

**William Arms, Geri Gay, Dan
Huttenlocher, Jon Kleinberg,
Michael Macy, & David Strang**

Cornell University



**Turning the Internet Archive into a New Cybertool for
Social Science Research**

NCeSS

30 June 06, Manchester, UK

Acknowledgements

- National Science Foundation Program on Next Generation CI Tools, \$2M over 2 years.
- Internet Archive, providing data and technical assistance
- Cornell University, \$300K “Social Science in the Age of Networks”

The Internet Archive

- Snapshots taken every two months for ≈ 10 years
 - Launched (and financed) by Brewster Kahle
 - Complete crawls, but some gaps (robots.txt, or owners request exclusion)
 - About 600 TByte (compressed)
 - Rate of increase is about 1 TByte/day (compressed)
 - Metadata contains format, links, [anchor text](#), file types
 - Organized to facilitate historical access to known URL (Wayback Machine)



Enter Web Address: http://

All

Take Me Back

Adv. Search Compare Archive Pages

searched for http://www.cornell.edu

301 Results

some duplicates are not shown. See all... notes when site was updated.

Search Results for Jan 01, 1996 - Apr 22, 2006

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
7 pages	3 pages	3 pages	17 pages	26 pages	19 pages	20 pages	118 pages	86 pages	0 pages	
Jun 05, 1997 *	Jan 19, 1998 *	Jan 25, 1999 *	Mar 03, 2000 *	Feb 24, 2001 *	Jan 24, 2002 *	Feb 01, 2003 *	Jan 25, 2004 *	Jan 28, 2005 *		
Jul 30, 1997 *	Jan 19, 1998 *	Feb 08, 1999 *	Mar 03, 2000 *	Feb 26, 2001 *	May 30, 2002 *	Feb 02, 2003 *	Feb 02, 2004 *	Feb 03, 2005 *		
Jul 30, 1997 *	Jul 09, 1998 *	Apr 27, 1999 *	May 11, 2000 *	Mar 01, 2001 *	Jun 05, 2002 *	Feb 20, 2003 *	Mar 26, 2004 *	Feb 04, 2005 *		
Oct 13, 1997 *			May 11, 2000 *	Mar 01, 2001 *	Sep 13, 2002 *	Mar 31, 2003 *	Apr 29, 2004 *	Feb 04, 2005 *		
Oct 13, 1997 *			May 11, 2000 *	Mar 01, 2001 *	Sep 26, 2002 *	Apr 04, 2003 *	Jun 05, 2004 *	Feb 04, 2005 *		
Dec 11, 1997 *			May 20, 2000 *	Mar 01, 2001 *	Sep 26, 2002 *	Apr 05, 2003 *	Jun 08, 2004 *	Feb 04, 2005 *		
Dec 11, 1997 *			Jun 16, 2000 *	Mar 02, 2001 *	Oct 02, 2002 *	Apr 21, 2003 *	Jun 10, 2004 *	Feb 04, 2005 *		
			Jun 21, 2000 *	Mar 02, 2001 *	Oct 08, 2002 *	May 23, 2003 *	Jun 11, 2004 *	Feb 05, 2005 *		
			Jul 07, 2000 *	Mar 02, 2001 *	Oct 16, 2002 *	Jun 10, 2003 *	Jun 12, 2004 *	Feb 05, 2005 *		
			Aug 15, 2000 *	Mar 02, 2001 *	Oct 22, 2002 *	Jun 18, 2003 *	Jun 14, 2004 *	Feb 06, 2005 *		
			Oct 14, 2000 *	Apr 01, 2001 *	Oct 26, 2002 *	Jun 22, 2003 *	Jun 14, 2004 *	Feb 06, 2005 *		
			Oct 18, 2000 *	Apr 05, 2001 *	Oct 29, 2002 *	Jul 26, 2003 *	Jun 16, 2004 *	Feb 07, 2005 *		
			Oct 19, 2000 *	Apr 28, 2001 *	Nov 04, 2002 *	Aug 06, 2003 *	Jun 16, 2004 *	Feb 08, 2005 *		
			Oct 19, 2000 *	Apr 28, 2001 *	Nov 18, 2002 *	Aug 13, 2003 *	Jun 18, 2004 *	Feb 08, 2005 *		
			Oct 27, 2000 *	Apr 28, 2001 *	Nov 22, 2002 *	Oct 05, 2003 *	Jun 19, 2004 *	Feb 09, 2005 *		
			Dec 03, 2000 *	Apr 29, 2001 *	Nov 25, 2002 *	Oct 28, 2003 *	Jun 23, 2004 *	Feb 09, 2005 *		
			Dec 17, 2000 *	May 07, 2001 *	Nov 25, 2002 *	Nov 25, 2003 *	Jun 24, 2004 *	Feb 11, 2005 *		
				May 15, 2001 *	Dec 02, 2002 *	Dec 11, 2003 *	Jun 25, 2004 *	Feb 12, 2005 *		
				May 20, 2001 *	Dec 08, 2002 *	Dec 17, 2003 *	Jun 26, 2004 *	Feb 13, 2005 *		
				May 22, 2001 *		Dec 26, 2003 *	Jun 27, 2004 *	Feb 14, 2005 *		
				May 24, 2001 *			Jun 27, 2004 *	Feb 15, 2005 *		
				May 25, 2001 *			Jun 28, 2004 *	Feb 16, 2005 *		
				Jun 02, 2001 *			Jun 29, 2004 *	Feb 17, 2005 *		



(Cornell's webpage on June 5, 1997)

[Academic Units](#) | [Admissions](#) | [Alumni, Parents, and Friends](#) | [Big Red Sports](#) | [Campus Tour](#)
coolsites@cornell | [CUinfo and Other Web Sites](#) | [Guest Book](#) | [News and Events](#)
[Research and Outreach](#) | [Searches and Directories](#) | [This is Cornell](#) | [Visiting the Campus](#)

[Class of 2001](#)

Have comments or questions [about this award-winning Web site](#)? Let us know in our [Guest Book](#).

