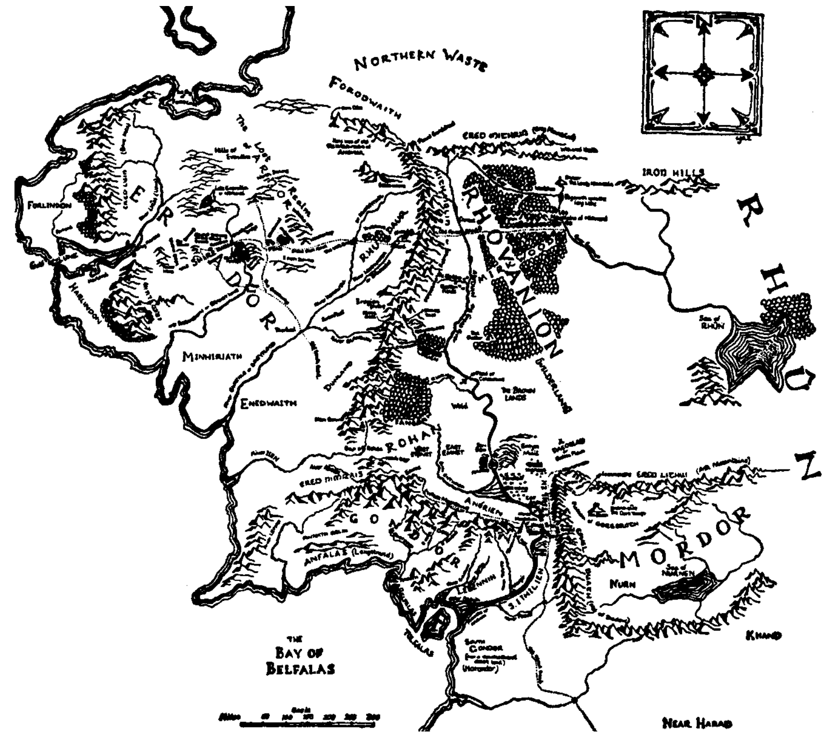
The Emerging Intersection of Social and Technological Networks

Jon Kleinberg

Cornell University

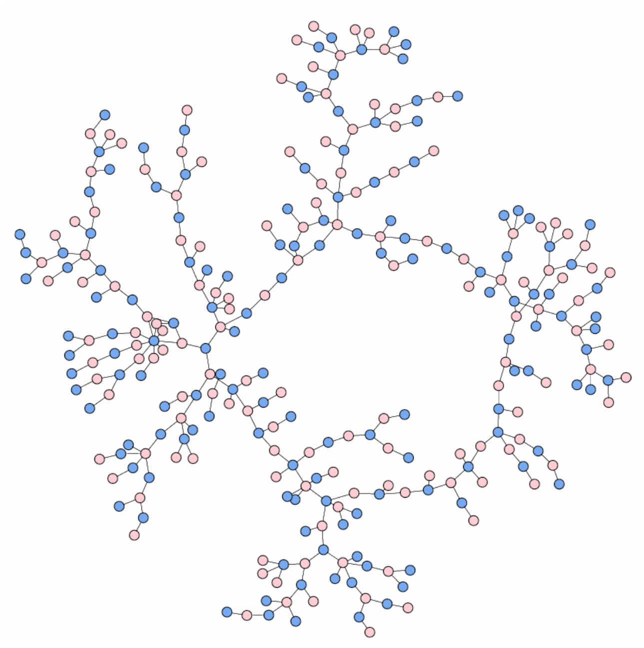
Networks as Phenomena

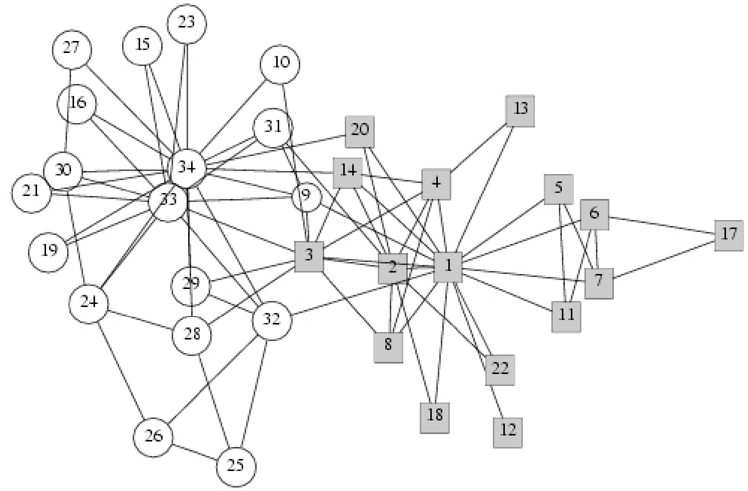
The emergence of 'cyberspace' and the World Wide Web is like the discovery of a new continent.
– Jim Gray,
1998 Turing Award address



- 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





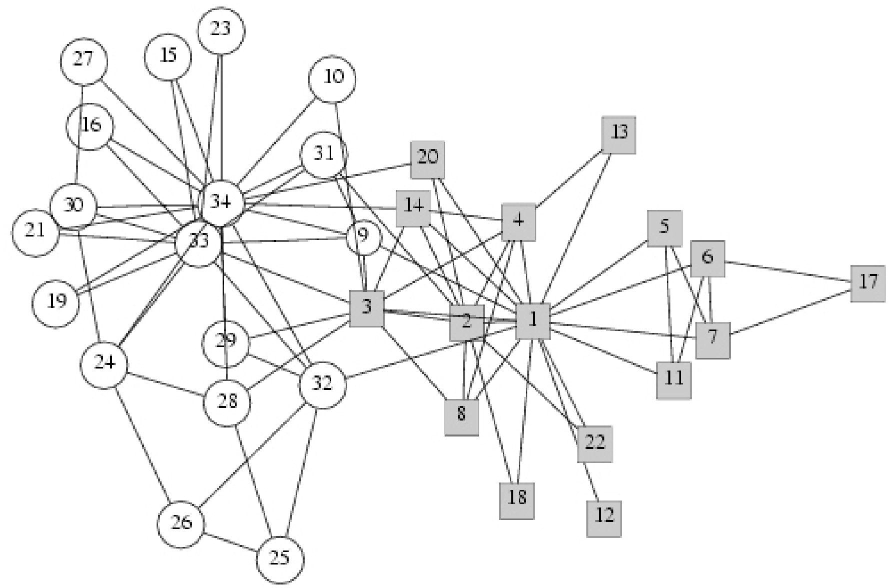
Social networks: friendships, contacts, collaboration, influence, organizational structure, economic institutions.

- 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

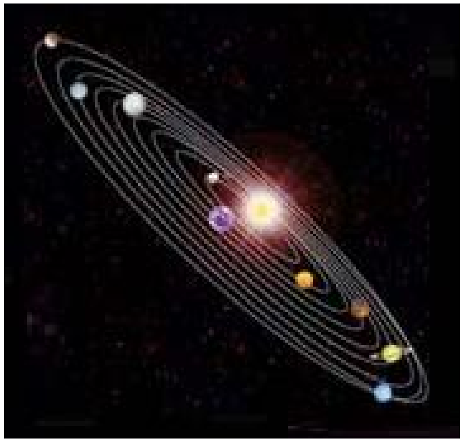


Two lines of research, looking for a meeting point.

- 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

Plan for the talk: two illustrations of this theme.

(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.

