Bursts, Cascades, and Hot Spots: A Glimpse of Some On-Line Social Phenomena at Global Scales

Jon Kleinberg

Cornell University

Including joint work with
Lars Backstrom, Larry Blume, David Crandall, David Easley, Dan Huttenlocher, Bobby Kleinberg, Ravi Kumar, Cameron Marlow, Éva Tardos, Johan Ugander.

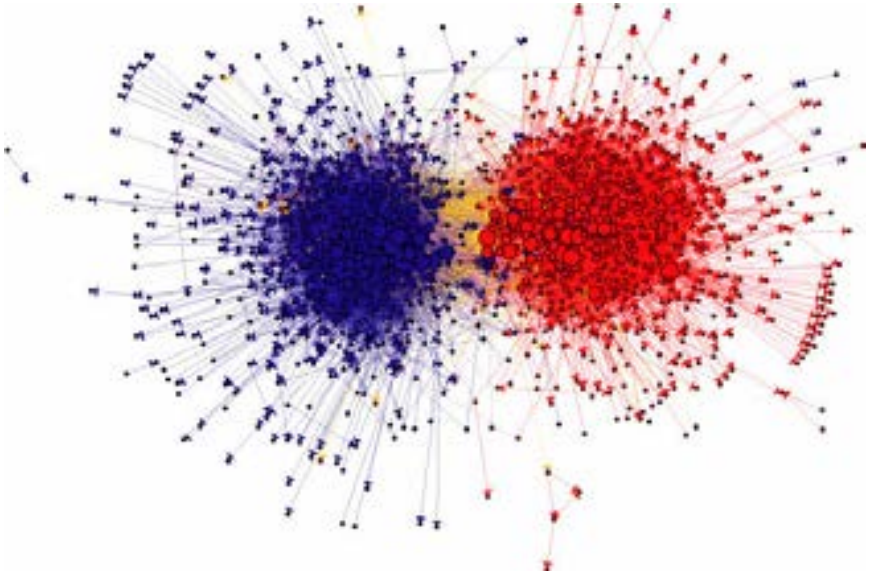
Networks as Phenomena

The emergence of 'cyberspace' and the World Wide Web is like the discovery of a new continent.

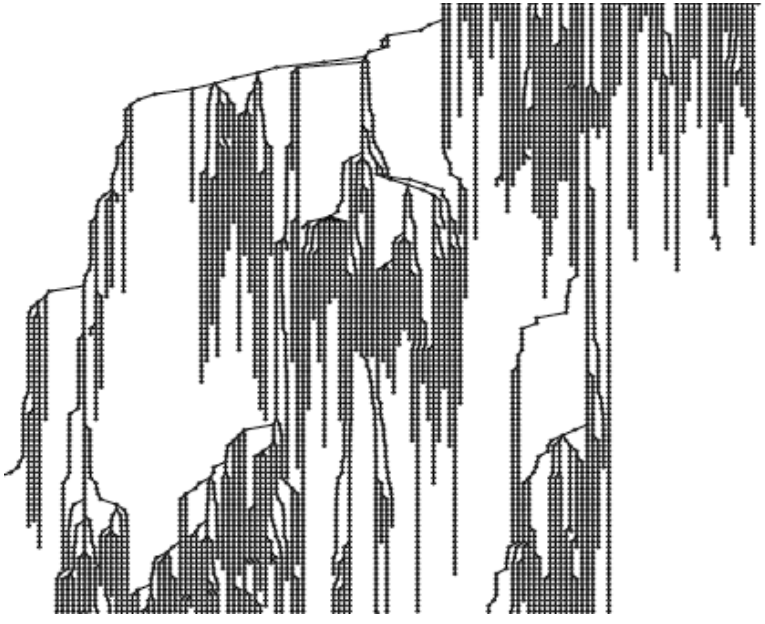
– Jim Gray,
1998 Turing Award address

- The on-line world as a phenomenon to be studied.
- The basic language for describing it is mathematical.

The Terrain of the On-Line World



Political blogs
(Adamic-Glance, 2005)



Anti-war chain letter (LibenNowell-Kleinberg 2008)

The terrain of the on-line world is geographic, social, and graph-theoretic.

- Combinatorial and probabilistic analyses of graphs form a central part of our understanding.

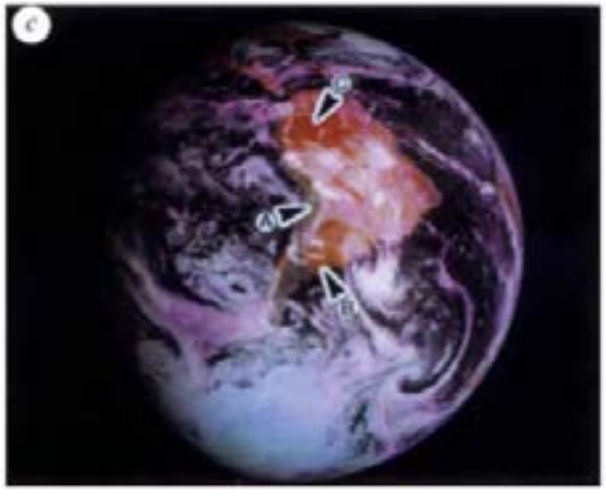
Is There Life on Earth?

A search for life on Earth from the Galileo spacecraft

Carl Sagan*, W. Reid Thompson*, Robert Carlson†, Donald Gurnett‡ & Charles Hord§

* Laboratory for Planetary Studies, Cornell University, Ithaca, New York 14853, USA
† Atmospheric and Cometary Sciences Section, Jet Propulsion Laboratory, Pasadena, California 91109, USA
‡ Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52242-1479, USA
§ Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado 80309, USA

In its December 1990 fly-by of Earth, the Galileo spacecraft found evidence of abundant gaseous oxygen, a widely distributed surface pigment with a sharp absorption edge in the red part of the visible spectrum, and atmospheric methane in extreme thermodynamic disequilibrium; together, these are strongly suggestive of life on Earth. Moreover, the presence of narrow-band, pulsed, amplitude-modulated radio transmission seems uniquely attributable to intelligence. These observations constitute a control experiment for the search for extraterrestrial life by modern interplanetary spacecraft.



A Portion of the Earth, as Seen from Flickr

Organize Traces of Human Behavior Around Hot Spots

Organize activities around "hot-spots" in space and time.

Use geo-tagged data from millions of people,
via photos, search engine queries, mobile devices

- [Backstrom-Kleinberg-Kumar-Novak 2008, Kennedy-Naaman 2008, Crandall-Backstrom-Huttenlocher-Kleinberg 2009]

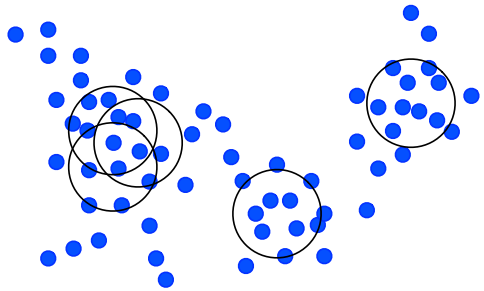
Hot-spot analysis

- Where are the hot-spots?
- How "intense" are they?
- What's distinctive in them?

Backstrom et al 2008, Kennedy-Naaman 2008, Crandall et al 2009

How Do We Find and Describe Hot-Spots?

Start with a local-search heuristic to find hot-spots.



Identifying Distinctive Features of a Hot-Spot.

- First: textual tags.
- Significance of a tag t at a hot-spot based on Bayes' Rule.
- Roughly: probability of seeing this density of photos w/ tag t, if tag were generated from world's background distribution?
- Next: identify distinctive photos at a location
  [Snavely-Seitz-Szeliski 2006, Kennedy-Naaman 2008]

