Graph Models

Class Algorithmic Methods of Data Mining
Program M. Sc. Data Science
University Sapienza University of Rome
Semester Fall 2015
Lecturer Carlos Castillo http://chato.cl/

Sources:

- Frieze, Gionis, Tsourakakis: “Algorithmic techniques for modeling and mining large graphs (AMAzING)” [Tutorial]
- Lada Adamic: Zipf, Power-Laws and Pareto [Tutorial]
- Giorgios Cheliotis: Social Network Analysis [Tutorial]

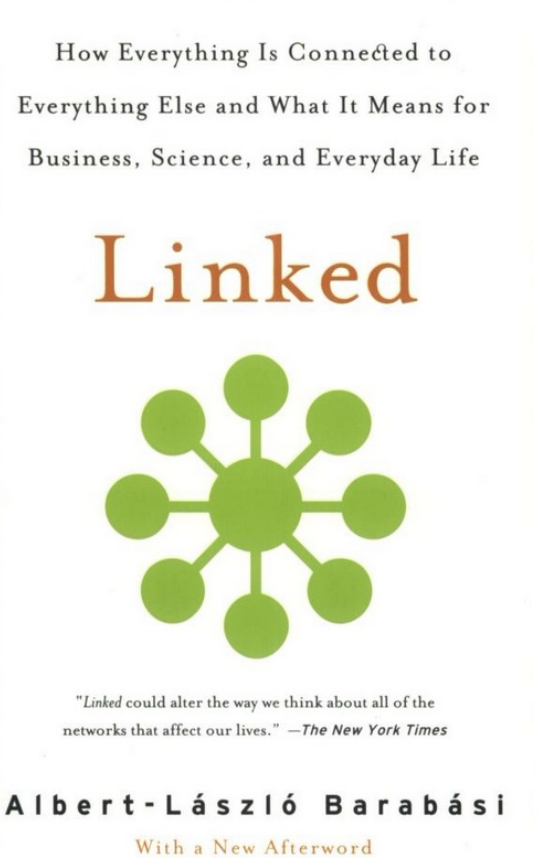
Network characteristics

- Static networks
  - Power-law degree distribution
  - Small diameter
- Time-evolving networks
  - Densification
  - Shrinking diameters

Heavy Tails

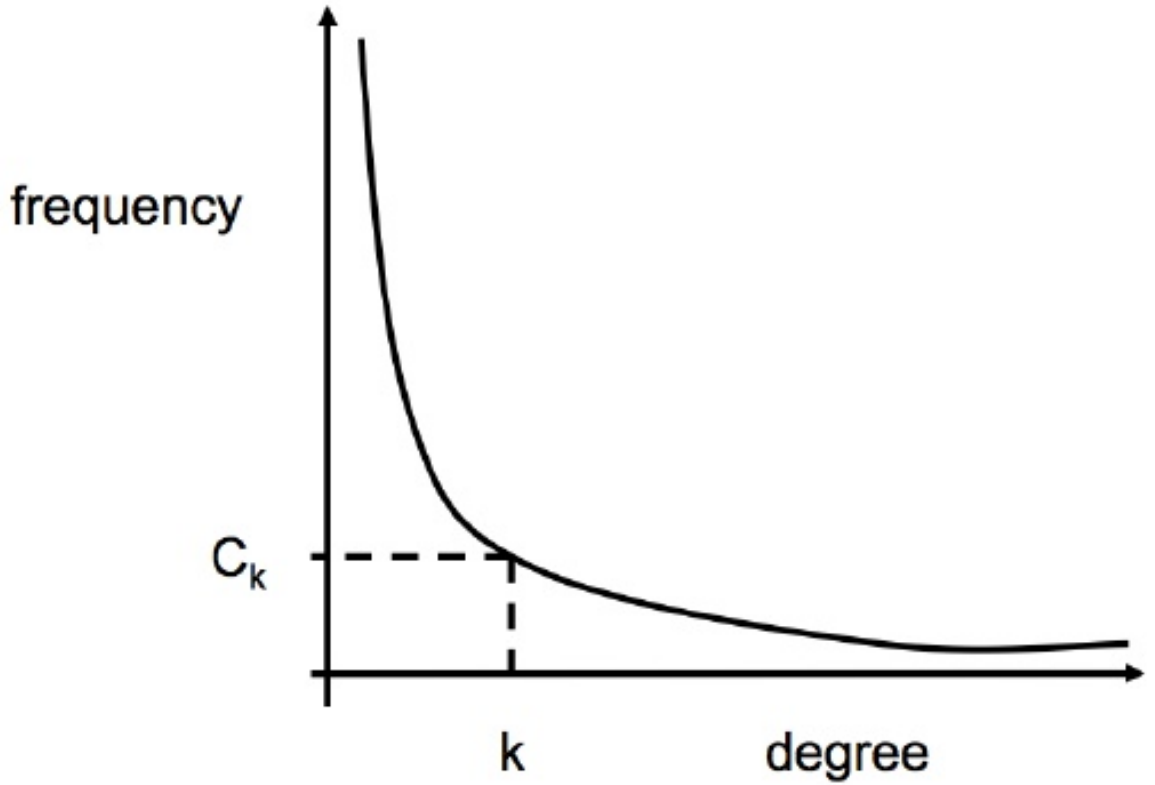
What do the proteins in our bodies, the Internet, a cool collection of atoms and sexual networks have in common? One man thinks he has the answer and it is going to transform the way we view the world.