URL: <http://www.info.cornell.edu/CUHomePage.html>

Last modified: 06/04/97

© 1997 Cornell University

U.S. Mailing Addresses: [Ithaca, New York 14853](#)

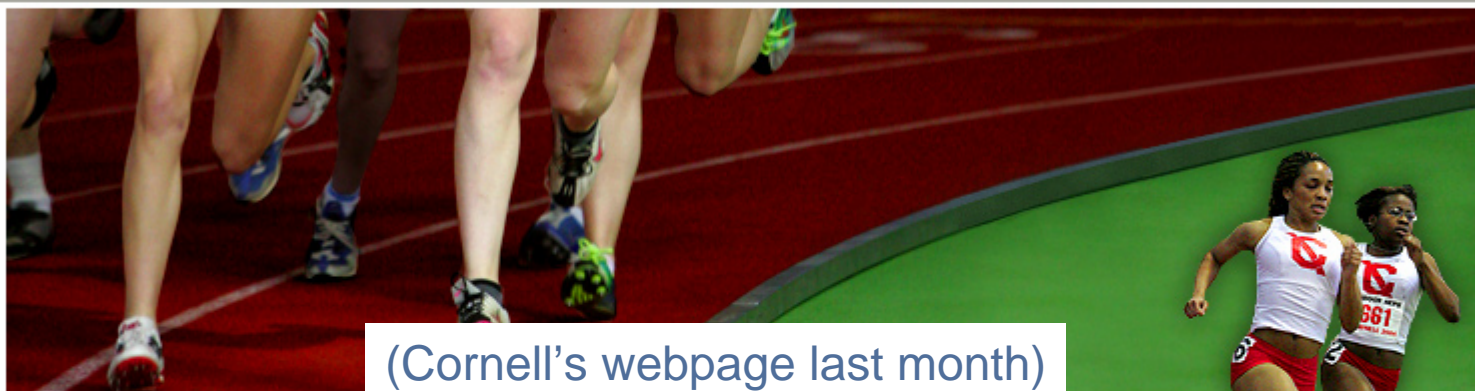


Cornell University

SEARCH: go

Pages People more options

Admissions Academics Research Outreach Collections Student Life Alumni



Big Red track and field at Barton Hall

"Any person...any study."
- Ezra Cornell, 1865

Welcome >



- [Colleges and Schools](#)
- [Facts about Cornell](#)
- [Diversity and Inclusiveness](#)

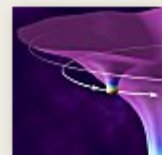
Visiting >

- [Visiting Cornell University](#)
- [Virtual Tours and Live Views](#)
- [Maps of Cornell](#)

News >



[Johnson Graduate School of Management dean to step down](#)
Robert Swieringa will leave at end of second term in June, 2007



[Hunt for gravity waves expands with new detectors](#)
Physicist Eanna Flanagan predicts what they should see



[Mitchell's late goal lifts No. 5 Cornell over No. 6 Princeton, 4-3](#)
Haswell leads Big Red with two goals

Events >



[Melvin Sparks Band](#)
Apr 22

[Queer Prom 2006: Roaring Back to the '20s!](#)
Apr 22

[Block & Bridle Student Livestock Show](#)

WebLab: the Way Forward Machine

- The Wayback Machine is useful for tracking a single URL.
- Not designed for network analysis.
- Imagine if the Internet Archive was put into a relational database?
- Unprecedented opportunities for analysis of social and information networks.

When It Rains, It Pours

- How to download, store, structure, and search Web-scale data?
 - Not traditional tabular databases with specified fields and relations.
 - New tools are needed to parse data into meaningful parts and structures.
 - Manual coding is beyond human capabilities
 - Privacy is a substantial concern

The Solution: Computational Social Science

- Bring together experts in computer, information, and social science
 - NLP
 - Machine learning
 - Data-mining
 - Privacy
 - Relational databases
 - Digital libraries
 - Graph theorists
 - Social network theory
 - Game theory
 - Experimental social psych
 - Organizational ecology
 - Agent-based modeling
 - Diffusion of innovation
 - Opinion dynamics

The Cornell Team

- William Arms,
Computer Science
- Geri Gay,
Communication
- Dan Huttenlocher,
Computer Science
- Jon Kleinberg,
Computer Science
- Michael Macy,
Sociology
- David Strang,
Sociology

From Archive to Database

- Copying IA to Cornell over Internet2 at 300-500 GB per day
- Plan to have 1/3 transferred by end of 2007
- Metadata stored in a relational format
 - allow users to specify download criteria
 - time frame (months, years, decade)
 - keywords
 - links
 - growth/decline
- Anticipate NLP and machine learning tools to train algorithms for intelligent analysis of page content.

Limitations of Web Data

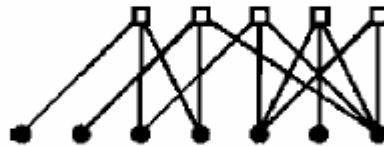
- Some pages were never collected, some are lost, and others are blocked
- Two month intervals cannot track things that spread/collapse quickly
- Digital divide, sampling bias
- Limited demographic data
- Interpretation of hyperlinks

So What Exactly is the Web?

- Hyperlinks are only the surface
 - Usually thought as the essence of the Web
 - Indicate
 - exposure to another site
 - status and authority (Kleinberg 1999)
 - trust (Davenport & Kronin 2000)
- But there are other kinds of ties as well...

Interlock and Affiliation Networks

Bi-partite Graph
(members and groups)

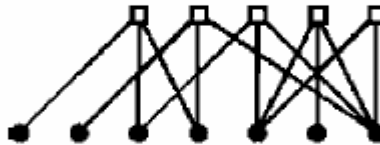


Interlock and Affiliation Networks

Group Interlock Network
(common members)



Bi-partite Graph
(members and groups)

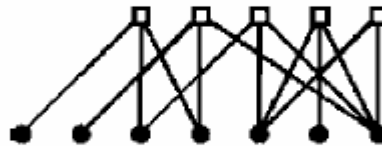


Interlock and Affiliation Networks

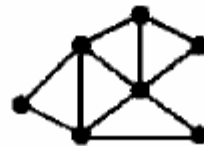
Group Interlock Network
(common members)



Bi-partite Graph
(members and groups)



Member Network
(shared affiliations)



Web Interlocks and Affiliations

- Ties induced from page content are similar to those in off-line networks.
 - Interlocks: people as links between communities
 - Affiliation: common participation in an event or community.

Possibilities for Web-based Research

- Diffusion of innovation:
 - What is the probability an individual will adopt a behavior from neighbors, depending on
 - number of neighbors who are adopters?
 - clustering/density/centrality of neighbors
 - Types of adoptions:
 - new technology or business practice (hotel wifi)
 - social movement, blog, candidate
 - rumor or fad or new term (“blog”)
 - on-line community

Finding the Dogs that Don't Bark

- Diffusion research tends to focus on innovations that spread successfully.
- What about those that
 - Were nipped in the bud?
 - Spread widely and then crashed?
- Run the tape back to the start and compare the take-off trajectories and early link structures of winners and losers.

