Carnegie Mellon

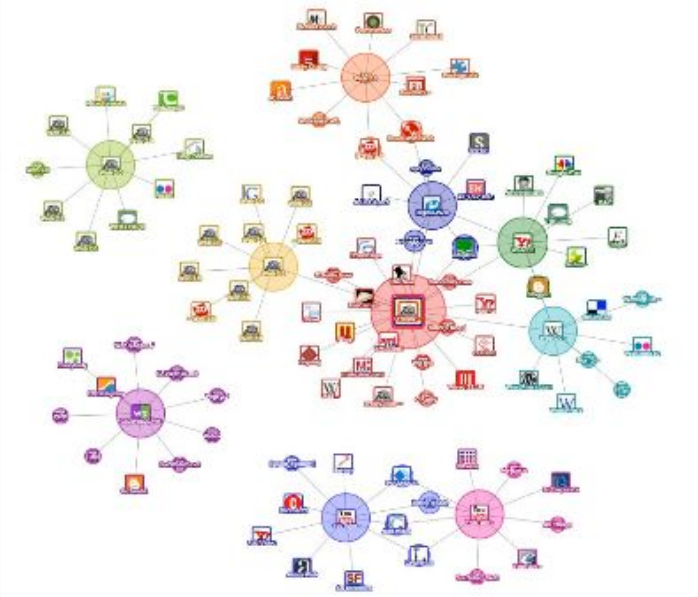
Graph Mining Techniques for Social Media Analysis

Mary McGlohon
Christos Faloutsos

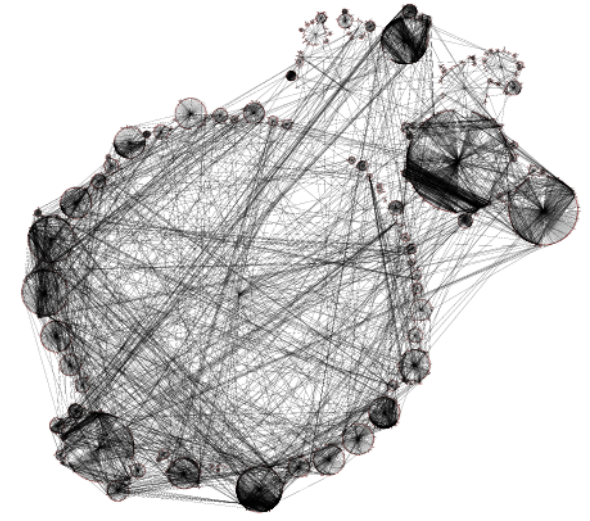
Carnegie Mellon

What is graph mining?

- Extracting useful knowledge (patterns, outliers, etc.) from structured data that can be represented as a graph.
- For our purposes, this is usually a social network.`



Facebook graph, via Touchgraph



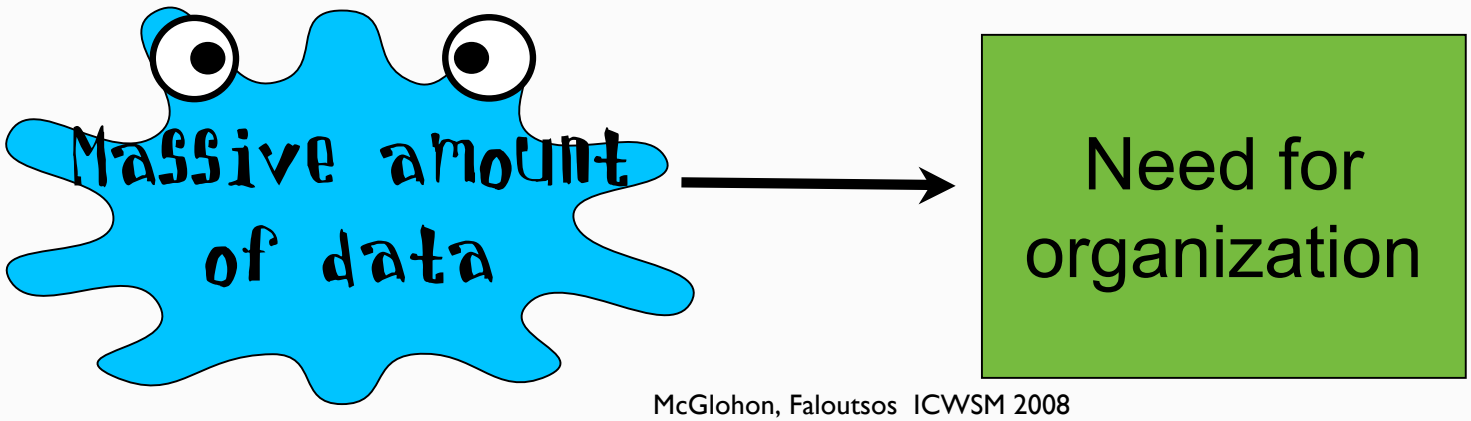
Livejournal, via Lehman and Kottler

What is graph mining?

- Example: Social media host tries to look at certain online groups and predict whether the group will flourish or disband.
- Example: Phone provider looks at cell phone call records to determine whether an account is a result of identity theft.

Why graph mining?

- Thanks to the web and social media, for the first time we have easily accessible network data on a large-scale.
- Understand relationships (links) as well as content (text, images).
- Large amounts of data raise new questions.



Motivating questions

- Q1: How do networks form, evolve, collapse?
- Q2: What tools can we use to study networks?
- Q3: Who are the most influential/central members of a network?
- Q4: How do ideas diffuse through a network?
- Q5: How can we extract communities?
- Q6: What sort of anomaly detection can we perform on networks?

Outline

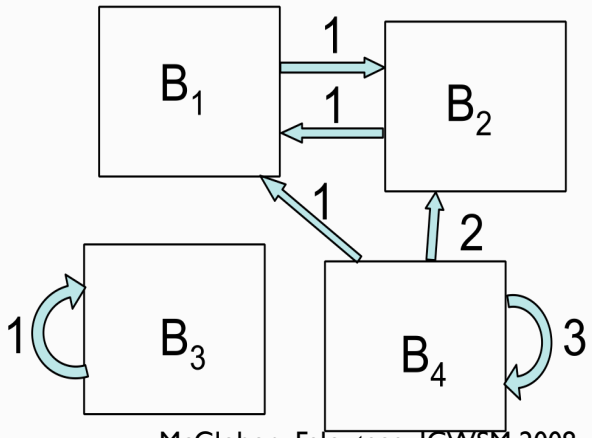
- Part 1: Q1: How do networks form, evolve, collapse?
  - Introduction to networks
  - Patterns, Laws
- Part 2: Q2: What tools can we use to study networks?
  - Q3: Ranking: Who are the most important members of a network?
- Part 3: Case studies
  - Q4: Diffusion: How do ideas diffuse through a network?
  - Q5: How can we extract communities?
  - Q6: What sort of anomaly detection can we perform?

Part 1 Outline

- Introduction to networks and 6 definitions
- Patterns
  - Diameter
  - Degree distribution
  - Connected components
  - Evolution over time

D1: Network

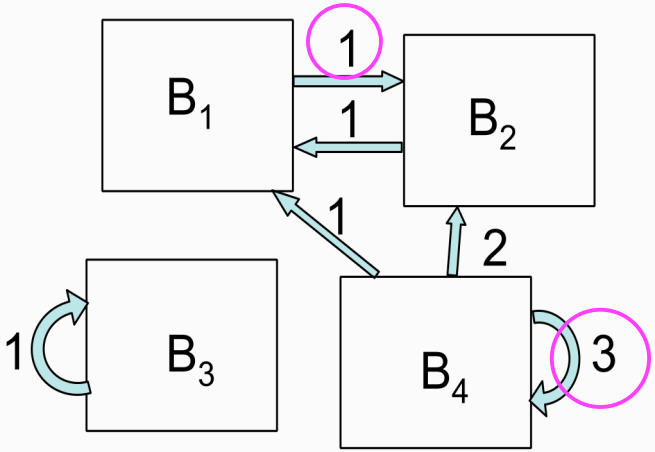
- A network is defined as a graph ***G=(V,E)***
  - ***V*** : set of vertices, or nodes.
  - ***E*** : set of edges.
- Edges may have numerical weights.



Carnegie Mellon

D2: Adjacency matrix

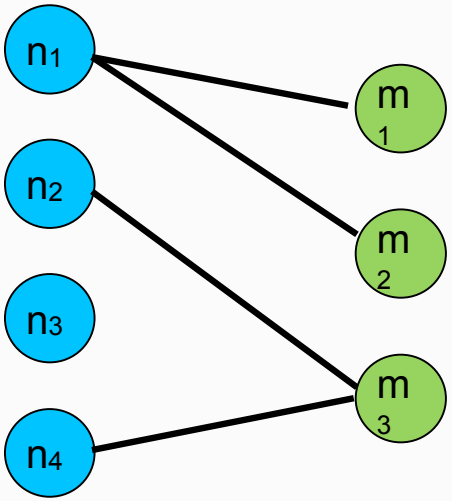
- To represent graphs, use adjacency matrix
- Unweighted graphs: all entries are 0 or 1
- Undirected graphs: matrix is symmetric



from \ *to*	B_1	B_2	B_3	B_4
B_1	0	1	0	0
B_2	1	0	0	0
B_3	0	0	1	0
B_4	1	2	0	3

D3: Bipartite graphs

- In a bipartite graph,
  - 2 sets of vertices
  - edges occur between different sets.
- If graph is undirected, we can represent as a non-square adjacency matrix.

