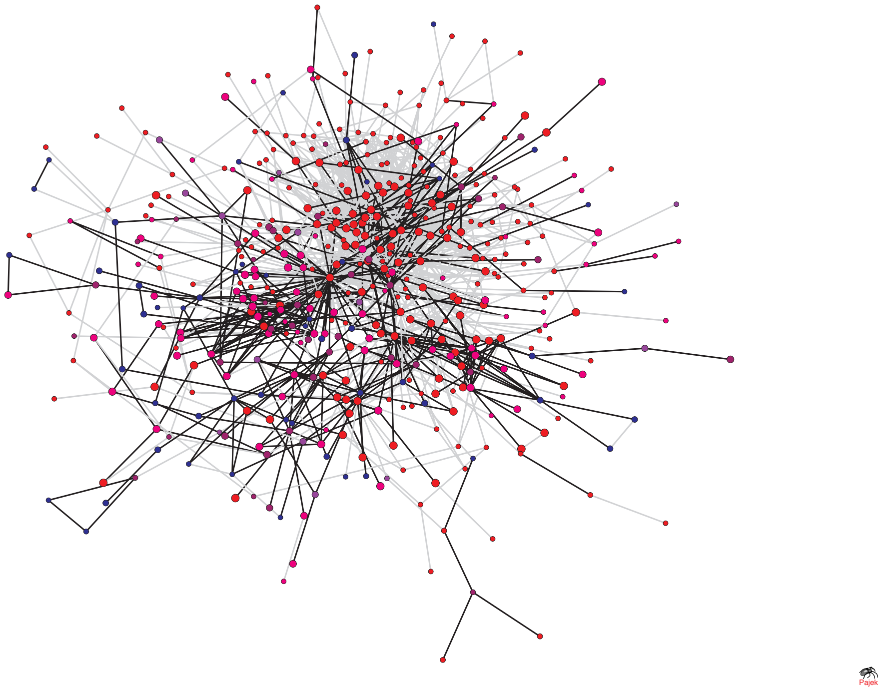
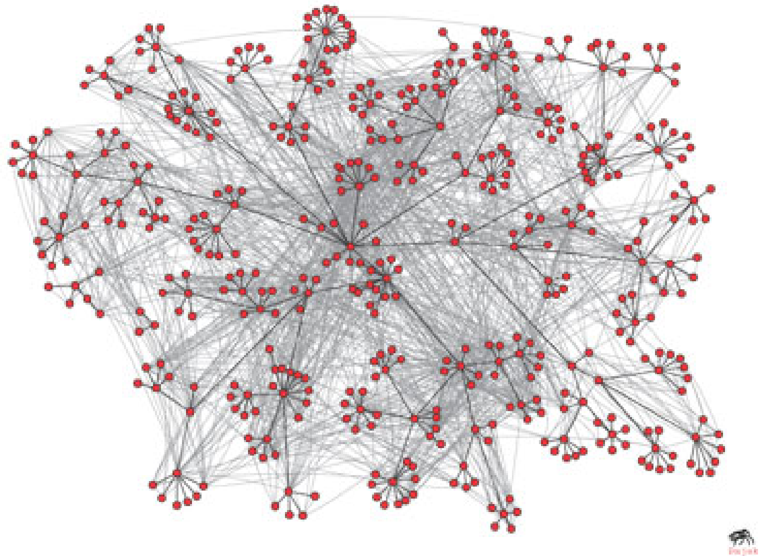
Algorithmic Models for Social Network Phenomena

Jon Kleinberg

Cornell University

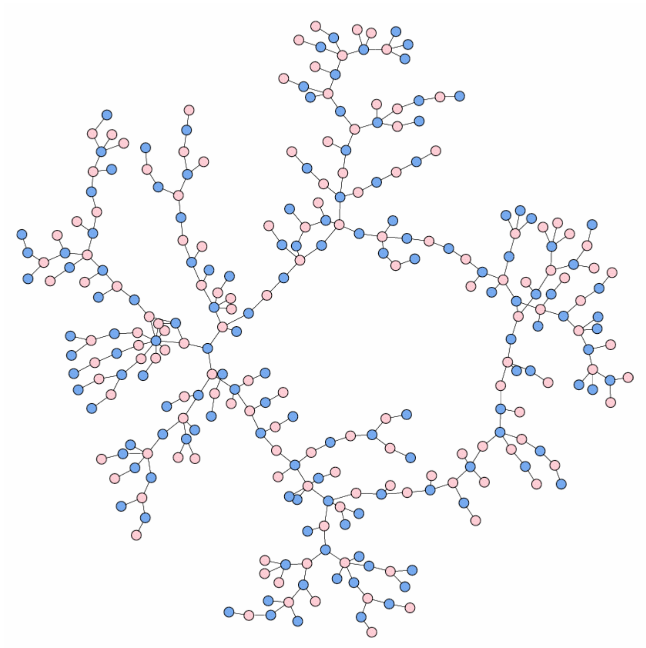
Networks as Phenomena

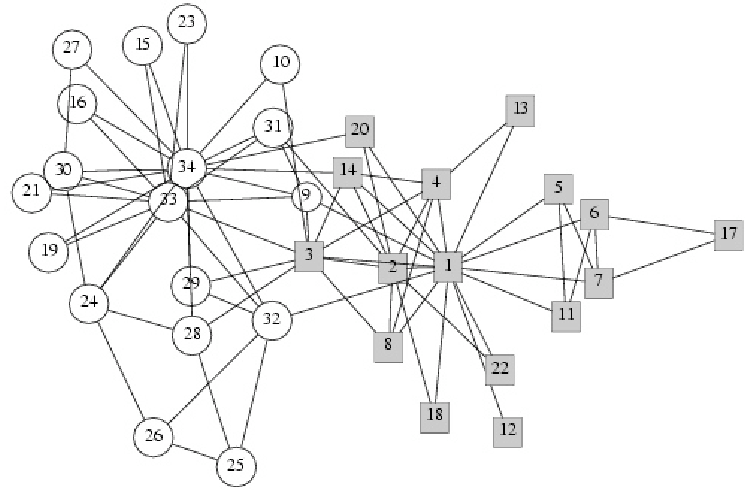




- Complex networks as phenomena, not just designed artifacts.
- What recurring patterns emerge, why are they there, and what are the consequences for computing and information systems?