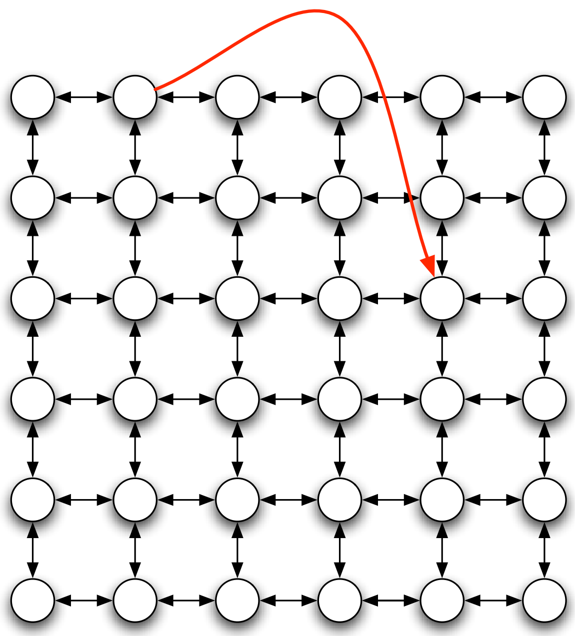
(3) Some further reflections on social interaction data.
- Modeling individuals vs. modeling populations

Small-World Networks

Milgram's small-world experiment (1967)

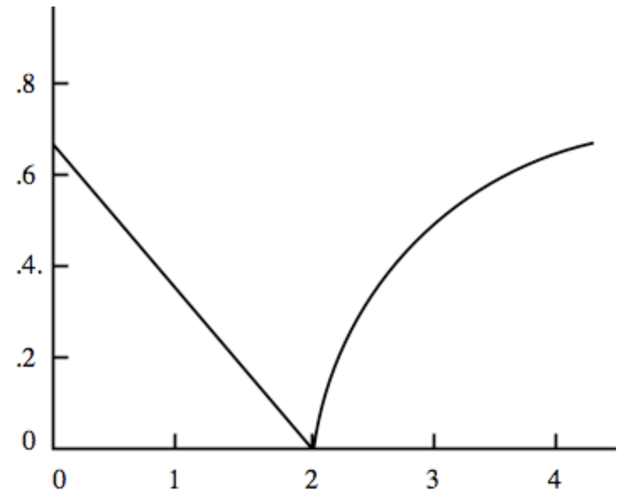
Choose a target in Boston, starters in Nebraska.
A letter begins at each starter, must be passed between personal acquaintances until target is reached.
Six steps on average $\longrightarrow$ six degrees of separation.

- 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

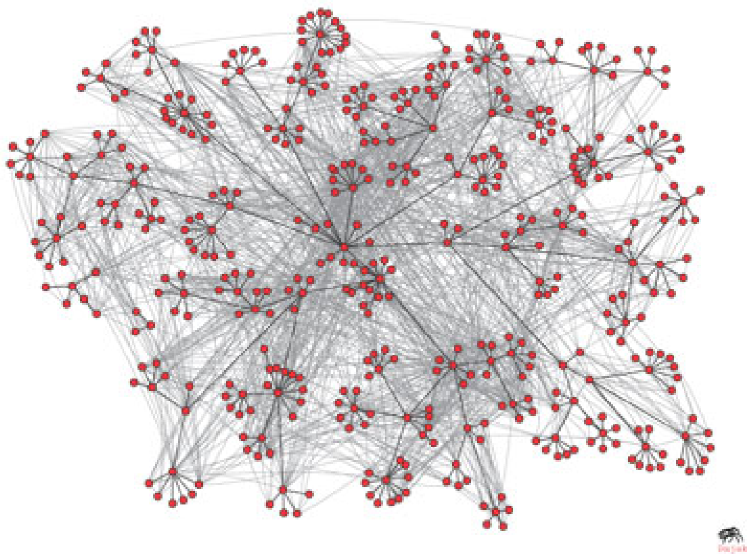
- Optimal exponent $\alpha = 2$: yields routing time $\sim c \log^2 n$.
- All other exponents yield $\sim n^{\varepsilon}$ for some $\varepsilon > 0$.



- Diameter at $\alpha = 2$ is $O(\log n)$; better routing via lookahead
  - [Fraigniaud-Gavoille-Paul '04, Lebhar-Schabanel '04, Manku-Naor-Wieder '04, Martel-Nguyen '04]
- Connections to long-range percolation in statistical physics
  - [Benjamini-Berger '01, Coppersmith-Gamarnik-Sviridenko '02, Biskup '04, Berger '06]
- Generalizations to random networks on different "scaffolds":
  - Trees, set systems [Kleinberg '01, Watts-Dodds-Newman '02]
  - Low tree-width, excl. minor [Fraigniaud '05, Abraham-Gavoille]
  - Doubling metrics [Slivkins '05, Fraigniaud-Lebhar-Lotker '06]

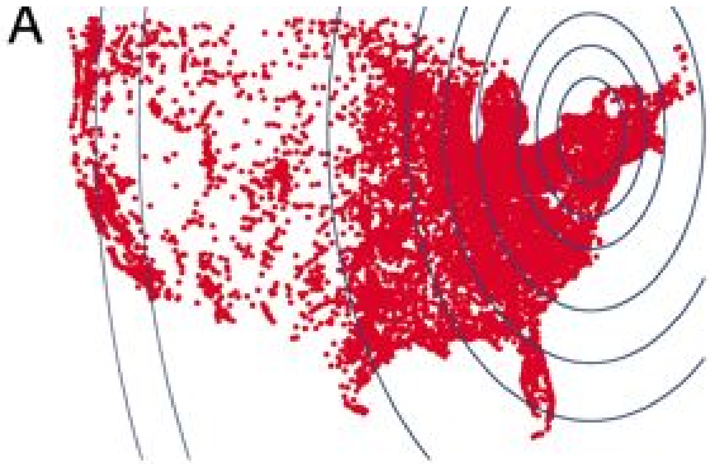
Social Network Data

- [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 $g^{-3/4}$.

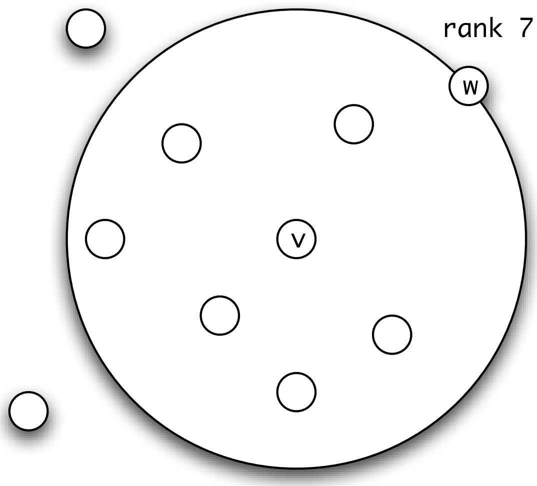
Geographic Data: LiveJournal

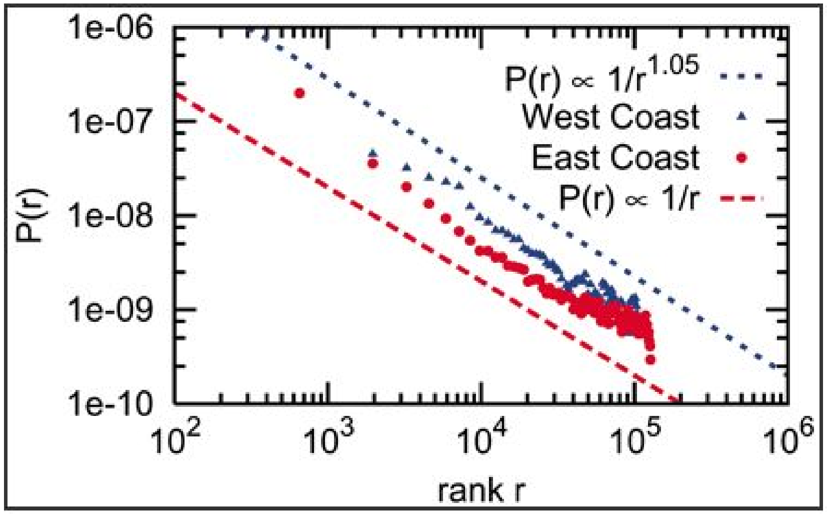


Liben-Nowell, Kumar, Novak, Raghavan, Tomkins (2005) studied LiveJournal, an on-line blogging community with friendship links.