New Scientist, 2002



Degree distribution

- C_k = number of elements with degree k



Power-law degree distribution

$$\mathrm{C}_k = ck^{-\gamma}$$

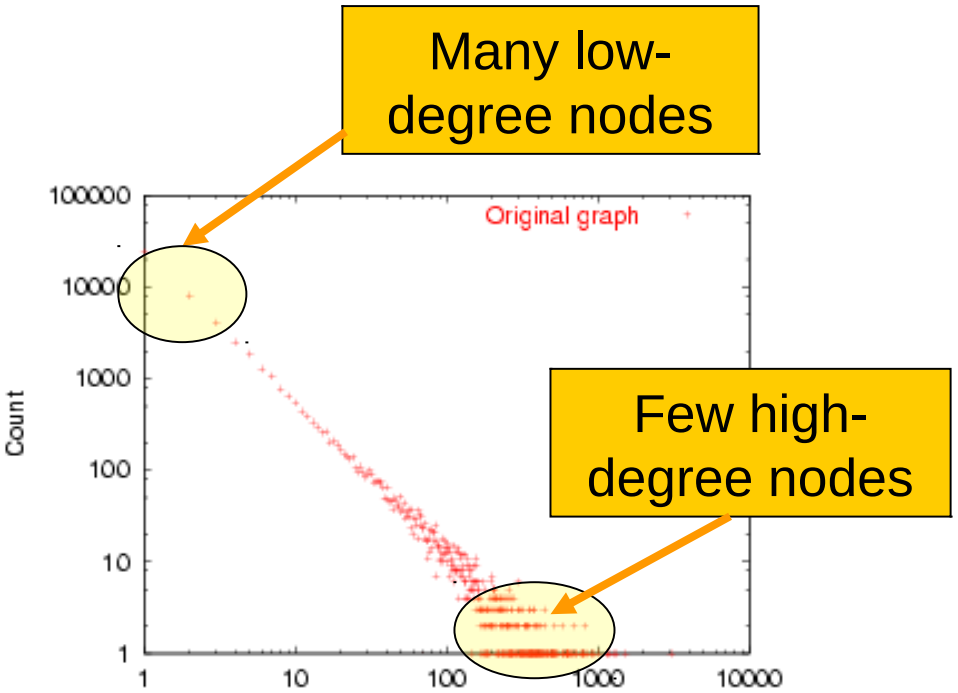
$$\log(C_k) = \log(c) - \gamma log(k)$$

- In a log-log plot, this looks like a straight line, descending with slope γ

Power-law degree distribution

- Power Law

Internet in December 1998

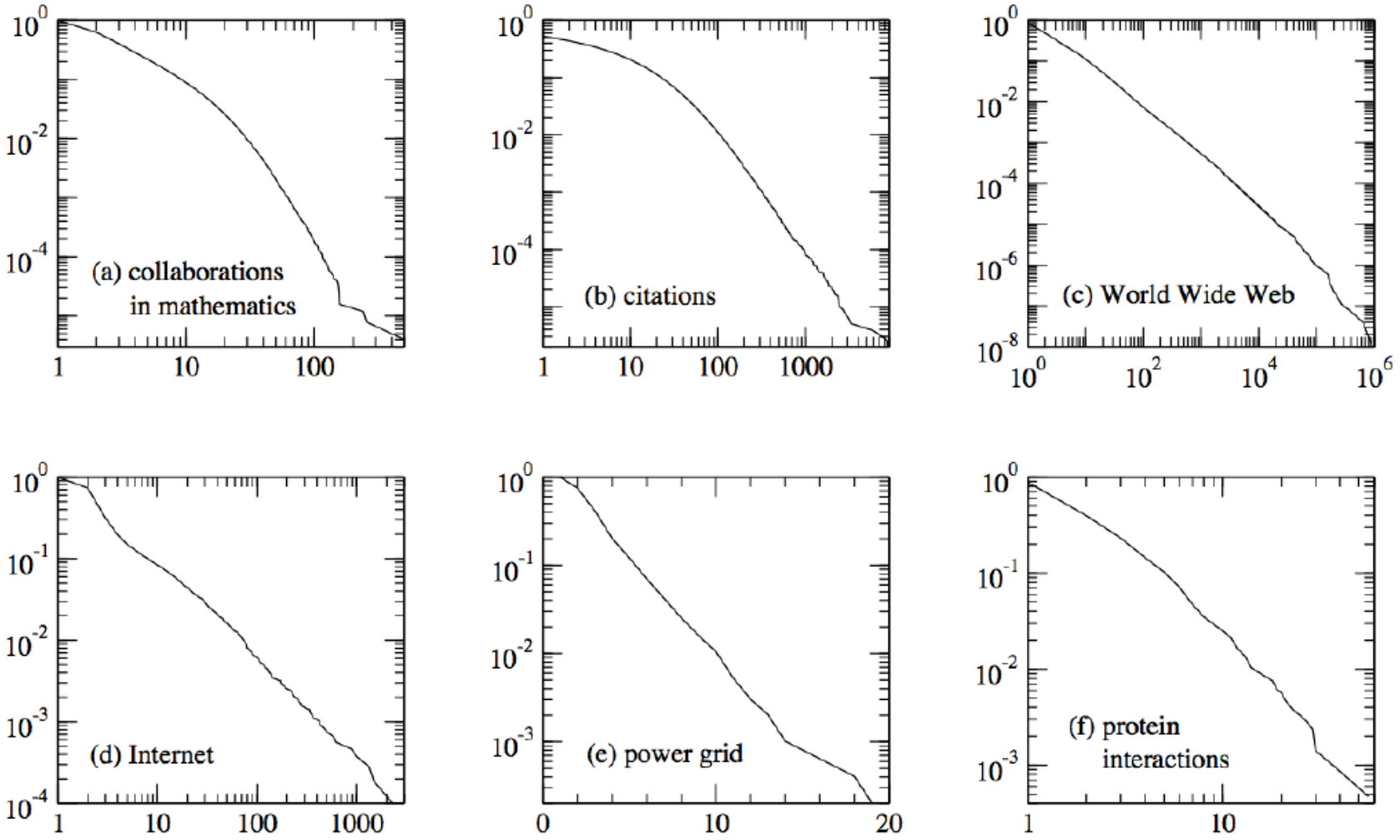


log(Count) vs. log(Degree)

$$C_k = ck^{-\gamma}$$

https://cs.stanford.edu/people/jure/pubs/powergrowth-kdd05.ppt

Power law appear in a wide variety of networks



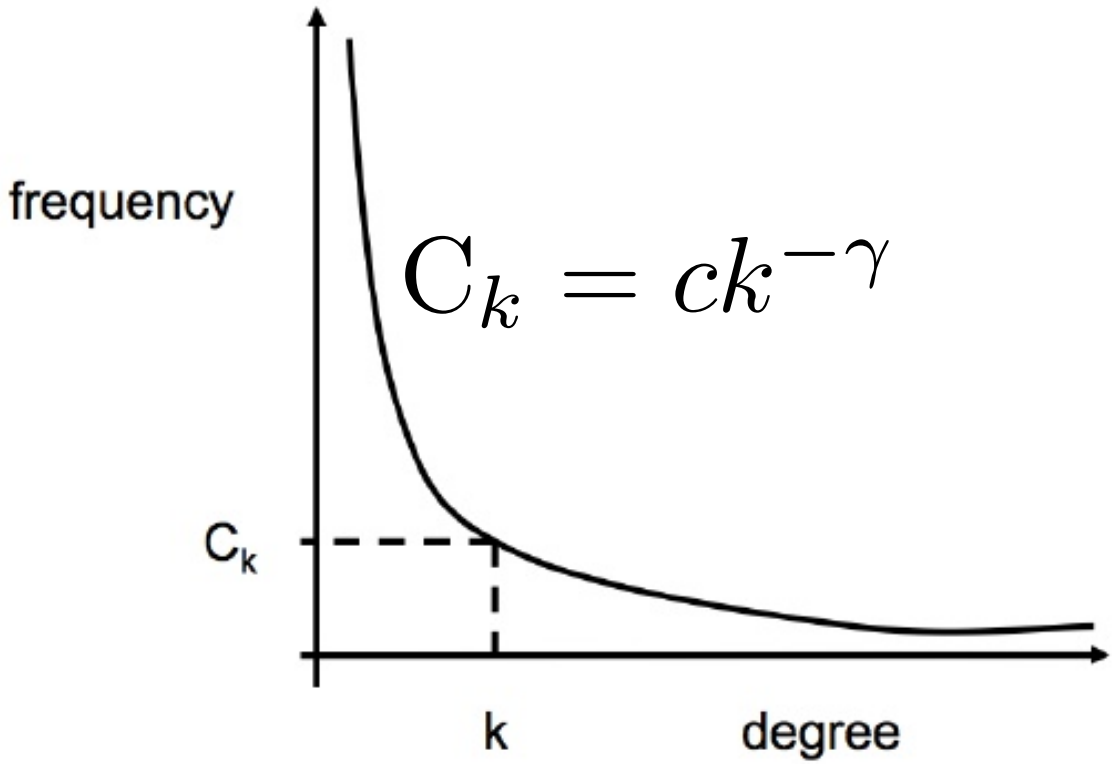
Example power-law networks

- Social networks
  - collaboration of movie actors in films
  - co-authorship by mathematicians of papers
- Internet router
- Web graphs
- Interbank payment networks
- Protein-protein interaction networks
- Semantic networks
- Airline networks

https://en.wikipedia.org/wiki/Scale-free_network#Examples

Try it!

- Create a dataset of numbers having a power-law degree distribution with gamma=1 (or show how to construct one)
- Create a graph having this sequence of degrees



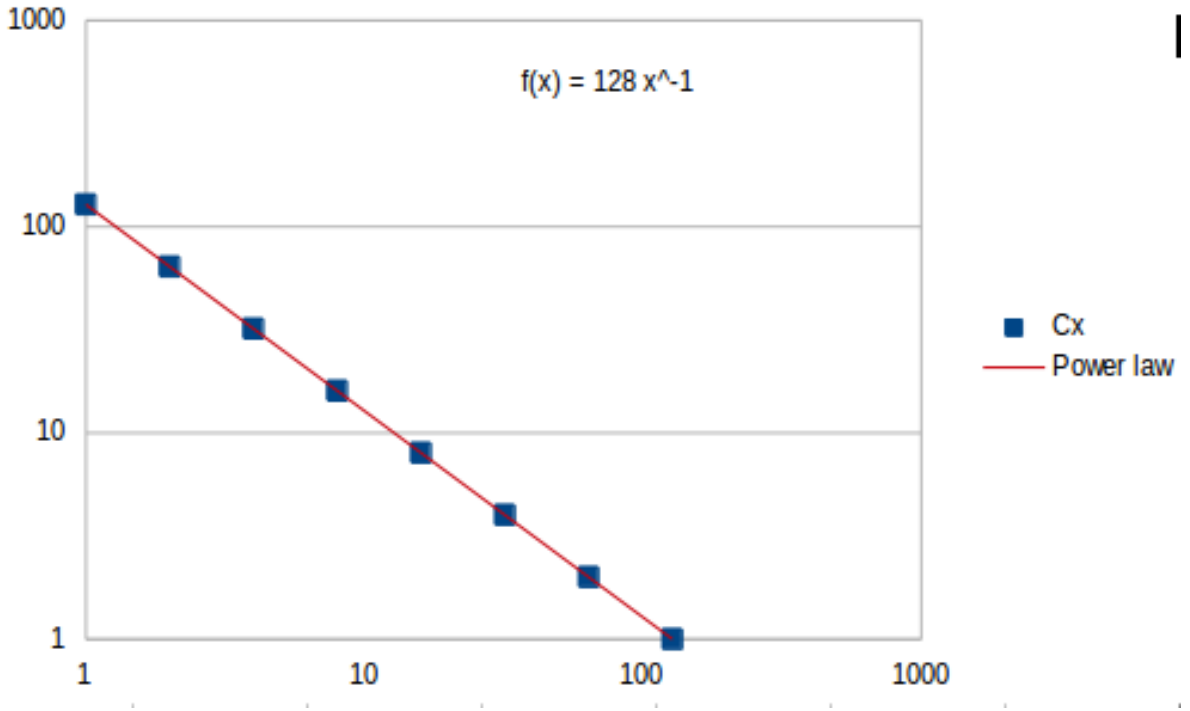
Estimating the exponent

- Option 1: draw the log-log plot and fit a line using least squares
- Option 2: Hill's estimator

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{x_i}{x_{min}} \right)$$

Example (in OpenOffice Calc)