Social and Technological Networks





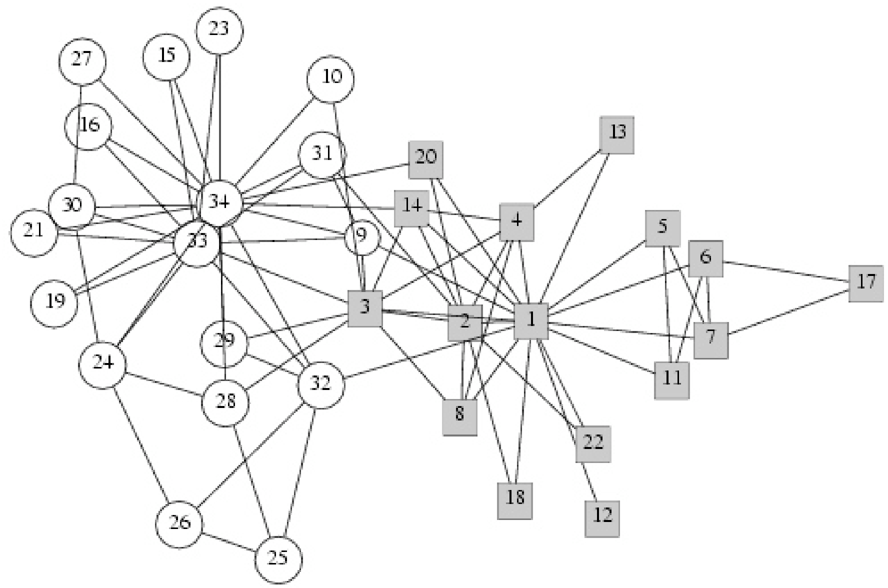
Social networks: friendships, contacts, collaboration, influence, organizational structure, economic institutions.

- Social and technological networks are intertwined:
  Web content, blogging, e-mail/IM, MySpace/Facebook/...
- New technologies change our patterns of social interaction.
- Collecting social data at unprecedented scale and resolution.

Rich Social Network Data

Traditional obstacle:
Can only choose 2 of 3.

- Large-scale
- Realistic
- Completely mapped

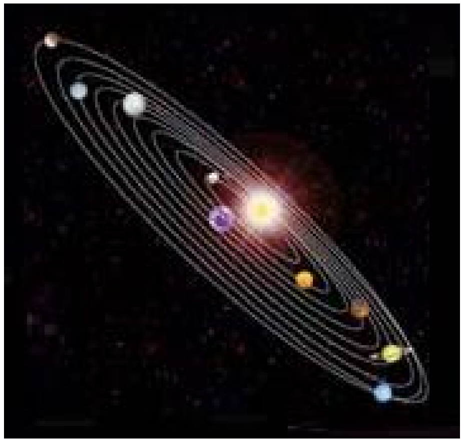


Two lines of research, looking for a meeting point.

- Social scientists engaged in detailed study of small datasets, concerned with social outcomes.
- Computer scientists discovering properties of massive network datasets that were invisible at smaller scales.

Modeling Complex Networks

We want Kepler's Laws of Motion for the Web.
– Mike Steuerwalt,
NSF KDI Workshop, 1998



Opportunity for deeper understanding of information networks and social processes, informed by theoretical models and rich data.

- Mathematical / algorithmic models form the vocabulary for expressing complex social-science questions on complex network data.
- Payoffs from the introduction of an algorithmic perspective into the social sciences.

Overview

(1) Small-world networks and decentralized search
- Stylized models expose basic patterns.
- Identifying the patterns in large-scale data.

(2) A problem that is less well understood at a large scale: diffusion and cascading behavior in social networks
- The way in which new practices, ideas, and behaviors spread through social networks like epidemics.
- Models from discrete probability, data from on-line communities, open questions in relating them.

(3) Privacy and anonymity in on-line data.
- The perils in using anonymized social network data.
- Attacks on anonymized networks using small identifiable subgraphs.

Small-World Networks

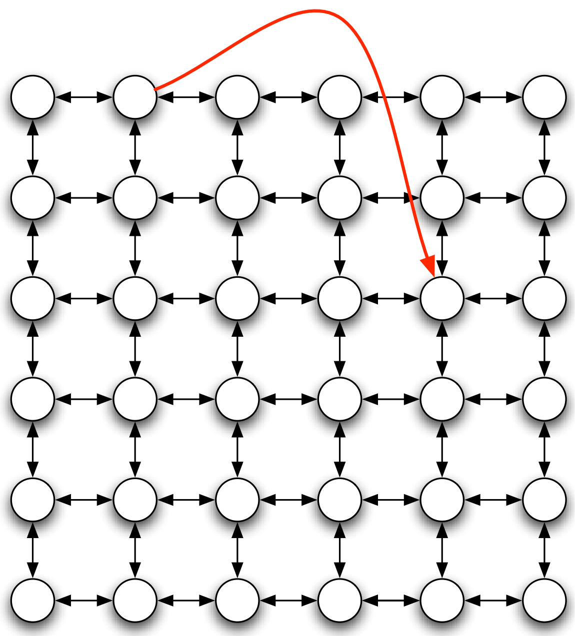
Milgram's small-world experiment (1967)

Choose a target in Boston, starters in Nebraska.

A letter begins at each starter, must be passed between personal acquaintances until target is reached.

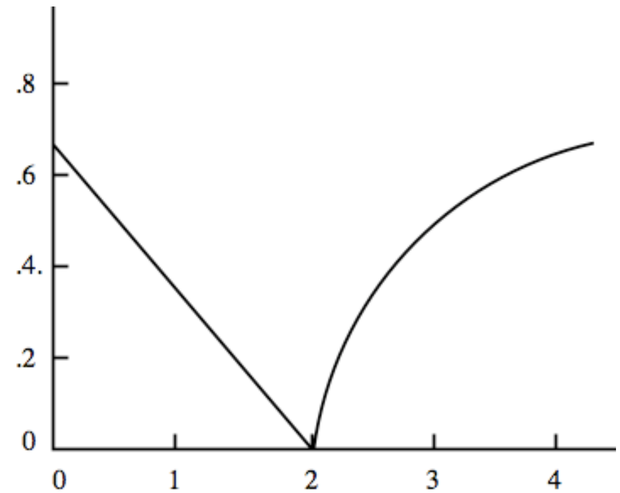
Six steps on average $\longrightarrow$ six degrees of separation.

- Routing in a (social) network: When is local information sufficient? [Kleinberg 2000]
- Variation on network model of Watts and Strogatz [1998].
- Add edges to lattice: u links to v with probability $d(u, v)^{-\alpha}$.



Small-World Models

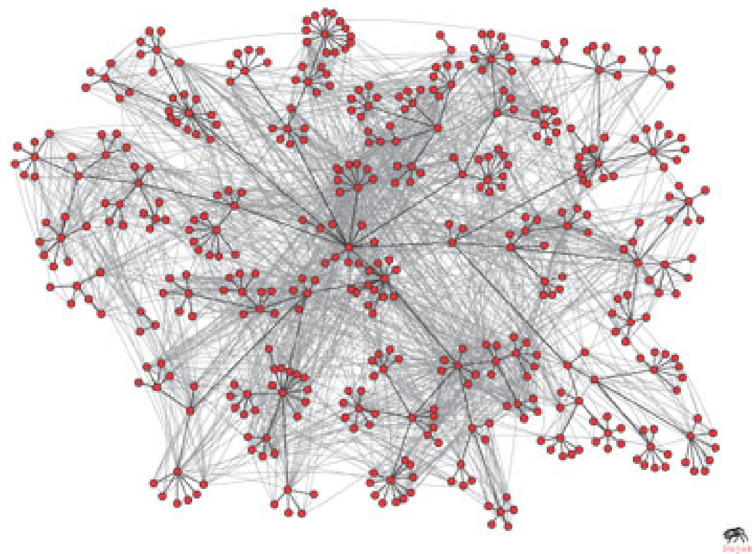
- Optimal exponent $\alpha = 2$: yields routing time $\sim c \log^2 n$.
- All other exponents yield $\sim n^{\varepsilon}$ for some $\varepsilon > 0$.