Burst Analysis

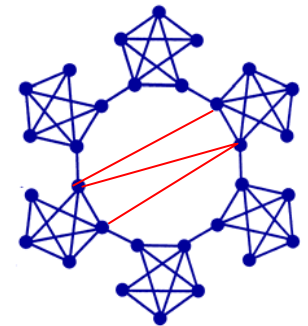
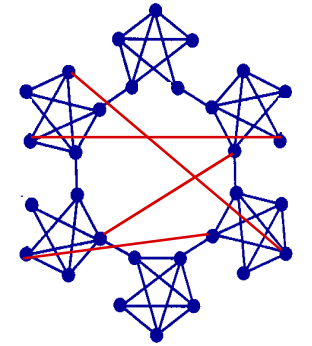
- Growth dynamics: gradual climb or stair steps?
 - How do bloggers become popular...
 - Response to a few widely-read postings?
 - Steady effort over months or years?
 - How do communities take off?
 - Spread of fads, urban legends, lingo (“Wifi”, “crunchy con”)

Threshold Effects

- Do network topologies that increase exposure to information also increase the willingness to act on it?
- Having information is not the same thing as acting on it.
 - Credibility, legitimacy, effectiveness tend to increase with the number of prior adopters.
 - Changing behavior requires social reinforcement from multiple sources.

Maybe It's Not Such a Small World After All?

- Computational experiments on graph perturbation
 - Paradoxically, “short cuts” facilitate *awareness* but impede *adoptions*.
 - Social reinforcement
 - Needed for behavior change
 - Requires overlapping clusters
 - Network bridges need to be wide, not long



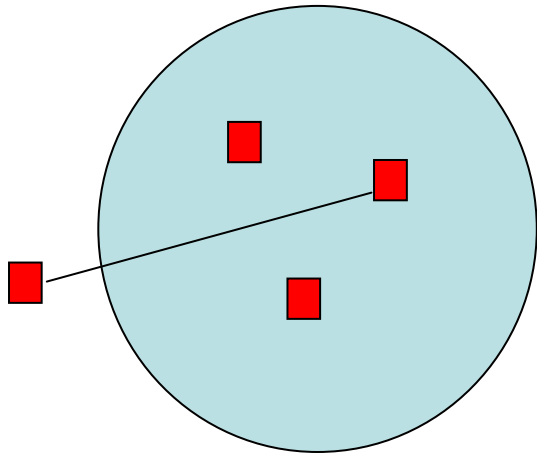
Find the Right Topology

- To spread information
 - Small worlds
 - Scale free, hub and spoke
- To change behavior
 - Overlapping clusters
 - Spatial networks (neighborhoods, dorms)

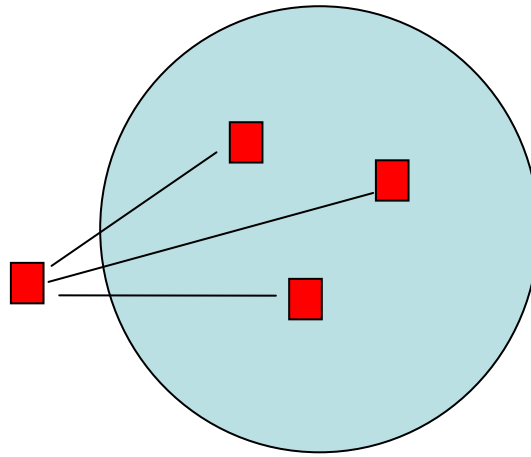
Growth of LJ Communities

- Backstrom, Kleinberg, Huttenlocher, Lan (2006):
“...processes by which communities in a social network come together, attract new members, and develop over time.”
 - 875 LJ communities
 - Individuals one degree removed
 - Joining as a function of
 - number of friends who are members
 - clustering among friends