x	Cx
1	128
2	64
4	32
8	16
16	8
32	4
64	2
128	1

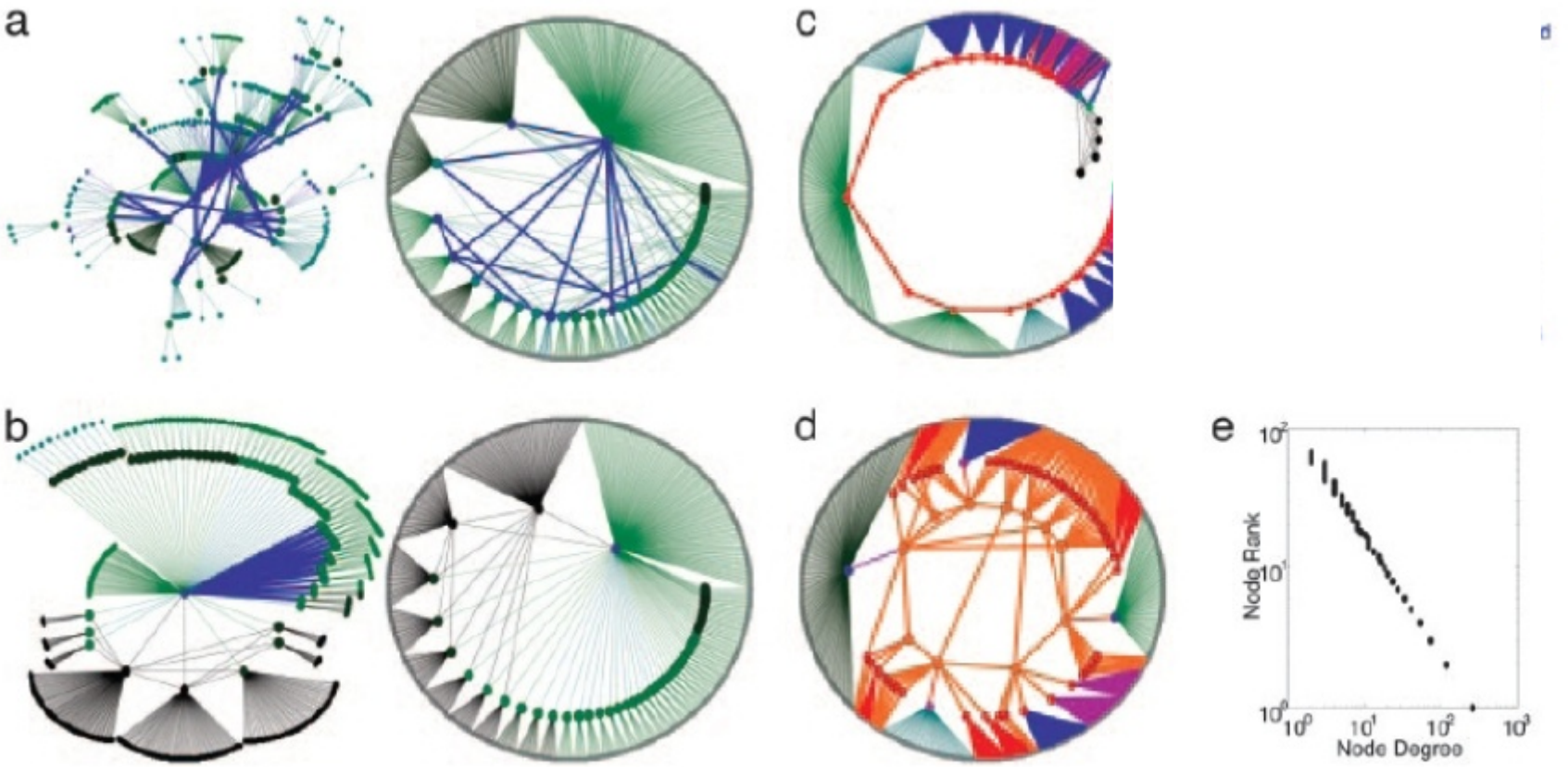


Actual value = 1.0
Hill's estimator = 1.31 in this case

http://chato.cl/2015/data_analysis/09_graph_models/hills-estimator-example.ods

Degree distribution is important but obviously isn't everything

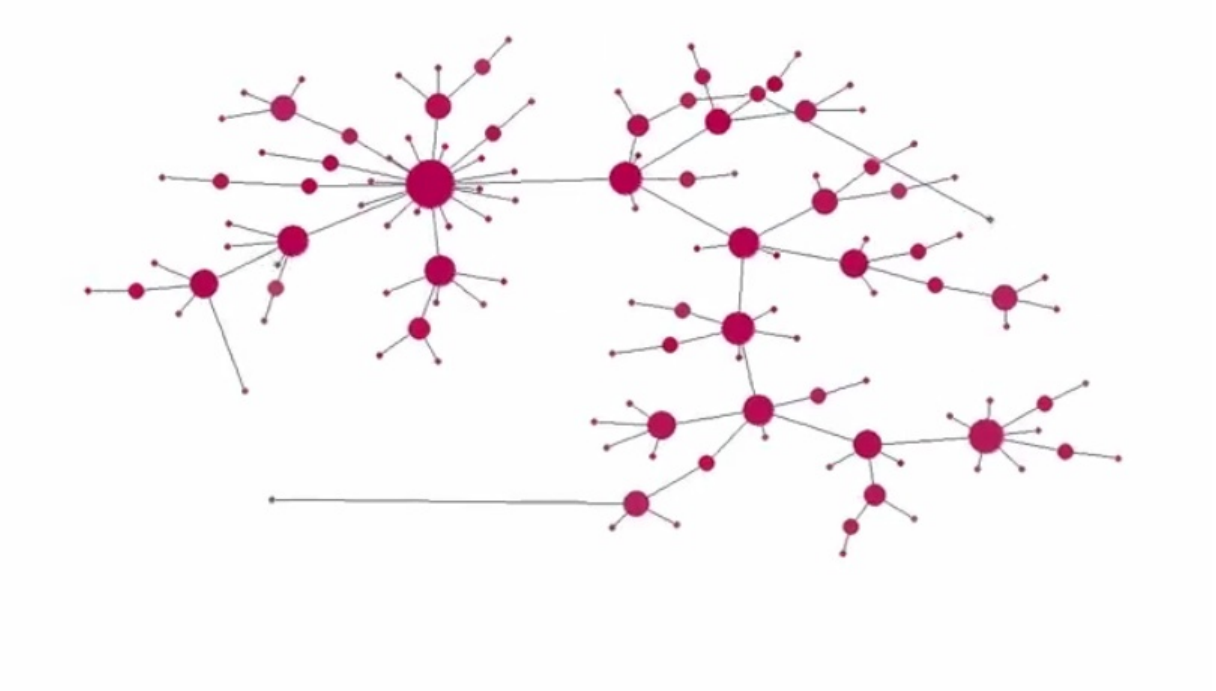
- All these graphs have the same number of nodes and degree sequence



How to obtain a power law?

- **Preferential attachment** is a frequently-used model for graph evolution
- At every time step, a new node arrives and **connects to existing nodes with probability proportional to their degree**

Example



https://www.youtube.com/watch?v=4GDqJVtPEGg

The copy model of graph evolution

Suggested as a model of WWW growth. Input: parameter α

At each timestep:

- create a new vertex t + 1
- choose an existing vertex u $\in V_{t-1}$ uniformly at random
- the i-th out-link of t + 1 is chosen as follows:
  - with probability α we select x $\in V_{t-1}$ uniformly at random, and
  - with probability $1 - \alpha$ it copies the i-th out-link of *u*

Produces power-law distribution AND a large number of bipartite cliques

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E.: Stochastic models for the web graph. FOCS 2000.

Why preferential attachment?

Popularity

We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective, measurable characteristics

Also known as 'the rich get richer'

Quality

We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention, faster

Also known as 'the good get better'

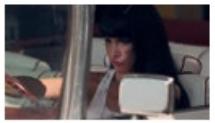
Mixed model

Among nodes of similar attributes, those that reach critical mass first will become 'stars' with many friends and followers ('halo effect')

May be impossible to predict who will become a star, even if quality matters

http://www.slideshare.net/gcheliotis/social-network-analysis-3273045

Why does this happen?

#1 *		**Ti Ho Voluto Bene Veramente** | Marco Mengoni
#2 ↓		**21 Grammi** | Fedez
#3 =		**What Do You Mean?** | Justin Bieber
#4 ↑		**Beautiful Disaster** | Fedez
#5 ↑		**La Vita Com'è** | Max Gazze
#6 ↓		**Roma - Bangkok (feat. Giusy Ferreri)** | Giusy Ferreri

MTV Italia 31-Oct-2015

An experiment

- *Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science 10 February 2006. Vol. 311 no. 5762 pp. 854-856 [link]*

Experimental conditions

	# of down loads	[Help] [Log off]	# of down loads		# of down loads
HARTSFIELD: "enough is enough"	20	GO MOREDCAI: "it does what its told"	12	UNDO: "while the world passes"	24
DEEP ENOUGH TO DIE: "for the sky"	17	PARKER THEORY: "she said"	47	UP FOR NOTHING: "in sight of"	13
THE THRIFT SYNDICATE: "2003 a tragedy"	20	MISS OCTOBER: "pink agression"	27	SILVERFOX: "gnaw"	17
THE BROKEN PROMISE: "the end in friend"	19	POST BREAK TRAGEDY: "florence"	14	STRANGER: "one drop"	10
THIS NEW DAWN: "the belief above the answer"	12	FORTHFADING: "fear"	24	FAR FROM KNOWN: "route 9"	18
NOONER AT NINE: "walk away"	6	THE CALEFACTION: "trapped in an orange peel"	20	STUNT MONKEY: "inside out"	46
MORAL HAZARD: "waste of my life"	8	52METRO: "lockdown"	17	DANTE: "lifes mystery"	14
NOT FOR SCHOLARS: "as seasons change"	27	SIMPLY WAITING: "went with the count"	16	FADING THROUGH: "wish me luck"	10
SECRETARY: "keep your eyes on the ballistics"	5	STAR CLIMBER: "tell me"	38	UNKNOWN CITIZENS: "falling over"	34
ART OF KANLY: "seductive intro, melodic breakdown"	10	THE FASTLANE: "til death do us part (i dont)"	31	BY NOVEMBER: "if i could take you"	20
HYDRAULIC SANDWICH: "separation anxiety"	20	A BLINDING SILENCE: "miseries and miracles"	17	DRAWN IN THE SKY: "tap the ride"	12
EMBER SKY: "this upcoming winter"	25	SUM RANA: "the bolshevik boogie"	15	SELSIUS: "stars of the city"	22
SALUTE THE DAWN: "i am error"	13	CAPE RENEWAL: "baseball warlock v1"	12	SIBRIAN: "eye patch"	14
RYAN ESSMAKER: "detour_(be still)"	14	UP FALLS DOWN: "a brighter burning star"	11	EVAN GOLD: "robert downey jr"	10
BEERBONG: "father to son"	12	SUMMERSWASTED: "a plan behind destruction"	17	BENEFIT OF A DOUBT: "run away"	38
HALL OF FAME: "best mistakes"	19	SILENT FILM: "all i have to say"	61	SHIPWRECK UNION: "out of the woods"	16