- Generalizations to random networks on different "scaffolds":
  - Trees, set systems, low-dimensional metrics [Kleinberg '01, Watts-Dodds-Newman '02, Slivkins '05, Fraigniaud-Lebhar-Lotker '06, Abraham-Gavoille '06]
- Relation to long-range percolation, structured random graphs
  - [Newman-Schulman'86, Aizenman-Chayes-Chayes-Newman'88, Bollobás-Chung '88, Benjamini-Berger '01, Coppersmith-Gamarnik-Sviridenko '02, Biskup '04, Berger '06]
- Connections to peer-to-peer algorithms
  - [Kempe-Kleinberg-Demers '01, Malkhi-Naor-Ratajczak '02, Aspnes-Diamadi-Shah '02, Zhang-Goel-Govindan '02, Manku-Bawa-Raghavan '03, Li et al. '05]

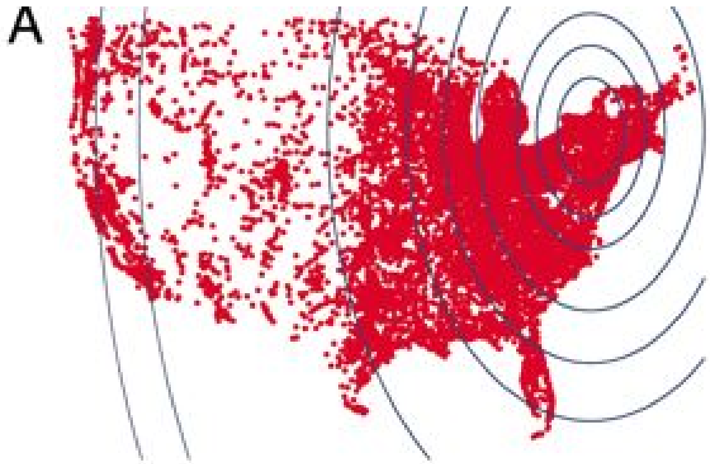
Social Network Data

- [Adamic-Adar 2003]: social network on 436 HP Labs researchers.
- Joined pairs who exchanged ≥ 6 e-mails (each way).



- Compared to "group-based" model [Kleinberg 2001]
  - Probability of link (v, w) prop. to $g(v, w)^{-\alpha}$, where $g(v, w)$ is size of smallest group containing v and w.
  - $\alpha = 1$ gives optimal search performance.
- In HP Labs, groups defined by sub-trees of hierarchy.
- Links scaled as $g^{-3/4}$.

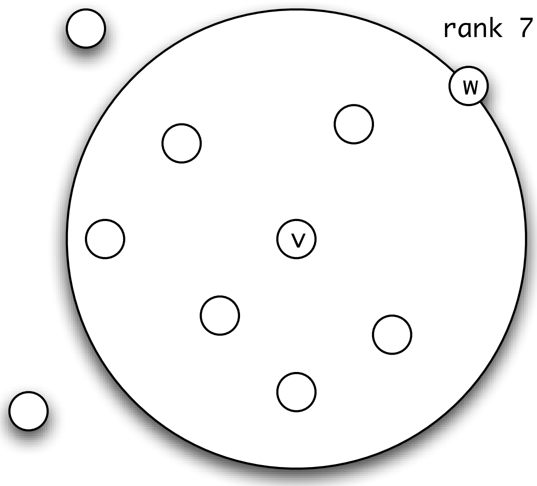
Geographic Data: LiveJournal



Liben-Nowell, Kumar, Novak, Raghavan, Tomkins (2005) studied LiveJournal, an on-line blogging community with friendship links.

- Large-scale social network with geographical embedding:
    - 500,000 members with U.S. Zip codes, 4 million links.
- Analyzed how friendship probability decreases with distance.
- Difficulty: non-uniform population density makes simple lattice models hard to apply.

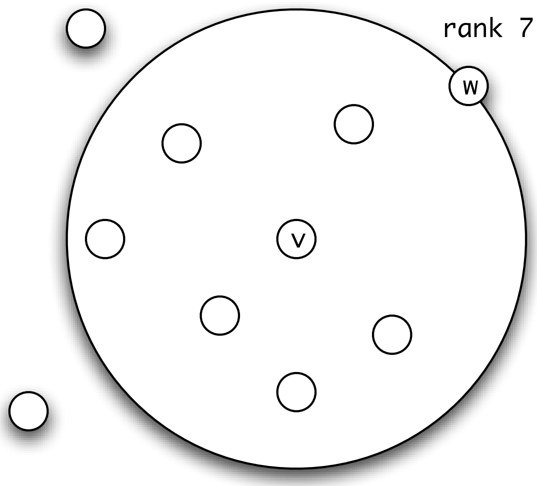
LiveJournal: Rank-Based Friendship

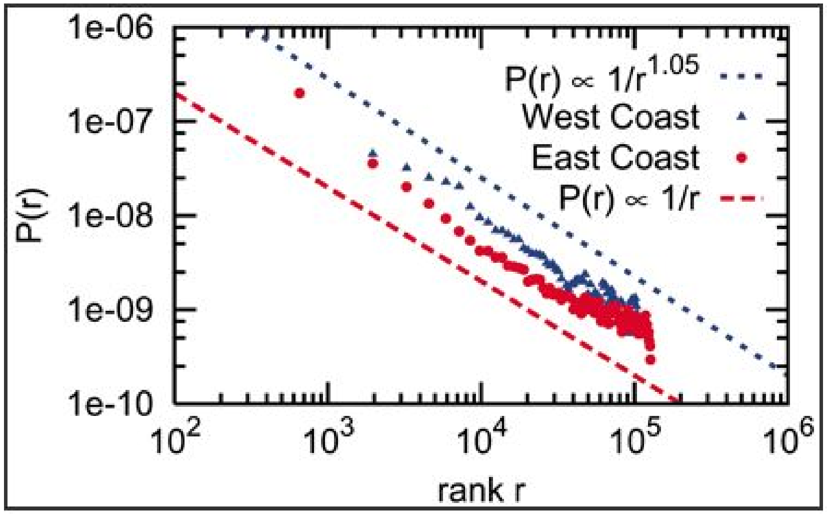


Rank-based friendship: <u>rank</u> of w with respect to v is number of people x such that $d(v, x) < d(v, w)$.

