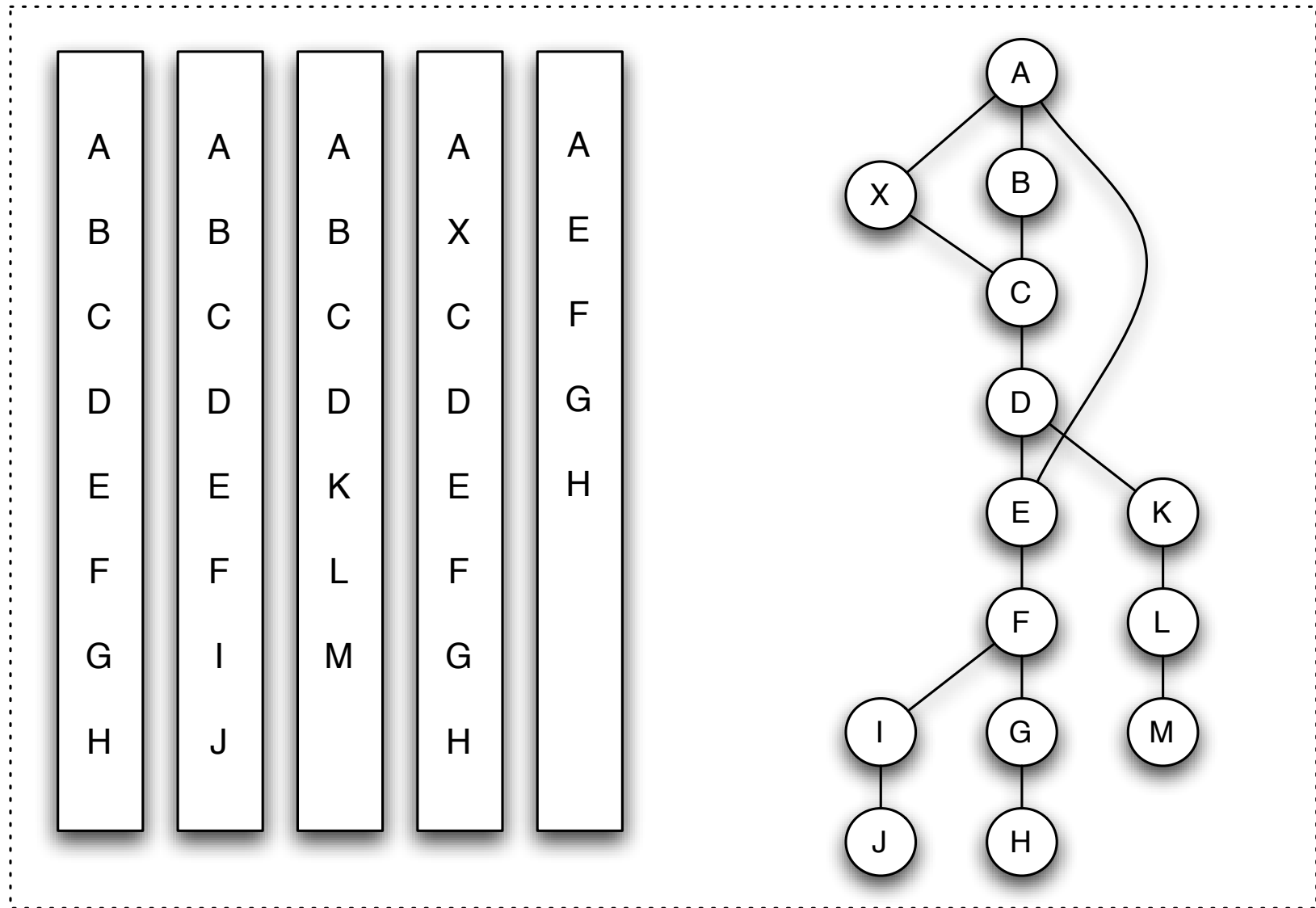
Assembling a Chain-Letter Tree



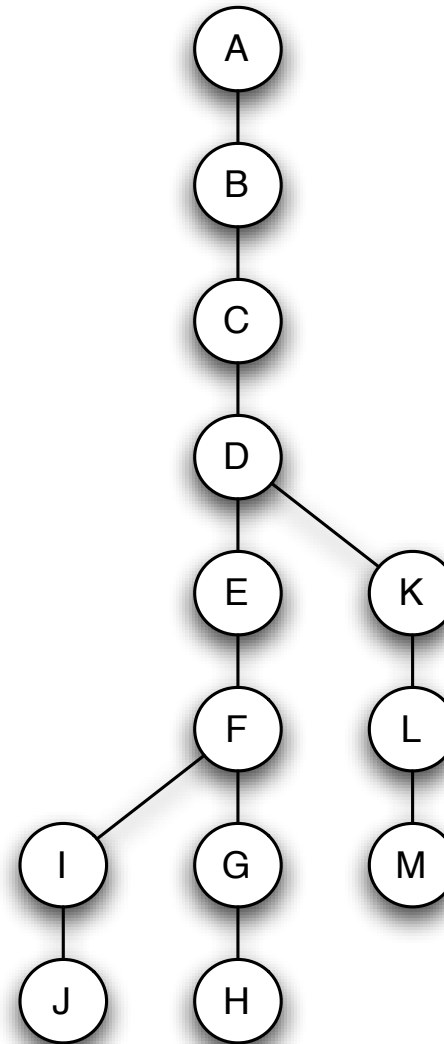
Assembling a Chain-Letter Tree

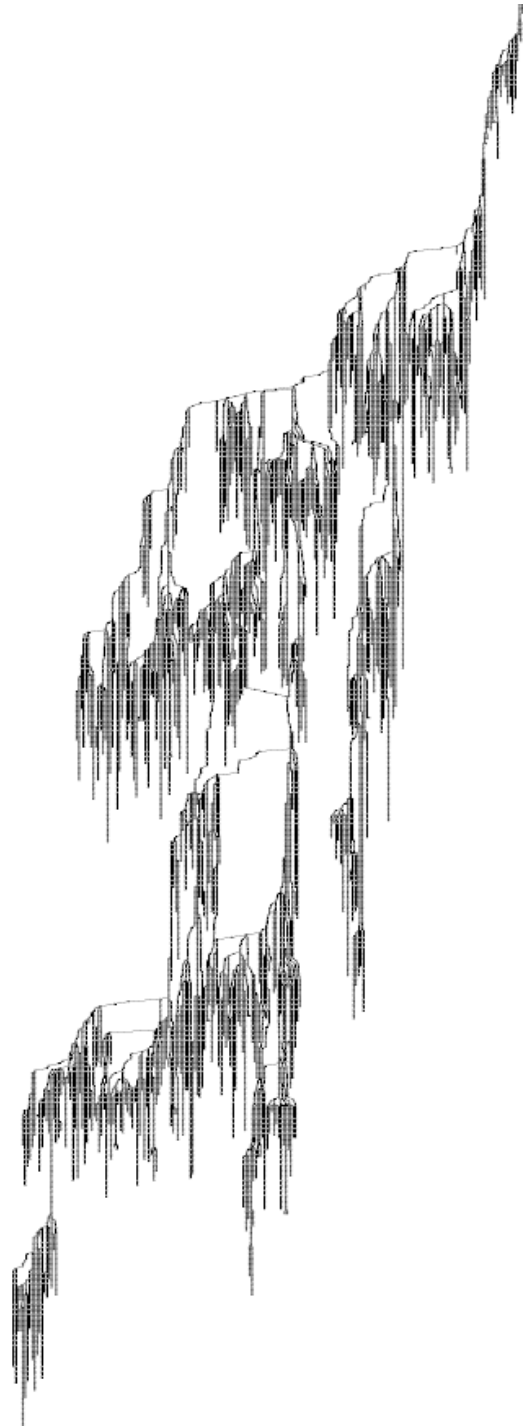


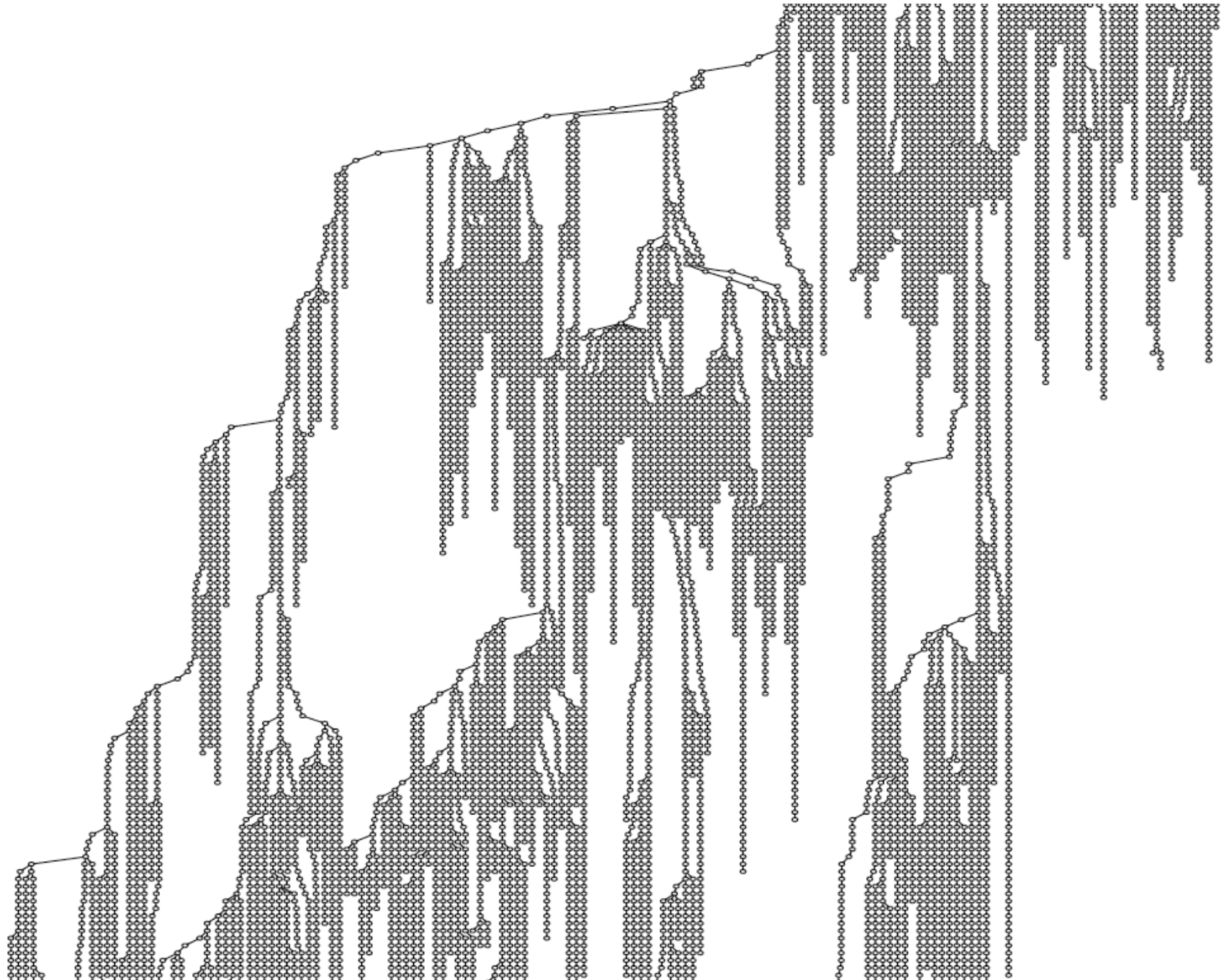