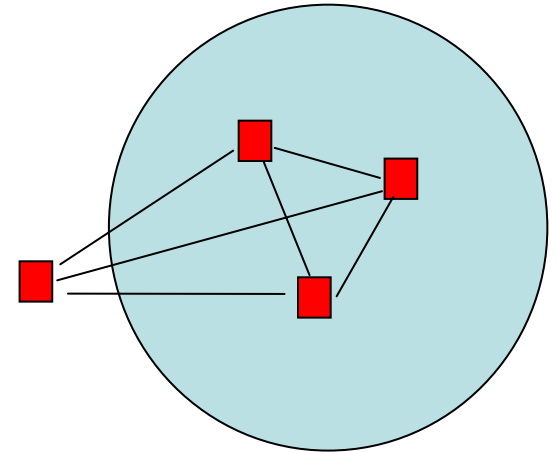
Number and connectedness of friends



A



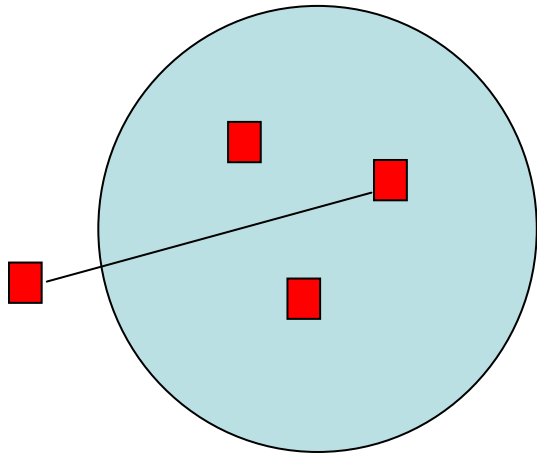
B



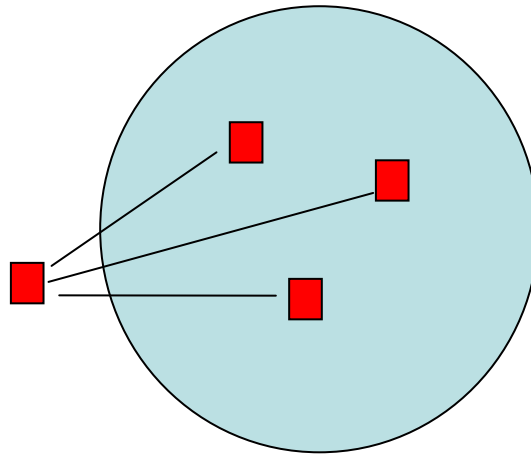
C

Time 1

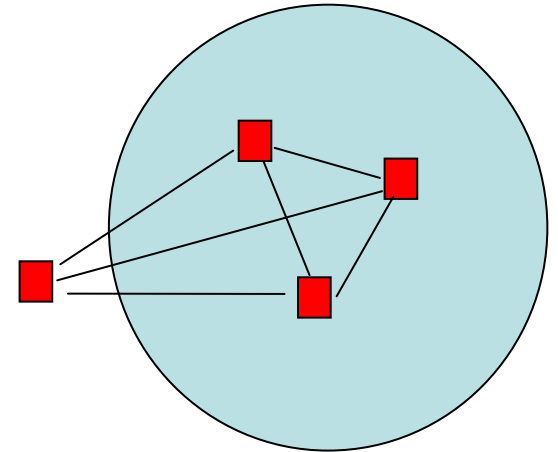
Number and connectedness of friends



A



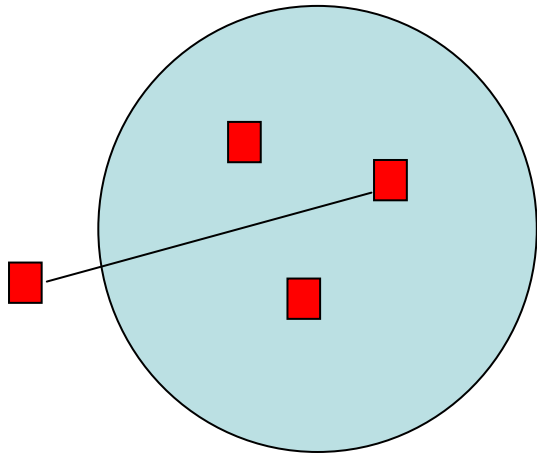
B



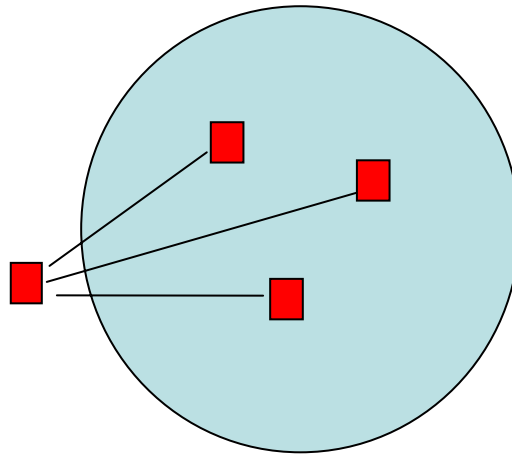
C

Time 1

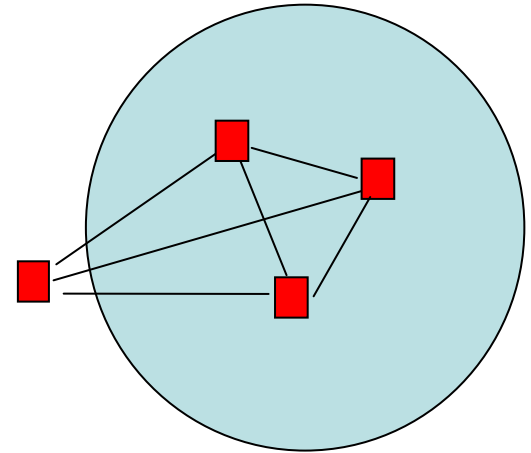
Number and connectedness of friends



A



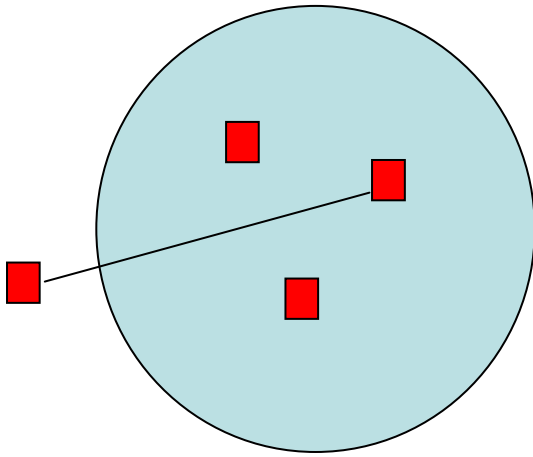