- Decentralized search with (essentially) arbitrary population density, when link probability proportional to $\mathrm{rank}^{-\beta}$.
- (LKNRT'05): Efficient routing when $\beta = 1$, i.e. $1/\mathrm{rank}$.
- Generalization of lattice result (diff. from set systems).

LiveJournal: Rank-Based Friendship



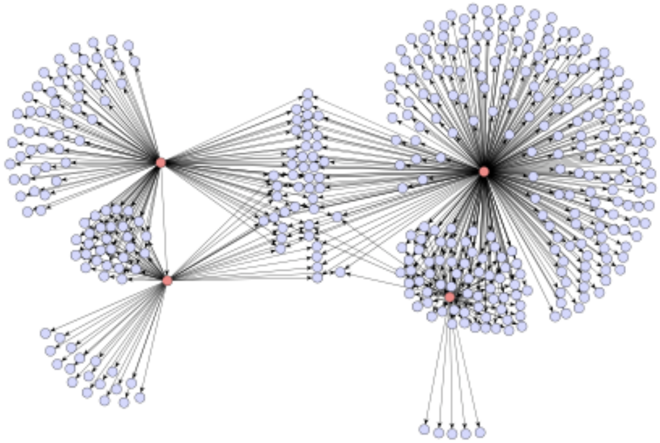


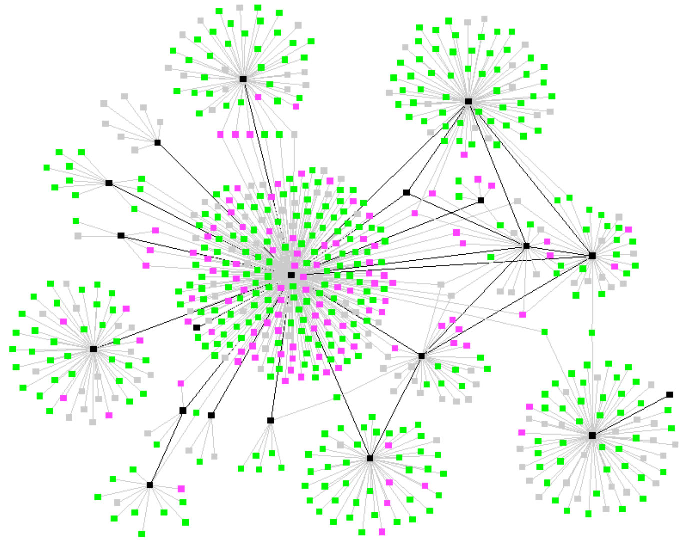
Rank-based friendship: rank of w with respect to v is number of people x such that $d(v, x) < d(v, w)$.

- Decentralized search with (essentially) arbitrary population density, when link probability proportional to $\mathrm{rank}^{-\beta}$.
- (LKNRT'05): Efficient routing when $\beta = 1$, i.e. $1/\mathrm{rank}$.
- Generalization of lattice result (diff. from set systems).

Punchline: LiveJournal friendships approximate $1/\mathrm{rank}$.

Diffusion in Social Networks





So far: focused search in a social network.

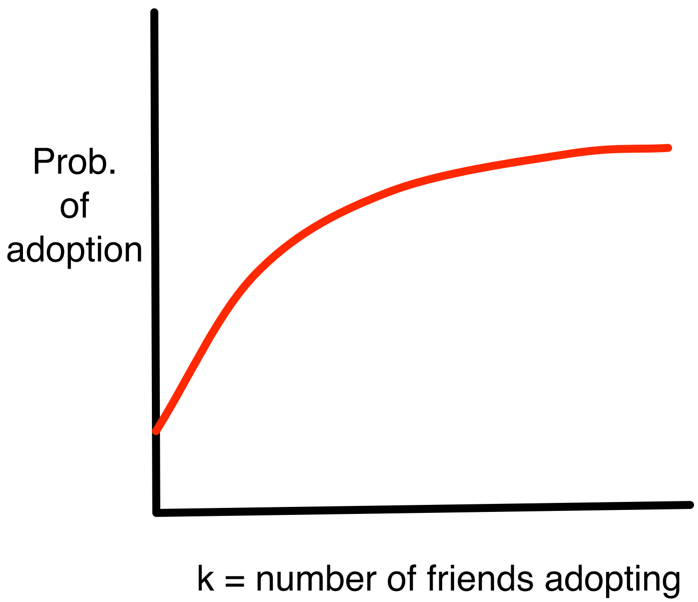
Now switch to diffusion, another fundamental social processs:
Behaviors that cascade from node to node like an epidemic.

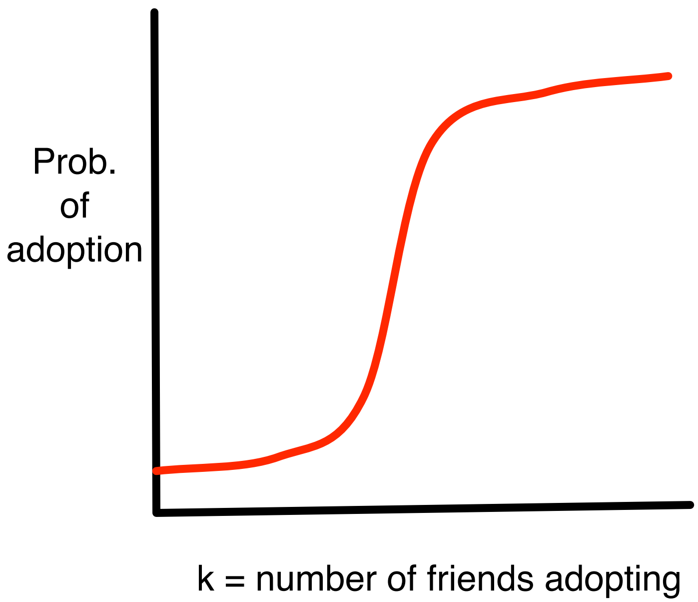
- News, opinions, rumors, fads, urban legends, ...
- Word-of-mouth effects in marketing, rise of new products.
- Changes in social priorities: smoking, recycling, ...
- Saturation news coverage; topic diffusion among bloggers.
- Localized collective action: riots, walkouts

Diffusion Curves

Basis for models: Probability of adopting new behavior depends on number of friends who have adopted.