Assembling a Chain-Letter Tree

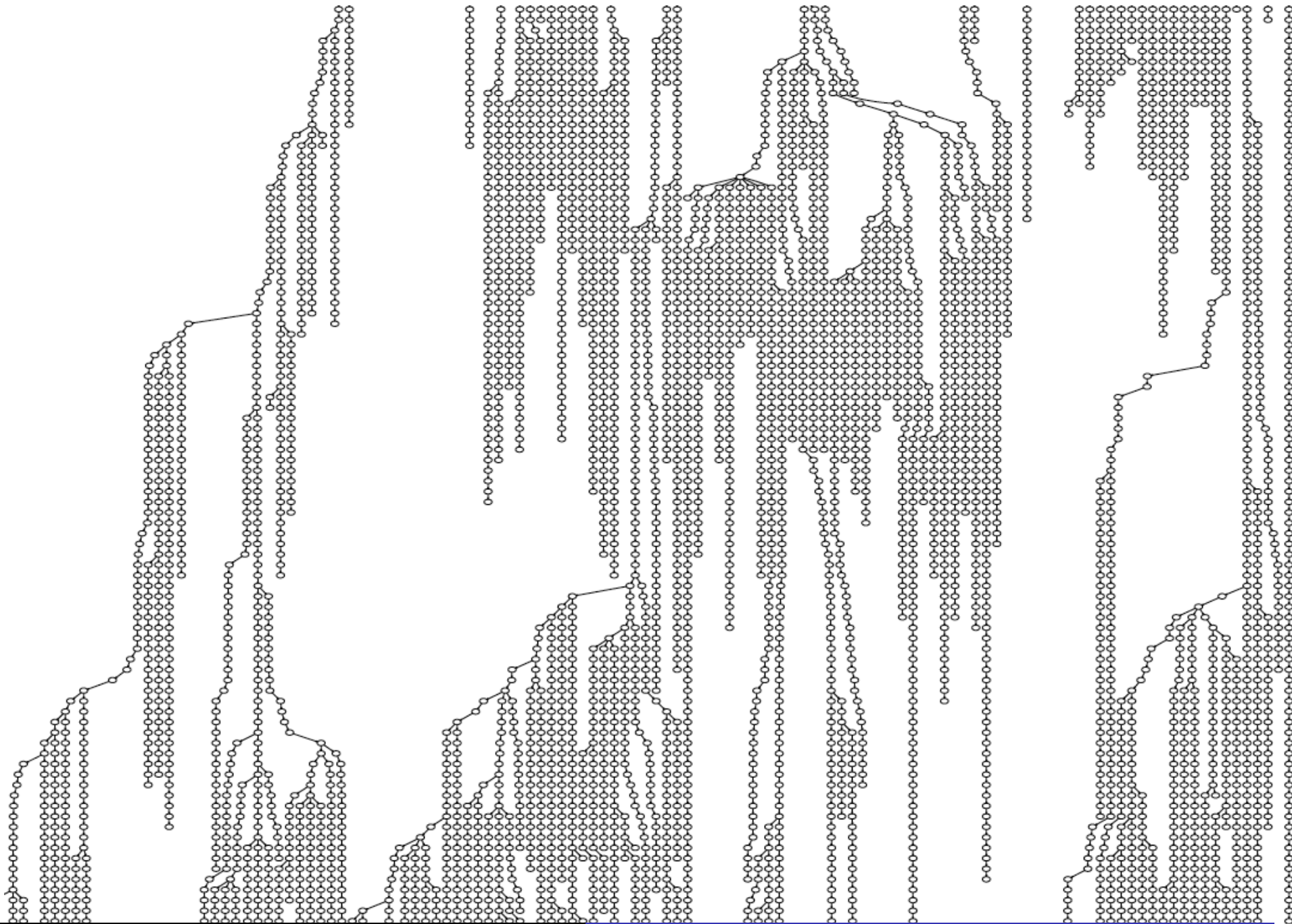


Assembling a Chain-Letter Tree



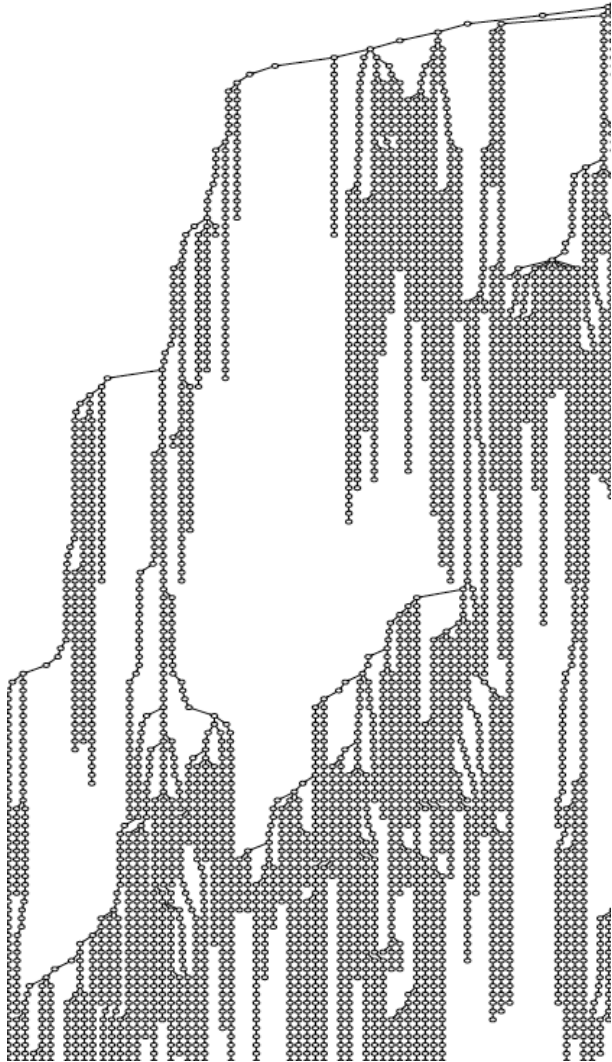








Modeling the Structure of the Tree



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.