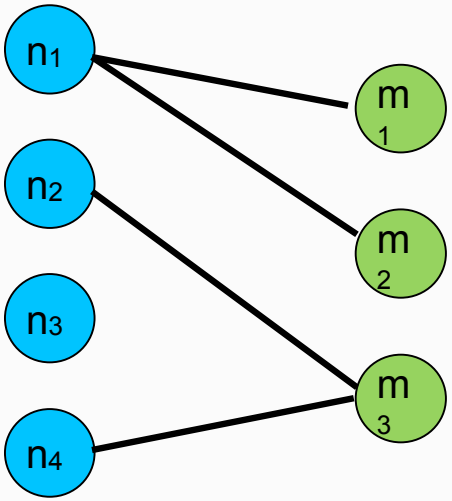


	m_1	m_2	m_3
n_1	1	1	0
n_2	0	0	1
n_3	0	0	0
n_4	0	0	1

Carnegie Mellon

D4: Components

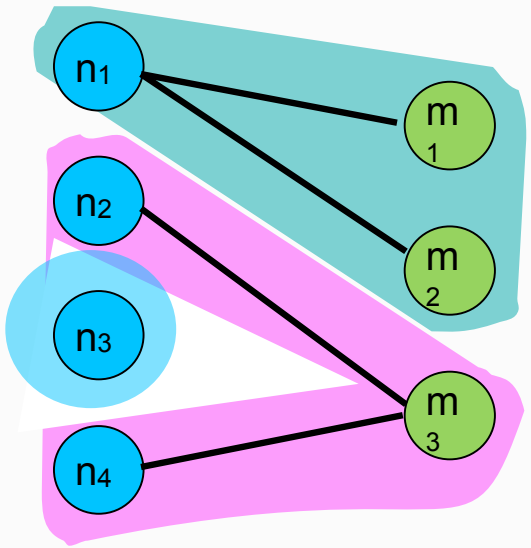
- Component: set of nodes with paths between each.



Carnegie Mellon

D4: Components

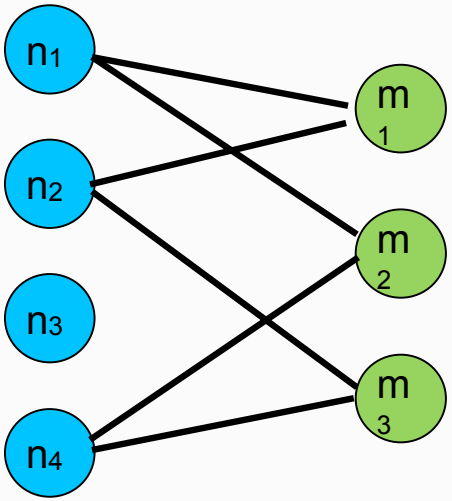
- Component: set of nodes with paths between each.
- We will see later that often real graphs form a giant connected component.



Carnegie Mellon

D5: Diameter

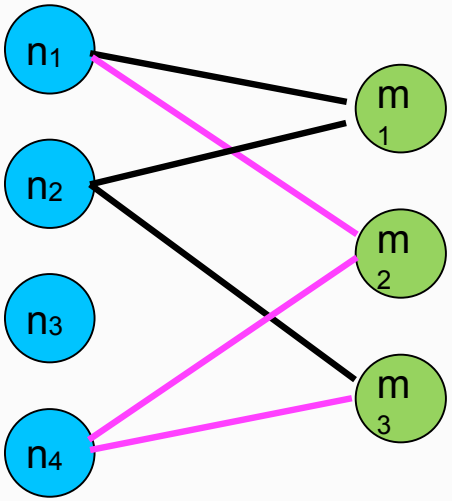
- Diameter of a graph is the “longest shortest path”.



Carnegie Mellon

D5: Diameter

- Diameter of a graph is the “longest shortest path”.

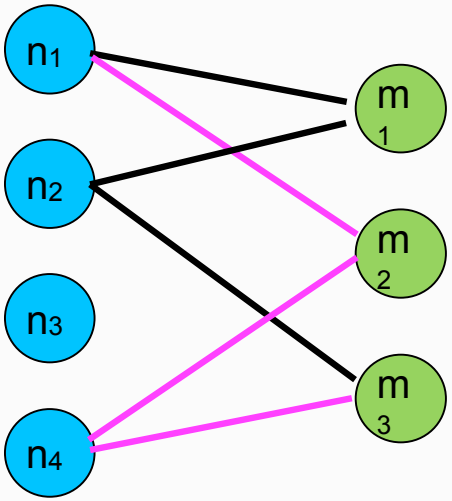


diameter=3

Carnegie Mellon

D5: Diameter

- Diameter of a graph is the "longest shortest path".
- We can estimate this by sampling.
- Effective diameter is the distance at which 90% of nodes can be reached.

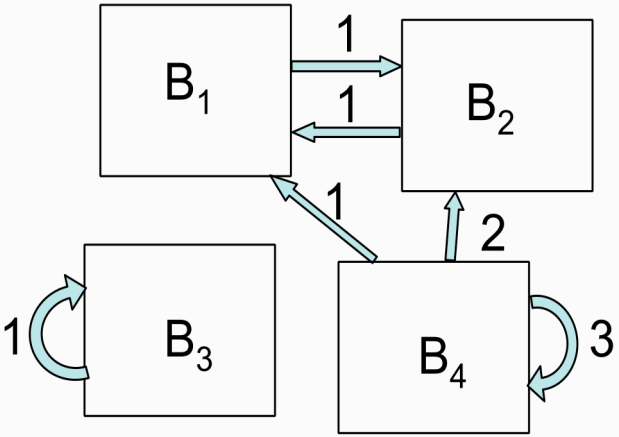


diameter=3

Carnegie Mellon

D6: Degree distribution

- We can find the degree of any node by summing entries in the (unweighted) adjacency matrix.

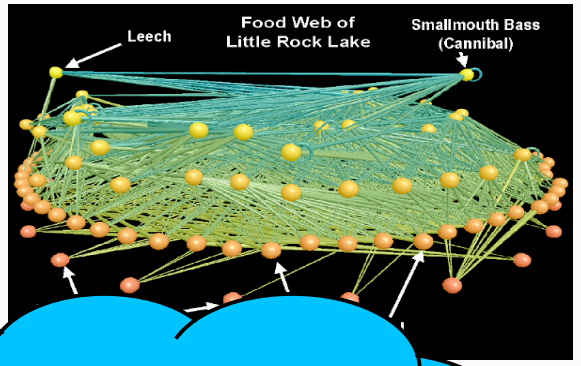


from \ to	B_1	B_2	B_3	B_4	out-degree
B_1	0	1	0	0	1
B_2	1	0	0	0	1
B_3	0	0	1	0	1
B_4	1	1	0	1	3
in-degree	2	2	1	1	

Carnegie Mellon

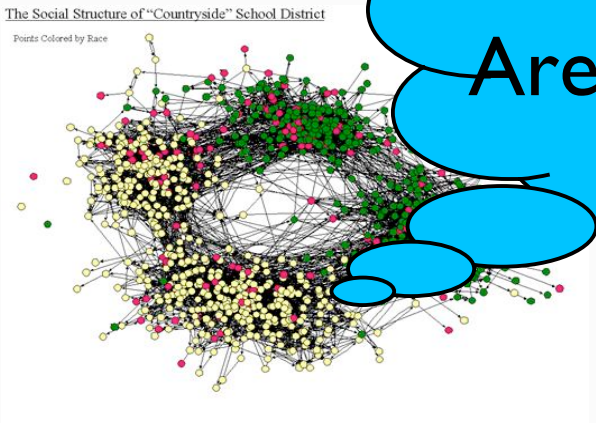
Some graphs

Internet Map [lumeta

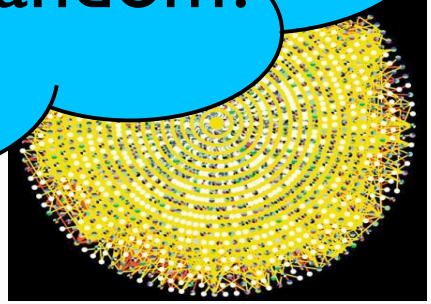


1]

Research question:
Are real graphs random?
(no)



Friendship Network [Moody '01]



Protein Interactions
[genomebiology.com]

Part 1 Outline

- Introduction to networks
- Patterns
  - Diameter
  - Degree distribution
  - Connected components
  - Evolution over time