- Bass 1969; Granovetter 1978; Schelling 1978





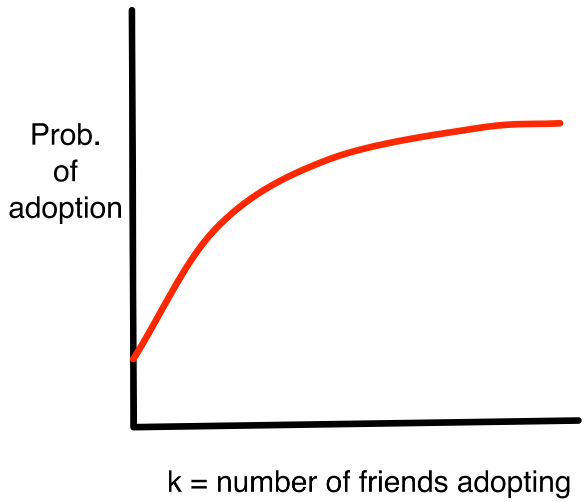
Key issue: qualitative shape of the diffusion curves.

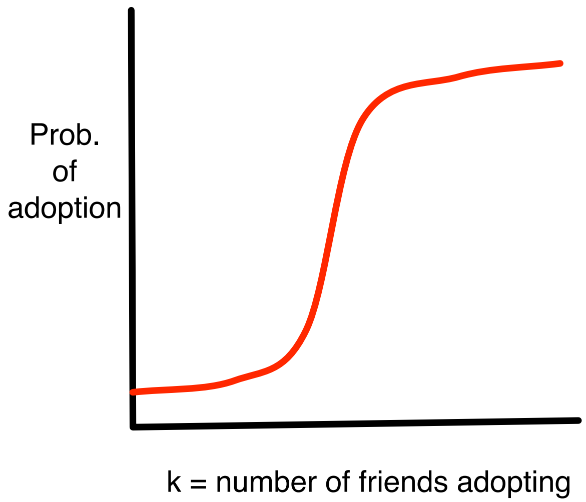
- Diminishing returns? Critical mass?

From individual-level model, can build network-level model:

- Run dynamics of contagion forward from initial "seed set."

Finding the Most Influential Set





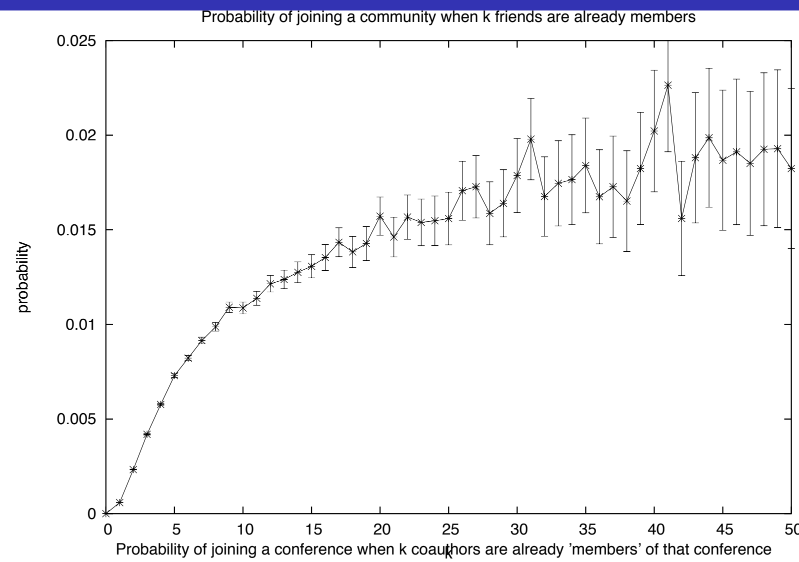
An algorithmic question [Domingos-Richardson 2001]:

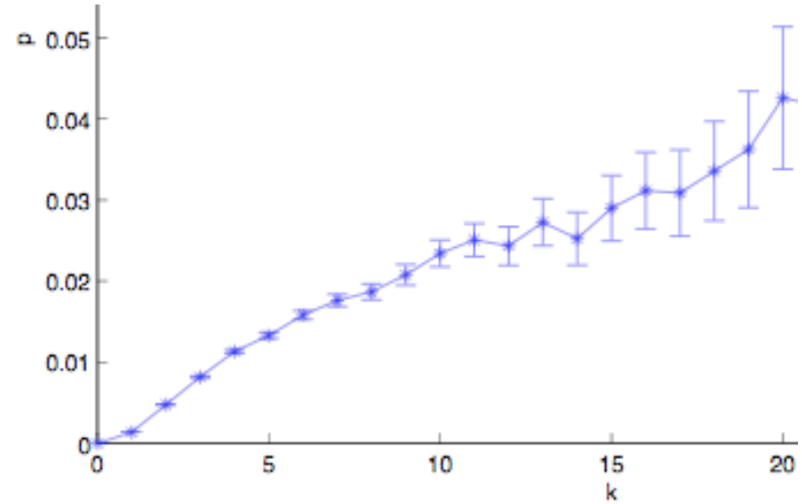
- If we can "seed" the new behavior at k nodes, and want to maximize the eventual spread, whom should we choose?

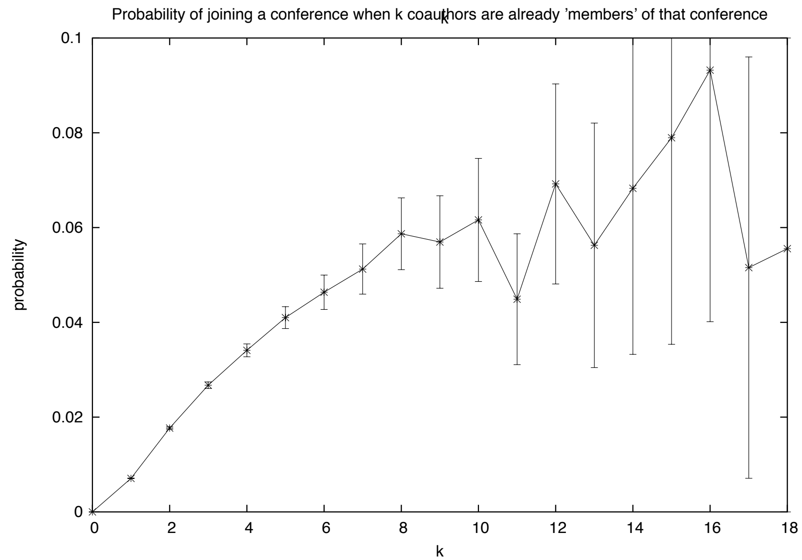
Computational complexity depends on diffusion curves.