- 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 $d(v, x) < d(v, w)$.

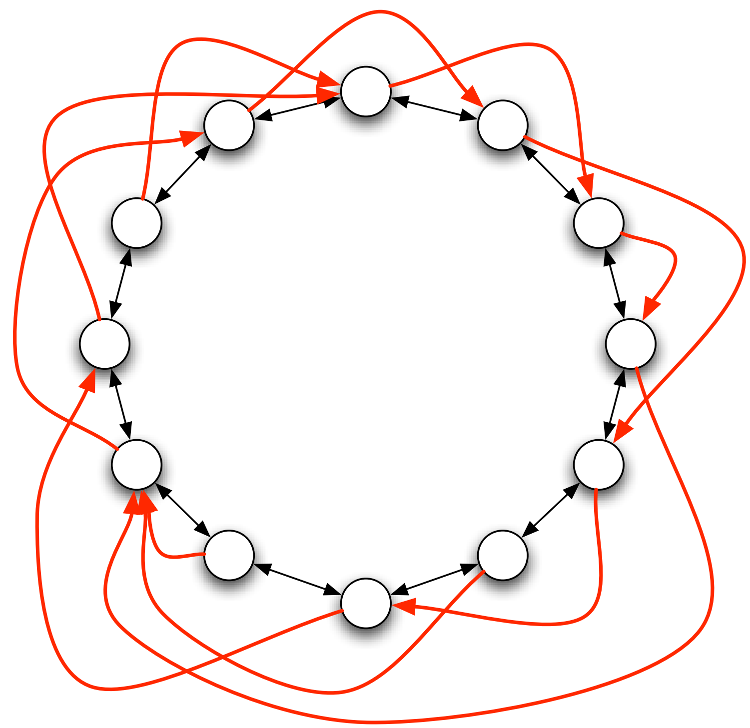
- Decentralized search with (essentially) arbitrary population density, when link probability proportional to $\mathrm{rank}^{-\beta}$.
- (LKNRT'05): Efficient routing when $\beta = 1$, i.e. $1/\mathrm{rank}$.
- Generalization of lattice result (diff. from set systems).

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

Open Question: Network Evolution

What causes a network to evolve toward searchability?

- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:

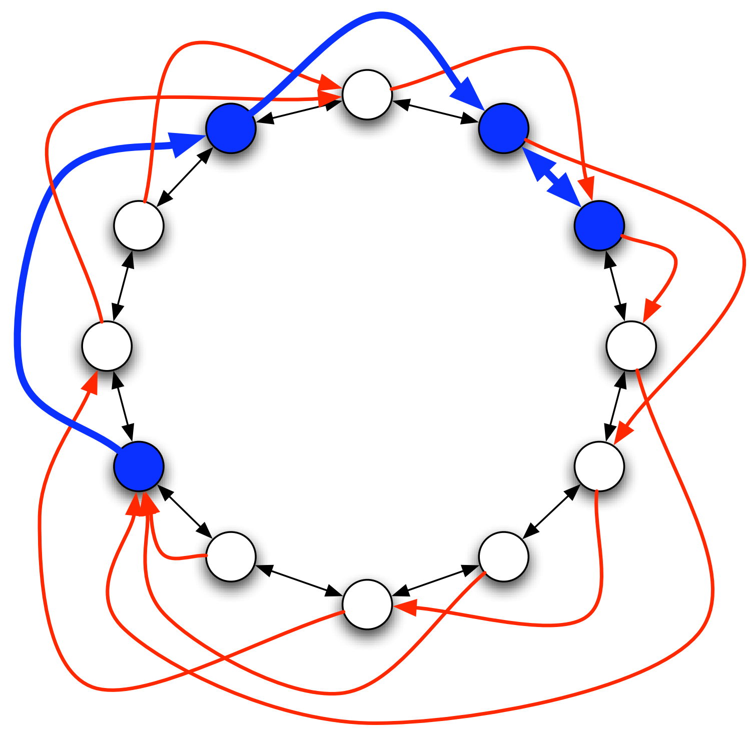


- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

Open Question: Network Evolution

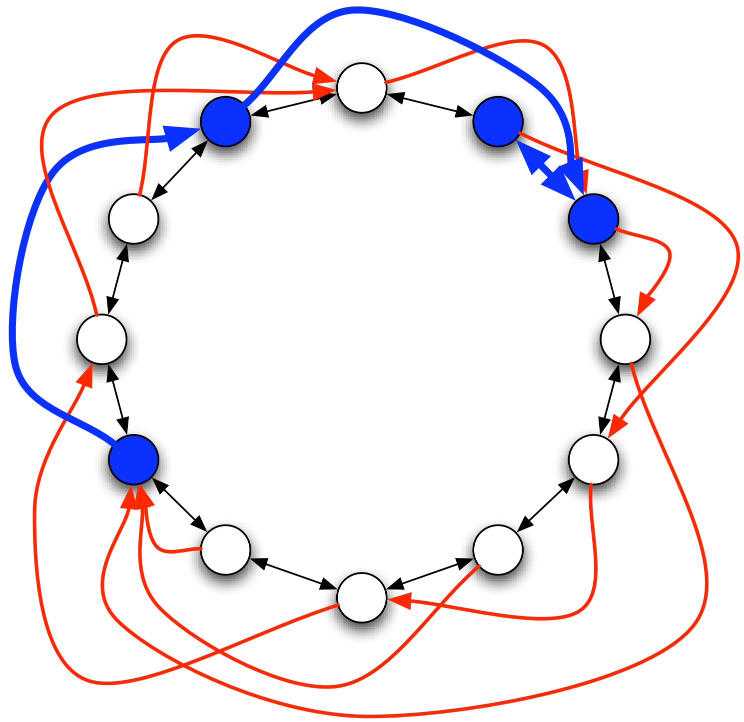
What causes a network to evolve toward searchability?

- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:



- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

Open Question: Network Evolution

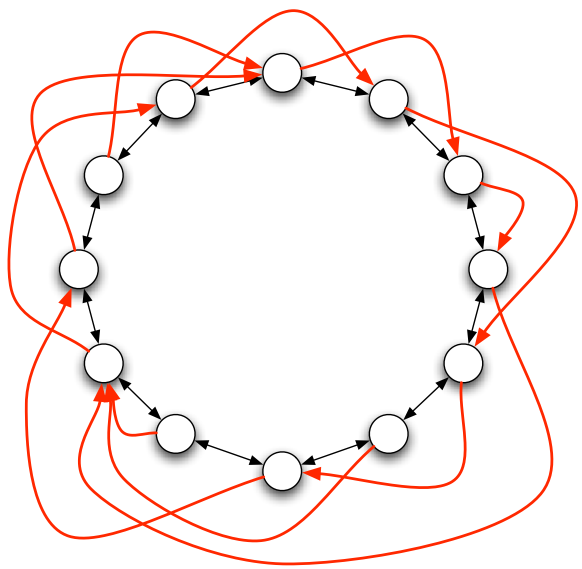


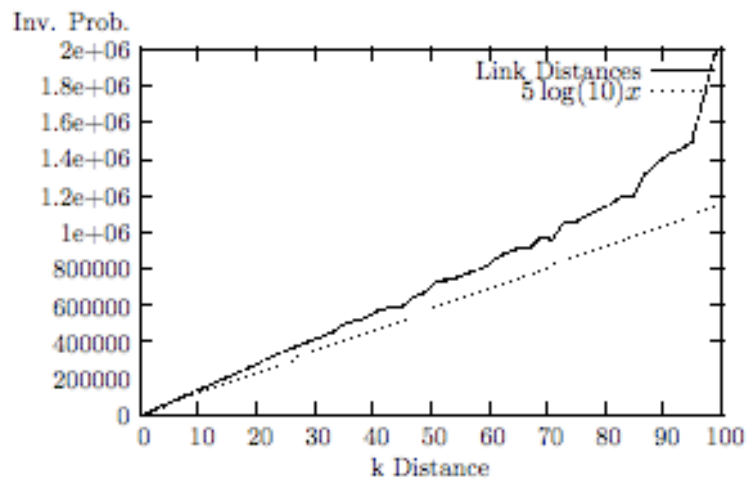
What causes a network to evolve toward searchability?

- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:

- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

Open Question: Network Evolution

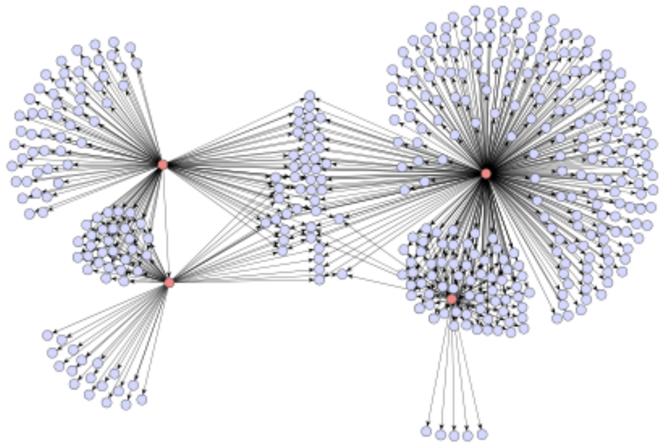


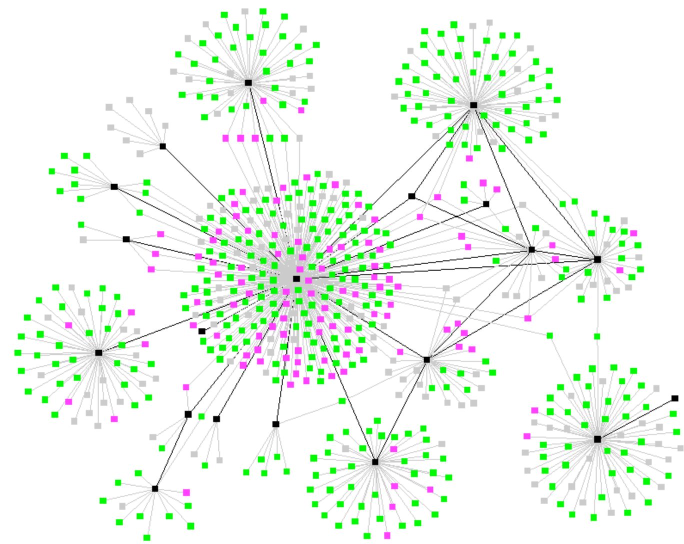


This defines a Markov chain on labeled graphs.
Conjecture [Sandberg-Clarke 2006]:

- At stationarity, distribution of distances spanned by long-range links is (close to) theoretical optimum for search.
- At stationarity, expected length of searches is polylogarithmic.
- Conjectures are supported by simulation.

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.

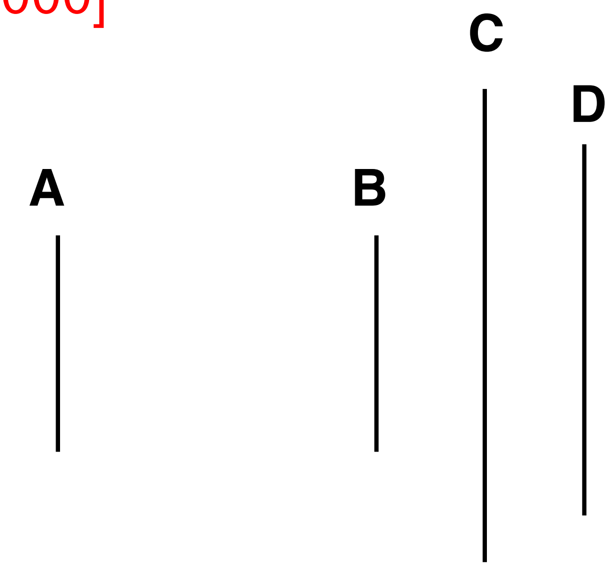
- News, opinions, rumors, fads, urban legends, ...
- Word-of-mouth effects in marketing, rise of new products.
- Changes in social priorities: smoking, recycling, ...
- Saturation news coverage; topic diffusion among bloggers.
- Localized collective action: riots, walkouts

Empirical Studies of Diffusion

Experimental and theoretical studies of diffusion have a long history in the social sciences

- Spread of new agricultural and medical practices [Coleman et al 1966]
- Media influence and two-stage flow [Lazarsfeld et al 1944]
- Modeling diffusion as a cascading sequence of strategy updates in a networked coordination game [Blume 1993, Ellison 1993, Young 1998, Morris 2000]

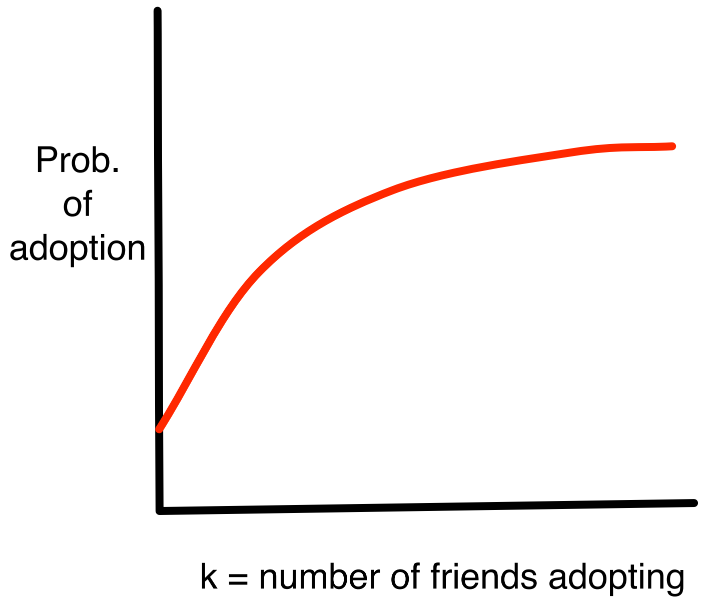
- Psychological effect of others' opinions. E.g.: Which line is closest in length to A? [Asch 1958]