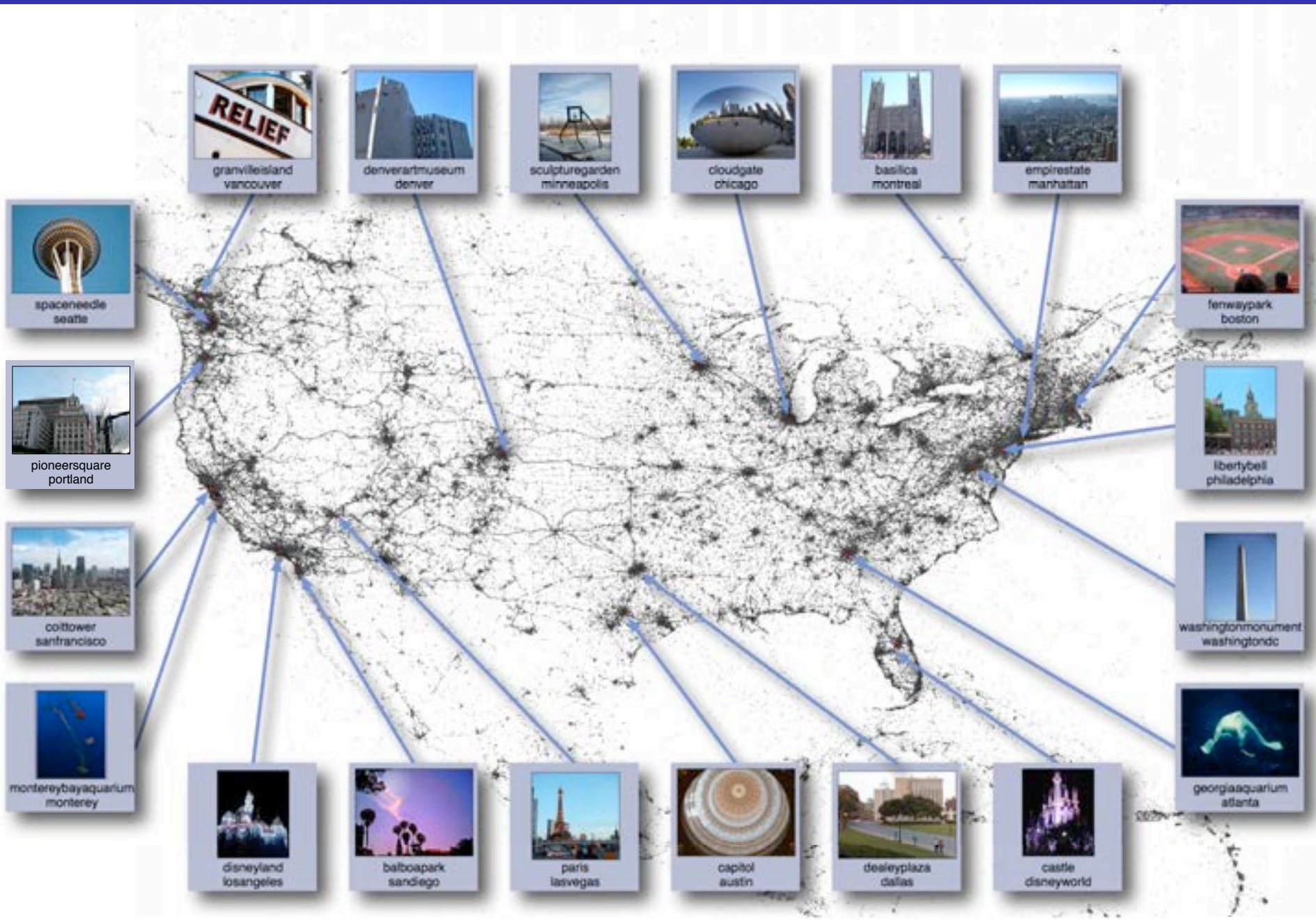
Try this for the global Flickr dataset, using two scales:

- 100 km: metropolitan scale
- 100 m: landmark scale

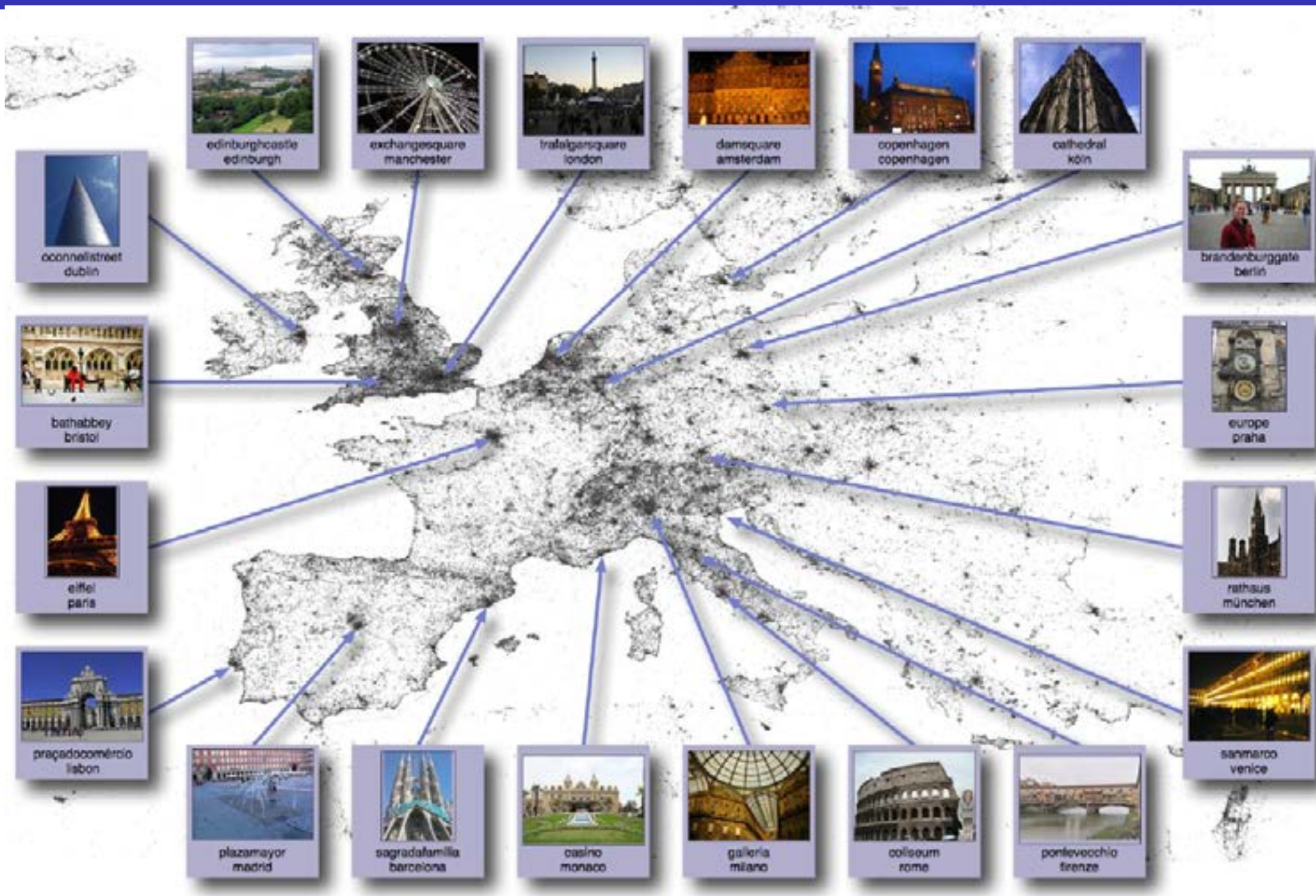
	Top landmark	2nd landmark	3rd landmark	4th landmark
1. **manhattan**	empirestate	timessquare	grandcentral	applestore
2. **london**	trafalgarsquare	tatemodern	eye	bigben
3. **sanfrancisco**	coittower	sealions	unionsquare	lombardstreet
4. **losangeles**	disneyland	hollywood	gettycenter	disneyhall
5. **paris**	eiffel	cathedral	sacrecoeur	pyramid
6. **washingtondc**	lincolnmemorial	monument	wwiimemorial	capitol
7. **chicago**	cloudgate	michiganavenue	gehry	artinstitute
8. **seattle**	spaceneedle	market	emp	library
9. **boston**	fenwaypark	trinitychurch	faneuilhall	publicgarden
10. **sandiego**	balboapark	sandiegozoo	seals	ussmidway
11. **amsterdam**	dam	annefrank	nieuwmarkt	museumplein
12. **rome**	colosseum	sanpietro	pantheon	fontanaditrevi
13. **barcelona**	sagradafamilia	parcguell	boqueria	casamil
14. **berlin**	brandenburggate	reichstag	potsdamerplatz	holocaustmemorial
15. **monterey**	montereybay	downtown	canneryrow	boardwalk
16. **lasvegas**	paris	bellagio	mgm	hooverdam
17. **toronto**	cntower	phillipssquare	dundassquare	rom
18. **vancouver**	granvilleisland	artgallery	aquarium	downtown
19. **firenze**	cathedral	pontevecchio	firenze	piazzadelcampo
20. **philadelphia**	libertybell	artmuseum	cityhall	jfkplaza
83. **stlouis**	gatewayarch	buschstadium	oldcourthouse	citymuseum