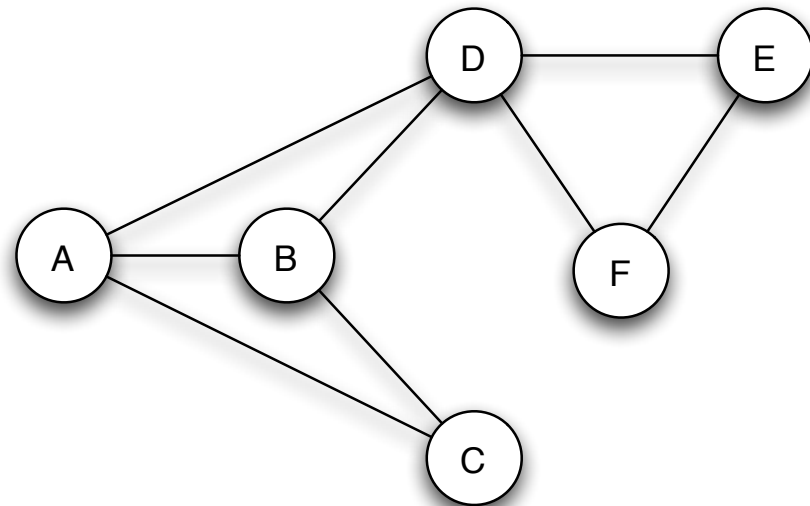
Some plausible models that can produce trees of this shape:

- (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?

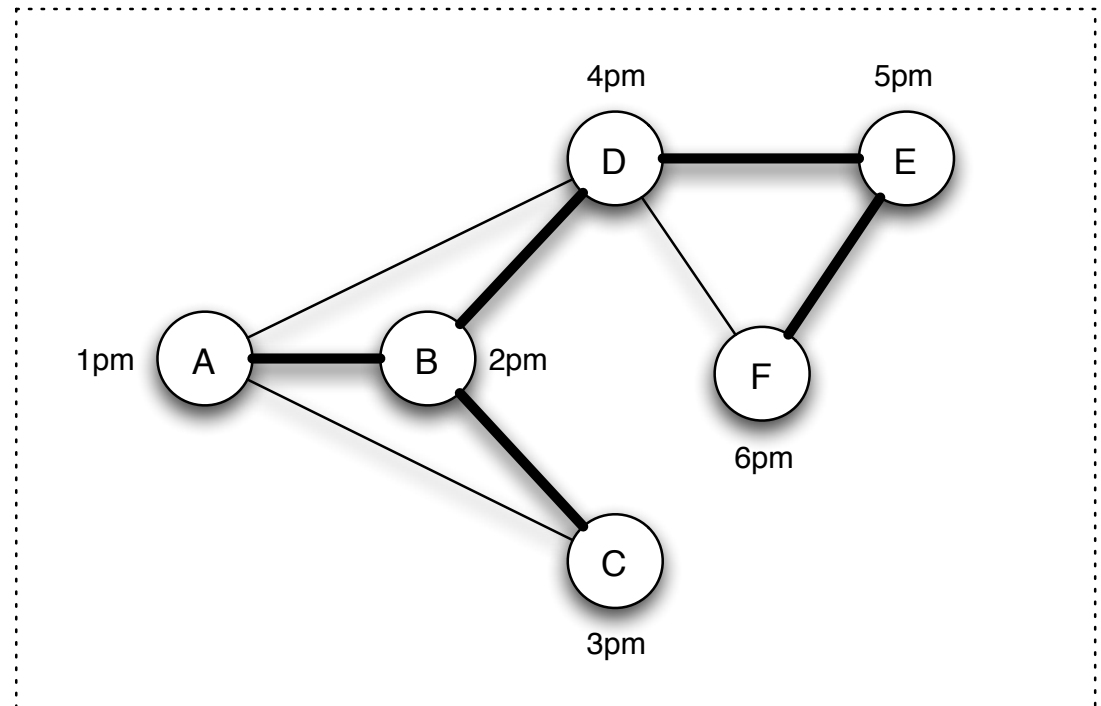
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

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

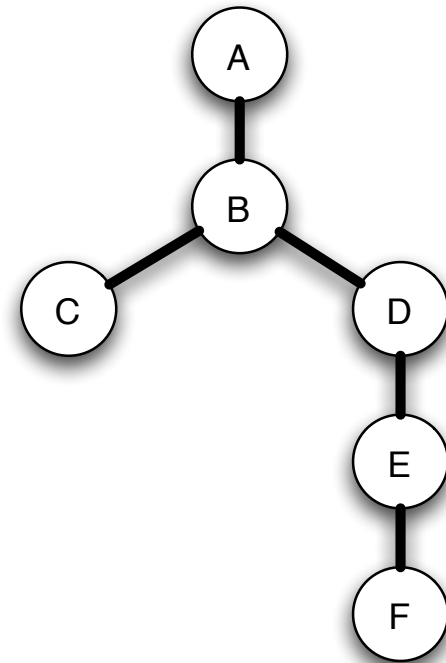
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

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

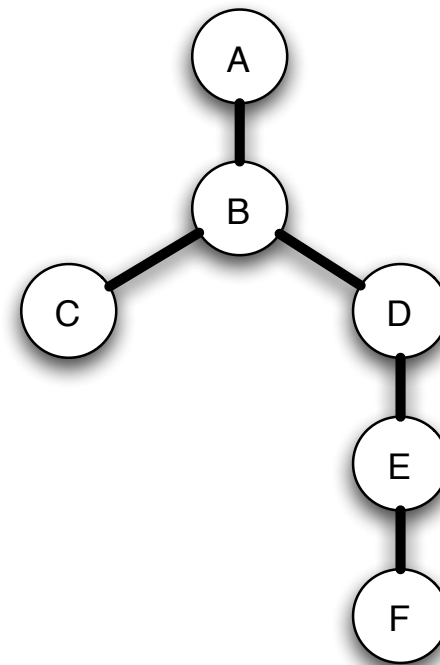
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



Why is the Tree So Deep and Narrow?

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

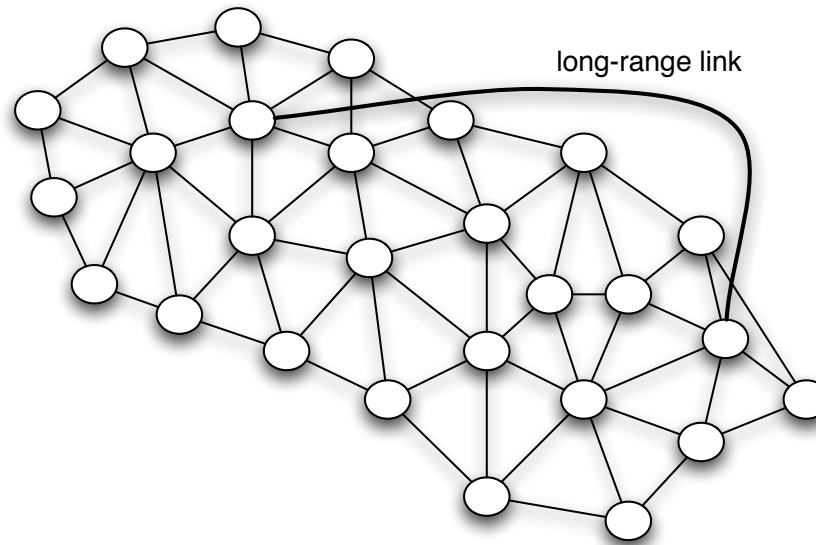
- Possible model based on timing.
- Assume nodes act on messages according to a delay distribution.



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.

Spatial Clustering and Thresholds

A second class of theories based on spatial ideas.



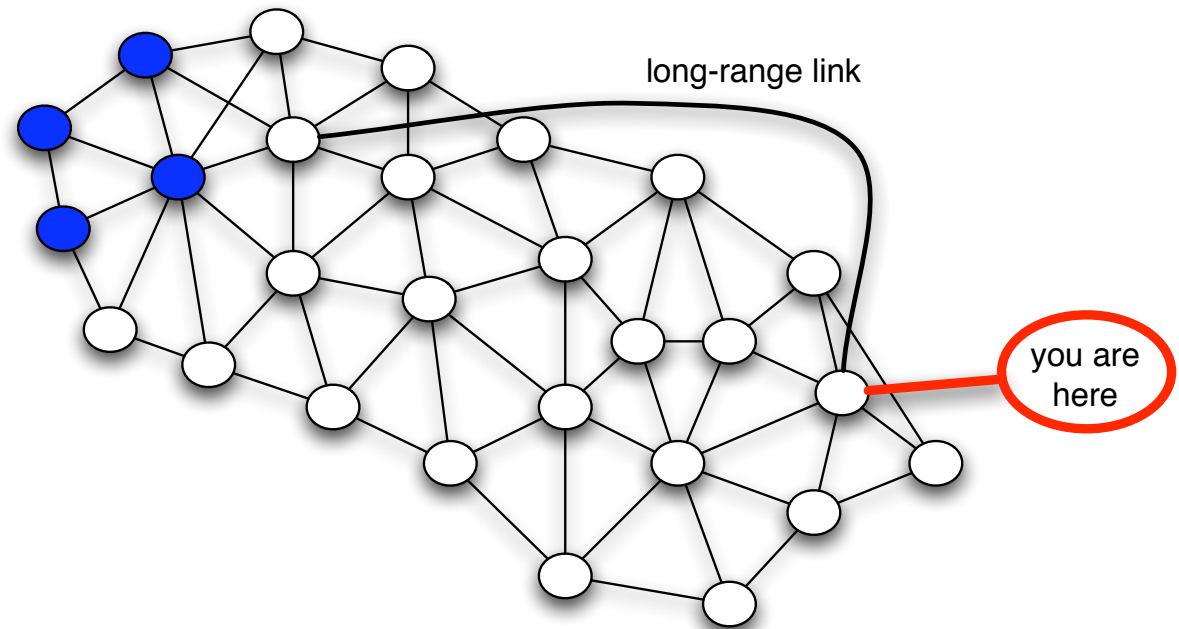
- Even in on-line social networks, most friends are geographically (and demographically) similar to you [McPherson et al. 2001, Liben-Nowell et al. 2005]
- Decision rules for acting may involve thresholds: e.g., you may need to see multiple friends advocating a cause before signing on [Granovetter 1978,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