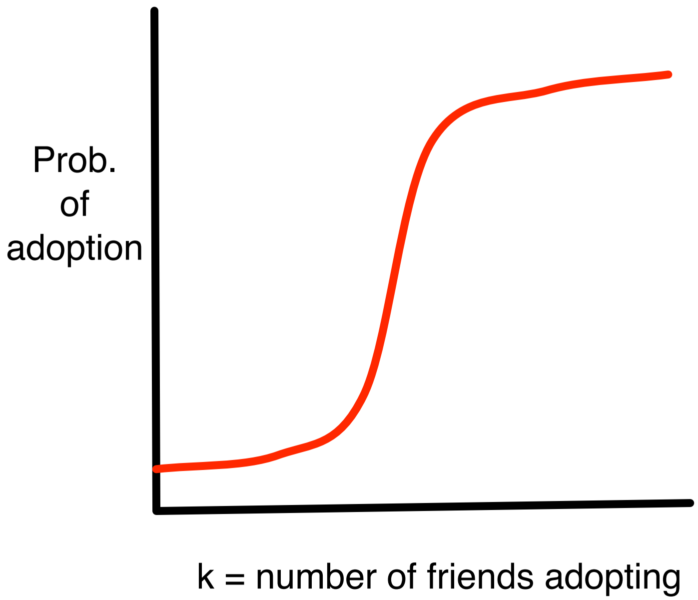


Diffusion Curves

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

- Bass 1969; Granovetter 1978; Schelling 1978





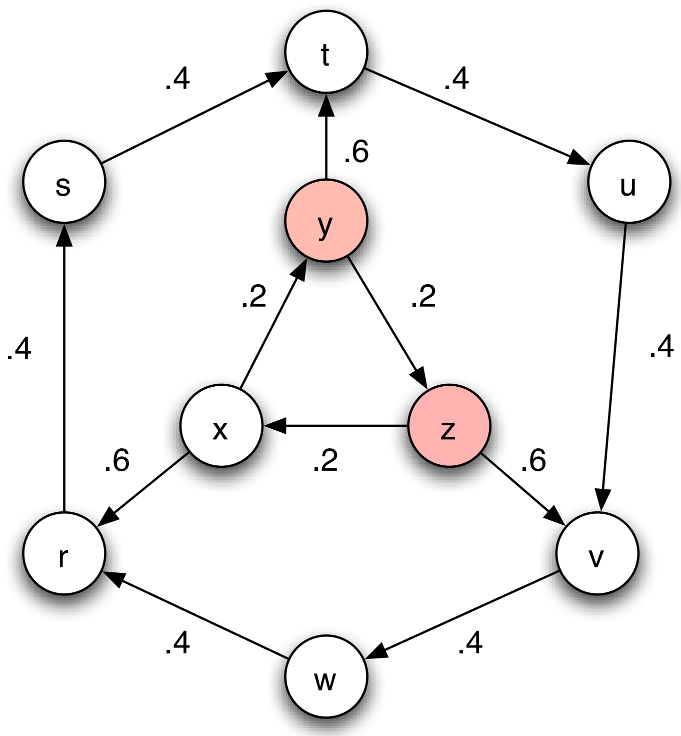
Build models for contact processes based on local behavior.

Key issue: qualitative shape of the diffusion curves.

- Diminishing returns? Critical mass?

A Simple Model: Independent Contagion

- Initially some nodes are active.
- Each edge (v, w) has probability p_{vw}.

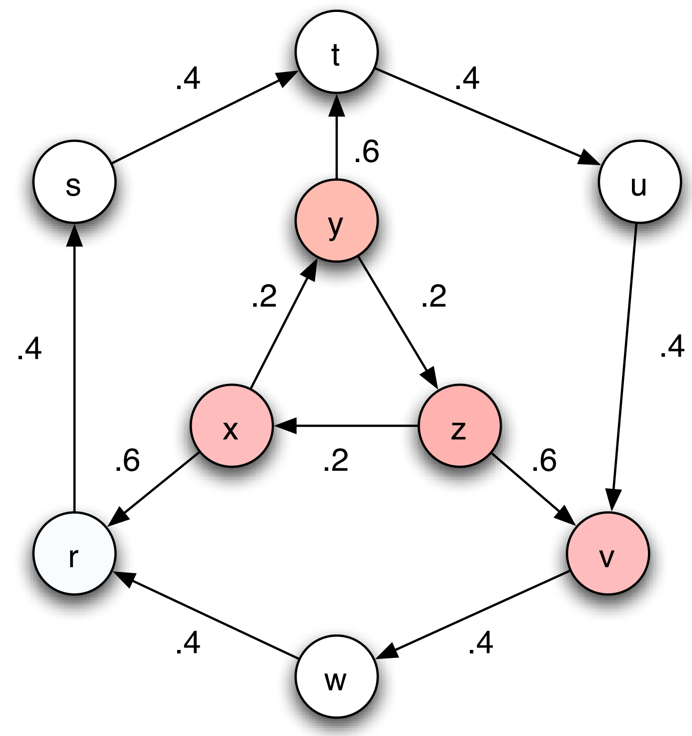


- v becomes active: chance to activate w with probab. p_{vw}.
- Activations spread through network.
- Let S = initial active set, $f(S)$ = exp. size of final active set.

Node don't "deactivate," though this is an easy modification.

A Simple Model: Independent Contagion

- Initially some nodes are active.
- Each edge (v, w) has probability p_{vw}.

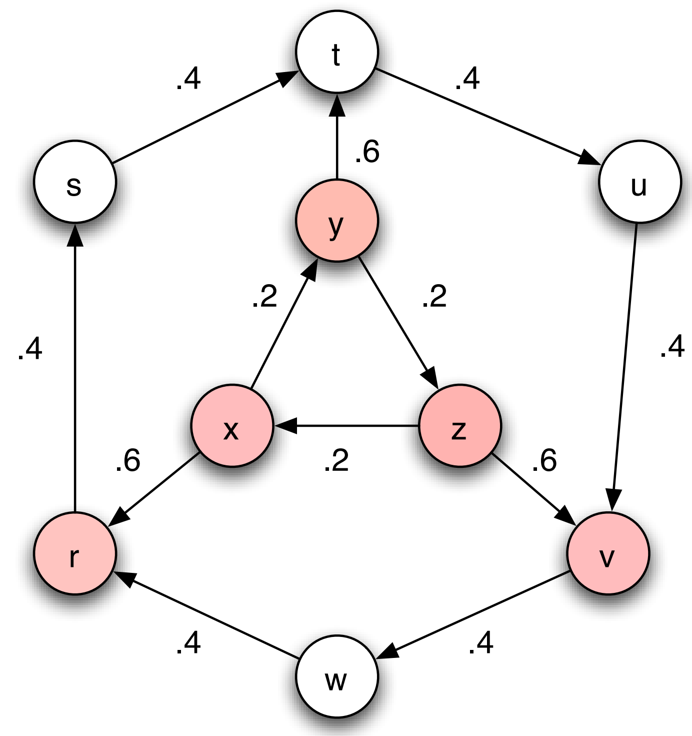


- v becomes active: chance to activate w with probab. p_{vw}.
- Activations spread through network.
- Let S = initial active set, $f(S)$ = exp. size of final active set.