B



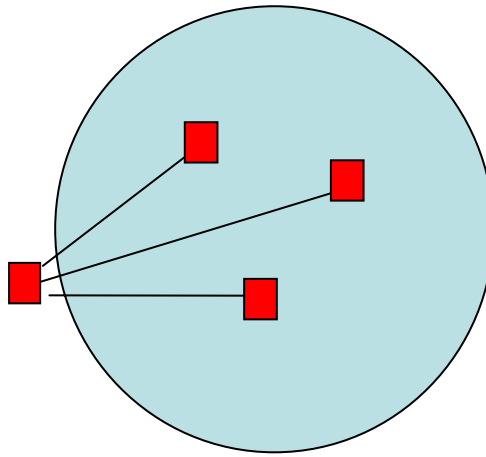
C

Time 2

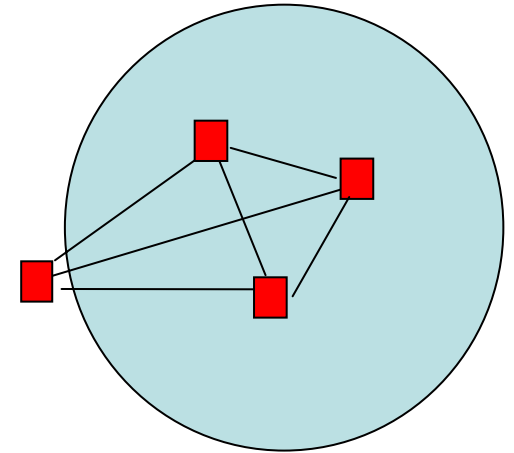
Number and connectedness of friends



A



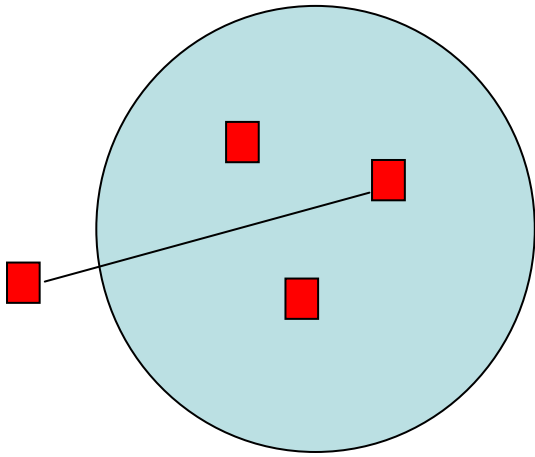
B



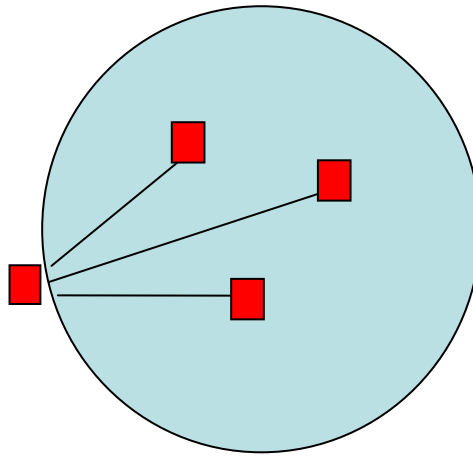
C

Time 3

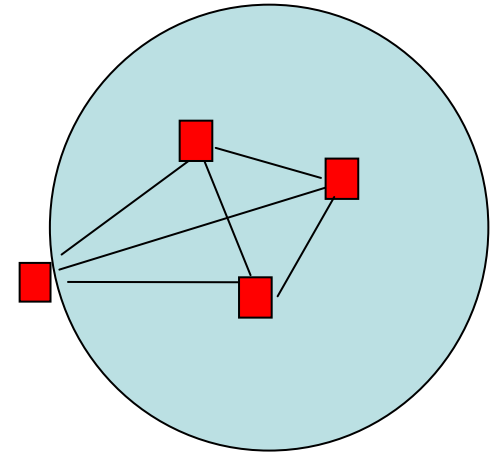
Number and connectedness of friends



A



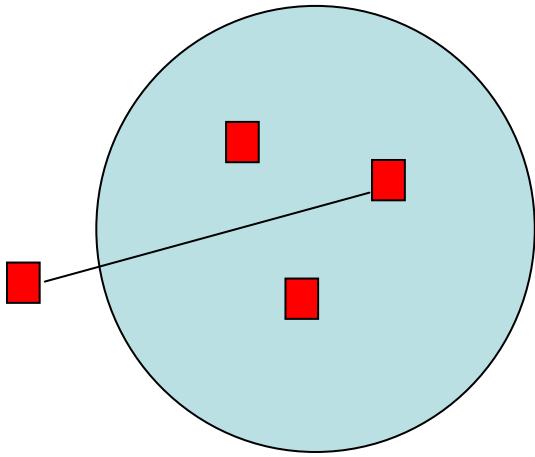
B



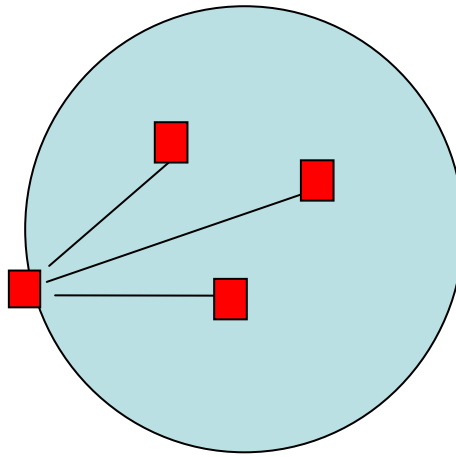
C

Time 4

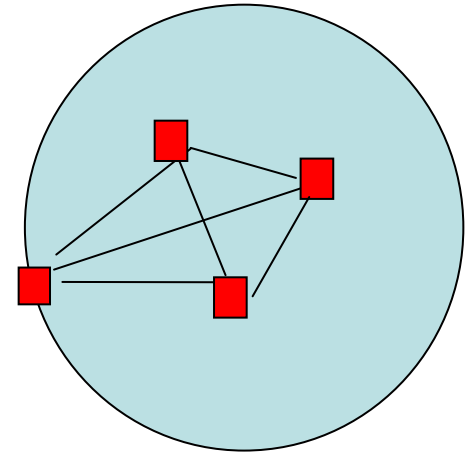