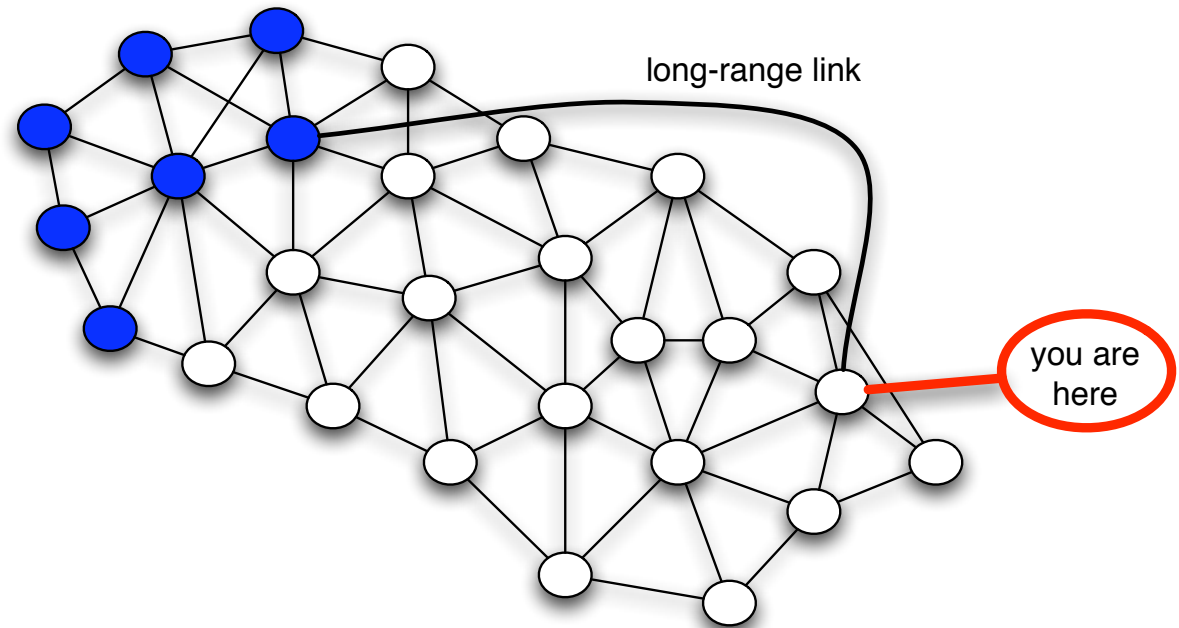


Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

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

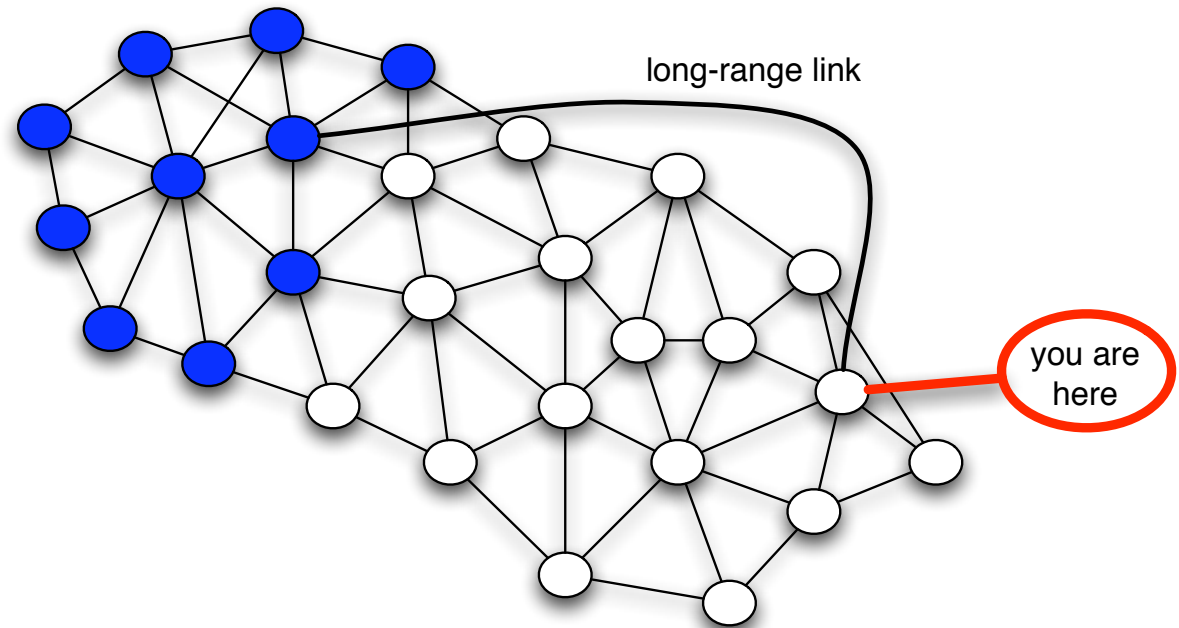


Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

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

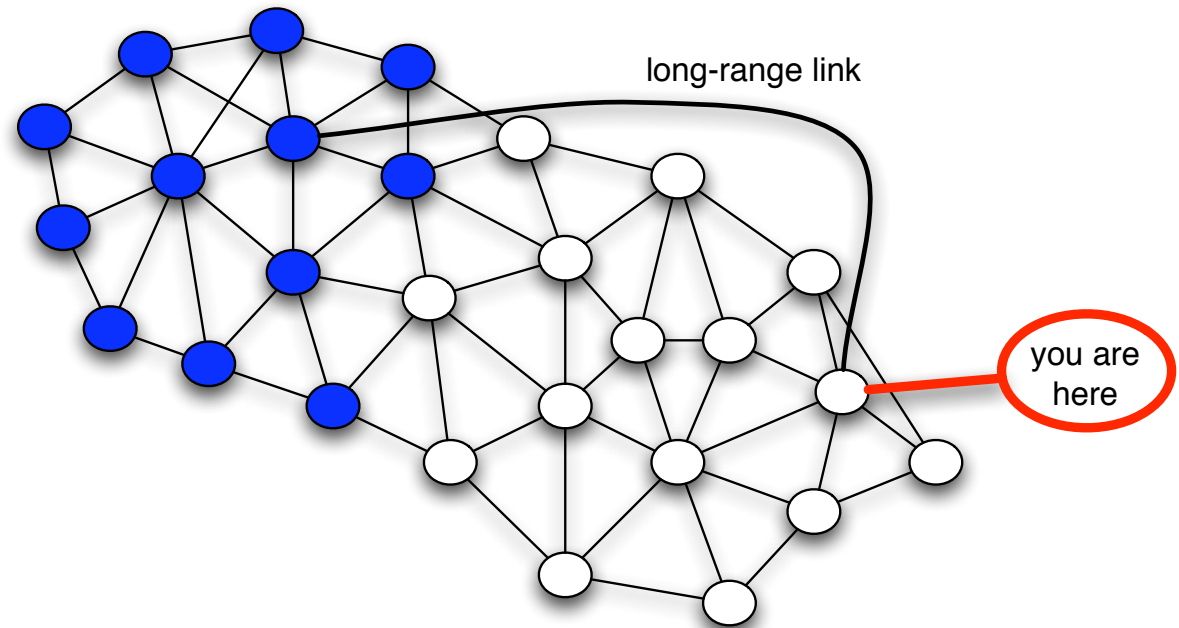


Spatial Clustering and Thresholds

Interaction of local structure and thresholds

[Centola-Macy 2007]