The Earth, as Seen from Flickr

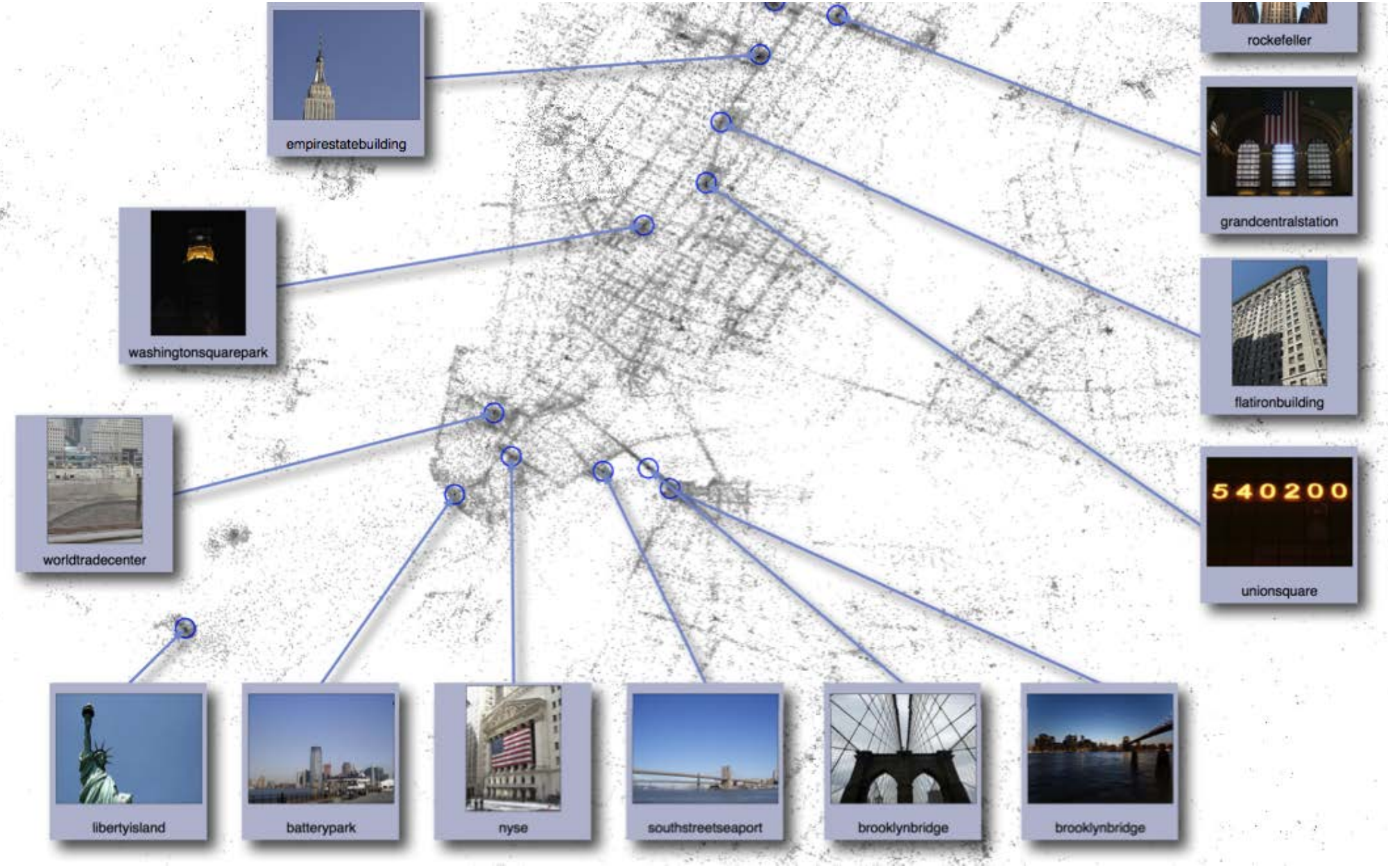
U.S. and Canada (Crandall et al 2009)



Europe


edinburghcastle
edinburgh
exchangesquare
manchester
trafalgarsquare
london
damsquare
amsterdam
copenhagen
copenhagen
cathedral
köln
oconnellstreet
dublin
brandenburggate
berlin
bathabbey
bristol
europe
praha
eiffel
paris
rathaus
münchen
praçadocomércio
lisbon
sanmarco
venice
plazamayor
madrid
sagradafamilia
barcelona
casino
monaco
galleria
milano
colloseum
roma
pontevecchio
firenze

Lower Manhattan

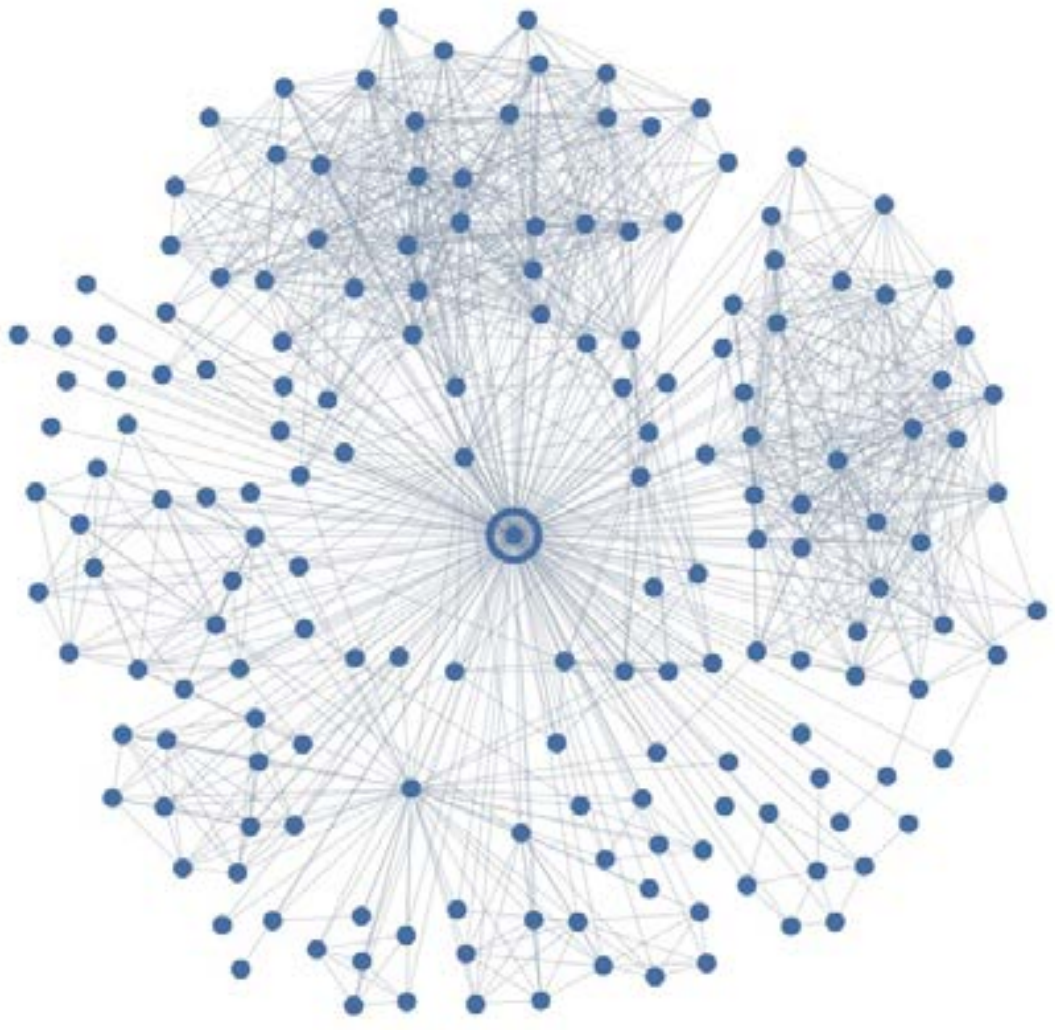


Bringing Network Structure Into the Picture

How do we discover the fine structure of a billion-node network?