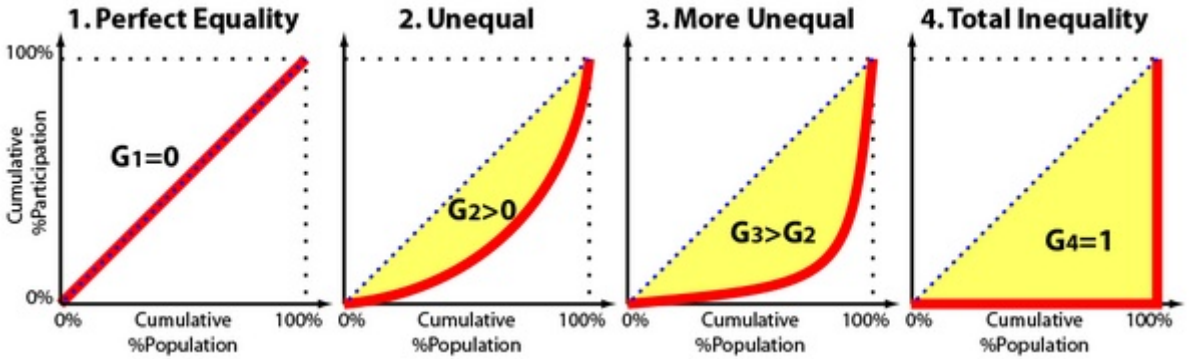
Experiment 1: random order with number of downloads

[Help] [Log off]	# of down loads
PARKER THEORY: "she said"	159
THE FASTLANE: "til death do us part (i dont)"	103
SELSIUS: "stars of the city"	62
STUNT MONKEY: "inside out"	56
BY NOVEMBER: "if i could take you"	55
FORTHFADING: "fear"	49
HYDRAULIC SANDWICH: "separation anxiety"	43
SILENT FILM: "all i have to say"	40
UNDO: "while the world passes"	36
BENEFIT OF A DOUBT: "run away"	32
A BLINDING SILENCE: "miseries and miracles"	27
MISS OCTOBER: "pink agression"	26
STAR CLIMBER: "tell me"	24
FAR FROM KNOWN: "route 9"	22
HALL OF FAME: "best mistakes"	21
EMBER SKY: "this upcoming winter"	19

Experiment 2: sorted by descending number of downloads

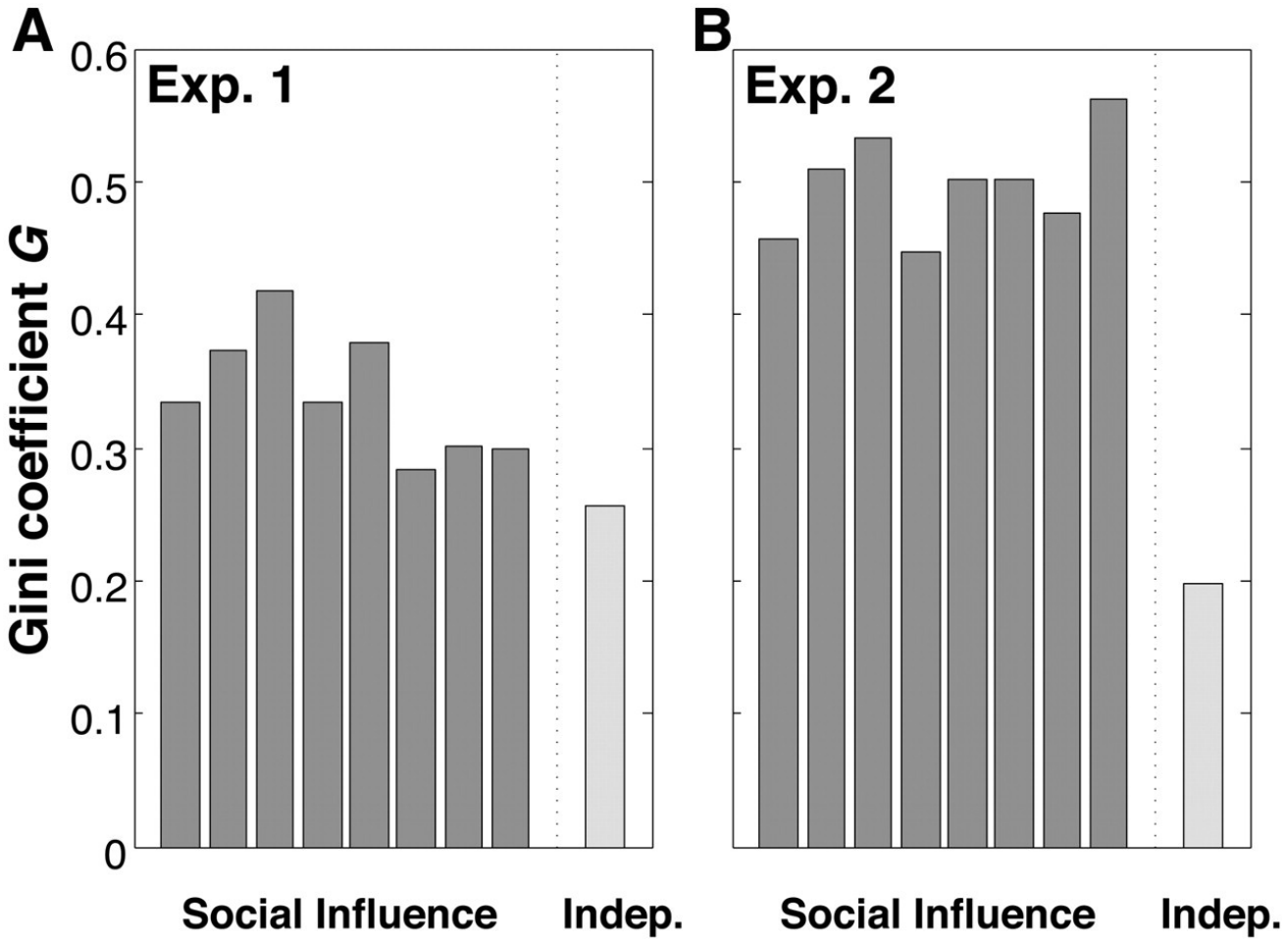
Control: random order *without* number of downloads

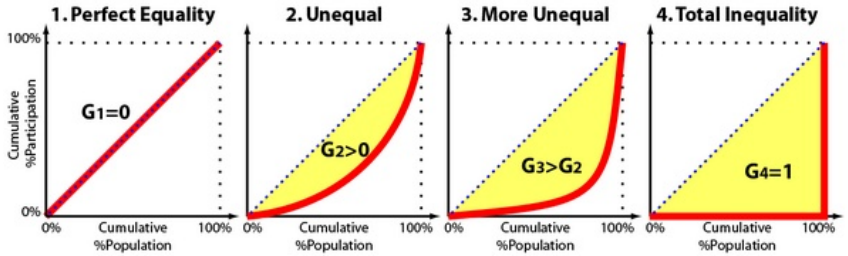
Evaluation metric: Gini Coefficient



Namibia 0.61, Chile 0.50, US 0.41, Italy 0.36, Spain 0.34, Norway 0.25

Results


A
Exp. 1
B
Exp. 2
Gini coefficient G
0.6
0.5
0.4
0.3
0.2
0.1
0
Social Influence
Indep.
Social Influence
Indep.

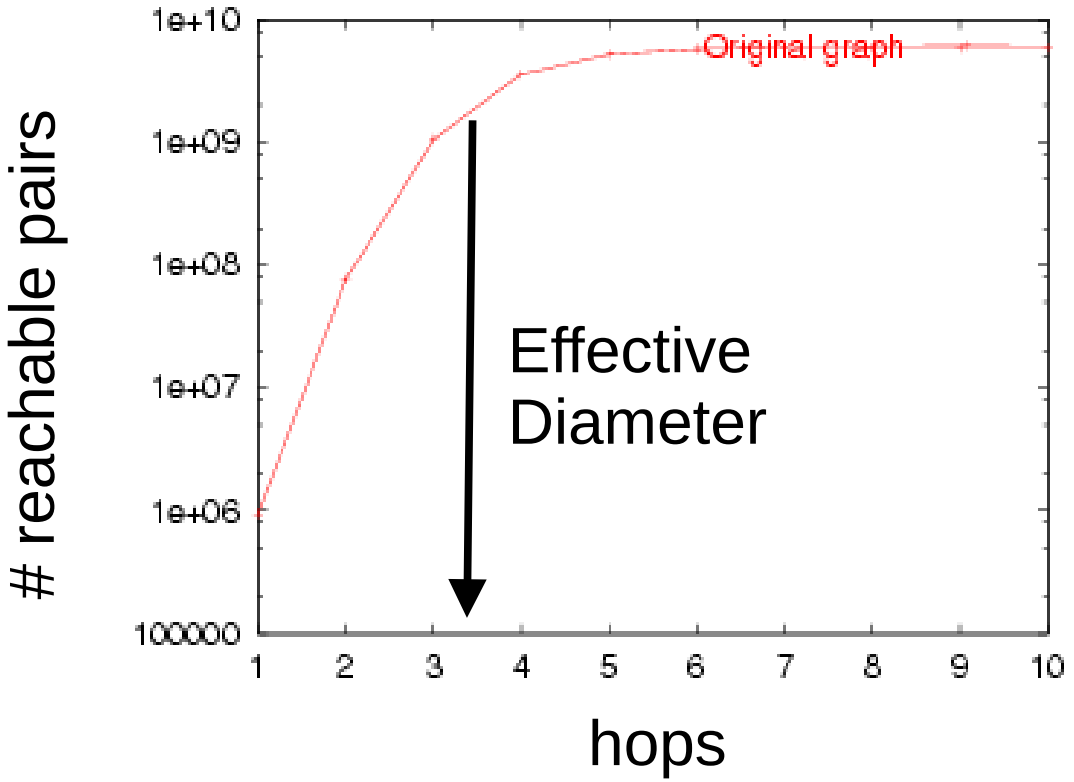

1. Perfect Equality
2. Unequal
3. More Unequal
4. Total Inequality
100%
0%
Cumulative %Participation
G1=0
G2>0
G3>G2
G4=1
0%
Cumulative %Population
100%