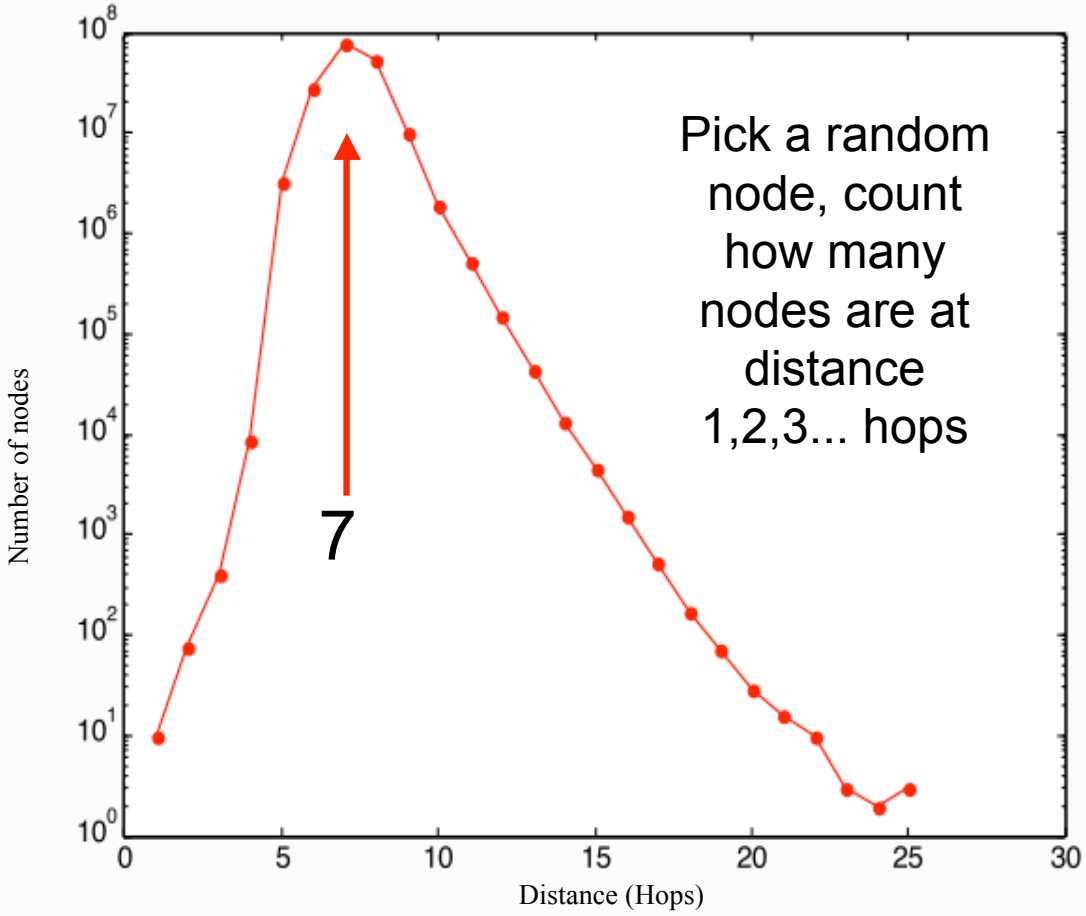
Small-world effect

- Graphs usually display small diameter.
- First demonstrated by Travers & Milgram in 1960.
  - Most of the time, distance was around 6.
- Similarly, real graphs we see have small diameter…

Carnegie Mellon

[Leskovec & Horvitz 07]

- Distribution of shortest path lengths
- Microsoft Messenger network
  - 180 million people
  - 1.3 billion edges
  - Edge if two people exchanged at least one message in one month period

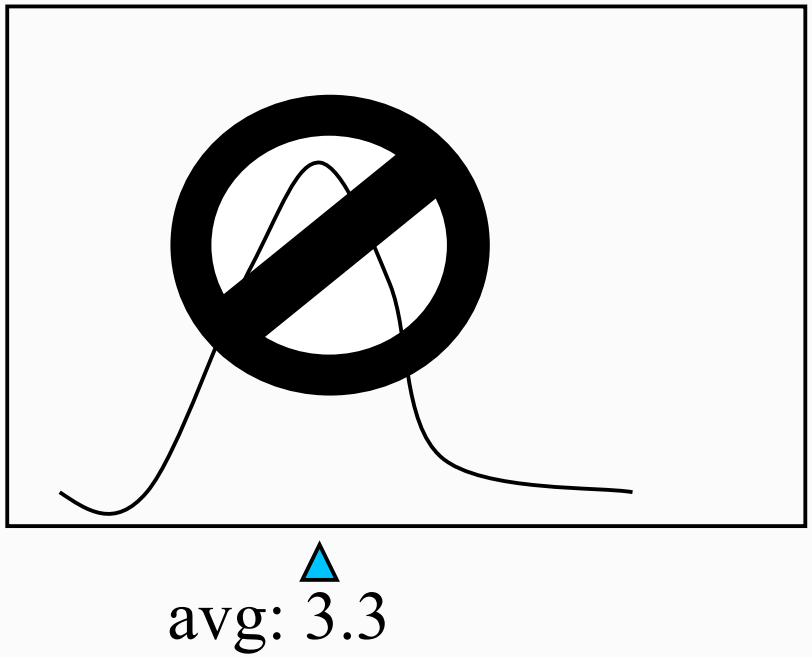


Part 1 Outline

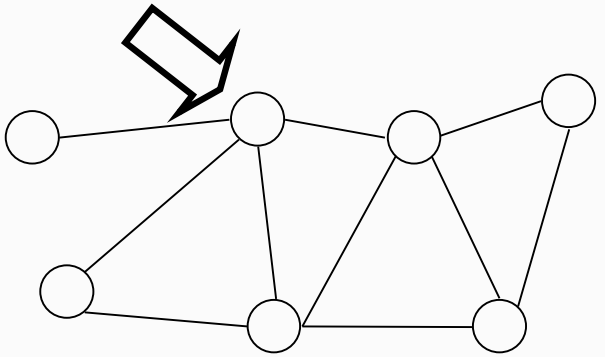
- Introduction to networks
- Patterns
  - Diameter: “small world effect”
  - Degree distribution
  - Connected components
  - Evolution over time

Degree distribution

Count vs. degree



- Suppose average degree is 3.3
- If we pick a node at random, can we guess its degree?
- In real graph, “mode” is 1!



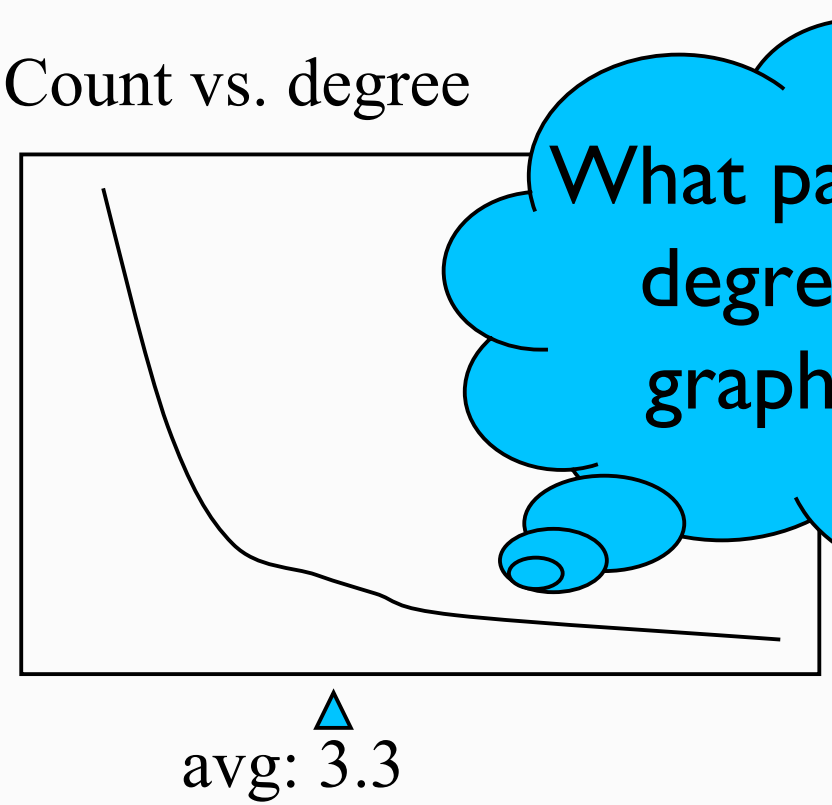
Degree distribution

- Suppose average degree is 3.3

ick a node at
can we
degree?

aph, "mode" is 1!

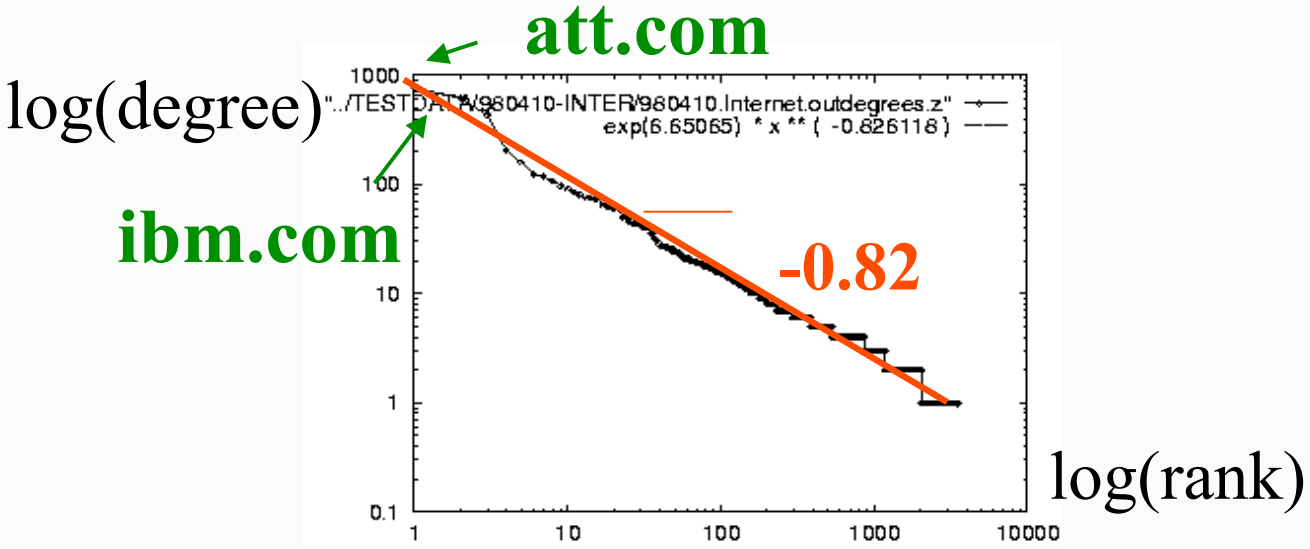
herefore, mean is "meaningless".