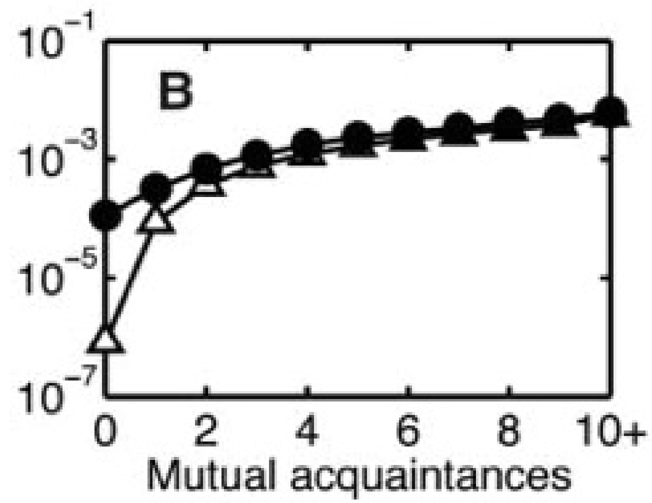
- Highly inapproximable with critical mass.
- With diminishing returns: constant-factor approximation [Kempe-Kleinberg-Tardos 2003, 2005; Mossel-Roch 2007]

Diffusion Curves









Joining a LiveJournal community [Backstrom et al. '06]

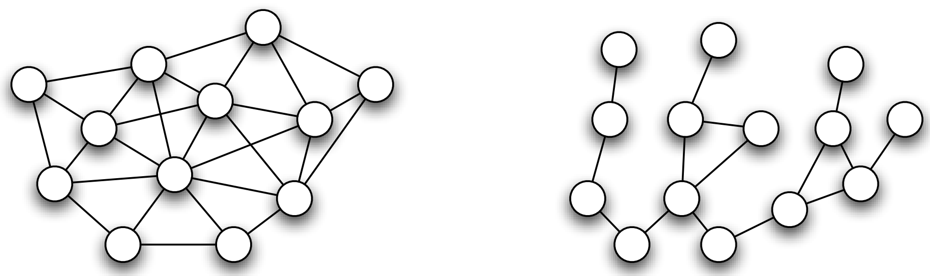
Authoring at a CS conference [Backstrom et al. '06]

Editing a Wikipedia article [Huttenlocher et al. '07]

Triadic closure in e-mail [Kossinets-Watts '06]

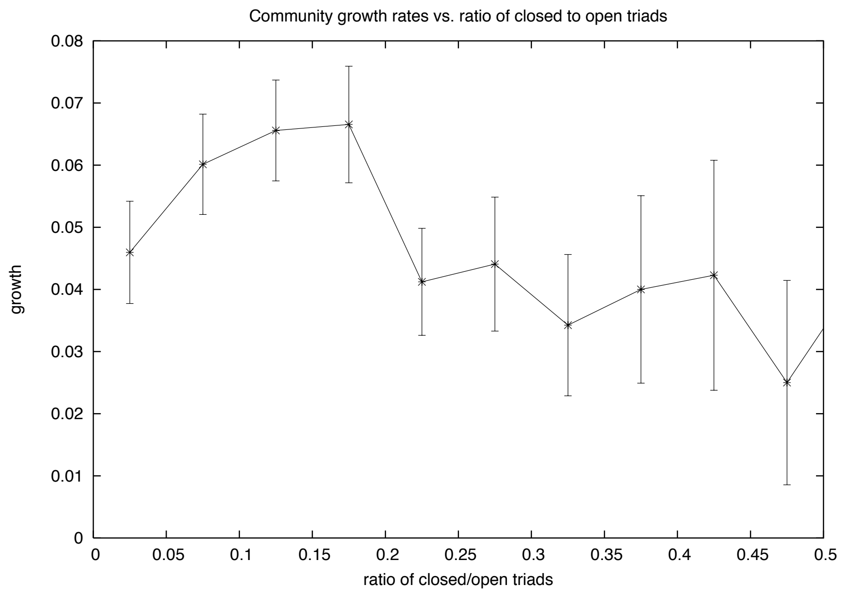
Toward a Notion of "Life Cycles"

How does a group's tendency to grow depend on its structural properties? [Backstrom et al. 2006]



- Define clustering = # triangles / # open triads.
- Look at growth from t_1 to t_2 as function of clustering.

- Groups with large clustering grow slower.
- Yet individuals are more likely to join when their friends in a group all know each other.



Diffusion in Computing and Information

- Diffusion of Topics [Gruhl et al 2004, Adar et al 2004]
  - News stories cascade through networks of bloggers and media
  - How should we track stories and rank news sources?
  - A taxonomy of sources: discoverers, amplifiers, reshapers, ...
- Building diffusion into the design of social media [Leskovec-Adamic-Huberman 2006, Kleinberg-Raghavan 2005]
  - Incentives to propagate interesting recommendations along social network links.
  - Simple markets based on question-answering and information-seeking.
- Predictive frameworks for diffusion
  - Machine learning models for the growth of communities [Backstrom et al. 2006]
  - Is a new idea's rise to success inherently unpredictable? [Salganik-Dodds-Watts 2006]

The Perils of Anonymized Data

Can accomplish a lot with public social network data.
But many interesting questions arise in private data:

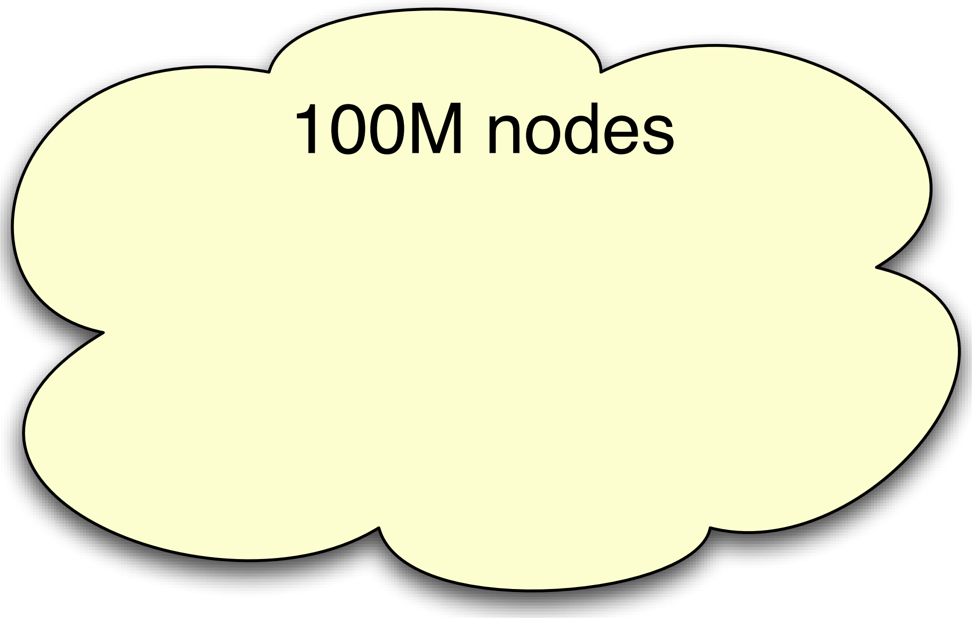
- E.g. E-mail, IM, voice, members-only communities.
- Standard approach to protecting the data: anonymize, replacing name at each node by a random string.
- After doing this, is it safe to release?