Graph diameter

Ways to characterize diameter

- *diameter:* largest shortest-path over all pairs.
- *effective diameter:* upper bound of the shortest path of 90% of the pairs of vertices.
- *average shortest path :* average of the shortest paths over all pairs of vertices.
- *characteristic path length :* median of the shortest paths over all pairs of vertices.
- *hop-plots :* plot of |Nh(u)|, the number of neighbors of u at distance at most h, as a function of h [Faloutsos et al., 1999]

Effective diameter

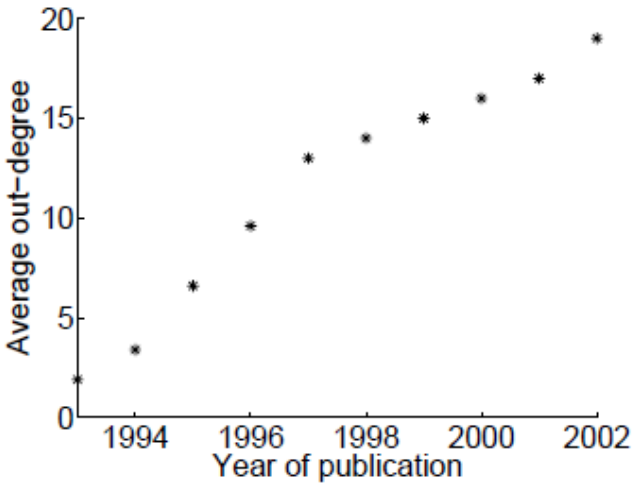


https://cs.stanford.edu/people/jure/pubs/powergrowth-kdd05.ppt

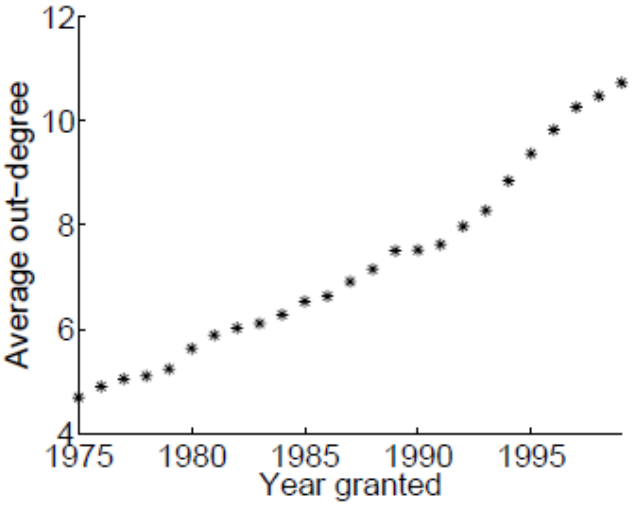
Temporal evolution of graphs

- *Jure Leskovec, Jon Kleinberg, and Christos Faloutsos: “Graphs over time: densification laws, shrinking diameters and possible explanations.” In KDD 2005. [DOI][Slides]*
- Two main findings:
  - Diameter tends to shrink
  - Average degree tends to increase

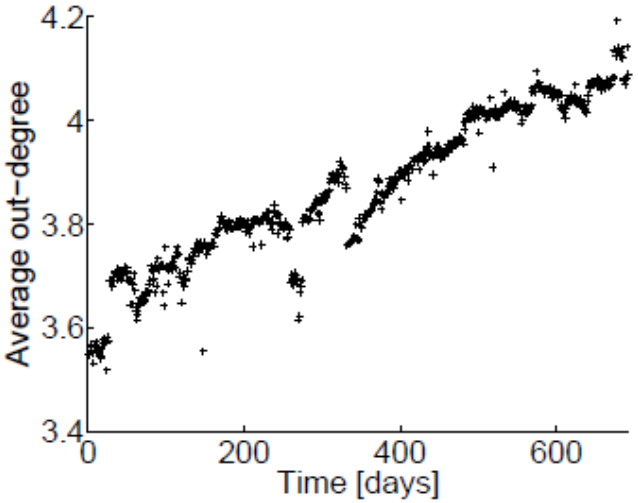
Average out-degree



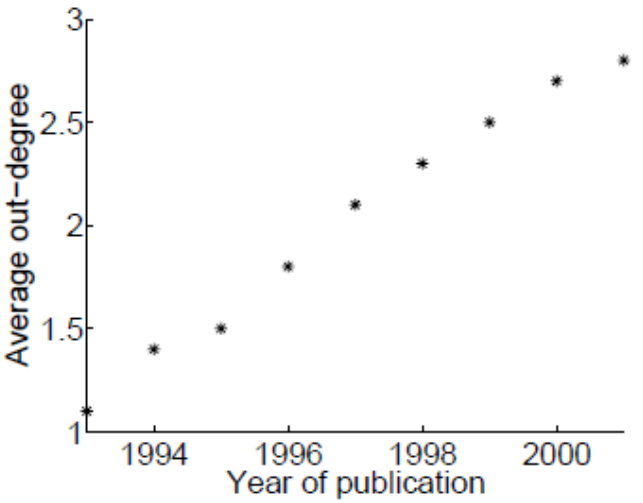
(a) arXiv



(b) Patents



(c) Autonomous Systems



(d) Affiliation network

Temporal Evolution of the Graphs

- Densification Power Law
  - networks are becoming denser over time
  - the number of edges grows faster than the number of nodes – average degree is increasing

$$E(t) \propto N(t)^a$$

or equivalently $\frac{\log(E(t))}{\log(N(t))} = const$

a … densification exponent

Slides on this section from: https://cs.stanford.edu/people/jure/pubs/powergrowth-kdd05.ppt

Graph Densification – A closer look

- Densification Power Law

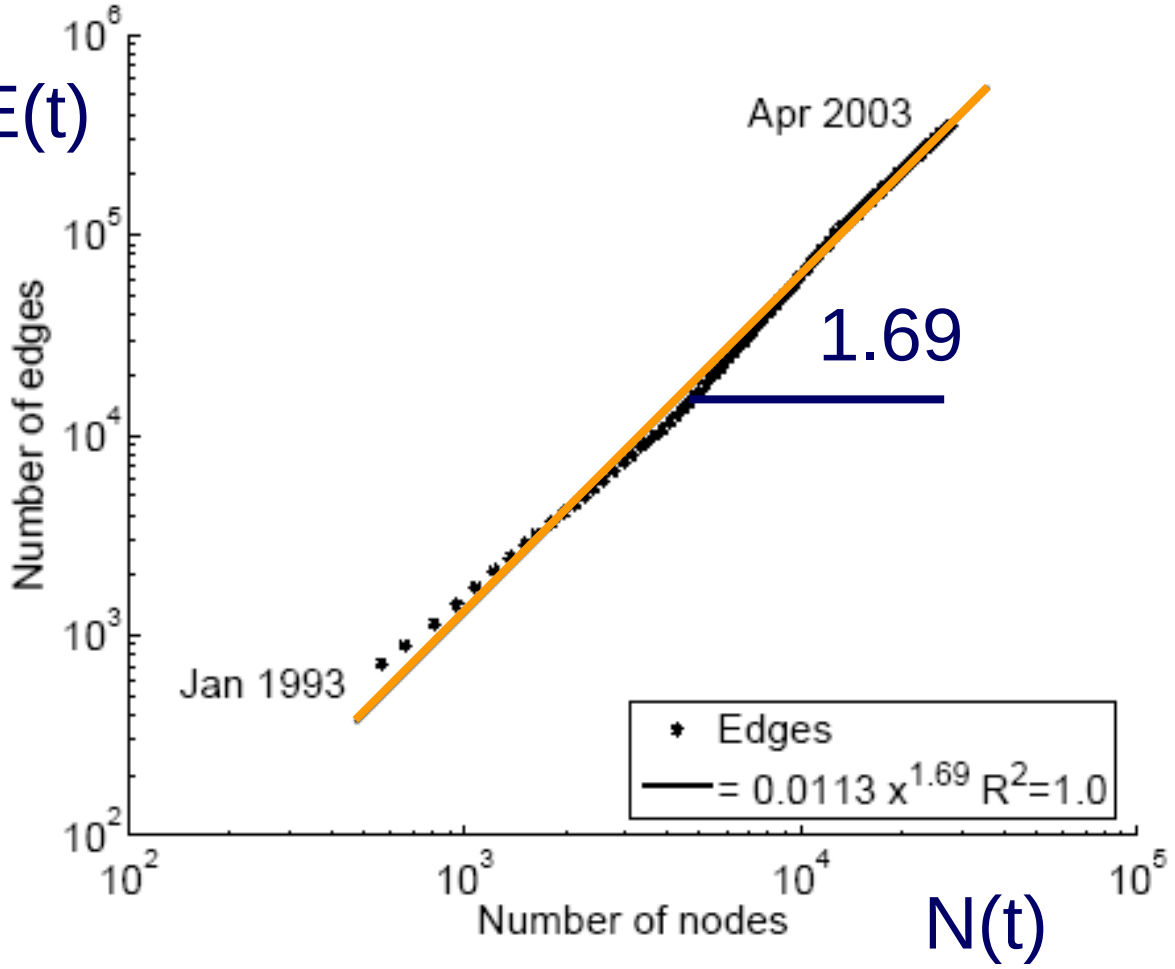
$$E(t) \propto N(t)^a$$

- Densification exponent: $1 \leq a \leq 2$:
  - a=1: linear growth – constant out-degree (assumed in the literature so far)
  - a=2: quadratic growth – clique

- Let's see the real graphs!