Power law degree distribution

- Measure with rank exponent R
- [SIGCOMM99]

internet domains



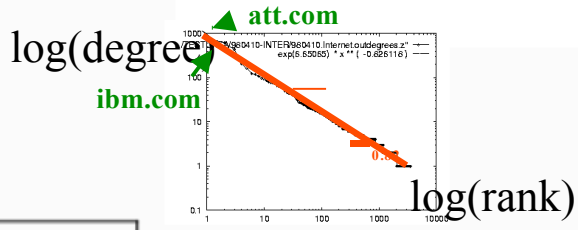
Power laws - discussion

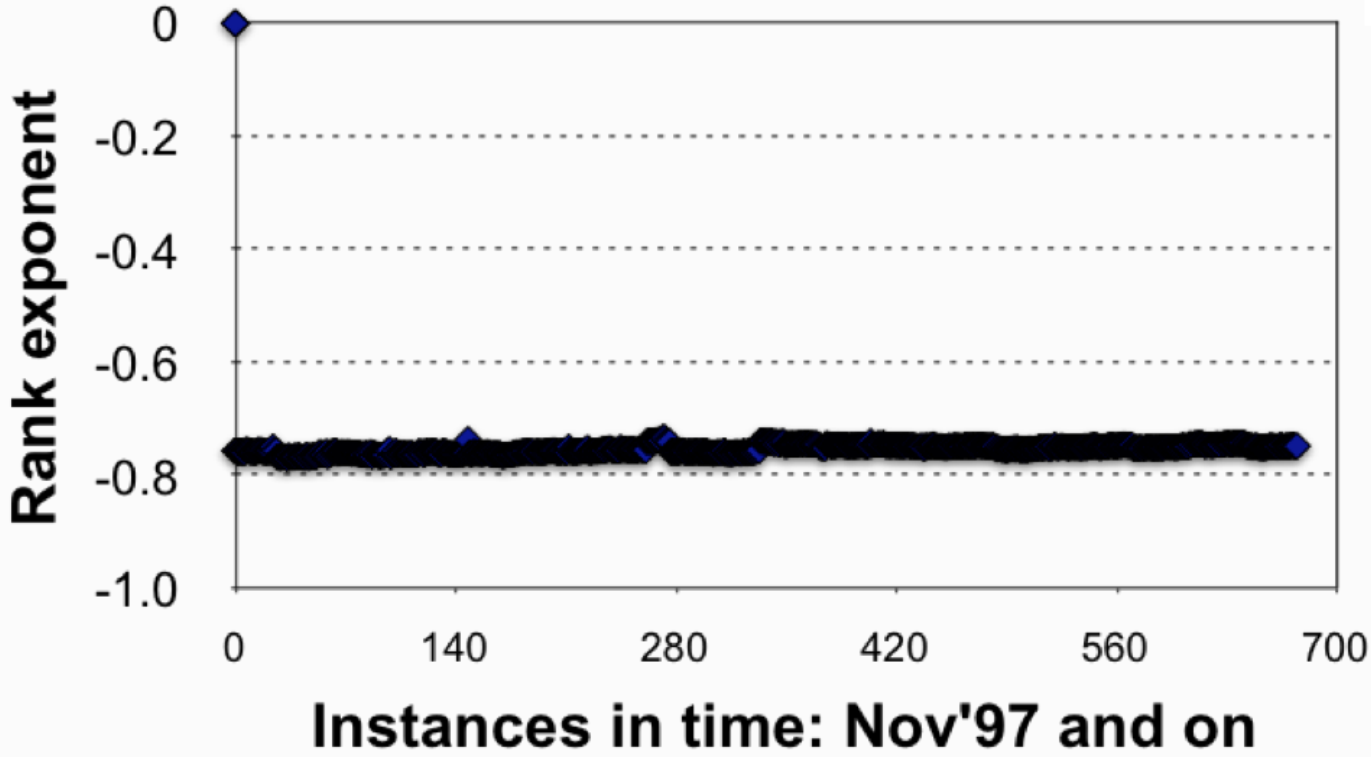
- Do they hold, over time?
- Do they hold on other graphs/domains?

Power laws - discussion

- Do they hold, over time?
  - Yes! for multiple years [Siganos+]
- Do they hold on other graphs/domains?
  - Yes!
  - Web sites and links [Tomkins+], [Barabasi+]
  - Peer-to-peer graphs (gnutella-style)
  - Who-trusts-whom (epinions.com)

Time Evolution: rank *R*



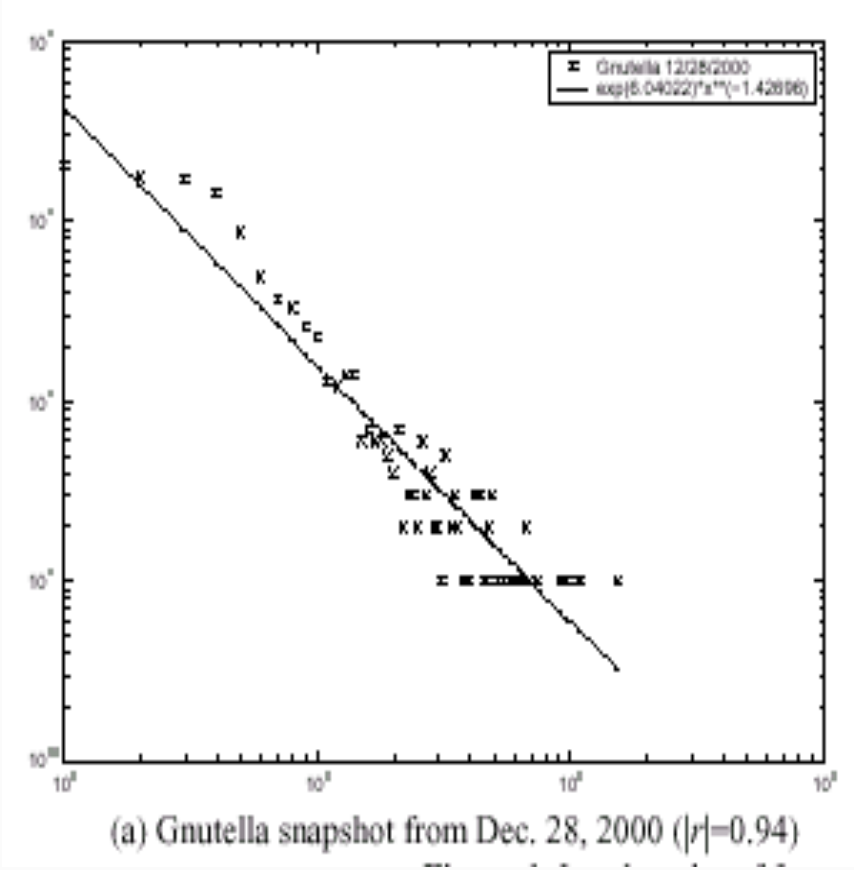


Domain level

- The rank exponent has not changed! [Siganos+]

The Peer-to-Peer Topology

count



(a) Gnutella snapshot from Dec. 28, 2000 ($|r|$=0.94)