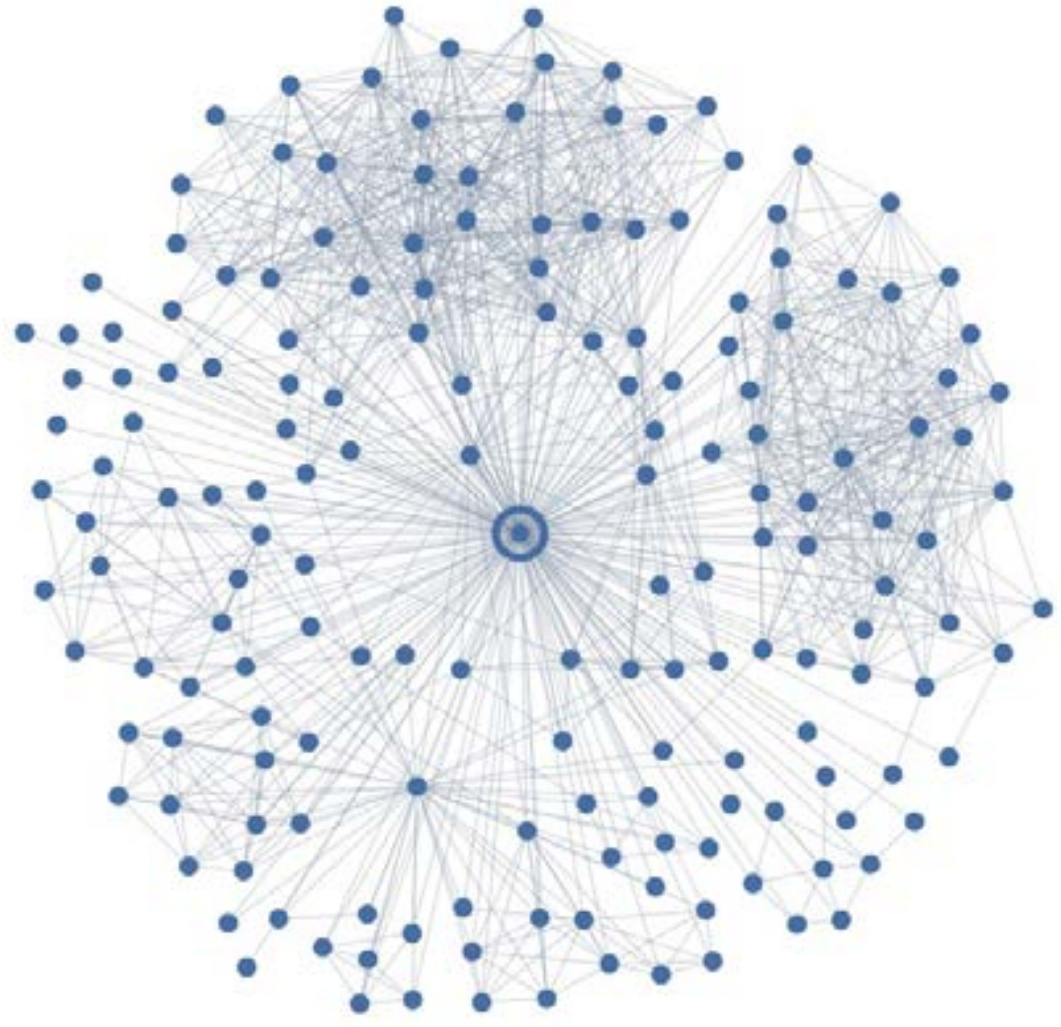
The strange geography of our collective social experience ...

Network structure via neighborhoods



Start with network neighborhoods

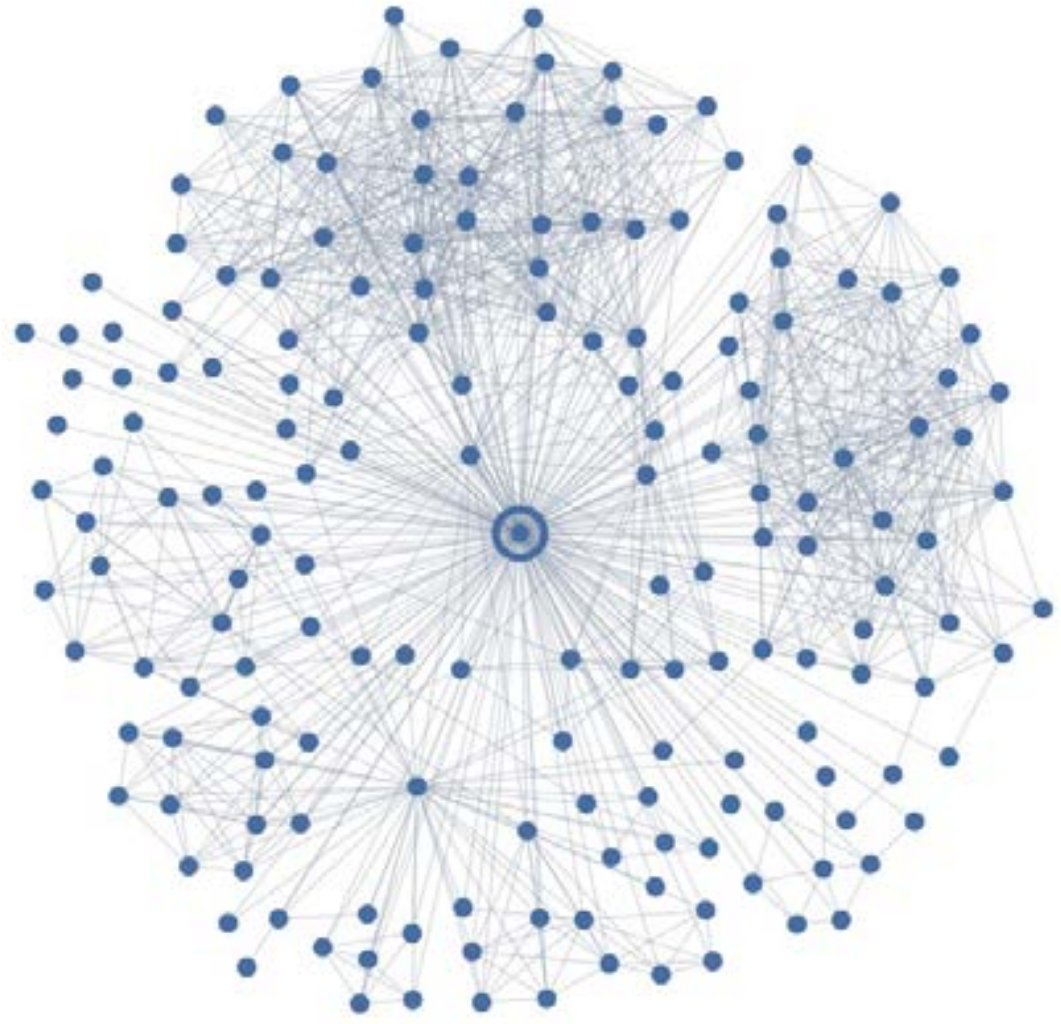
Network structure via neighborhoods



Start with network neighborhoods

- Think of Facebook not as a billion-node network, but instead as a collection of a billion (relatively dense) small networks.

Network structure via neighborhoods



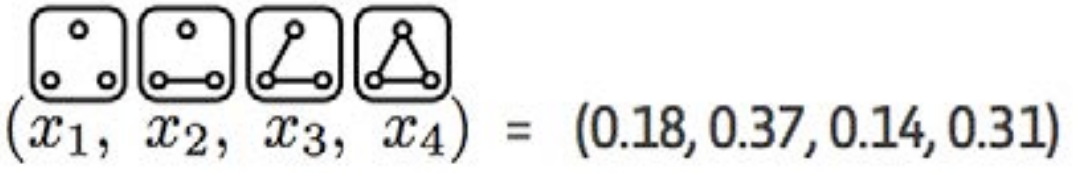
Start with network neighborhoods

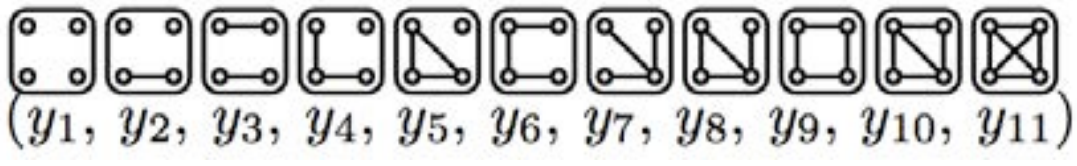
- Describe neighborhood G by vector of subgraph frequencies: For small k, and each k-node graph H, let $f_G(H) =$ frac. of k-node sets inducing H.

Characterizing neighborhoods

$f_G(H)$ = frac. of k-node sets in G that induce H.