Node don't "deactivate," though this is an easy modification.

A Simple Model: Independent Contagion

- Initially some nodes are active.
- Each edge (v, w) has probability p_{vw}.

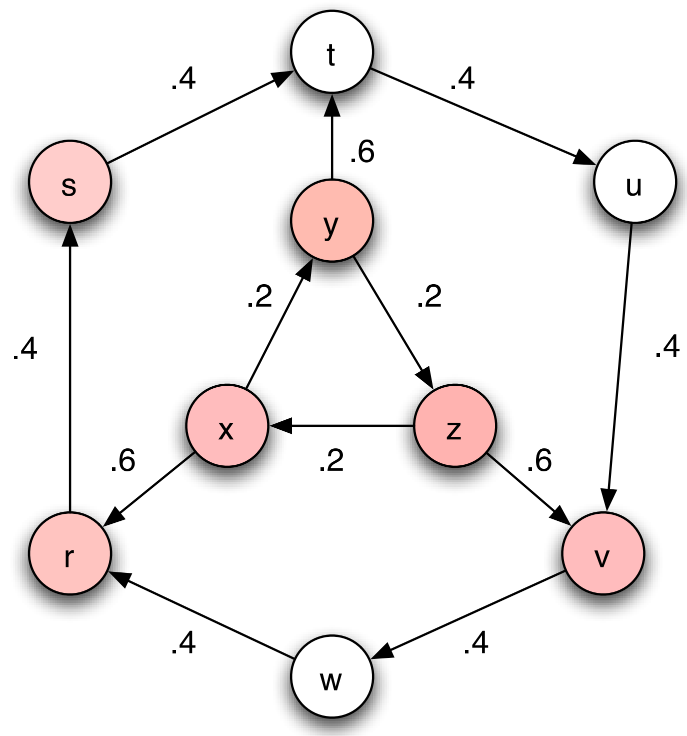


- v becomes active: chance to activate w with probab. p_{vw}.
- Activations spread through network.
- Let S = initial active set, $f(S)$ = exp. size of final active set.

Node don't "deactivate," though this is an easy modification.

A Simple Model: Independent Contagion

- Initially some nodes are active.
- Each edge (v, w) has probability p_{vw}.



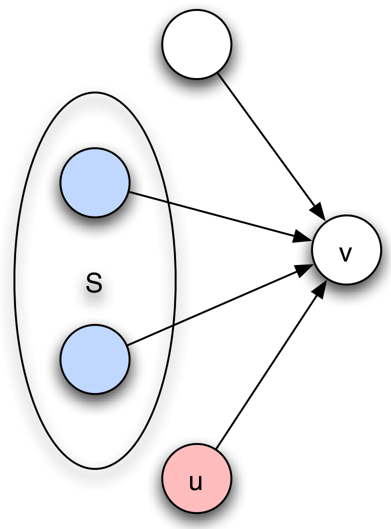
- v becomes active: chance to activate w with probab. p_{vw}.
- Activations spread through network.
- Let S = initial active set, $f(S)$ = exp. size of final active set.

Node don't "deactivate," though this is an easy modification.

A General Contagion Model

Kempe-Kleinberg-Tardos 2003,
Dodds-Watts 2004:

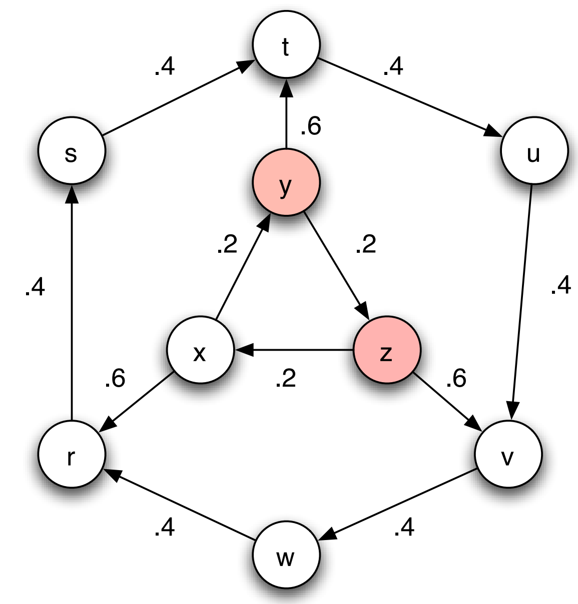
- When u tries to influence v: success based on set of nodes S that already tried and failed.
- Success functions $p_v(u, S)$.



- Independent contagion: $p_v(u, S) = p_{uv}$.
- Threshold: $p_v(u, S) = 1$ if $|S| = k$; else $p_v(u, S) = 0$.
- Diminishing returns: $p_v(u, S) \geq p_v(u, T)$ if $S \subseteq T$.

The Most Influential Subset

Most influential set of size k: the k nodes producing largest expected cascade size if activated. [Domingos-Richardson 2001]



As a discrete optimization problem:

$$\max_{S \text{ of size } k} f(S).$$

NP-hard and highly inapproximable.

- Inapproximability proof relies on critical mass.
- With diminishing returns: constant-factor approximation [Kempe-Kleinberg-Tardos 2005]

An Approximation Result

Diminishing returns: $p_v(u, S) \geq p_v(u, T)$ if $S \subseteq T$.

- Hill-climbing: repeatedly select maximum marginal gain.
- Performance guarantee: within $(1 - \frac{1}{e}) \sim 63\%$ of optimal [Kempe-Kleinberg-Tardos 2005].
- Analysis: diminishing returns at individual nodes implies diminishing returns at a "global" level.
  - Cascade size $f(S)$ grows slower and slower as S grows. f is submodular: if $S \subseteq T$ then

    $$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T).$$

  - Can then use results of Nemhauser-Wolsey-Fisher 1978 on approximate maximization of submodular functions.
- Open: For how general a model is $f(S)$ submodular, or at least well-approximable?

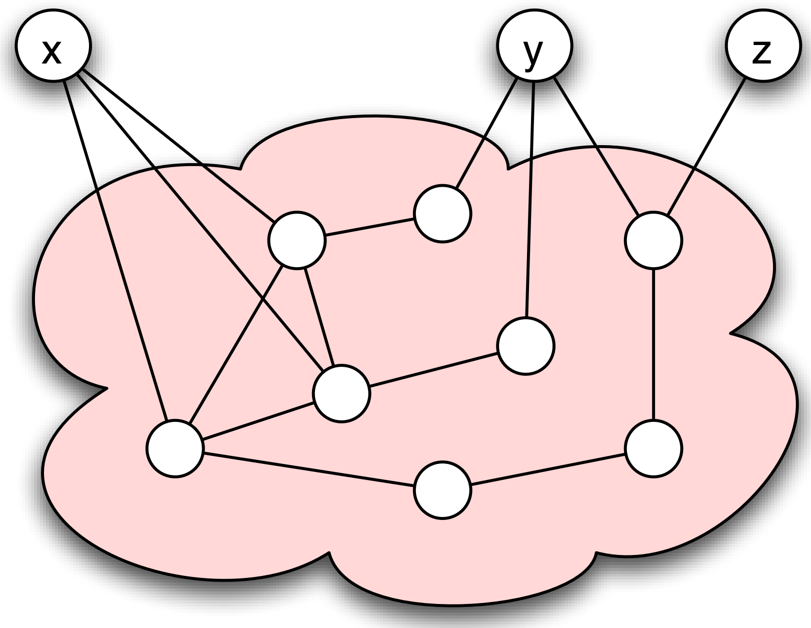
Empirical Analysis of Diffusion Curves

What do real diffusion curves look like?