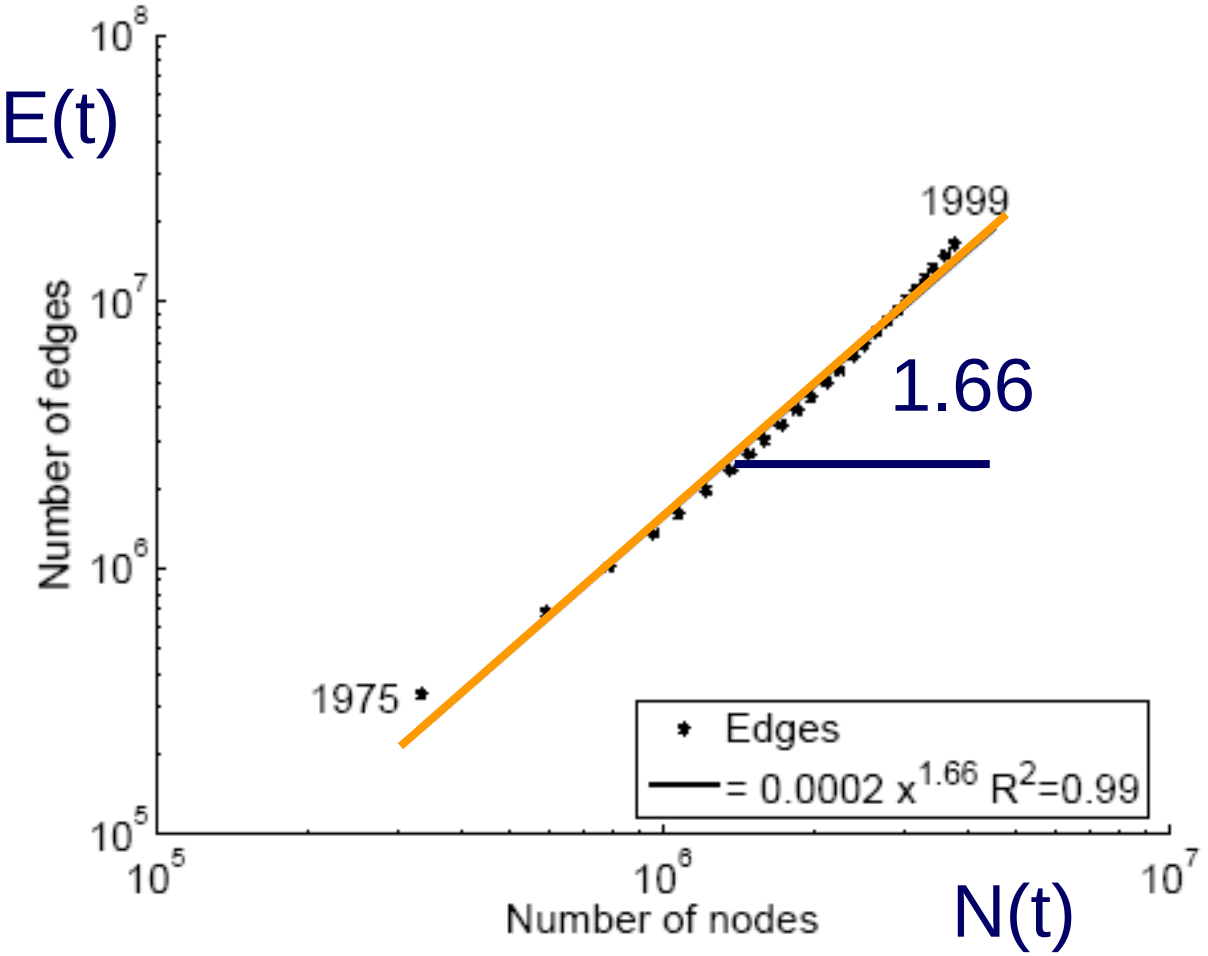
Densification – Physics Citations

- Citations among physics papers
- 1992:
  - 1,293 papers, 2,717 citations
- 2003:
  - 29,555 papers, 352,807 citations
- For each month *M*, create a graph of all citations up to month *M*



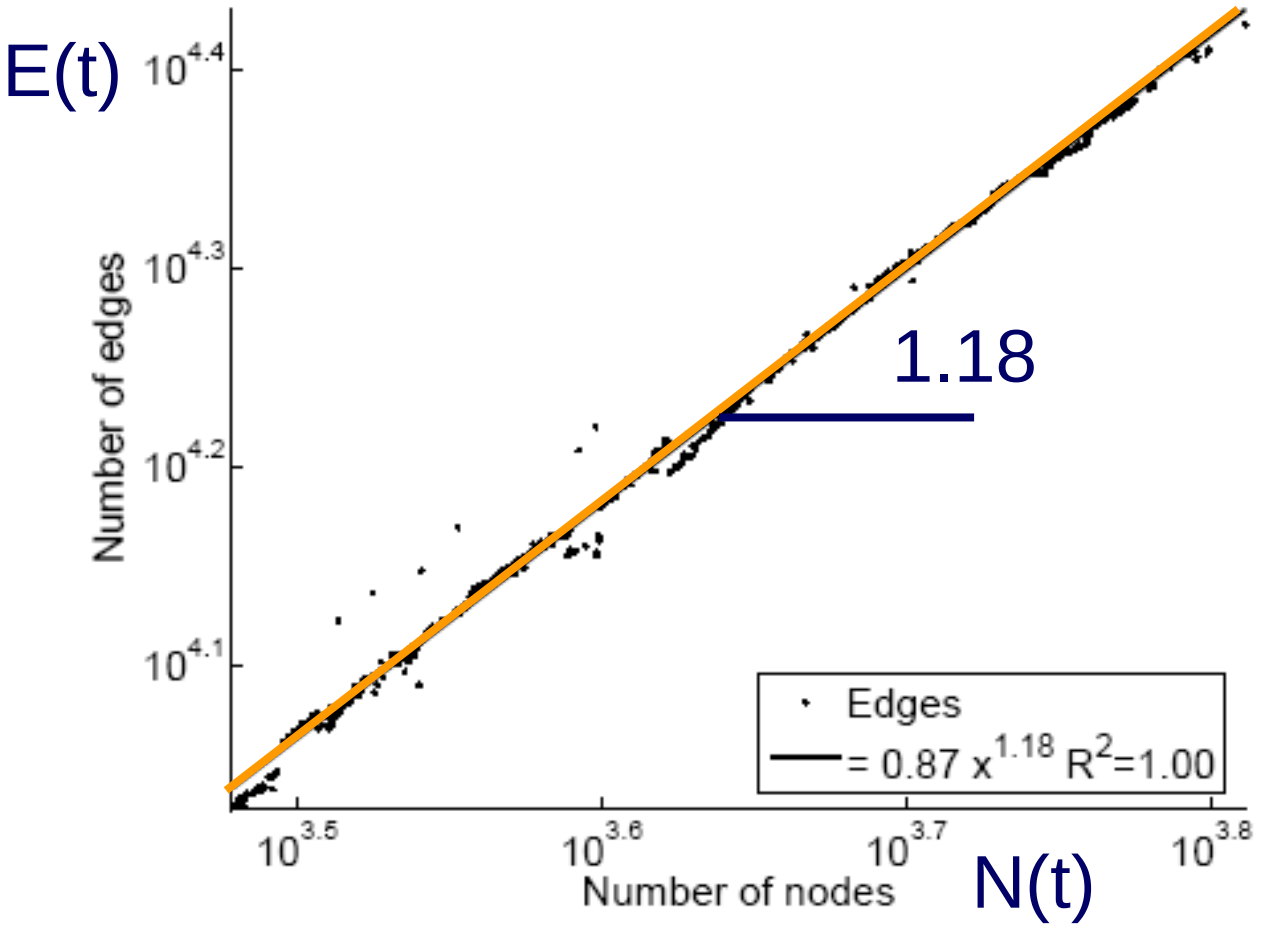
Densification – Patent Citations

- Citations among patents granted
- 1975
  - 334,000 nodes
  - 676,000 edges
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint



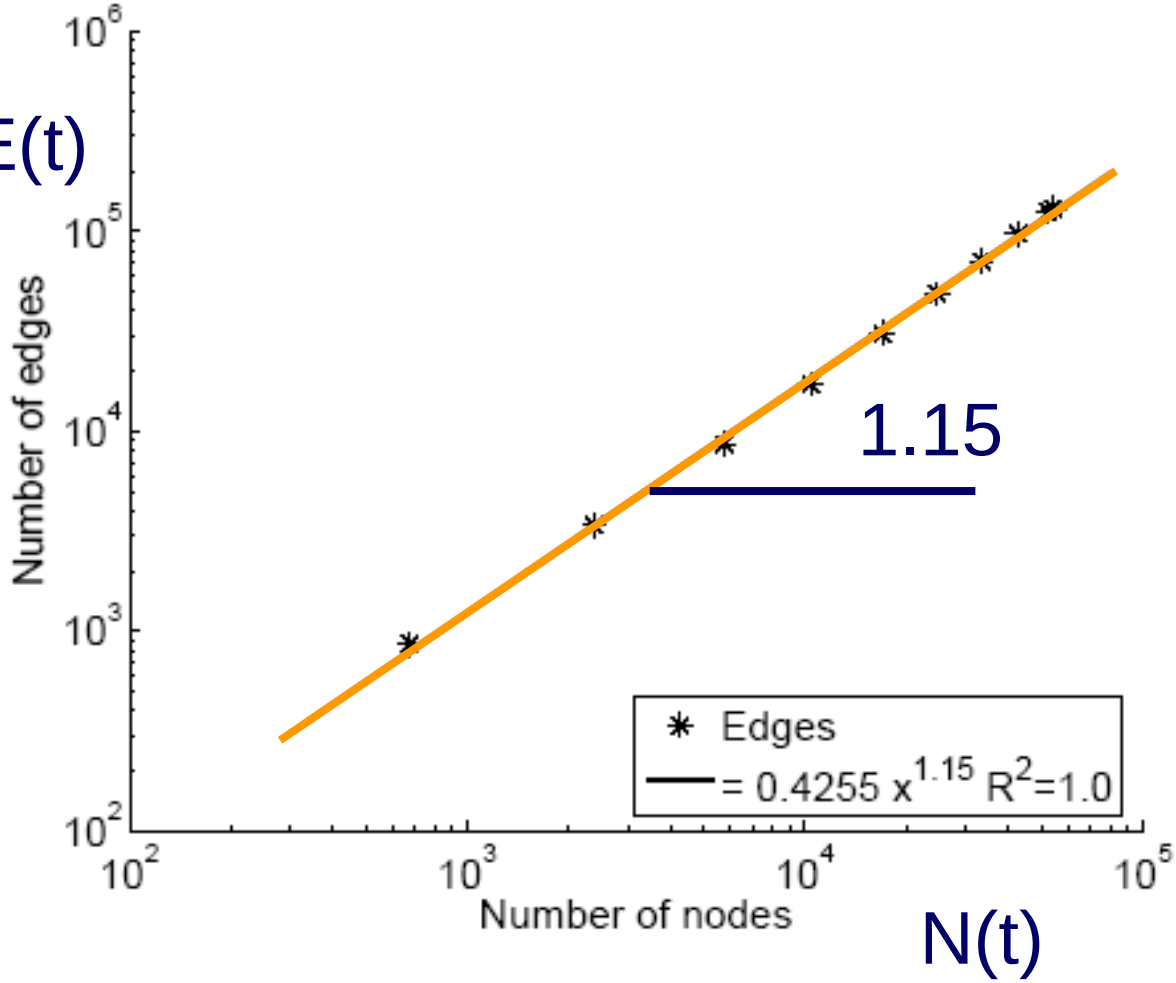
Densification – Autonomous Systems

- Graph of Internet
- 1997
  - 3,000 nodes
  - 10,000 edges
- 2000
  - 6,000 nodes
  - 26,000 edges
- One graph per day