Number and connectedness of friends



A



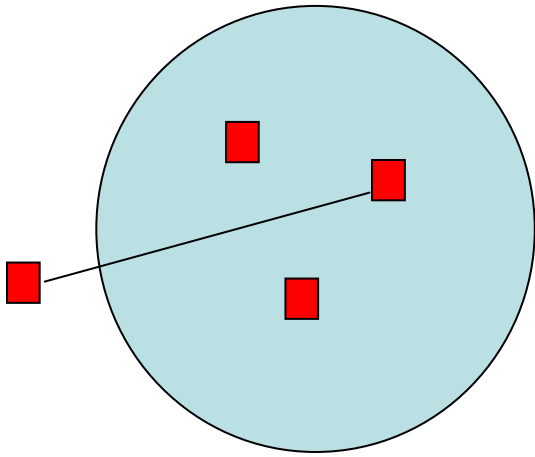
B



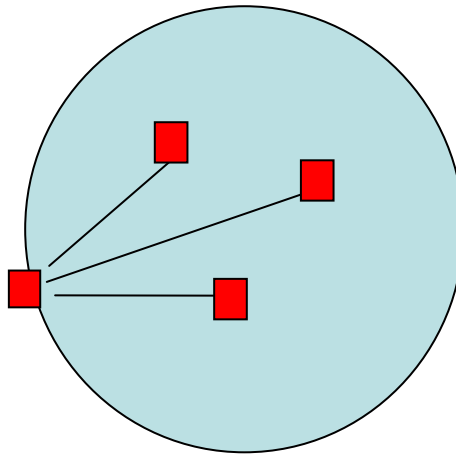
C

Time 5

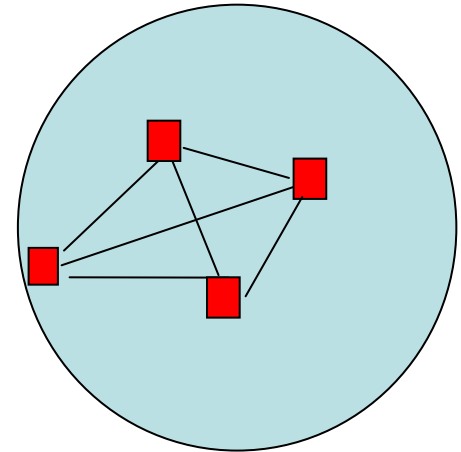
Number and connectedness of friends



A



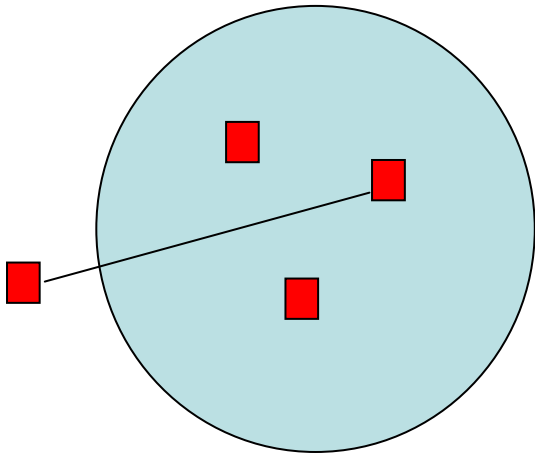
B



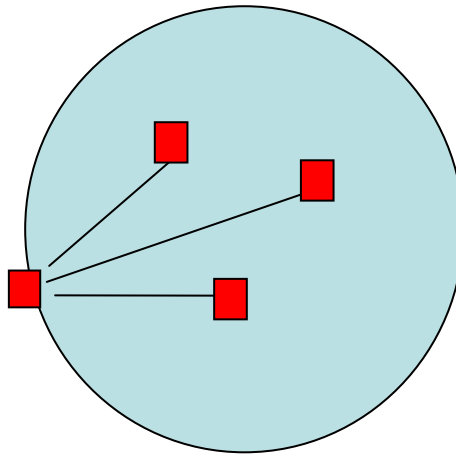
C

Time 6

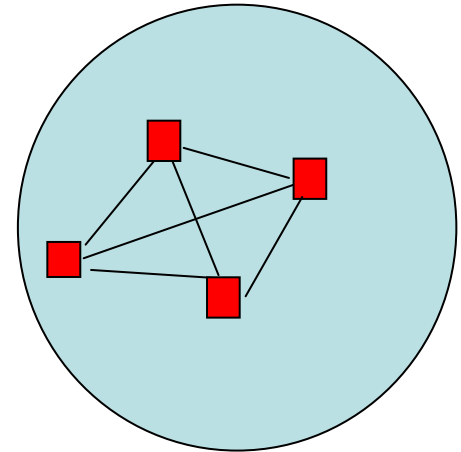
Number and connectedness of friends



A



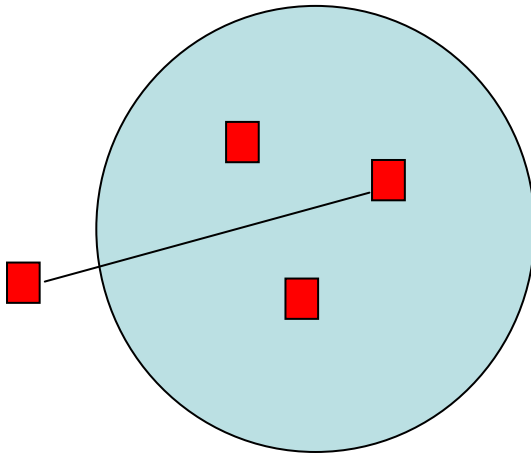
B



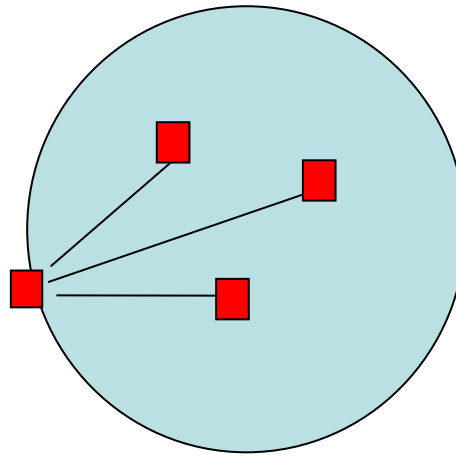
C