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]
- The AOL query logs ["This was a screw-up, and we're angry and upset about it." —AOL press release, 7 August 2006]

But what about just the unlabeled nodes and edges of a social network?

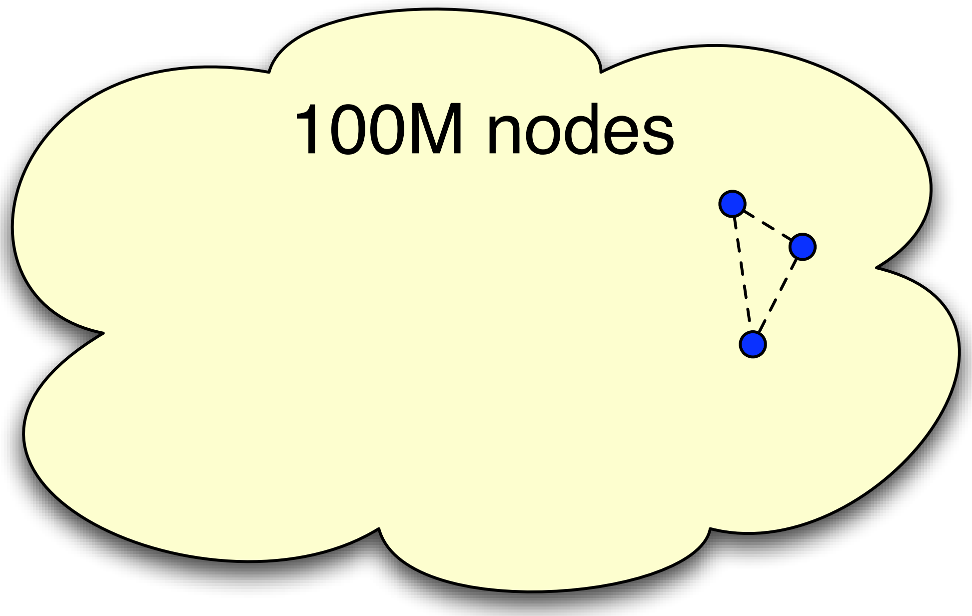
An Attack



Scenario from Backstrom-Dwork-Kleinberg 2007:

> Suppose a big company were going to release an anonymized communication graph on 100 million users.

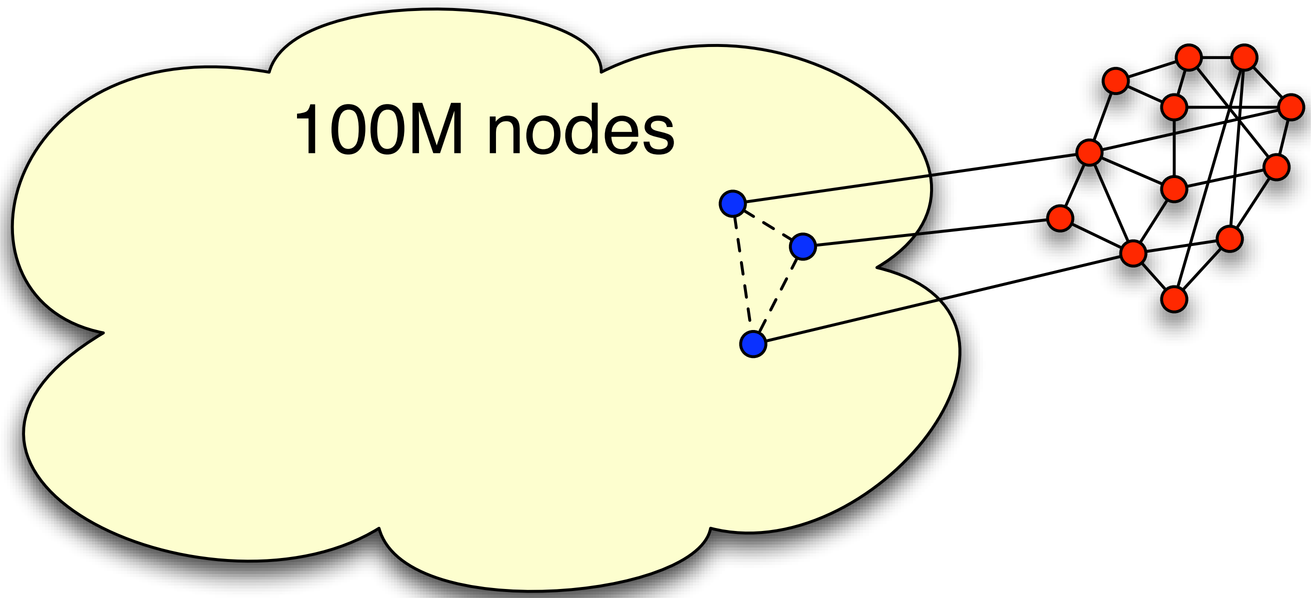
An Attack



An attacker chooses a small set of b user accounts to "target":

Goal is to learn edge relations among them.

An Attack

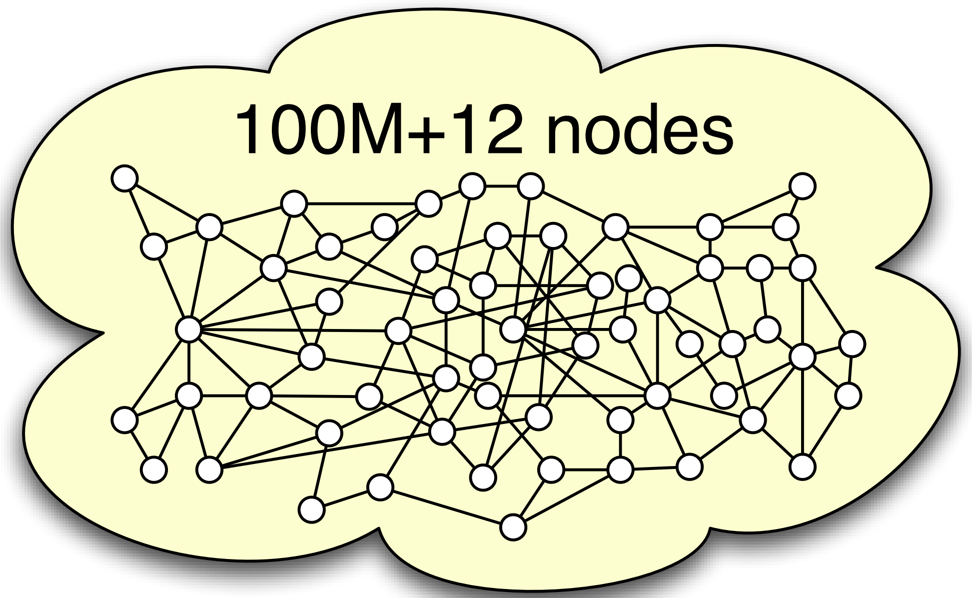


Before dataset is released:

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

Attach this new subgraph H to targeted accounts.