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

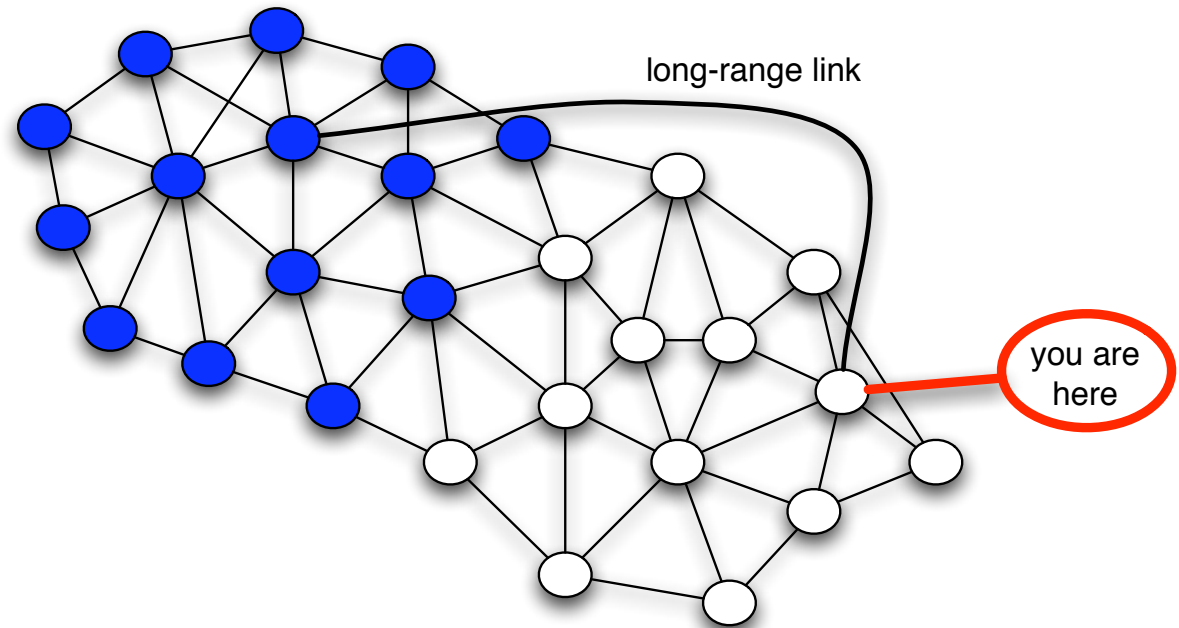


Spatial Clustering and Thresholds

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.

Are Cascades Inherently Unpredictable?

Salganik-Dodds-Watts 2006

- Music download site: obscure songs, 14,000 visitors.
- Users first previewed songs, then rated, then optionally downloaded.
- Varied the information available:
 - Just song titles
 - Titles and download statistics (feedback from other users).
- Version with feedback: ran 8 “parallel universes.”
- Feedback produced greater inequality, and top songs varied across parallel runs.



Protecting Privacy in Social Network Data

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

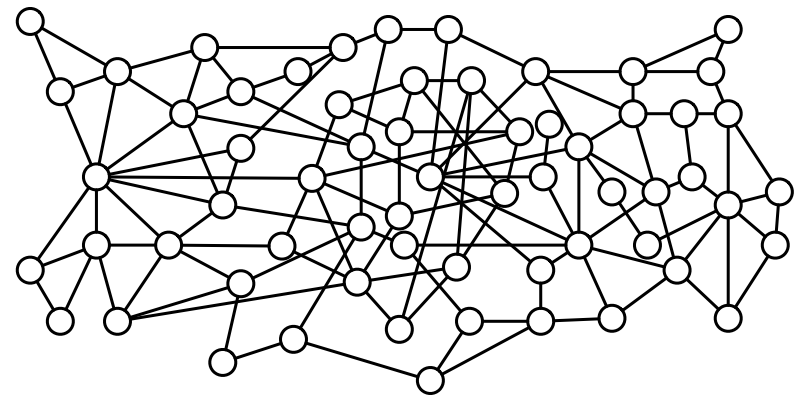
- Current safeguards based on anonymization: replace node names with random IDs.

With more detailed data, anonymization has run into trouble:

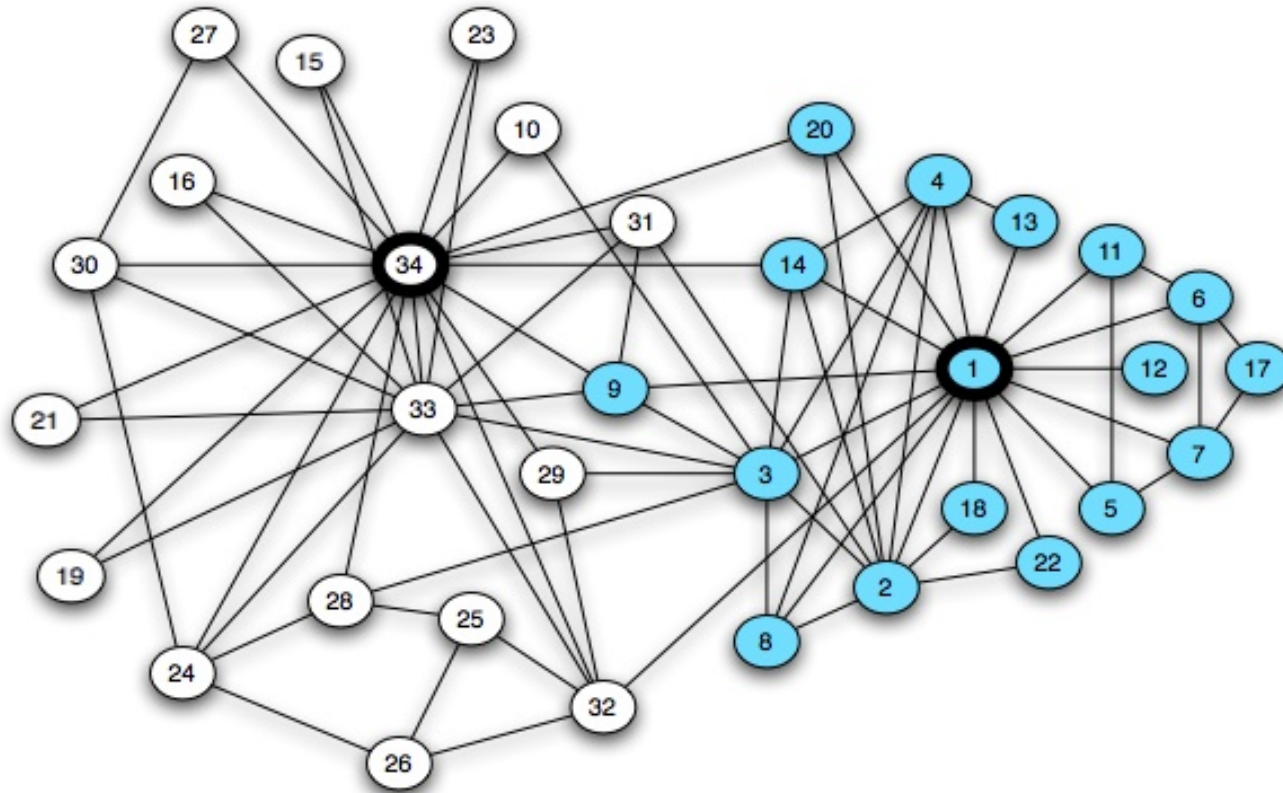
- Identifying on-line pseudonyms by textual analysis [Novak-Raghavan-Tomkins 2004]
- De-anonymizing Netflix ratings via time series [Narayanan-Shmatikov 2006]
- Search engine query logs: identifying users from their queries.