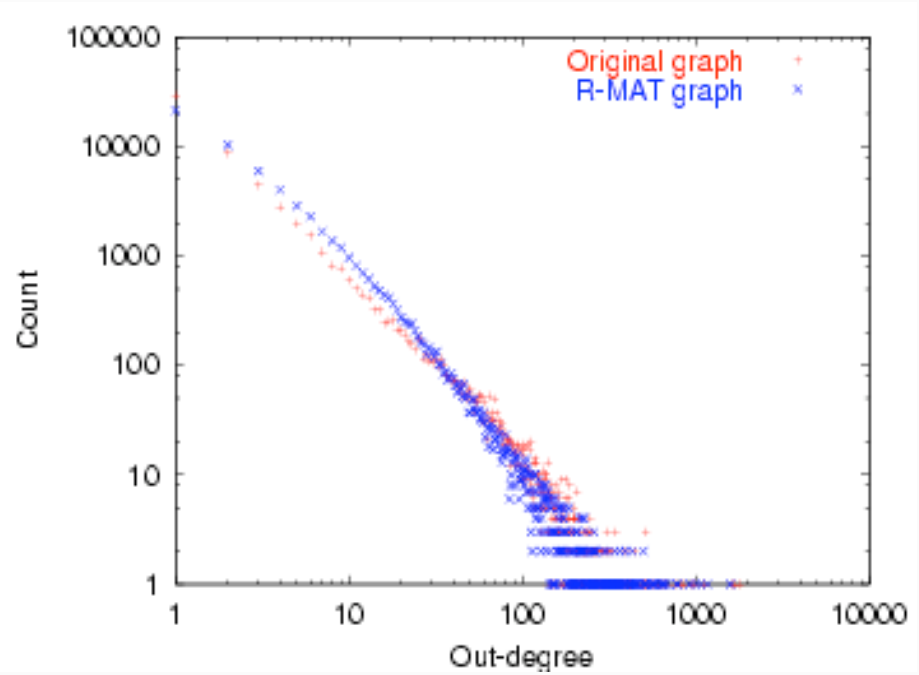
[Jovanovic+]

degree

- Number of immediate peers (= degree), follows a power-law

epinions.com

count



(out) degree

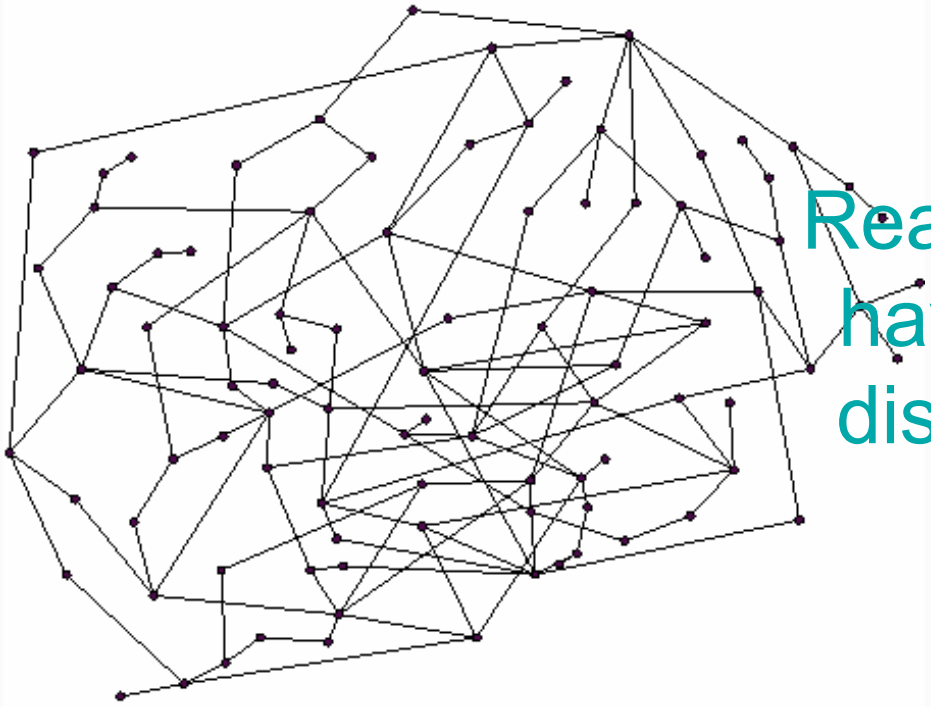
- who-trusts-whom [Richardson + Domingos, KDD 2001]

Part 1 Outline

- Introduction to networks
- Patterns
  - Diameter: “small world effect”
  - Degree distribution: power law
  - Connected components
  - Evolution over time

Components

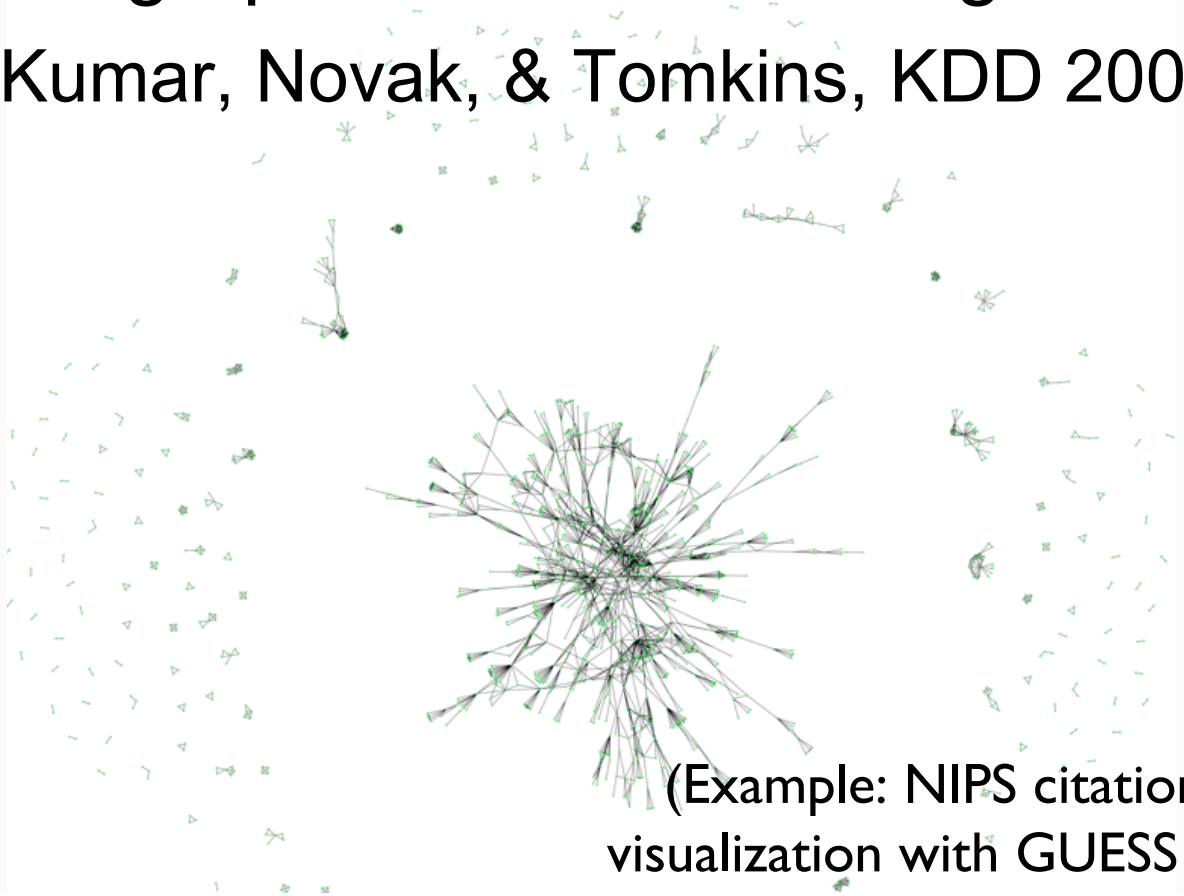
- Basic graph generator, Erdos-Renyi
  - For *n* vertices, connect any two IID with probability *p*.
- Many provable properties, including emergence of a giant connected component.



Real graphs do not have E-R degree distribution, but...

Giant connected component

- Nearly all real networks have a giant connected component (GCC) emerge!
- Often SM graphs have “middle region”
  - See [Kumar, Novak, & Tomkins, KDD 2006]



(Example: NIPS citation graph, visualization with GUESS [Adar 06]

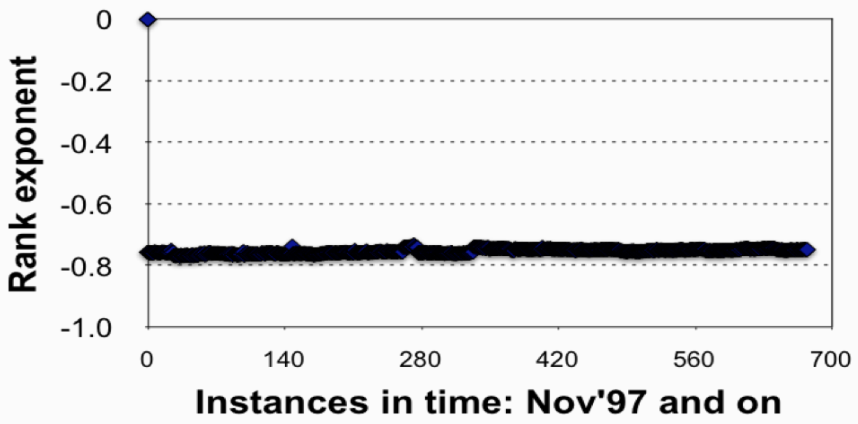
Part 1 Outline

- Introduction to networks
- Patterns
  - Diameter: “small world effect”
  - Degree distribution: power law
  - Connected components: giant CC
  - Evolution over time

Carnegie Mellon

Motivating questions

- How do graphs evolve?
- Degree-exponent seems constant - any other consistent patterns?



Carnegie Mellon

Evolution of diameter?

- Prior analysis, on power-law-like graphs, hints diameter slowly increasing with time.

  diameter ~ O(log(N)) or

  diameter ~ O(log(log(N)))

- Slowly increasing with network size
- What is happening, in reality?