An Attack

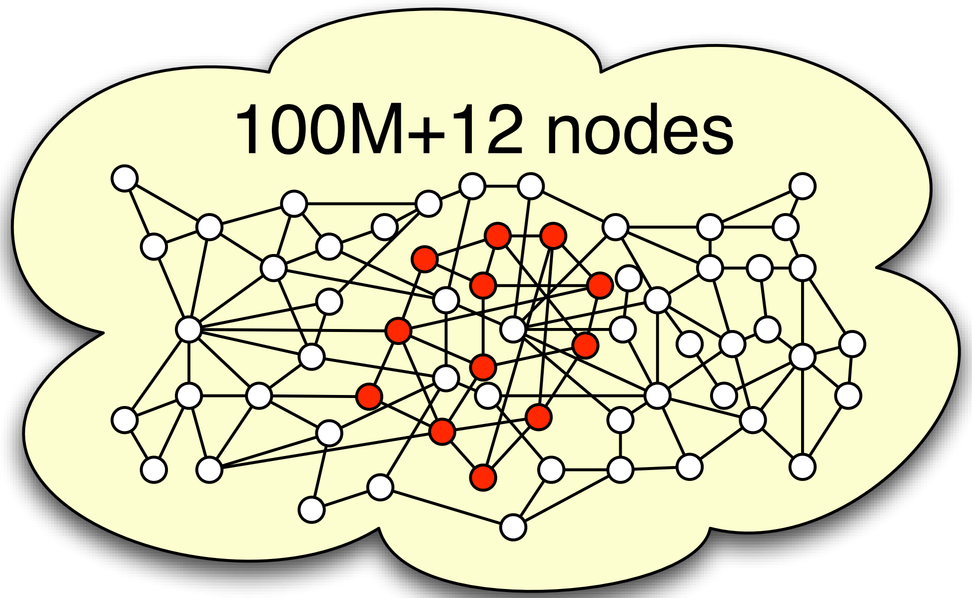


When anonymized dataset is released, need to find H.

Why couldn't there be many copies of H in the dataset?

Isn't subgraph isomorphism supposed to be a hard problem?

An Attack

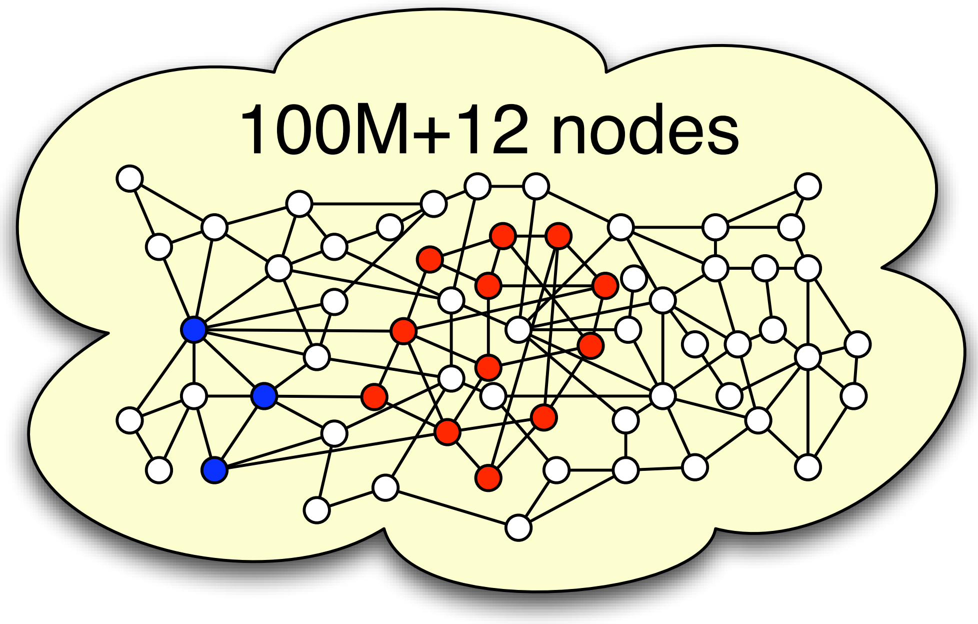


If H is random and of size $(2+\varepsilon)\log n$, then:

It's unique with high probability
(cf. Erdös's (1947) non-constructive Ramsey bound).

Brute-force search tree for H has near-linear size, since H is small and random.

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 H on $(2+\varepsilon)\log n$ nodes.
  Can compromise $\Theta(\log^2 n)$ targeted 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:

- Create H on $c\sqrt{\log n}$ nodes.
  Can compromise $(\frac{1}{2}-\varepsilon)c\sqrt{\log n}$ targeted nodes.
- Reconstruct from Gomory-Hu tree: break apart G along small cuts; find H as a "contiguous" piece.

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.

The Perils of Anonymized Data

What's the conclusion from this?

- Doesn't apply to social network data that's already public; orthogonal to issues of legal/contractual safeguards.
- But widespread release of an anonymized social network? Danger: you don't what someone's hidden in there. (And passive attacks don't even require advance planning.)
- Interesting direction: 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])

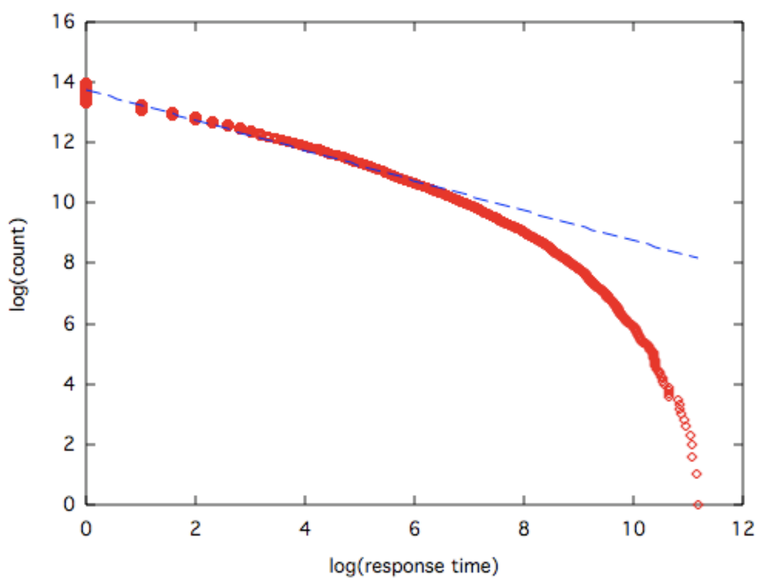
Reflections: Toward a Model of You

Further direction: from populations to individuals

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



Reflections: Interacting in the On-Line World

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.
- Trade-offs between rich data and individual privacy will remain an issue.
- Models based on algorithmic ideas will be crucial in understanding these developments.