Time 7

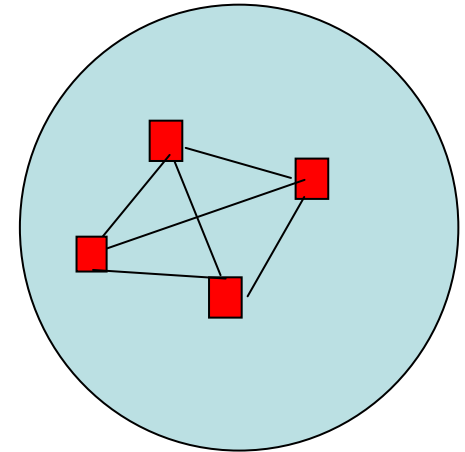
Number and connectedness of friends



A



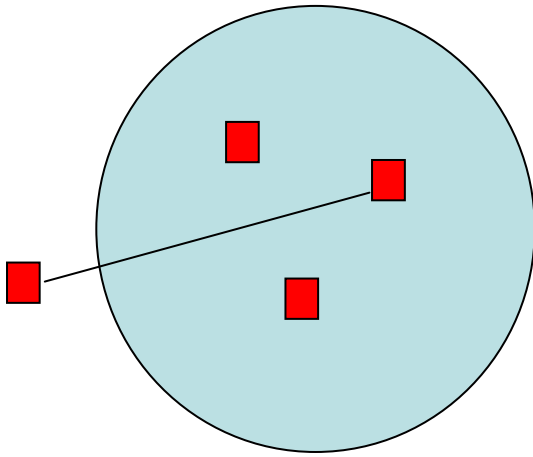
B



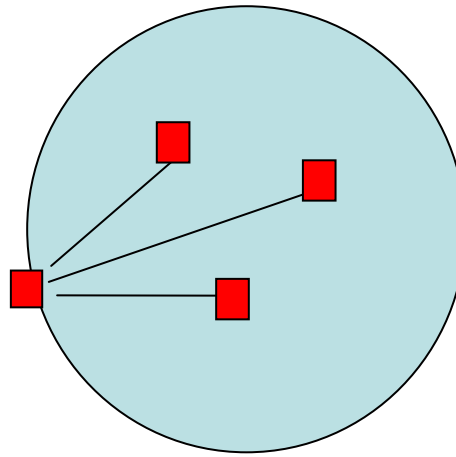
C

Time 8

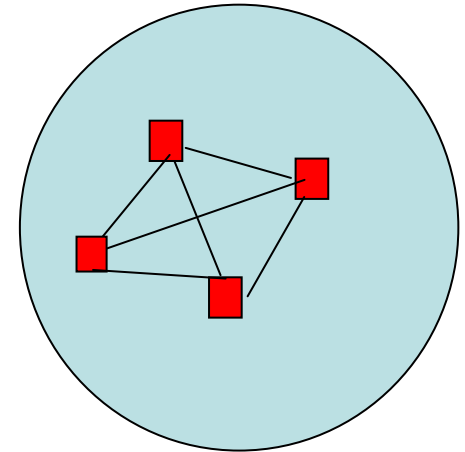
Number and connectedness of friends



A



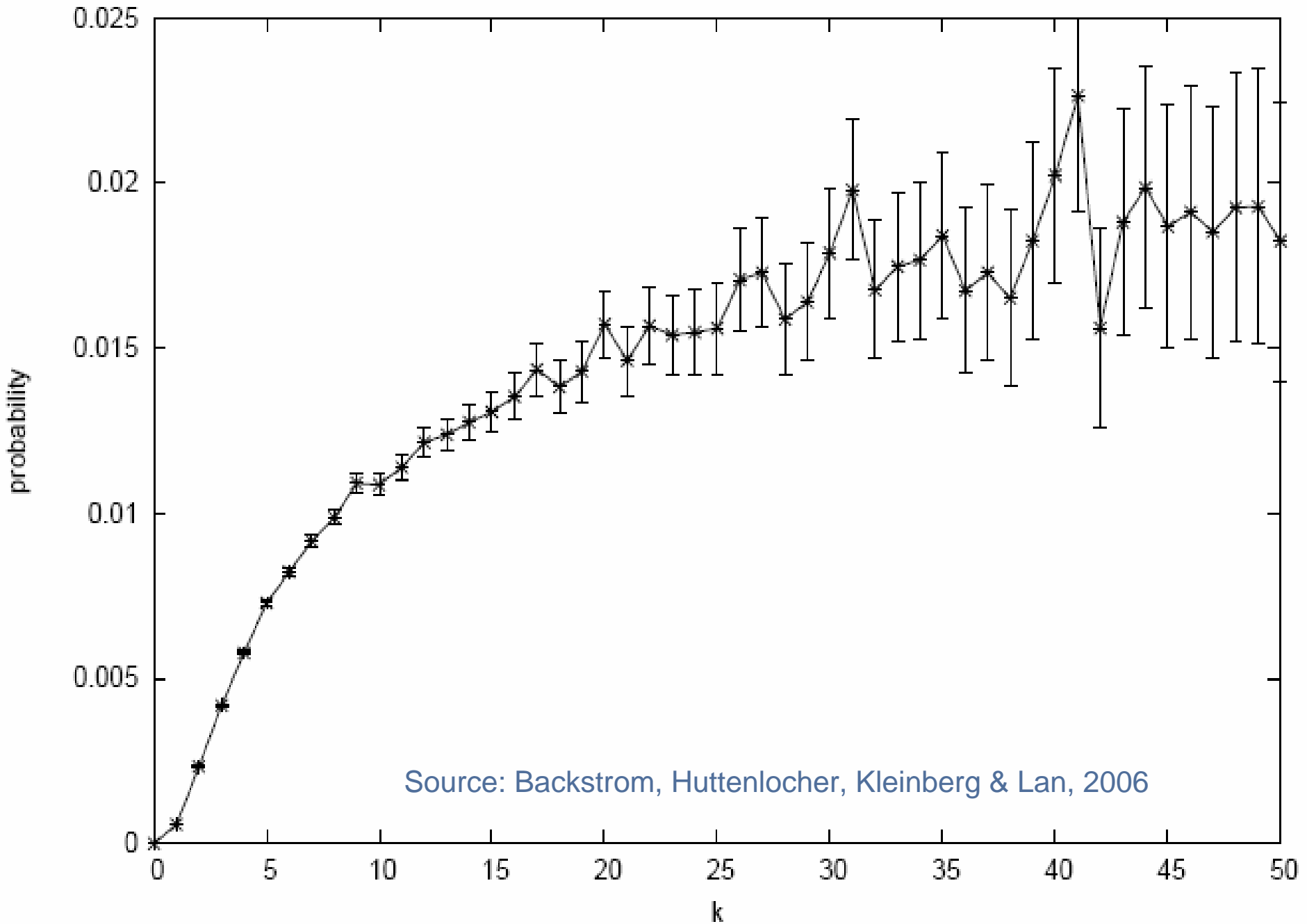
B



C

Time 8

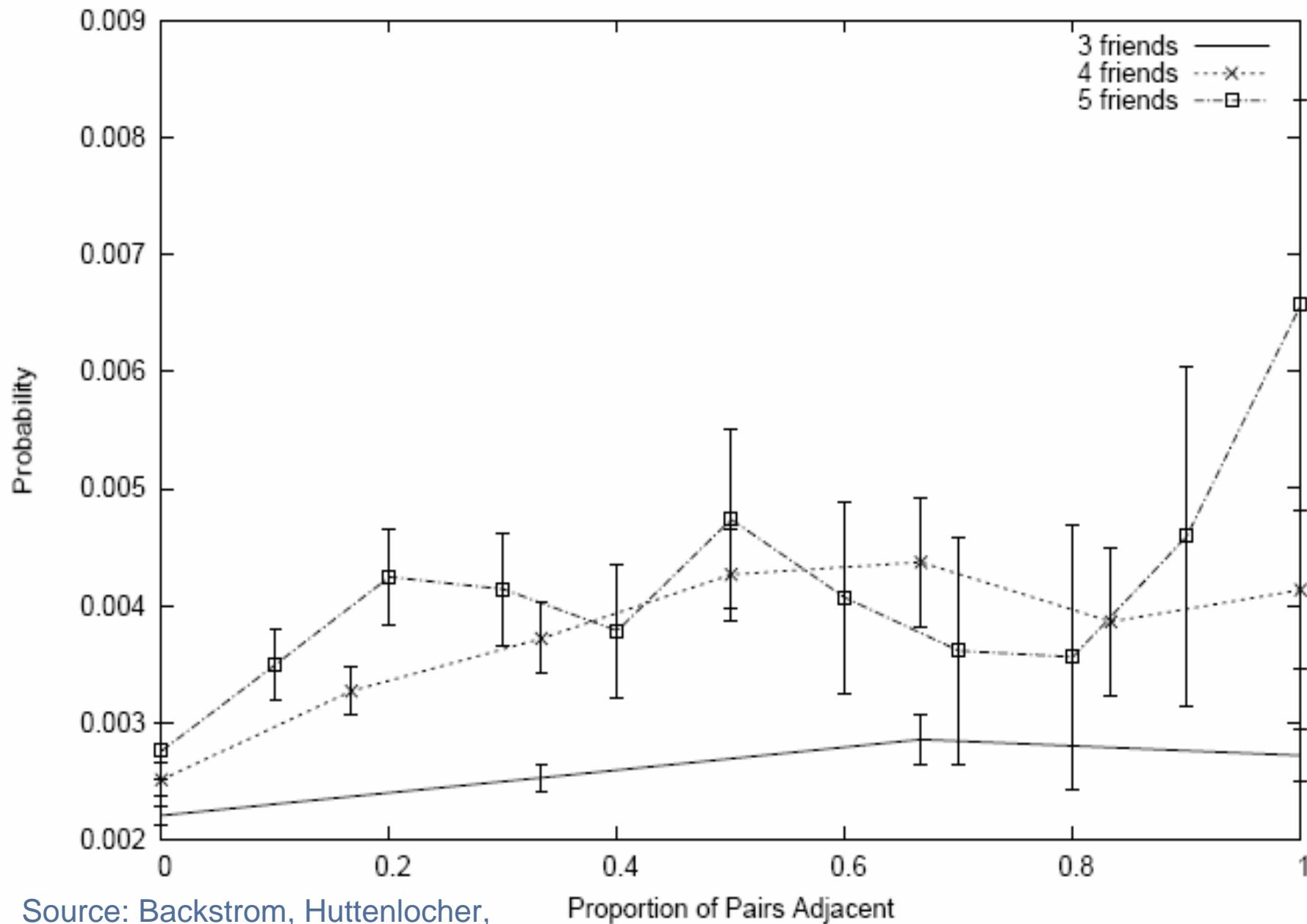
Probability of joining a community when k friends are already members



Is “Essential Redundancy” an Oxymoron?