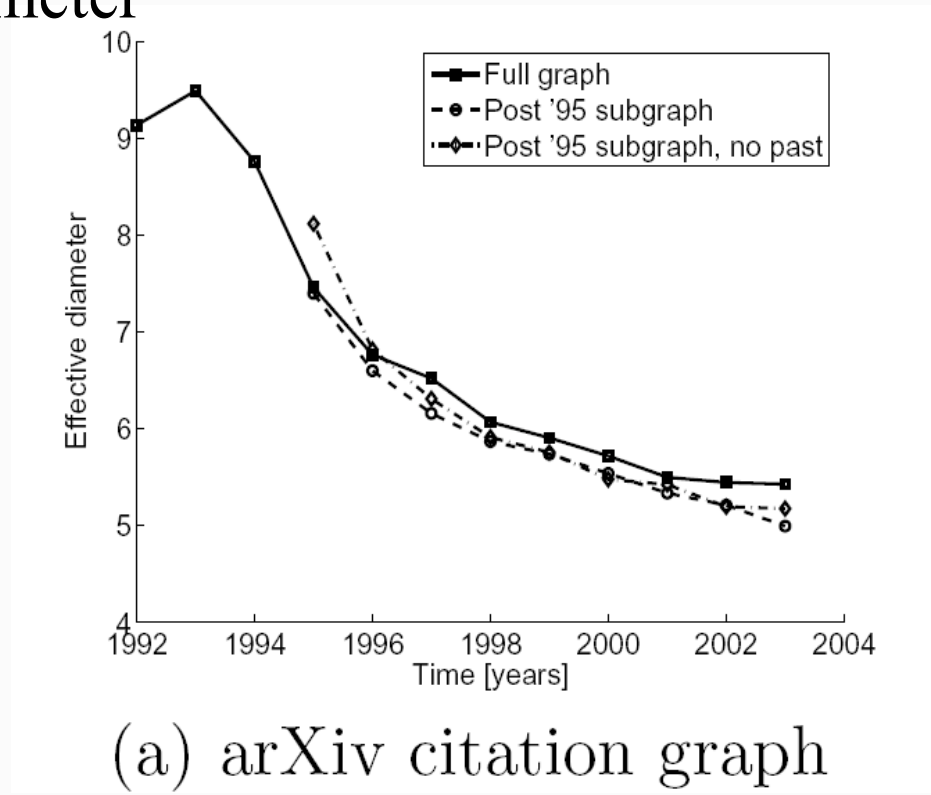
Diameter shrinks, toward a constant value!

Shrinking diameter

[Leskovec, Faloutsos, Kleinberg KDD 2005]

- Citations among physics papers
- 11yrs; @ 2003:
  - 29,555 papers
  - 352,807 citations
- For each month *M*, create a graph of all citations up to month *M*

diameter



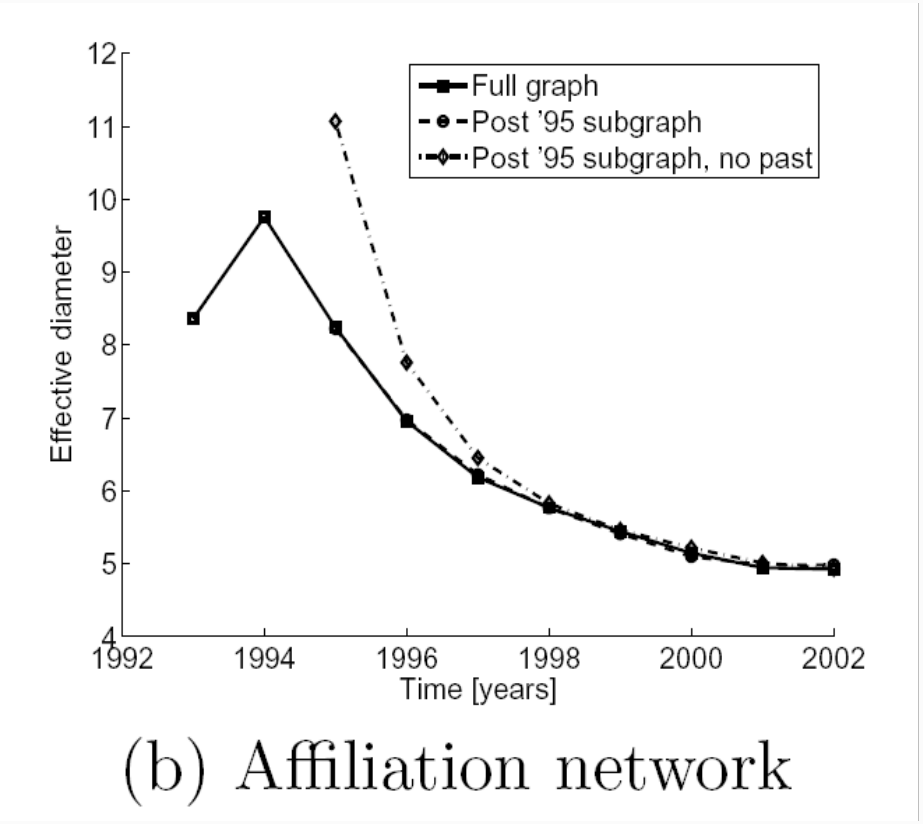
(a) arXiv citation graph

time

Carnegie Mellon

Shrinking diameter

- Authors & publications
- 1992
  - 318 nodes
  - 272 edges
- 2002
  - 60,000 nodes
- 20,000 authors
- 38,000 papers
  - 133,000 edges

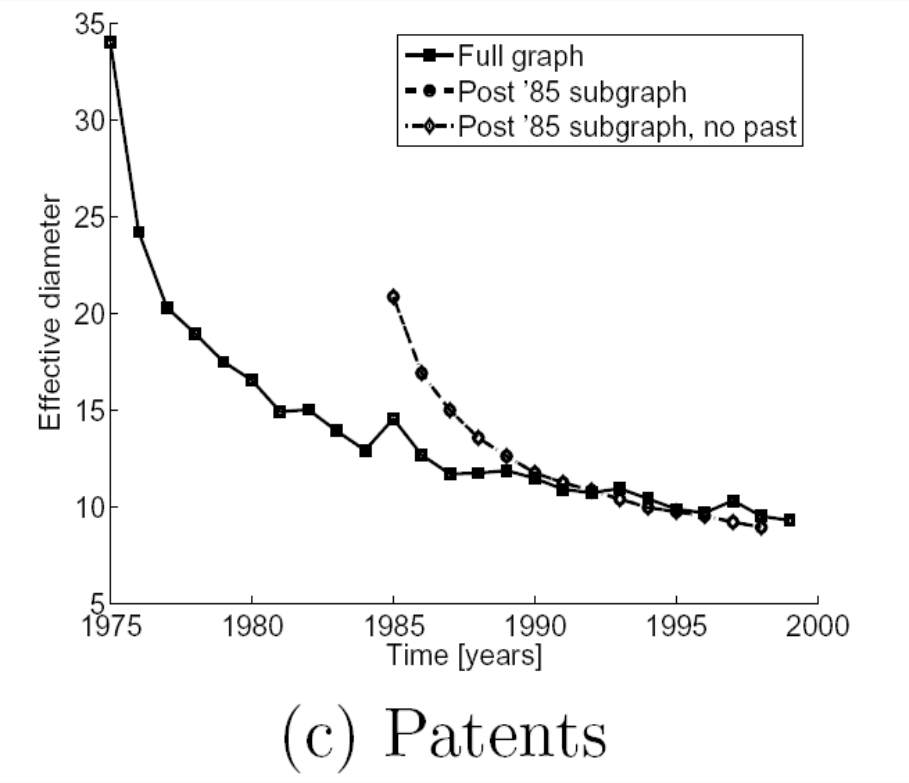


(b) Affiliation network

Carnegie Mellon

Shrinking diameter

- Patents & citations
- 1975
  - 334,000 nodes
  - 676,000 edges
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint



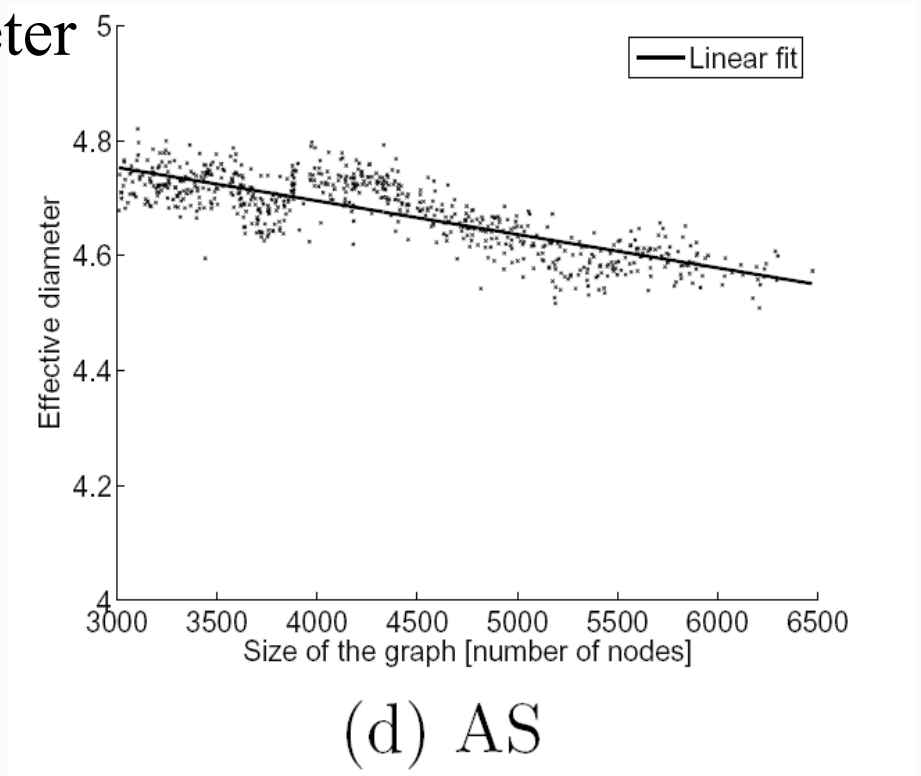
(c) Patents

Carnegie Mellon

Shrinking diameter

- Autonomous systems
- 1997
  - 3,000 nodes
  - 10,000 edges
- 2000
  - 6,000 nodes
  - 26,000 edges
- One graph per day

diameter



(d) AS

N

Temporal evolution

- N(t) nodes; E(t) edges at time t
- Suppose that

$$N(t+1) = 2 * N(t)$$

- What is your guess for

$$E(t+1) =? 2 * E(t)$$

Temporal evolution

- N(t) nodes; E(t) edges at time t
- Suppose that

$$N(t+1) = 2 * N(t)$$

- What is your guess for

$$E(t+1) =? 2 * E(t)$$

- Edges over-double!

Temporal evolution

- Growth of edges obeys power law with:

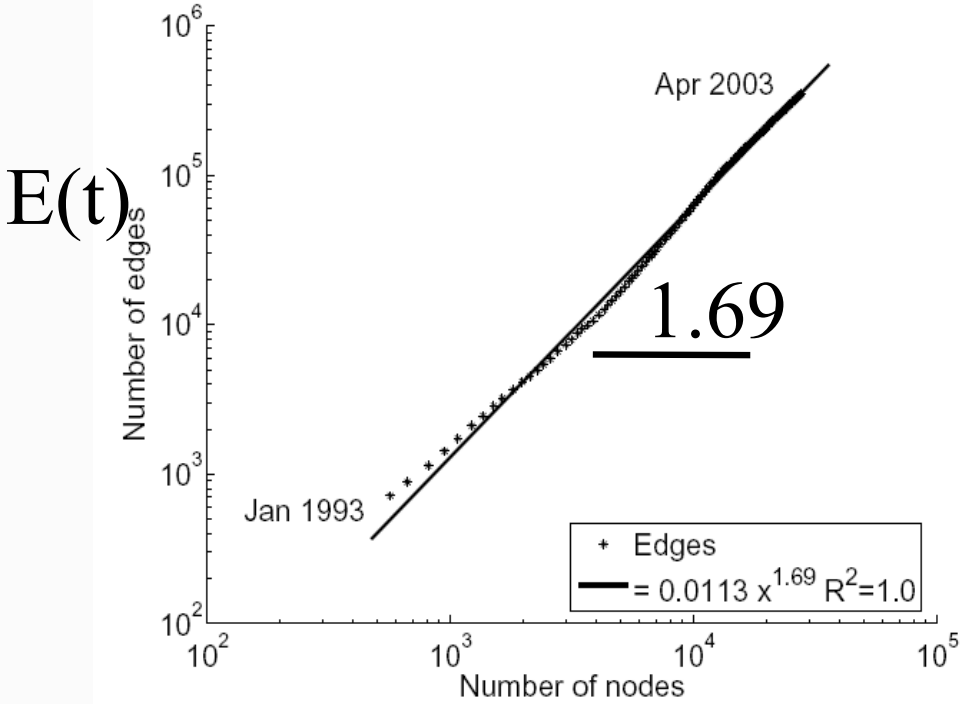
$$E(t) \sim N(t)^a \quad \text{for all } t$$

where $1<a<2$

Carnegie Mellon

Densification Power Law

ArXiv: Physics papers and their citations



E(t)

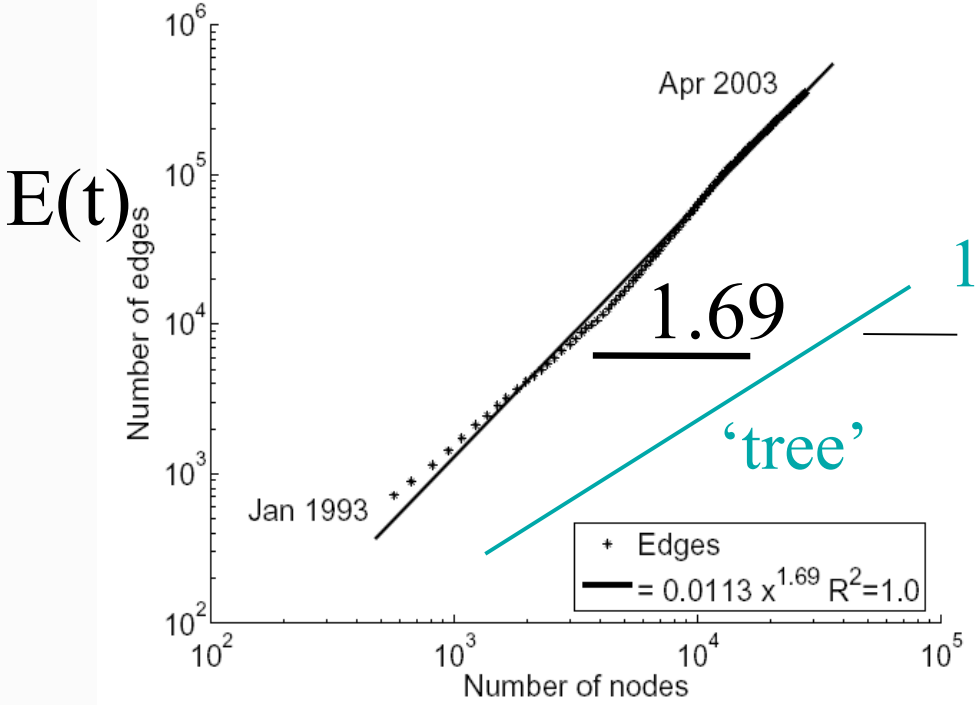
(a) arXiv

N(t)

Carnegie Mellon

Densification Power Law

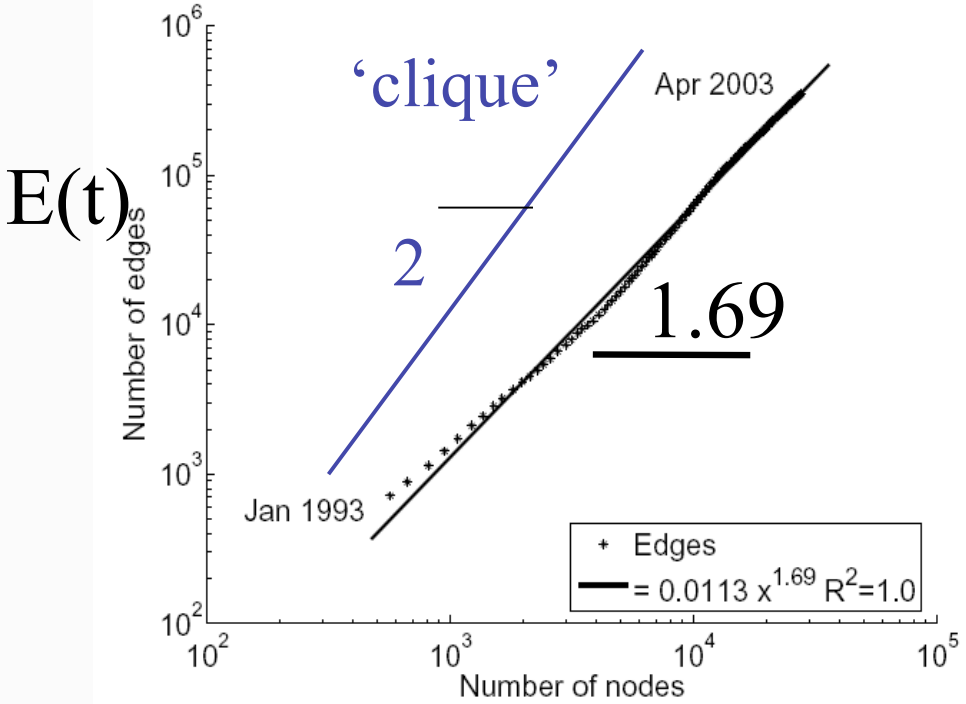
ArXiv: Physics papers and their citations