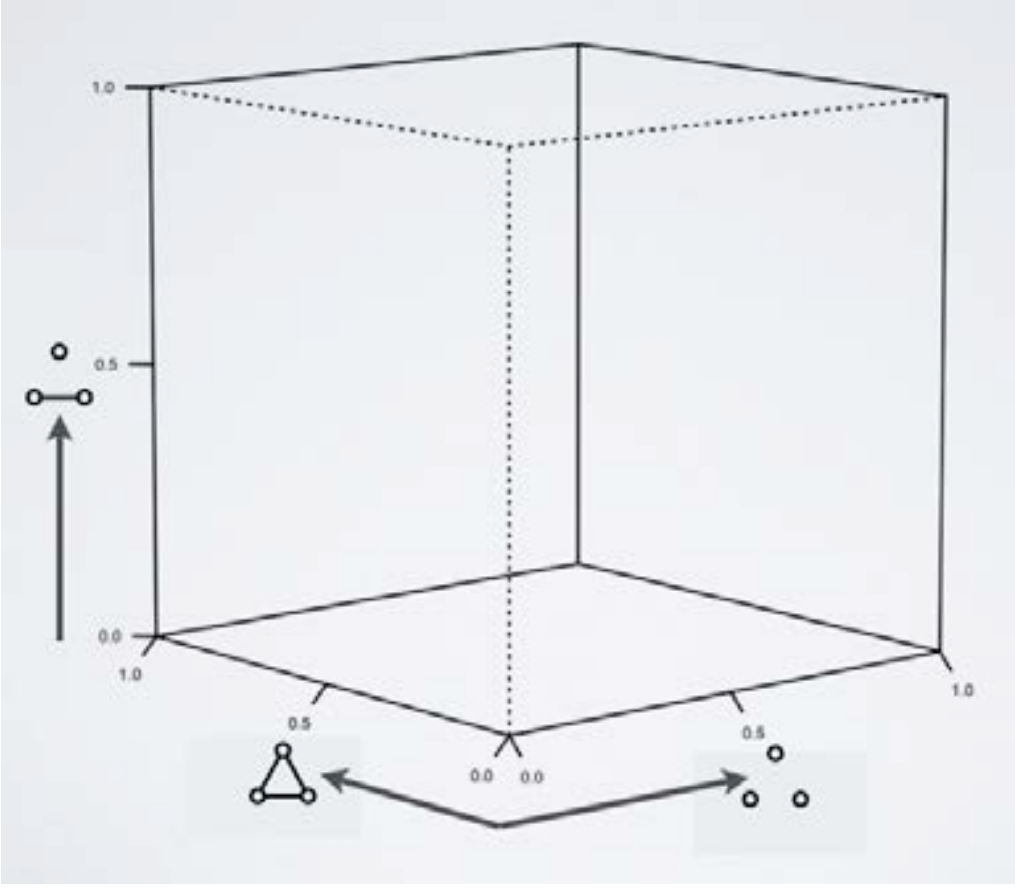
- Triad census: Davis-Leinhardt 71
- Network motifs: Milo et al 02
- Frequent subgraph mining: Yan-Han 02, Kuramochi-Karypis 04
- Subgraph homomorphism density: Borgs et al 06
- Characterizing neighborhoods: Ugander et al 13





The geography of Facebook neighborhoods

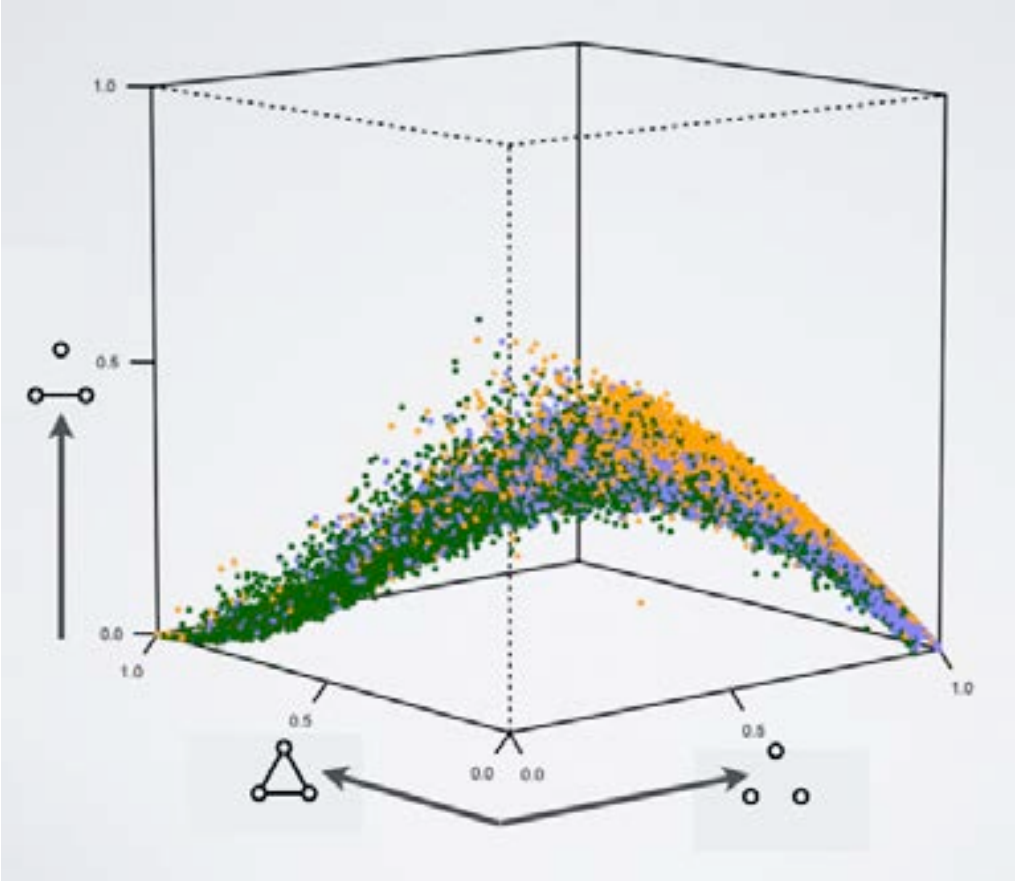
Axes: triad frequencies



The geography of Facebook neighborhoods

Axes: triad frequencies

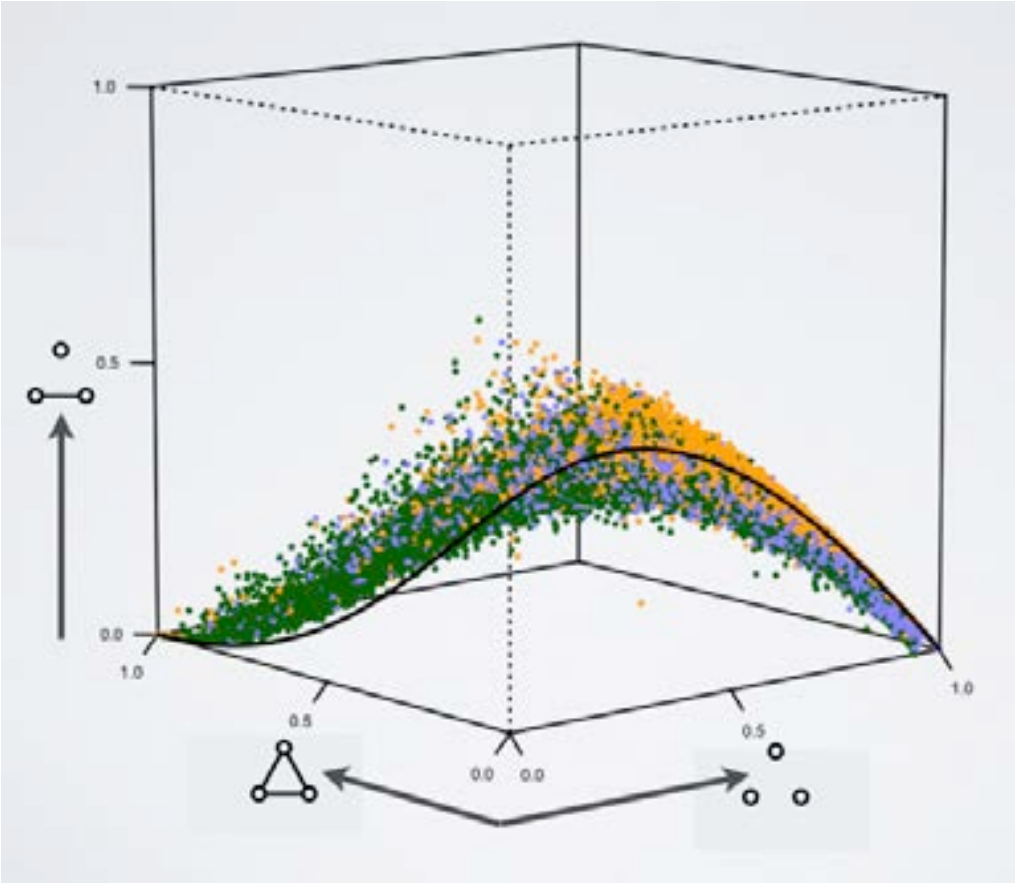
- "Coastlines:" freq of 1-edge triad is $\leq 3/4$.
- Unpopulated areas: freq of 2-edge triad never close to 3/4 in real life.
- Full feasible region would imply [Razborov 2007].