- Challenge: large datasets where diffusion can be observed.
- Need social network links and behaviors that spread.

Backstrom-Huttenlocher-Kleinberg-Lan, 2006:

- Use social networks where people belong to explicitly defined groups.
- Each group defines a behavior that diffuses.
- Probability of joining, based on friends?



Networks with Explicit Groups

LiveJournal

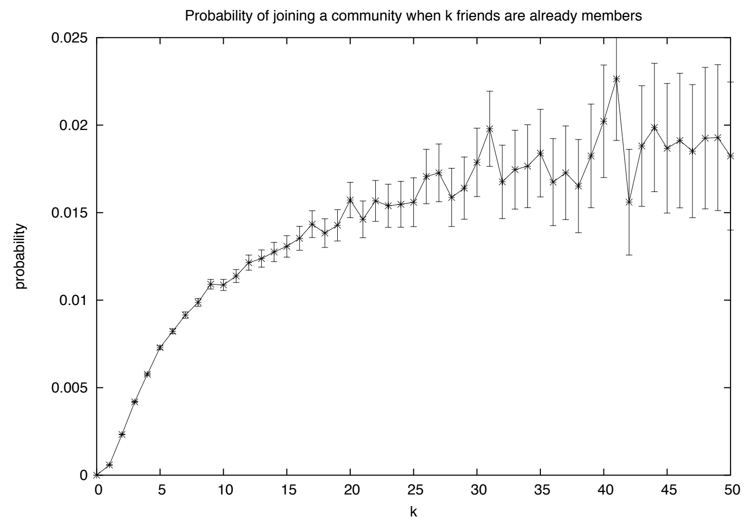
- On-line blogging community with friendship links and user-defined groups.
- Over a million users update content each month.
- Over 250,000 groups to join.

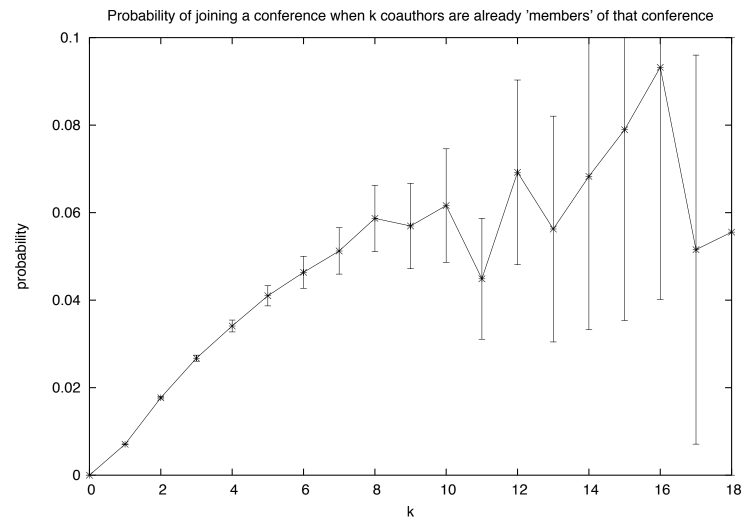
DBLP

- Database of CS papers: co-author links and conferences.
- 100,000 authors; 2000 conferences.
- You "join" a conference by publishing a paper there.

What do the diffusion curves look like in these two settings?

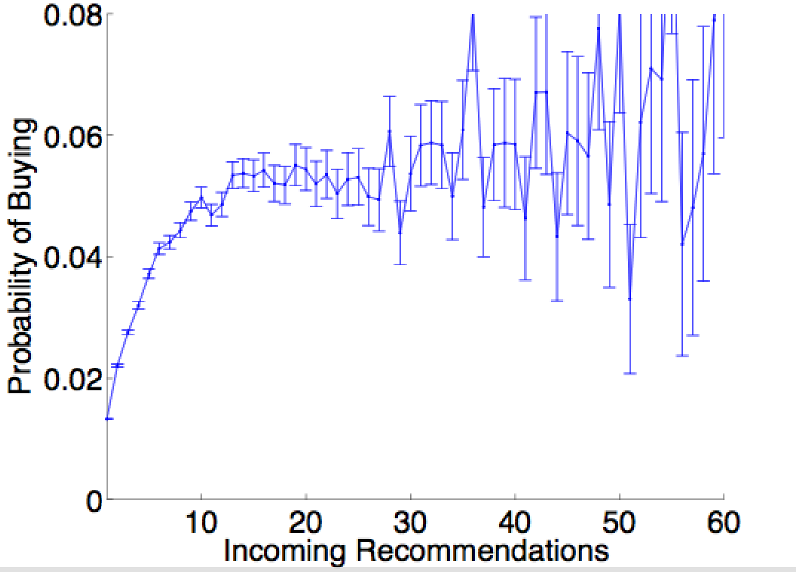
LiveJournal and DBLP Diffusion

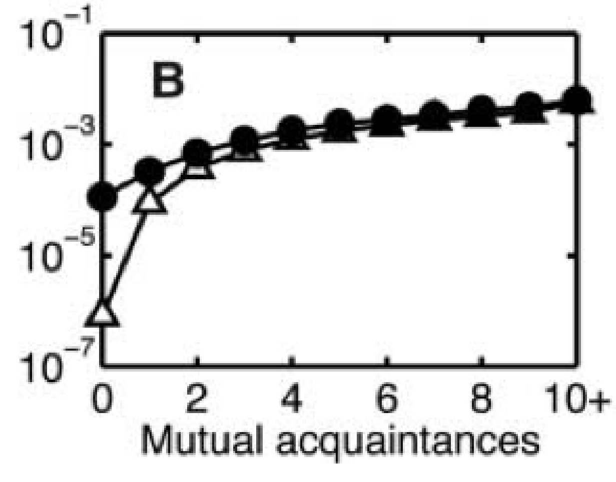




- Mainly diminishing returns.
- But both curves turn upward for $k = 0, 1, 2$.
- LiveJournal curve particularly smooth; fits $f(x) = \epsilon \log x$. Roughly half billion pairs (u, C) where user u is one step from community C.

Recommendation and Email Diffusion





Leskovec-Adamic-Huberman, 2006

- Recommendation program at large on-line retailer.
- Prob. of purchase as function of # of recommendations.

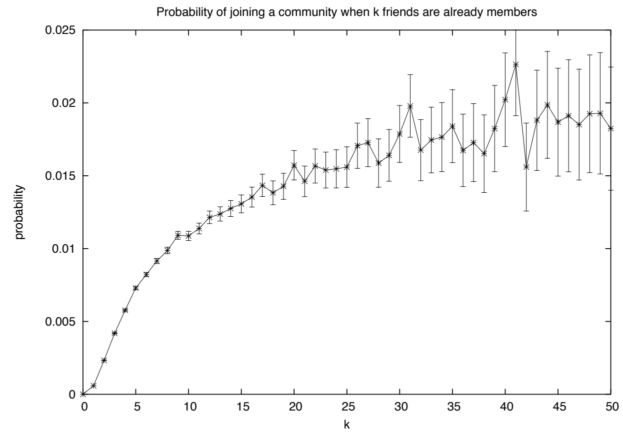
Kossinets-Watts, 2006

- Email network at large university.
- Prob. of link as function of # of shared acquaintances.