- “Strength of Weak Ties”
 - Closed triads
 - If A is friends with B and C, B and C are friends
 - leads to information redundancy
 - Open triads increase the odds of hearing about the community
- Or is “redundancy” necessary to change behavior?

Probability of joining a community versus adjacent pairs of friends in the community



Source: Backstrom, Huttenlocher, Kleinberg & Lan, 2006



Network Dynamics

- Extend this idea to changes in network structure.
- What is the probability page i will link to page j , as a function of
 - the number of pages linked to both i and j ?
 - the network structure of these pages?

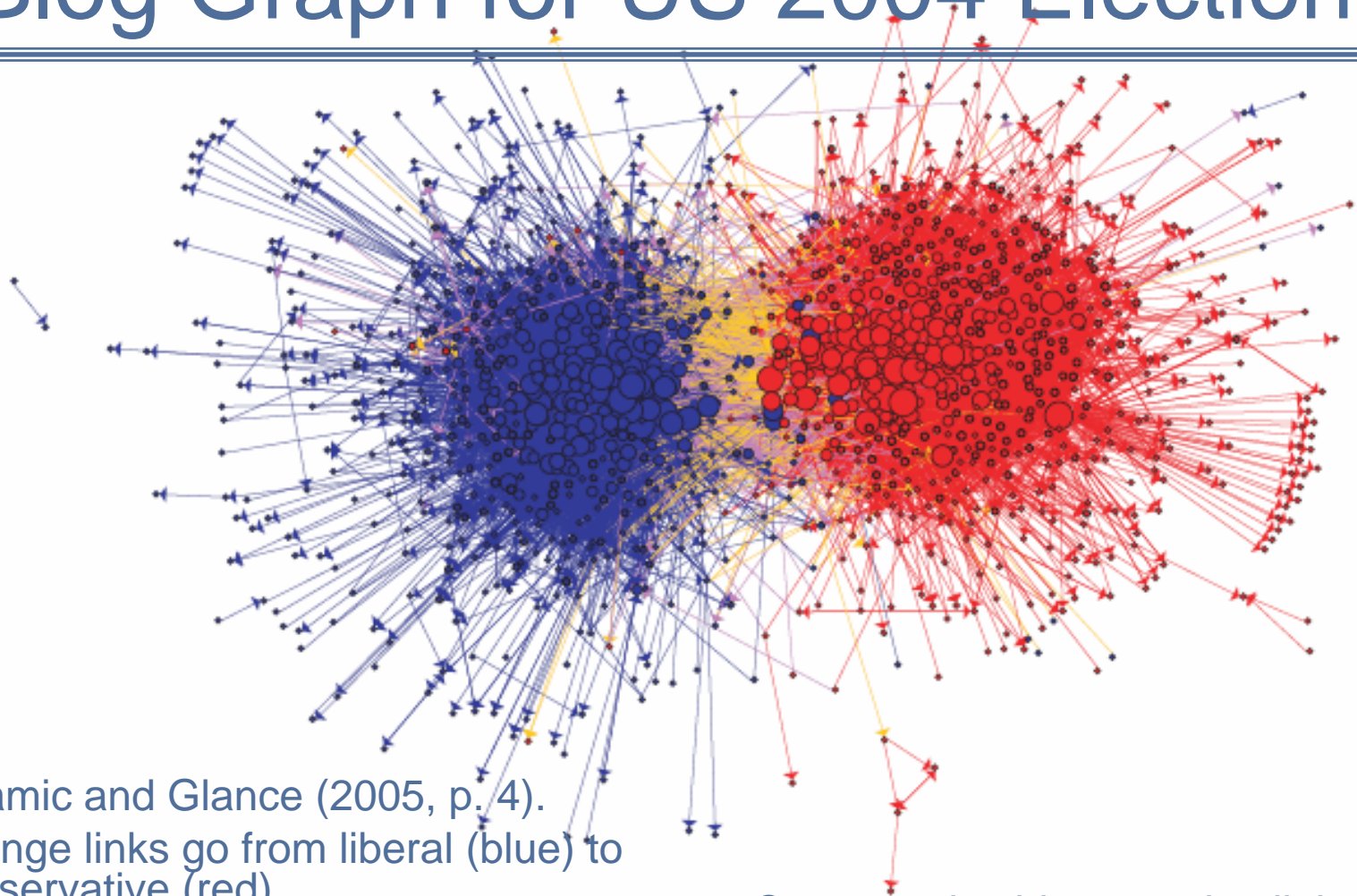
Self-organizing Communities

- Online communities are historically unique
 - No recognizable demographic and social identities
 - Few spatial and cultural constraints on interaction
- Test-beds to study emergence, spread, and enforcement of norms
 - Are norms imposed top-down via formal institutional arrangements (e.g. moderators)
 - or can they also self-organize through bottom-up interactions among users?

The Spread of Opinions and Beliefs

- News group threads, blogs, epinions
- Locate the structural positions of opinion leaders in networks of directed messages.
- NLP tools can “read” movie reviews  
 - Extending same methods to political views [Thomas, Pang, and Lee, EMNLP 2006]
 - Train using on-line text by politicians or blog authors whose positions are known

Blog Graph for US 2004 Election



- Adamic and Glance (2005, p. 4).
- Orange links go from liberal (blue) to conservative (red)
- Purple: from conservative to liberal.
- The size of each blog = degree

Conservative blogs tend to link to one another more frequently than do their liberal counterparts

A Parting Shot ...

- Some network methodologists claim we don't need to study large networks...
- Surprising discovery by Leskovec, Kleinberg and Faloutsos (2005)
- As arXiv citation network grows
 - densifying (not constant density)
 - diameter is decreasing (not increasing)

... & a Parting Thought

- Our immediate goal is to learn how to study diffusion by mining the Web
- The hidden agenda: a center for computational social science that links the talents, tools, and expertise of social, computer, and information scientists.