The geography of Facebook neighborhoods

Axes: triad frequencies

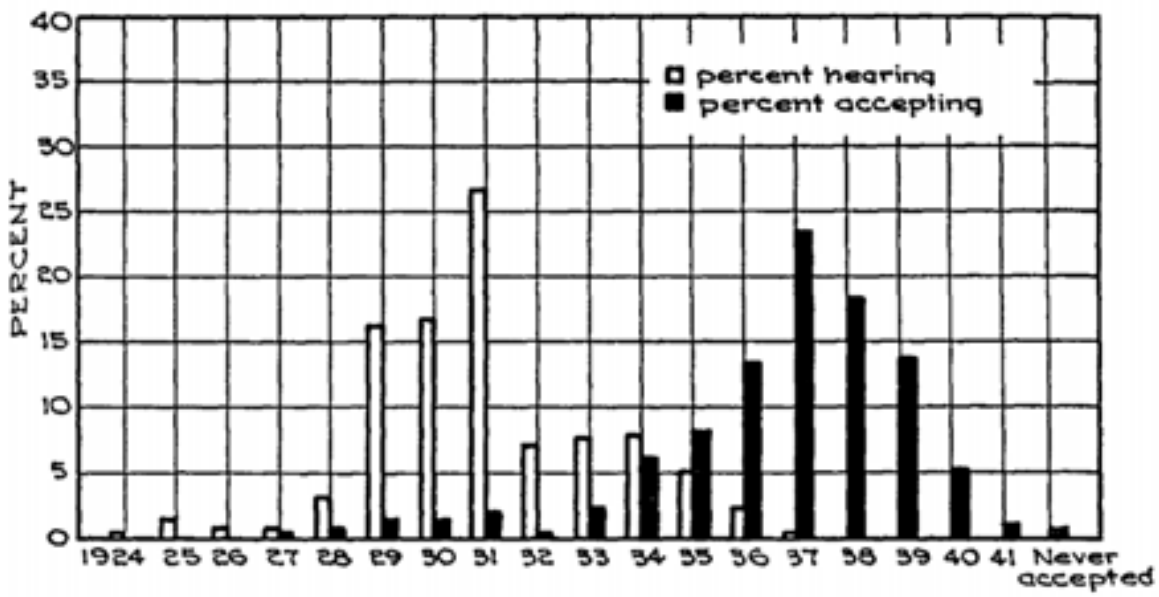
- "Coastlines:" freq of 1-edge triad is $\leq 3/4$.
- Unpopulated areas: freq of 2-edge triad never close to 3/4 in real life.
- Full feasible region would imply [Razborov 2007].



$G_{n,p}$ is the "river" that runs through the points.

- With deviations based on triadic closure and clustering.

Diffusion and Contagion



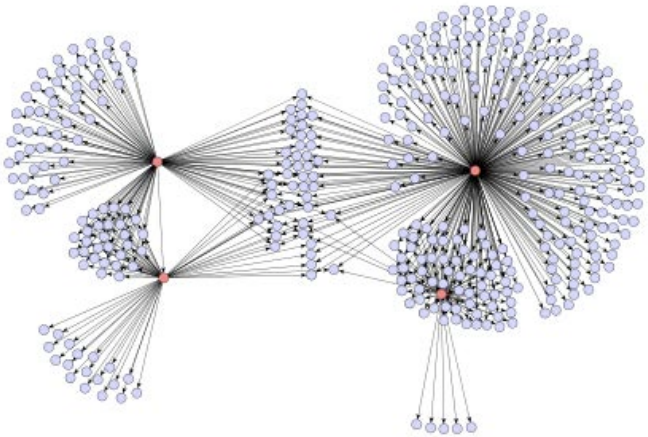
A basic "transport mechanism" for these systems:

- Cascading behavior through a network

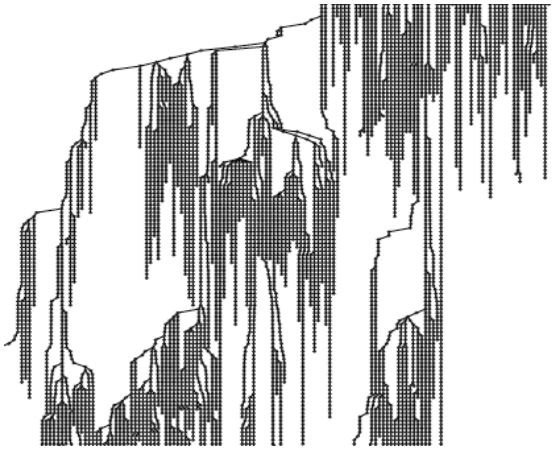
Long history of research in social contagion:

- Agricultural, medical innovations [Ryan-Gross 1943, Coleman et al 1966]
- Media influence and two-stage flow [Lazarsfeld et al 1944]
- "Virality" of news, rumors, marketing strategies, political messages, ...
- Cascading failures in engineered and financial systems.

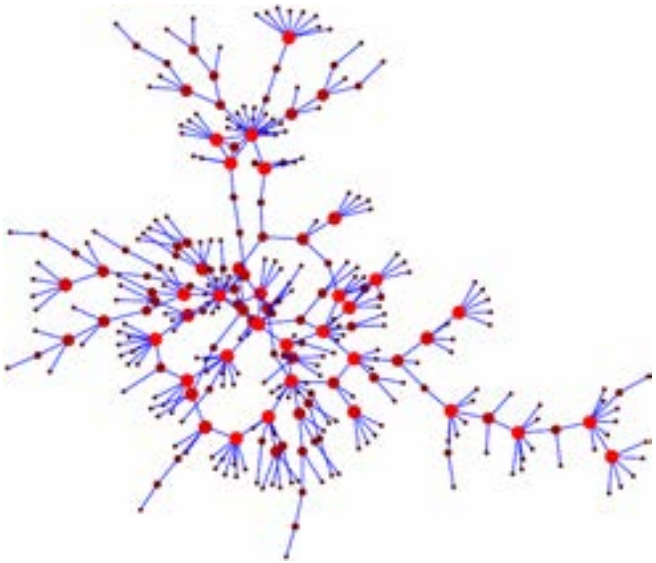
Contagion in Networks



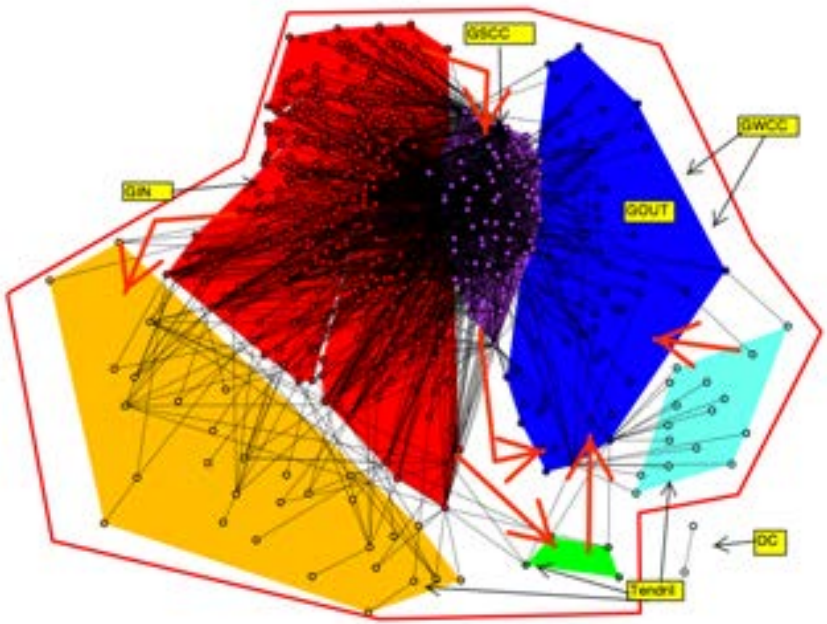
Book recommendations (Leskovec et al 2006)



Anti-war chain letter (LibenNowell-Kleinberg 2008)



Networks of sexual contacts (Potterat et al 02)

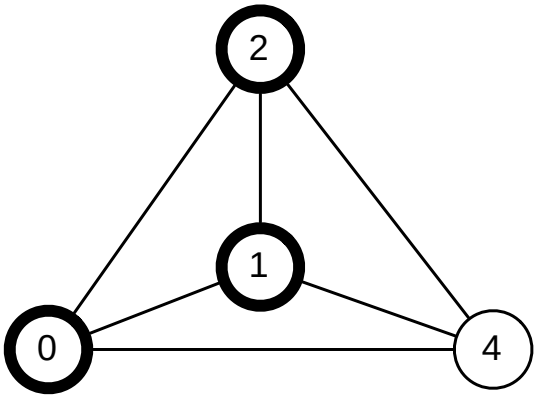


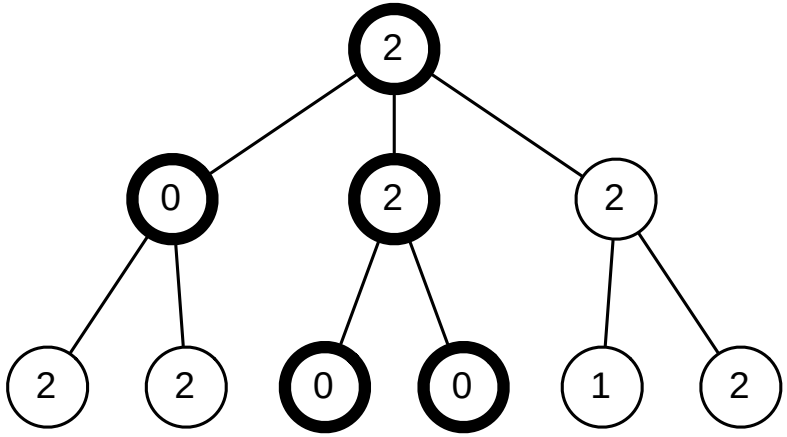
Overnight loans among financial institutions (Bech and Atalay, 2008)

Threshold Contagion

Each node v chooses a threshold $f(v)$ at the start, from a distribution μ over $\{0, 1, 2, \ldots, d+1\}$.