Caveats

What we're measuring (e.g. for LJ)

- Snapshot of everyone's state relative to each group at time t_1.
- Which of these groups had people joined at time $t_2 > t_1$?



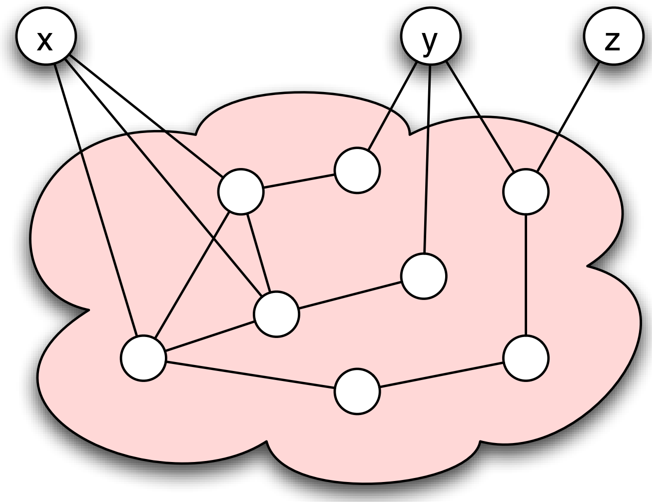
Challenge: Infer an operational model.

- At time t_1, we see the behavior of node v's friends.
- When did v become aware of their behavior?
  When did this translate into a decision by v to act?
  How long after this decision did v act?
- Much of the problem: modeling the asynchrony.

More subtle features

Dependence on number of friends:
a first step toward general prediction.

- Given network and v's position in it at t_1, estimate probability v will join a given group by t_2.
- Number of friends in community is only one of many possible features.



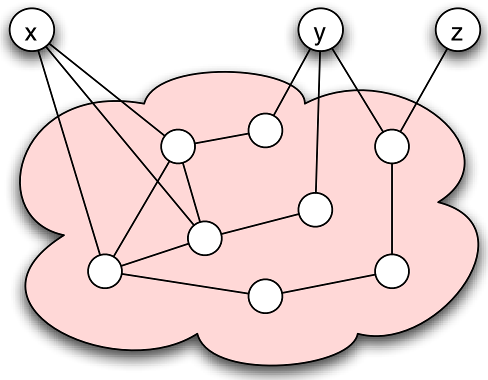
When formulated as a probability estimation problem, connectedness of friends emerges as a significant feature.

- x and y each have three friends in group.
- x's friends are all connected; y's friends are independent.
- Who is more likely to join?

Connectedness of friends

Competing sociological theories

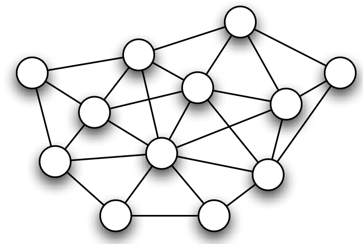
- Informational argument [Granovetter '73]
- Social capital argument [Coleman '88]

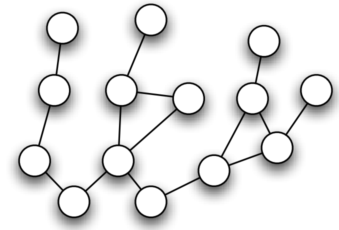


- Informational argument: unconnected friends give independent support.
- Social capital argument: safety/trust advantage in having friends who know each other.
- In LiveJournal, joining probability increases significantly with more connections among friends in group.

A Puzzle

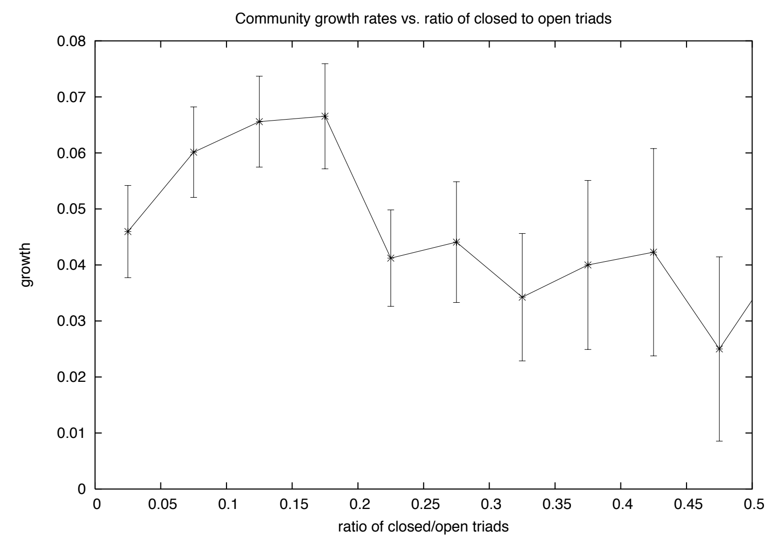
If connectedness among friends promotes joining, do highly "clustered" groups grow more quickly?





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

- Groups with large clustering grow slower.
- But not just because clustered groups had fewer nodes one step away.



Further Directions for Diffusion

- 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, ...
- 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]
- 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.

Recommendation Incentive Networks

Recall: recommendation incentive program at large on-line retailer [Leskovec-Adamic-Huberman'06, Leskovec-Singh-Kleinberg'06]

- With each purchase of a product, you can e-mail a recommendation of the product to friends.
- If one of them buys it, you both get a discount.

Theoretical models and analysis for such systems largely open.

- Adds a third component to word-of-mouth marketing models.
  - Direct advertising to full population
  - Targeted approach to influential nodes
  - Incentives to reduce "friction" on links between nodes.
- How to optimally trade off among (1), (2), and (3)?
  How does this depend on properties of the product/idea being marketed?
- How do different strategies affect the types of cascading behavior that result?

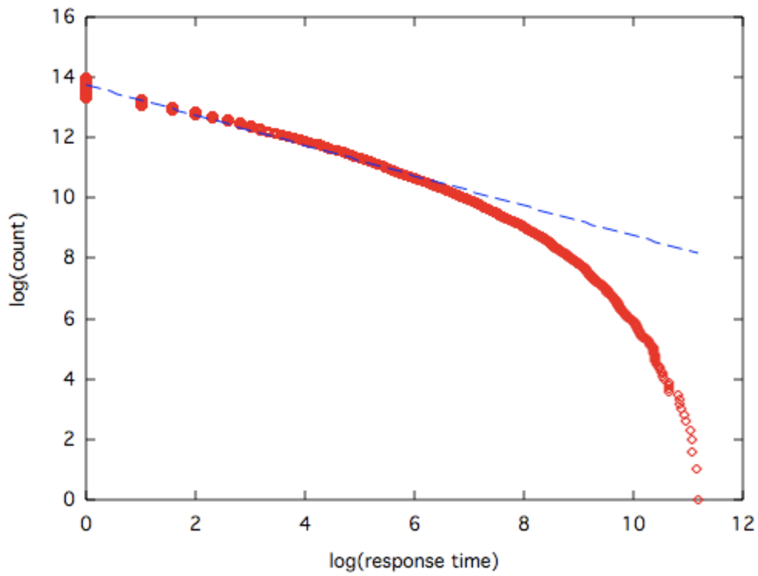
Final Reflections: Toward a Model of You

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 within t days (conditioned on answering at all)?

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



Final 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.
- Models based on algorithmic ideas will be crucial in understanding these developments.