Densification – Affiliation Network

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

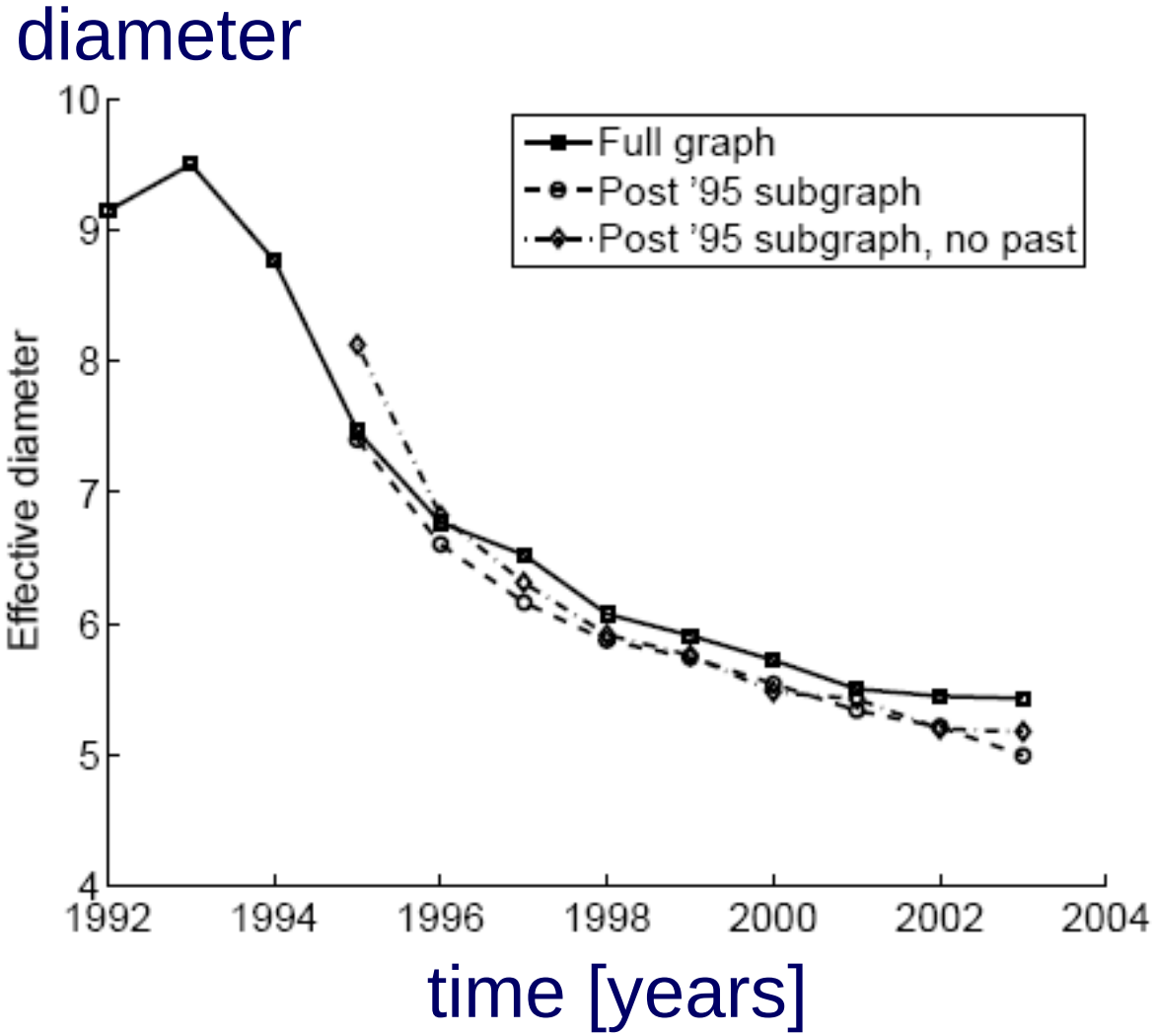


Effective diameter over time

- As the network grows, distances among nodes slowly decrease...

Diameter – ArXiv citation graph

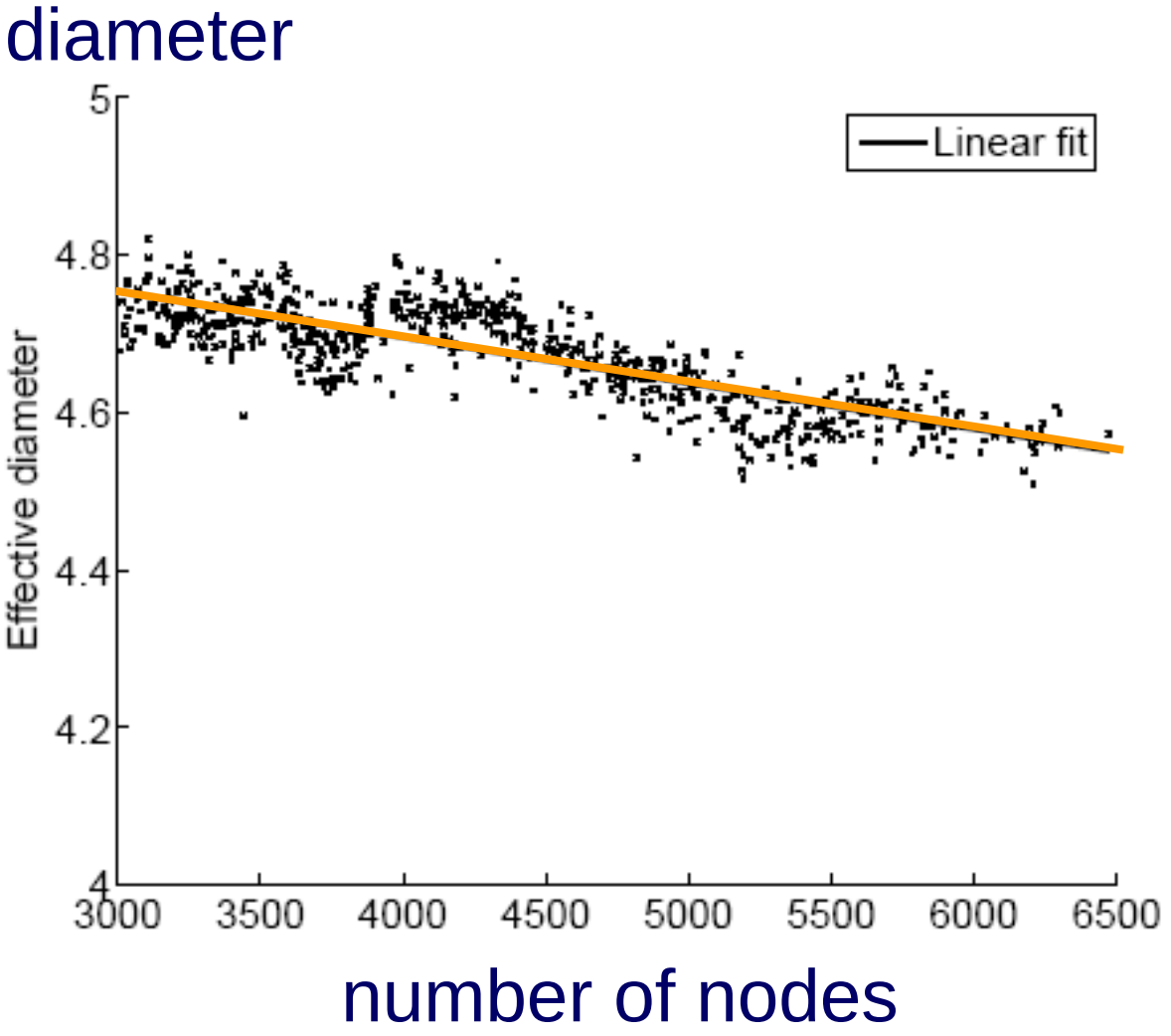
- Citations among physics papers
- 1992 –2003
- One graph per year



Slides on this section from: https://cs.stanford.edu/people/jure/pubs/powergrowth-kdd05.ppt