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

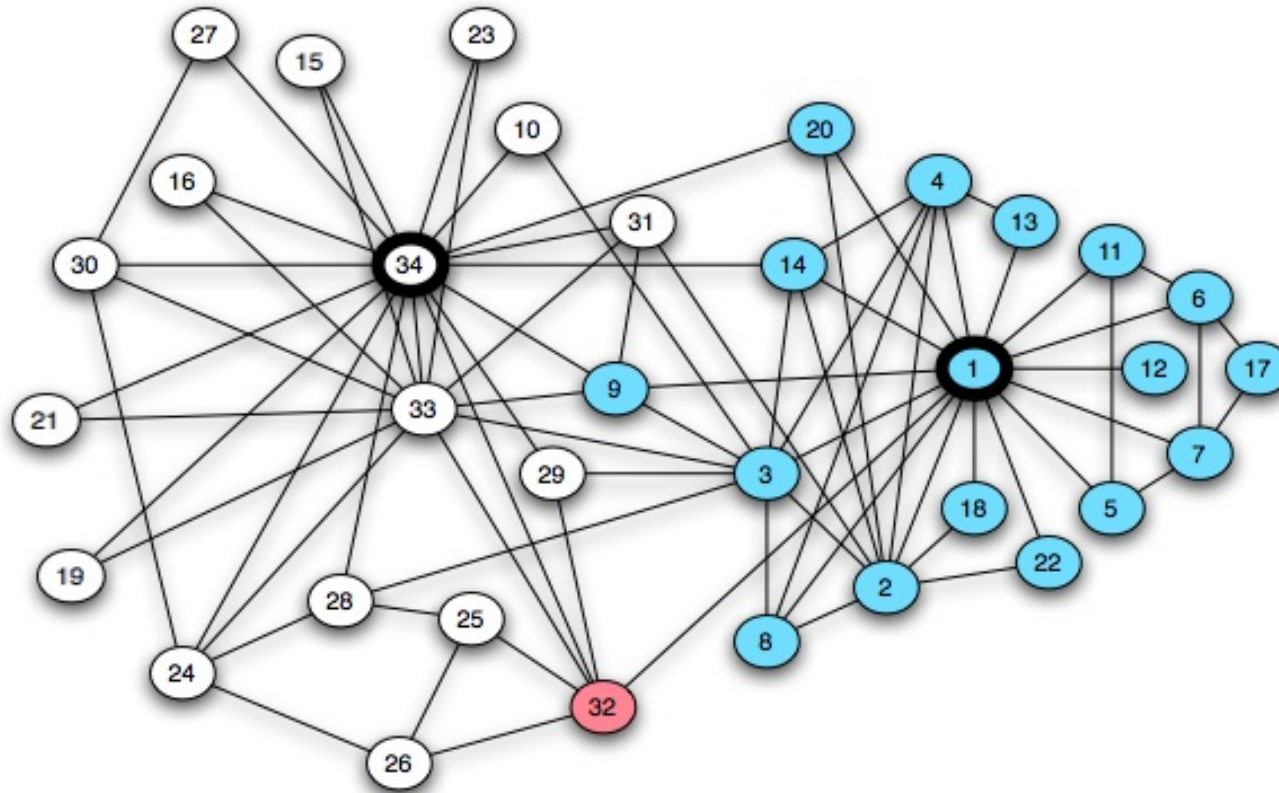
- E.g. no text, time-stamps, or node attributes
- Just a graph with nodes numbered $1, 2, 3, \dots, n$.



An Example of What Can Go Wrong

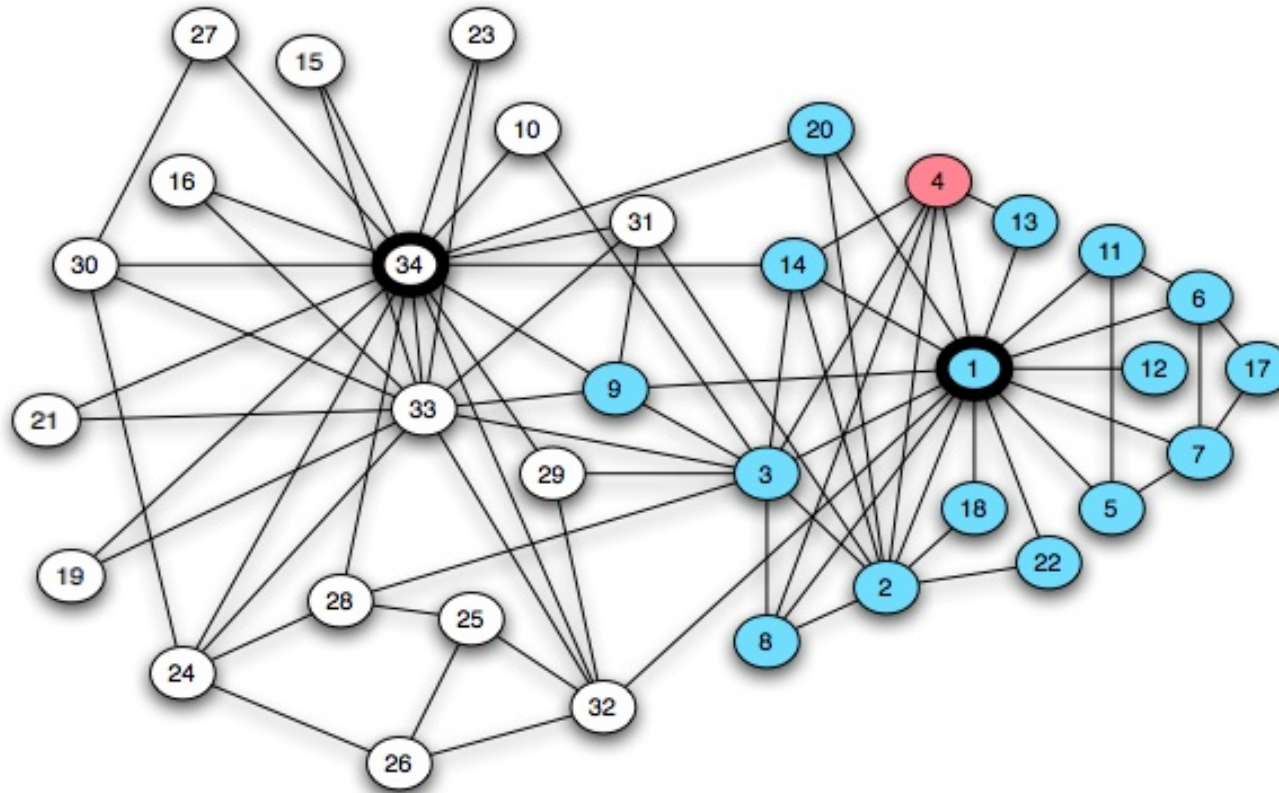


An Example of What Can Go Wrong



- 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.

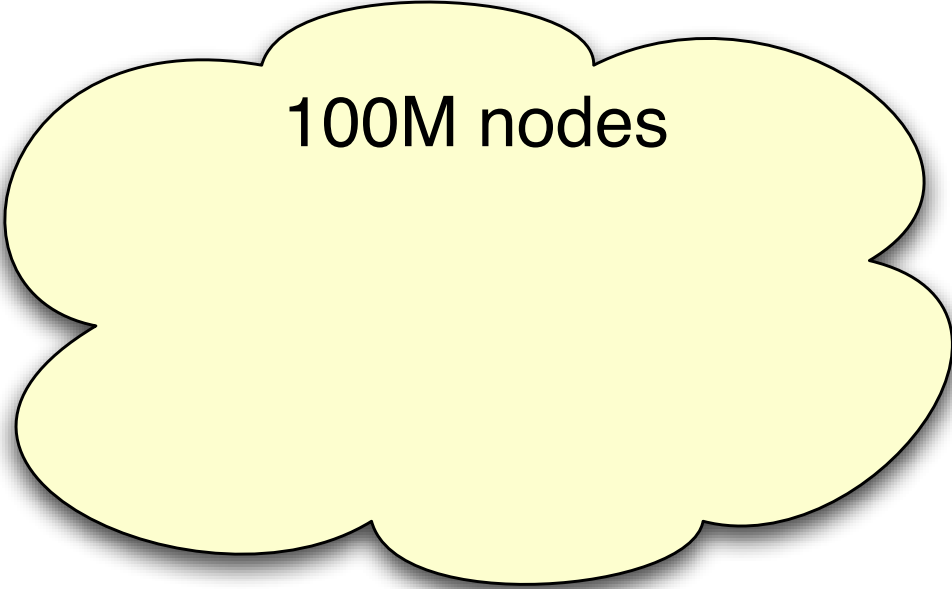
Template for an active attack on an anonymized network
[Backstrom-Dwork-Kleinberg 2007]

- 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

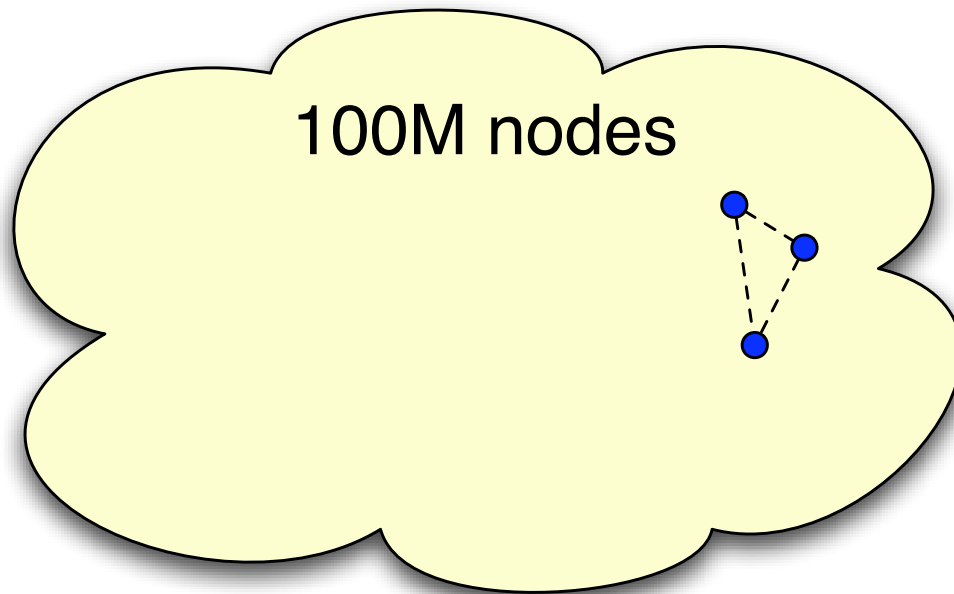


100M nodes

Scenario:

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

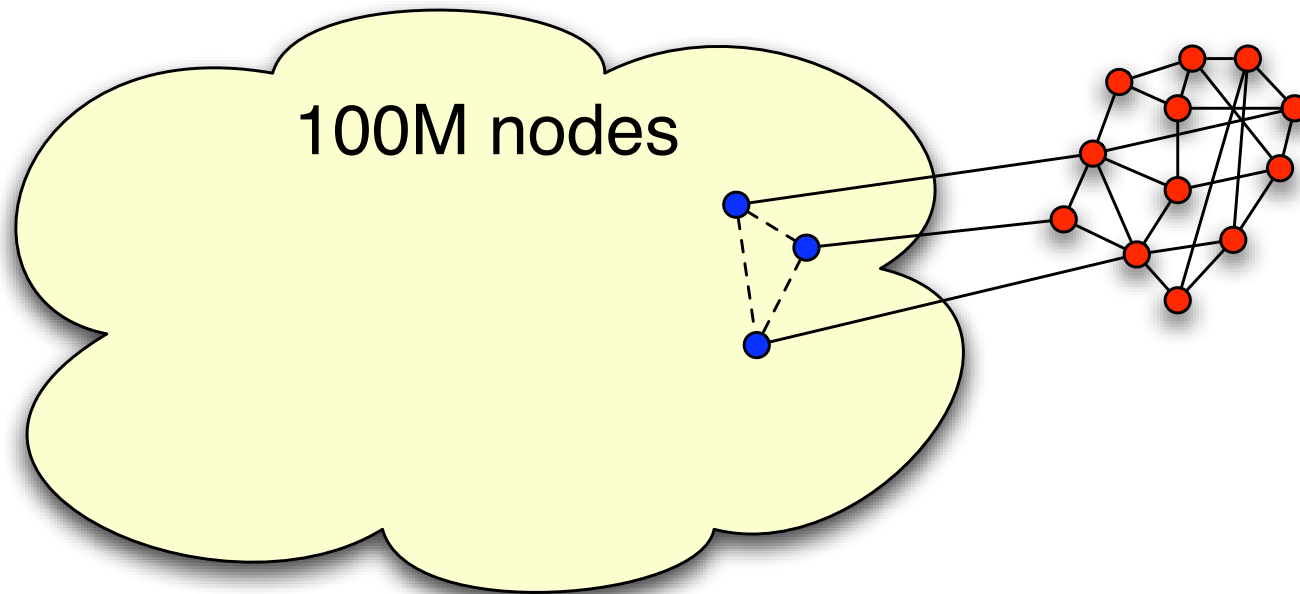
An Attack



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

Goal is to learn edge relations among them.

An Attack

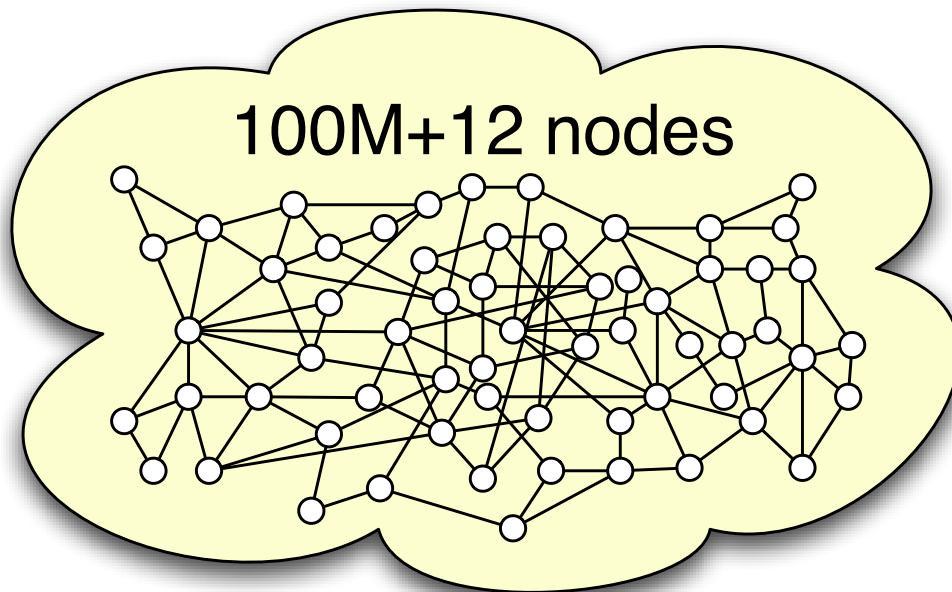


Before dataset is released:

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

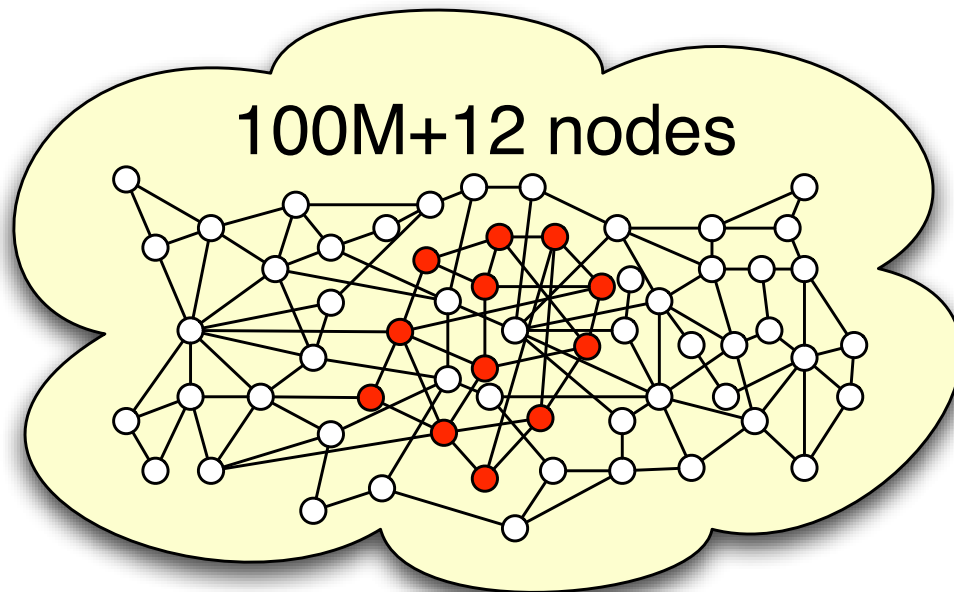


When anonymized dataset is released, need to find H .

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

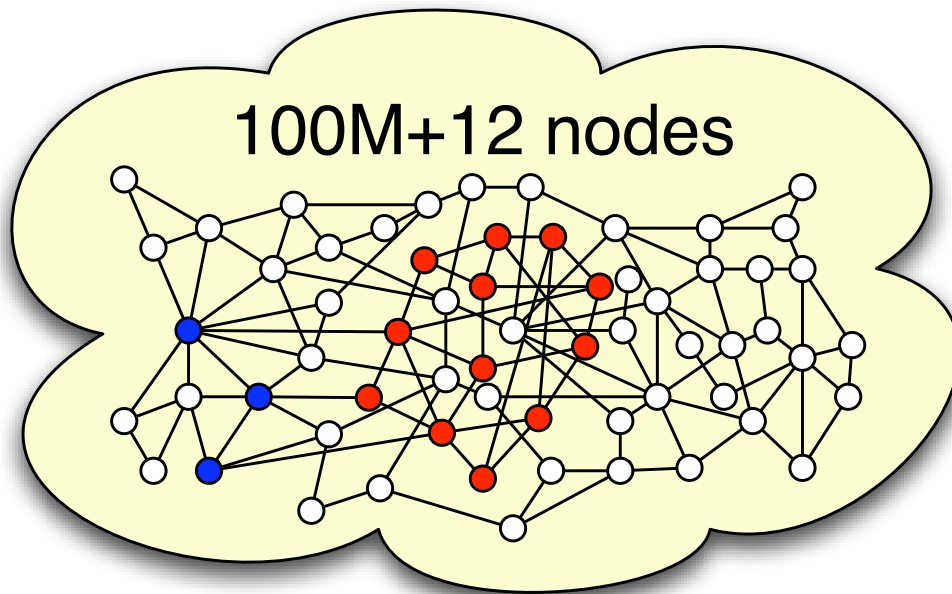


In fact,

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

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

An Attack

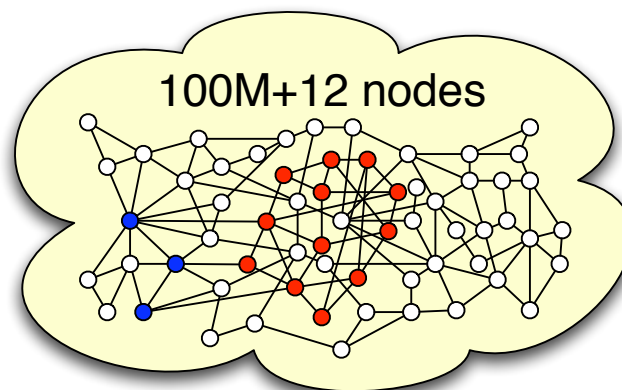


Once H is found:

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

Privacy Challenges for Social Network Data

Basic result: Can begin breaching privacy by creating roughly $2 \log n$ new nodes.