Carnegie Mellon

Densification Power Law

ArXiv: Physics papers and their citations

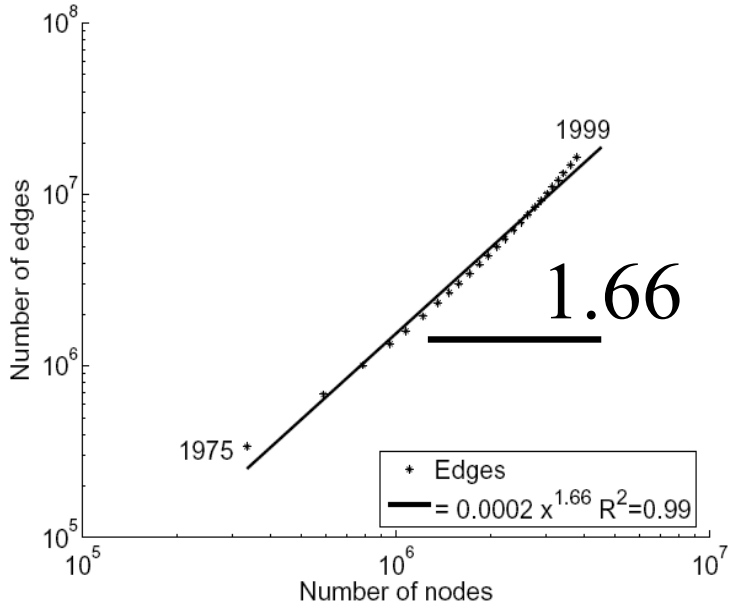


Carnegie Mellon

Densification Power Law

U.S. Patents, citing each other

E(t)



(b) Patents

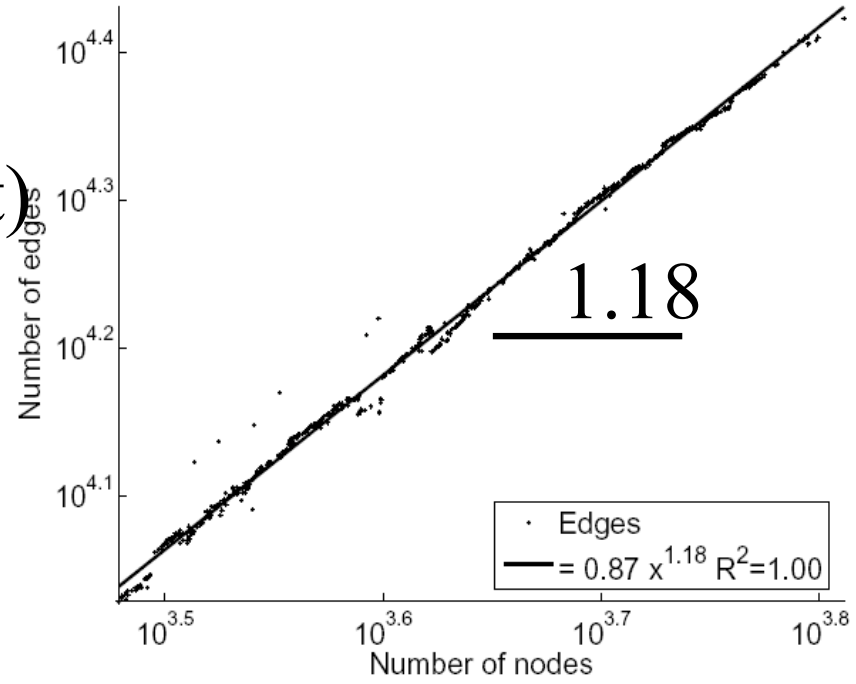
N(t)

Carnegie Mellon

Densification Power Law

Autonomous Systems

E(t)



(c) Autonomous Systems

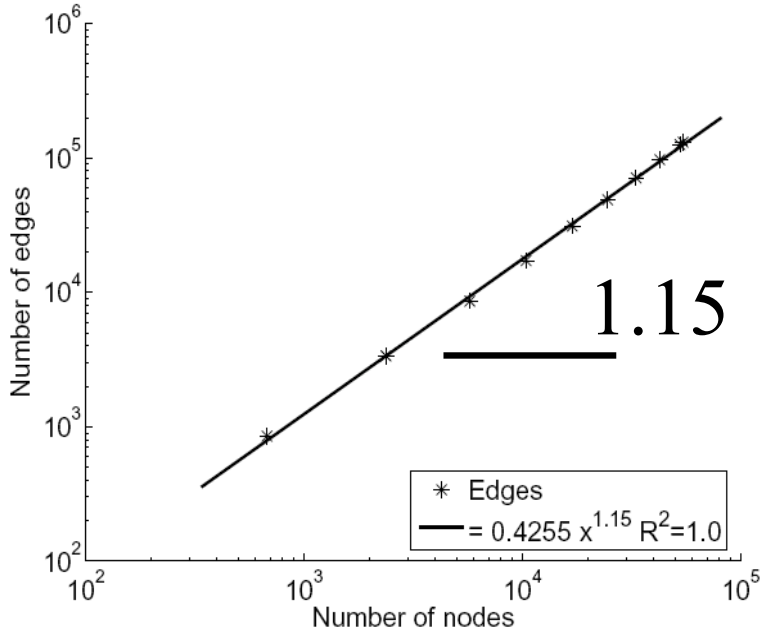
N(t)

Carnegie Mellon

Densification Power Law

ArXiv: authors & papers

E(t)



(d) Affiliation network

N(t)

Carnegie Mellon

Part 1 Outline

- Introduction to networks
- Patterns
  - Diameter: "small world effect"
  - Degree distribution: power law
  - Connected components: giant CC
  - Evolution over time: Shrinking diameter, Densification power law

Another big question

- Q: How can we generate realistic networks?

Another big question

- Q: How can we generate realistic networks?
- A: Answering this question fully would require another tutorial. ☺
- Some models are preferential attachment (Barabasi et. al.), copying model (Kleinberg et. al.), “winners don’t take all” (Pennock et. al.), Kronecker multiplication (Leskovec et. al.).

Bibliography: Part 1

- Power Laws and patterns
  - Albert, R. & Barabasi, A. (2002), 'Statistical mechanics of complex networks', *Reviews of Modern Physics* **74**, 47.
  - Barabasi, A. L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509--512.
  - Domingos, P. & Richardson, M. (2001), Mining the network value of customers, *in* 'KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining', ACM Press, New York, NY, USA, pp. 57--66.
  - Faloutsos, M.; Faloutsos, P. & Faloutsos, C. (1999), 'On Power-law Relationships of the Internet Topology', *SIGCOMM*, 251-262.

Bibliography: Part 1

- Jovanovic, M. A. (2001), 'Modeling Large-scale Peer-to-Peer Networks and a Case Study of Gnutella ', Master's thesis, University of Cincinatti.
- Kossinets, G. & Watts, D. J. (2006), 'Empirical Analysis of an Evolving Social Network', *Science* **311**(5757), 88--90.
- Kumar, R.; Novak, J.; Raghavan, P. & Tomkins, A. (2004), 'Structure and evolution of blogspace', *Commun. ACM* **47**(12), 35--39.
- Kumar, R.; Novak, J.; Raghavan, P. & Tomkins, A. (2003), On the bursty evolution of blogspace, *in* 'WWW '03: Proceedings of the 12th international conference on World Wide Web', ACM Press, New York, NY, USA, pp. 568--576.
- Kumar, R.; Novak, J. & Tomkins, A. (2006), Structure and evolution of online social networks, *in* 'KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowedge Discover and Data Mining', pp. 611--617

Carnegie Mellon

Bibliography: Part 1

- Leland, W.; Taqqu, M.; Willinger, W. & Wilson, D. (1994), 'On the Self-Similar Nature of Ethernet Traffic', *IEEE Transactions on Networking* **2**(1), 1-15.
- Leskovec, J.; Kleinberg, J. & Faloutsos, C. (2005), Graphs over time: densification laws, shrinking diameters and possible explanations, *in* 'KDD '05.
- Milgram, S. (1967), 'The small-world problem', *Psychology Today* **2**, 60--67.
- Newman, M. E. J. (2005), 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323.
- Siganos, G.; Faloutsos, M.; Faloutsos, P. & Faloutsos, C. (2003), 'Power laws and the AS-level internet topology.', *IEEE/ACM Trans. Netw.* **11**(4), 514-524.
- Zipf, G. (1949), *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*, Addison Wesley, Cambridge, Massachusetts.

Bibliography: Part 1

- Models
  - Barabasi, A. (2005), 'The origin of bursts and heavy tails in human dynamics', *Nature* **435**, 207.
  - Erdos, P. & Renyi, A. (1960), 'On the evolution of random graphs', *Publ. Math. Inst. Hungary. Acad. Sci.* **5**, 17-61.
  - Granovetter, M. (1978), 'Threshold Models of Collective Behavior', *Am. Journal of Sociology* **83**(6), 1420--1443.
  - Leskovec, J.; Faloutsos, C. Scalable modeling of real graphs using Kronecker Multiplication. ICML 2007.
  - Vazquez, A.; Oliveira, J. G.; Dezso, Z.; Goh, K. I.; Kondor, I. & Barabasi, A. L. (2006), 'Modeling bursts and heavy tails in human dynamics', *Physical Review E* **73**, 036127.
  - Watts, D. J. & Strogatz, S. H. (1998), 'Collective dynamics of 'small-world' networks.', *Nature* **393**(6684), 440--442.

Bibliography: Part 1

- Visualization example
  - Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006
  - Katharina Anna Lehmann, Stephan Kottler, Michael Kaufmann. Visualizing Large and Clustered Networks. Graph Drawing 2006
  - Touchgraph: http://www.touchgraph.com/TGFacebookBrowser.html

Carnegie Mellon