- v will be affected as soon as it has $f(v)$ affected neighbors.

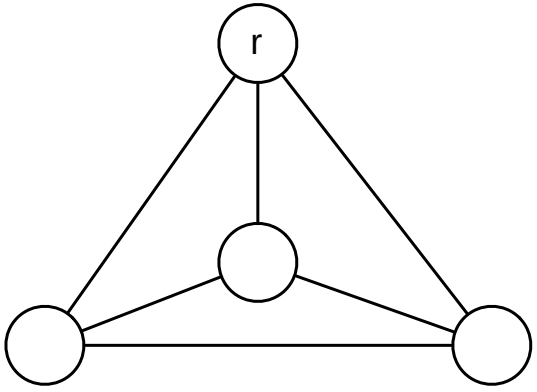


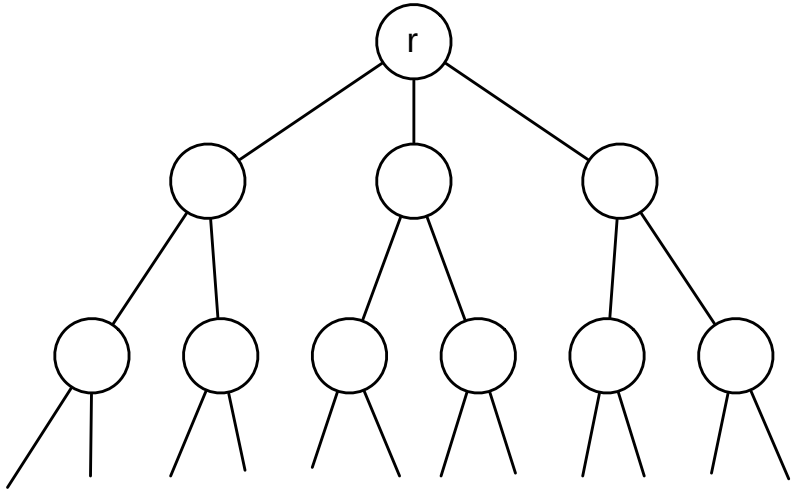


Despite simple formulation, a challenging model to analyze.

- Special-case results for diminishing thresholds ($\mu(1) \geq \mu(2) \geq \cdots$) [Kempe-Kleinberg-Tardos 03, Mossel-Roch 07].
- Special-case results when graph G is a tree [Dodds-Watts 04], lattice [Cox-Durrett 91], clique [Granovetter 78, Schelling 78], or small-world network [Ebrahimi-Ghasemiesfeh-Gao 13].
- Analysis for arbitrary graphs [Blume-Easley-Kleinberg-Kleinberg-Tardos 11].

Cascading Failures with Thresholds



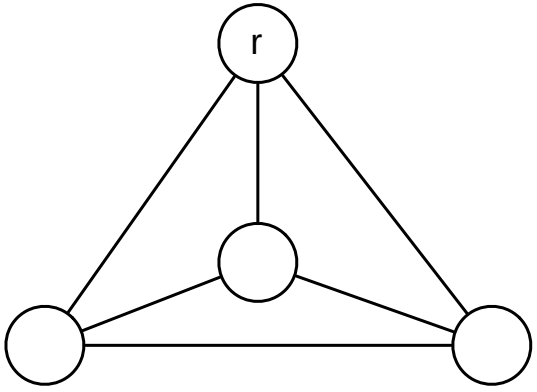


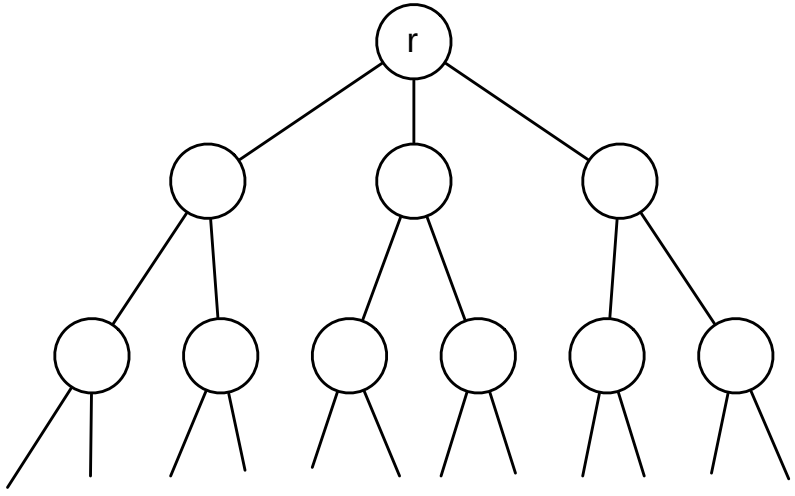
Which networks are least susceptible to cascading failures?

- Take edge density out of consideration: set of all graphs where each node has d neighbors.
- Choose μ from set of all distributions on $\{0, 1, 2, \ldots, d+1\}$.
- Risk of G = maximum failure probability of any node in G when thresholds are drawn from μ.

Given μ, which graphs have the lowest risk?

Testing One's Intuition about Failure-Resilience





Intuition from epidemiology:

- Dangerous to belong to a large connected component: the clique K_{d+1} is a resilient graph.

Intuition from financial markets:

- Want diversity among neighbors, uncorrelated shocks: the tree T_d is a resilient graph.

These two forms of intuition are in direct opposition to each other.

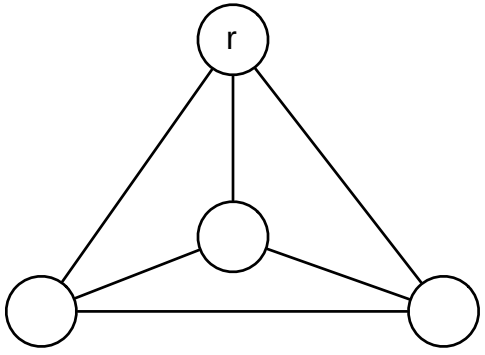
Cliques vs. Trees

To get further insight into the model:
Let's test these intuitions on distributions of the form

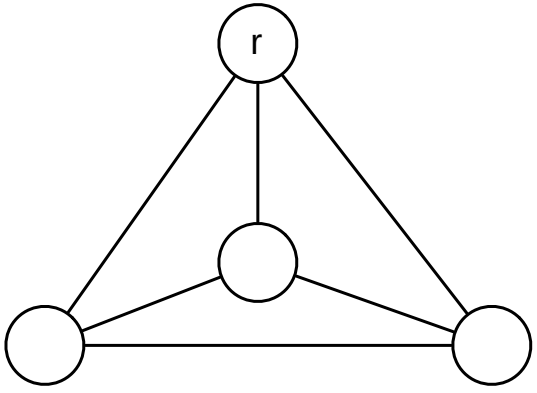
$$(\mu(0), \mu(1), \mu(2)) = (s, t, 1 - s - t)$$

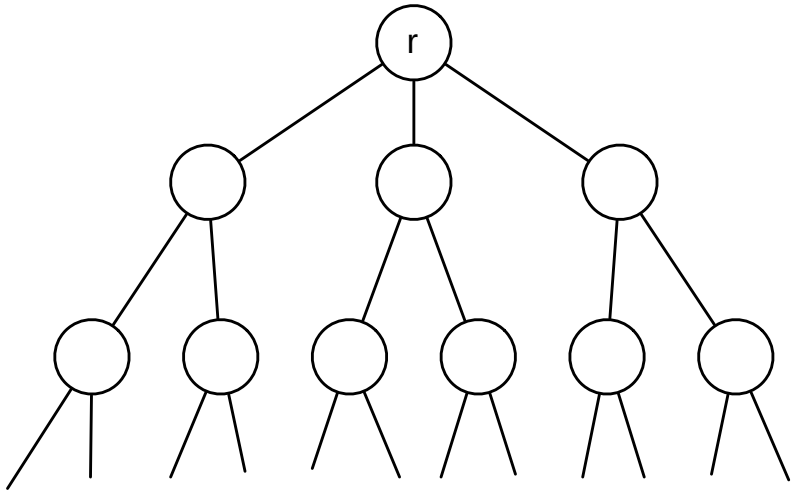
where s is fixed and small, and t varies. (All thresholds above 2 have prob. 0.)

- With threshold distribution $(s, 1 - s, 0)$, r's failure prob. is monontonic in the size of its component.
- The clique is uniquely optimal.