- Analysis uses ideas from Ramsey Theory.
- In experiments on 4.4 million-node LiveJournal graph, 7-node graph H can compromise 70 targeted nodes (and hence ~ 2400 edge relations).

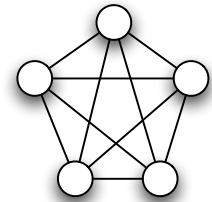
Passive attacks:

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

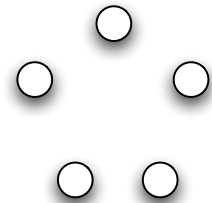
Why is H Unique? Ideas from Ramsey Theory

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:

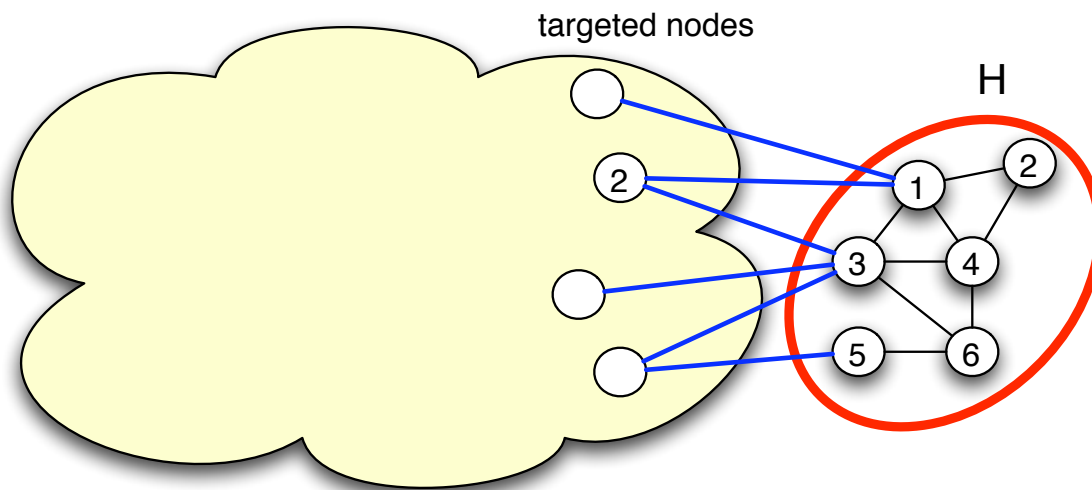
- Build random n -node graph, include each edge with prob. $\frac{1}{2}$.
- There are $< n^k$ sets of k nodes; each is a clique or independent set with probability $2^{-\binom{k}{2}} \approx 2^{-k^2/2}$.
- Product $n^k \cdot 2^{-k^2/2}$ upper-bounds probability of any clique or indep. set; it drops below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

Why is H Unique? Ideas from Ramsey Theory

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

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

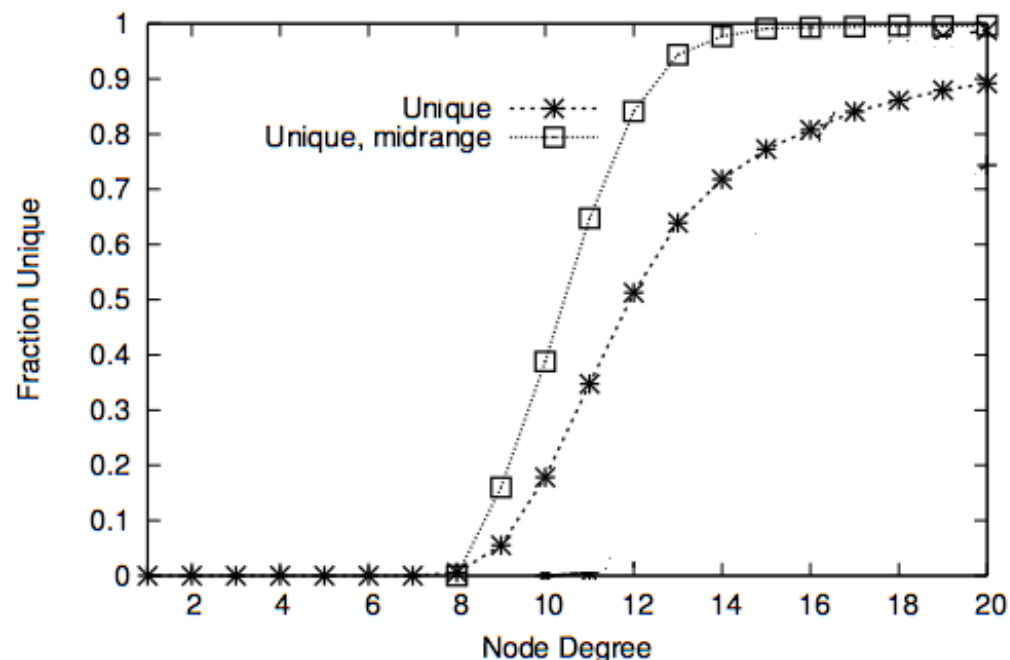
But main calculation starts from same premise.



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

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.
 - May be difficult to obfuscate network effectively (e.g. [Dinur-Nissim 2003, Dwork-McSherry-Talwar 2007])
 - Interactive mechanisms for network data may be possible (e.g. [Dwork-McSherry-Nissim-Smith 2006])
- Recent proposals specifically aimed at protecting
 - social networks [Hay et al '07, Zheleva et al '07, Korolova et al '08]
 - query logs [Adar '07, Kumar et al '07, Xiong-Agichtein '07]
- Further issues
 - Privacy-utility trade-offs.
 - Even without overt attacks, increasingly refined pictures of individuals begin to emerge.

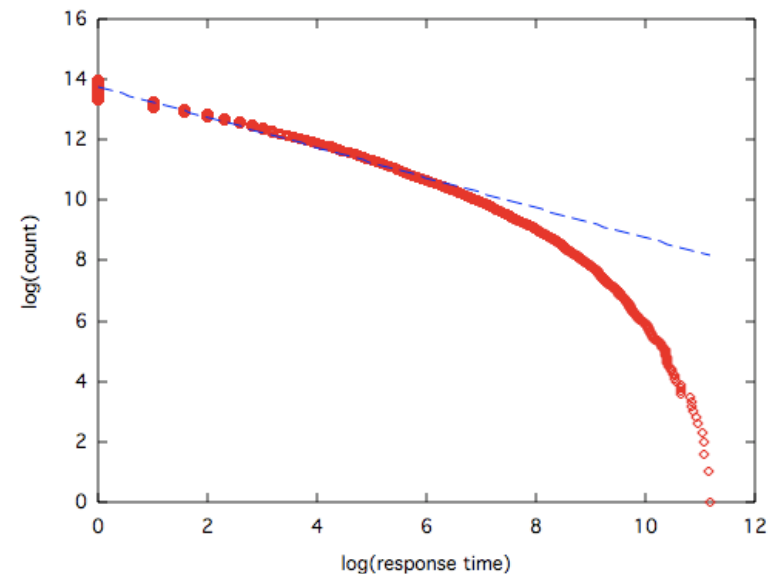
Final Reflections: Toward a Model of You

Further direction: from populations to individuals

- Distributions over millions of people leave open several possibilities:
 - Individual are highly diverse, and the distribution only appears in aggregate, or
 - Each individual personally follows (a version of) the distribution.
- Recent studies suggests that sometimes the second option may in fact be true.

Example: what is the probability that you answer a piece of e-mail t days after receipt (conditioned on answering at all)?

- Recent theories suggest $t^{-1.5}$ with exponential cut-off [Barabasi 2005]



Final Reflections: Glimpses into Massive Networks

MySpace is doubly awkward because it makes public what should be private. It doesn't just create social networks, it anatomizes them. It spreads them out like a digestive tract on the autopsy table. You can see what's connected to what, who's connected to whom.

– Toronto Globe and Mail, June 2006.

- Social networks — implicit for millenia — are increasingly being recorded at arbitrary resolution and browsable in our information systems.
- Your software has a trace of your activities resolved to the second — and increasingly knows more about your behavior than you do.
- Will all of this actually help us understand ourselves and each other any better?