Cliques vs. Trees



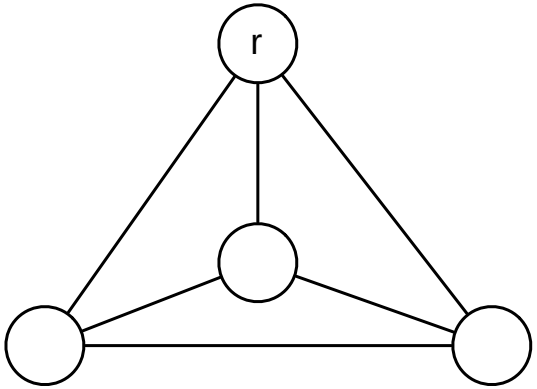


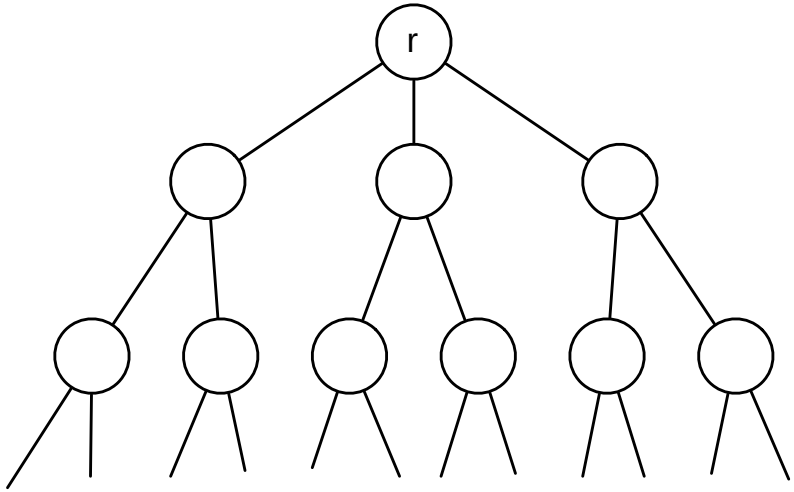
Recall $(\mu(0), \mu(1), \mu(2)) = (s, t, 1 - s - t)$.

A first result:

- There exist s, t (both small, with t larger than s) so that the tree T_d has lower risk than the clique K_{d+1}.
- Qualitative point: very different kinds of graphs are safer against different kinds of contagion processes.

Cliques vs. Trees





Tree:

$$s + \binom{d}{2} s^2 + dst + \text{terms of degree} \geq 3 \cdots$$

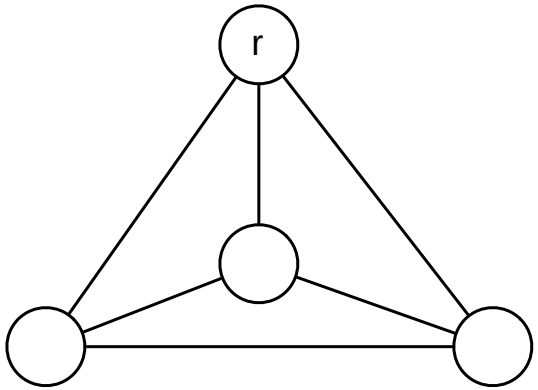
Clique:

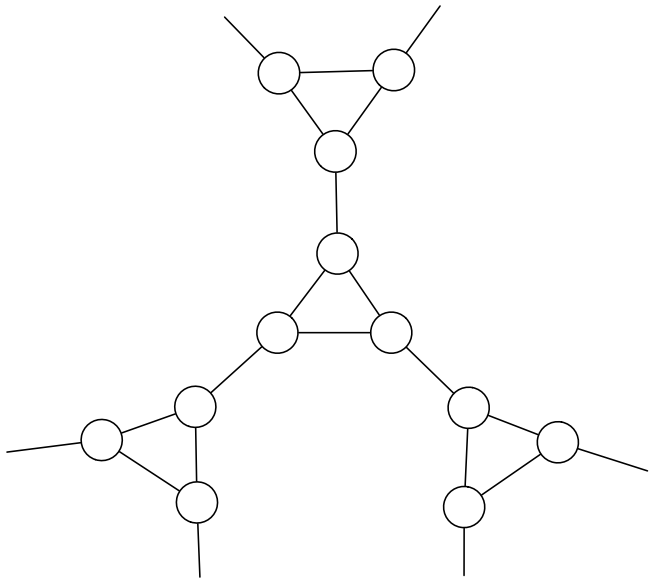
$$s + \binom{d}{2} s^2 + dst + \binom{d}{2} st + \text{terms of degree} \geq 3 \cdots$$

Now let $s, t \to 0$ so higher-order terms are negligible.

(Justification becomes subtle.)

Sufficient Sets





Recall:

- Set of all graphs with d neighbors per node.
- Set of all distributions μ on $\{0, 1, 2, \ldots, d+1\}$.

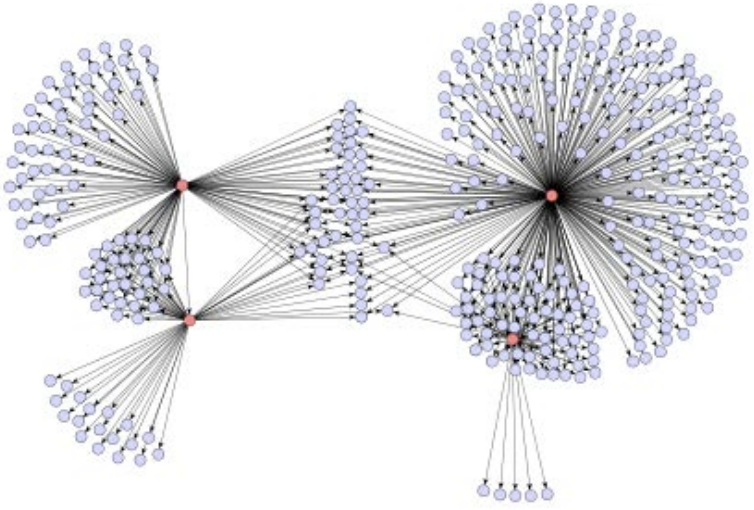
Call a set of graphs $\mathcal{H}$ "sufficient" if:

- For every distribution μ, some graph in $\mathcal{H}$ achieves minimum risk over all d-regular graphs.

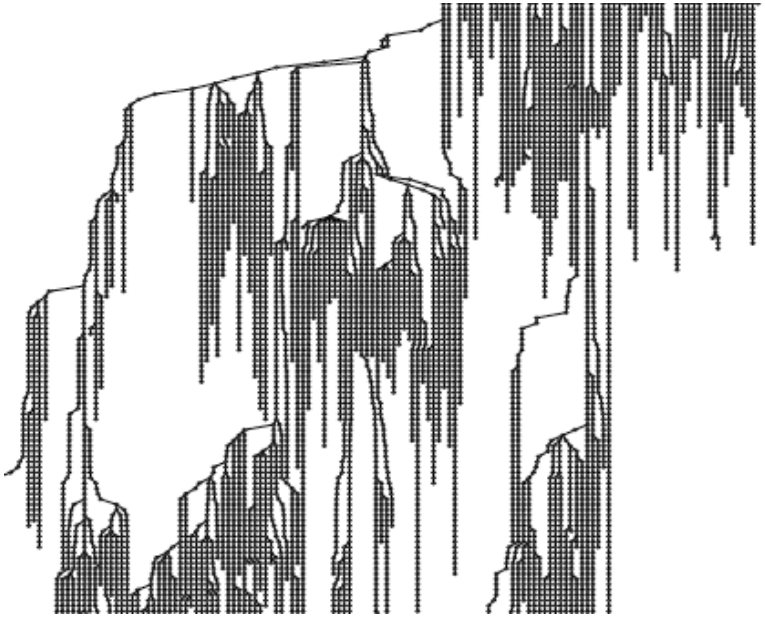
Question:

- Is there a finite sufficient set for family of all d-regular graphs?

The Flow of Information



Book recommendations (Leskovec et al 2006)



Anti-war chain letter (LibenNowell-Kleinberg 2008)

- "Open" vs. "closed" neighborhoods a crucial distinction in contagion on Facebook [Ugander-Backstrom-Marlow-Kleinberg 2012].
- Incentives to propagate information: e.g. Query incentive networks [Kleinberg-Raghavan 2005], DARPA Network Challenge [Pickard et al 2011], Bitcoin [Babaioff et al 2012].
- Simultaneous evolution of network structure and behavior [Holme-Newman 2006, Durrett et al 2012].

Final Reflections

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 being recorded at high resolution.
- What is the right framework for capturing the structures and phenomena that we see?
- What are the dangers of stockpiling this much personal data?
- An opportunity for fundamental mathematical models to inform the next steps on all these questions.