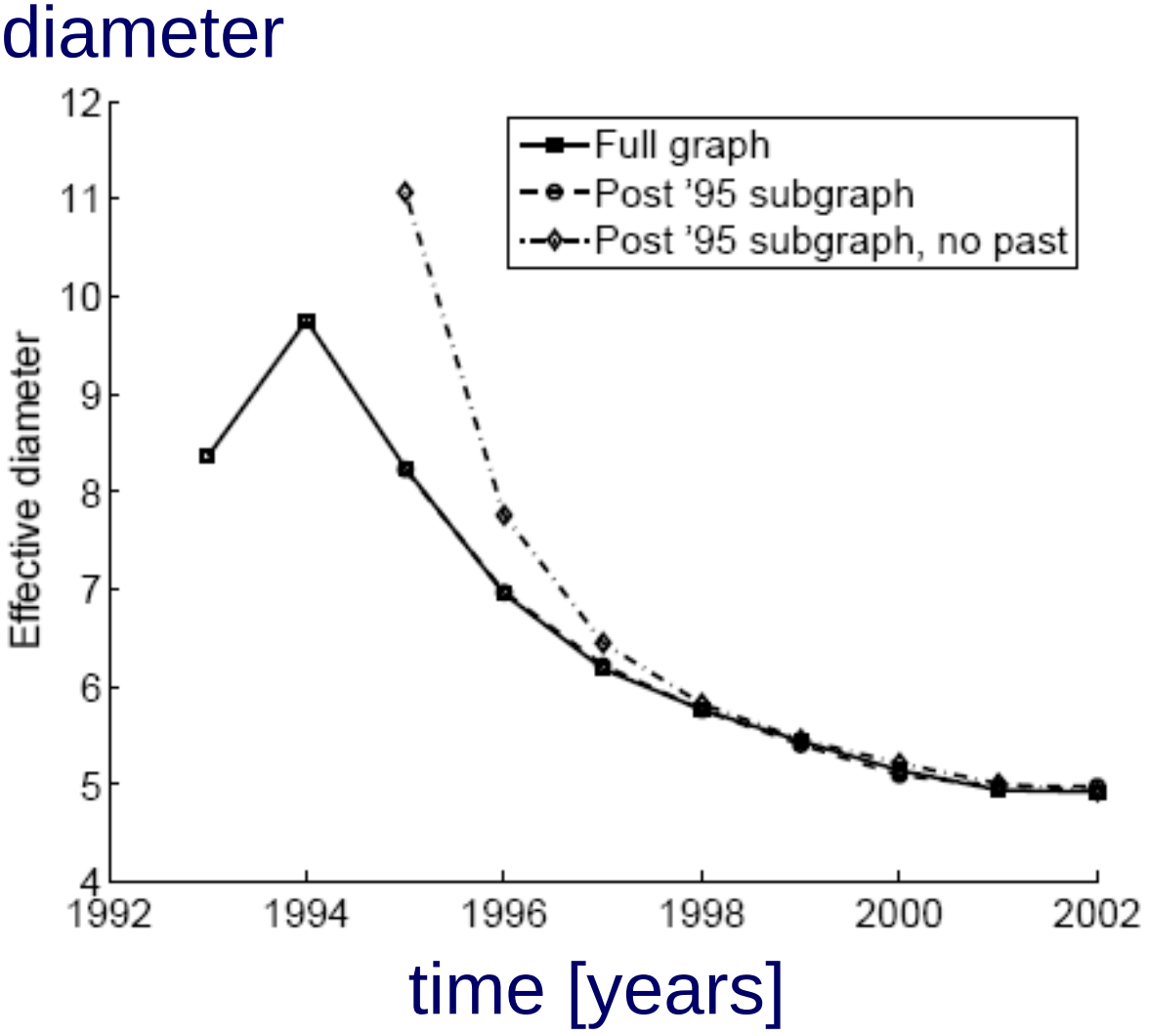
Diameter – "Autonomous Systems"

- Graph of Internet
- One graph per day
- 1997 – 2000



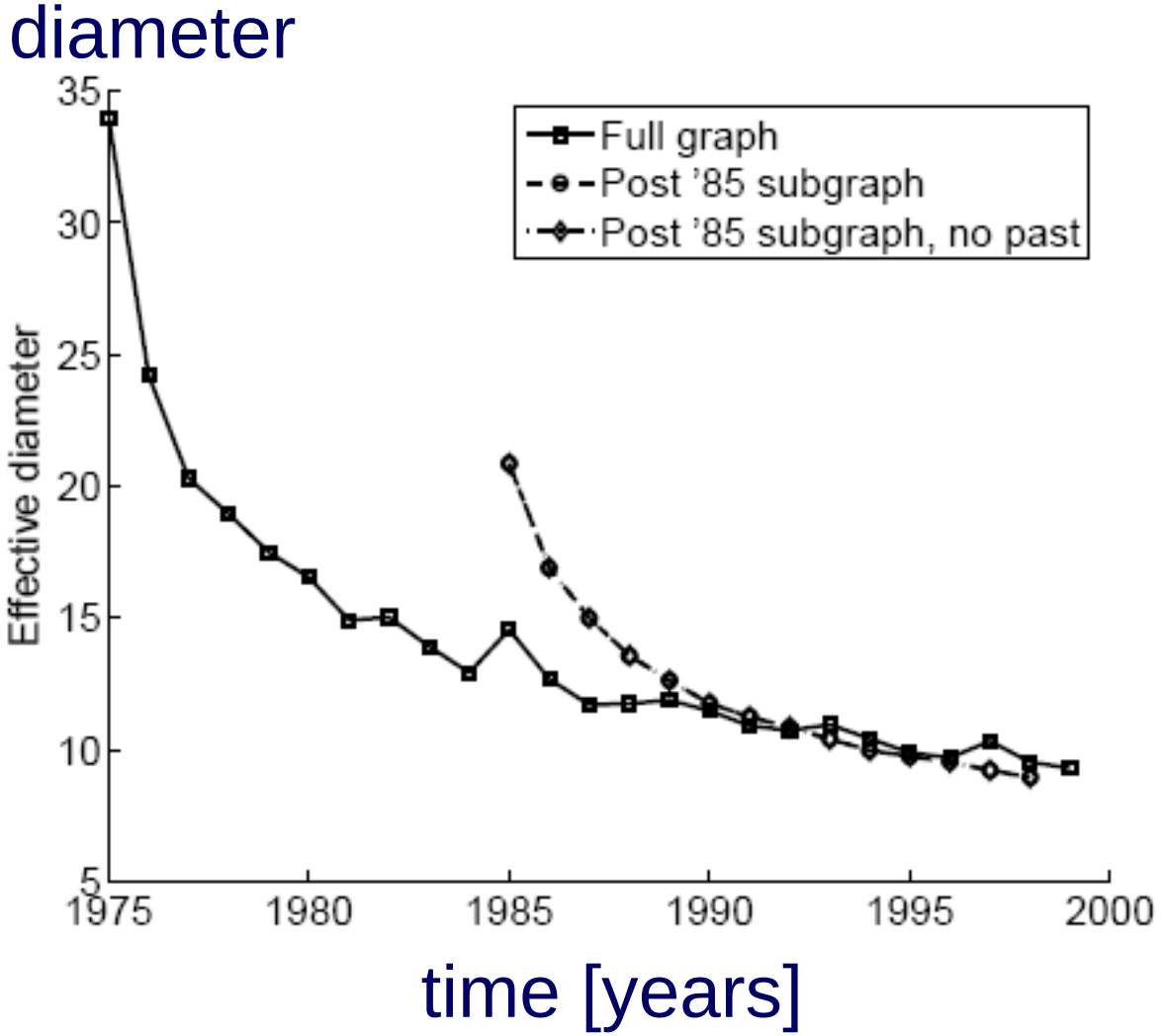
Diameter – “Affiliation Network”

- Graph of collaborations in physics – authors linked to papers
- 10 years of data



Diameter – “Patents”

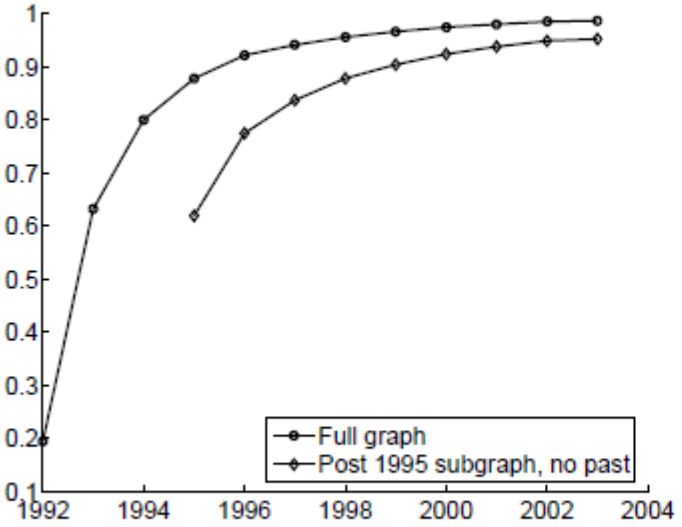
- Patent citation network
- 25 years of data



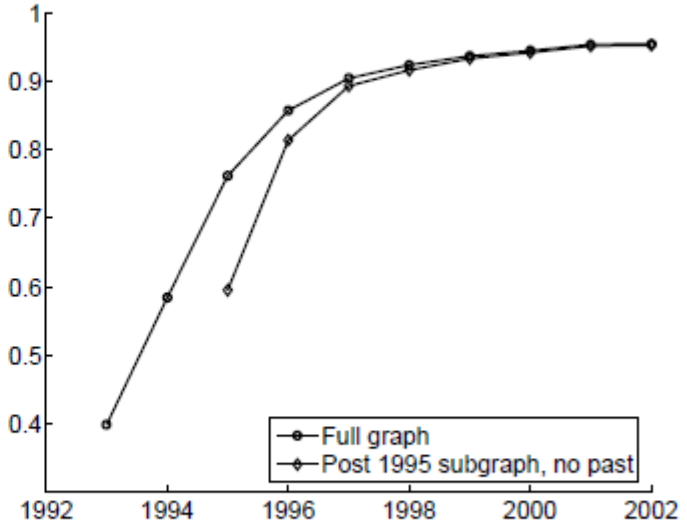
Other characteristics of graphs

- Giant connected component size
- Assortativity

Size of giant connected component as a proportion of number of nodes



(a) arXiv citation graph



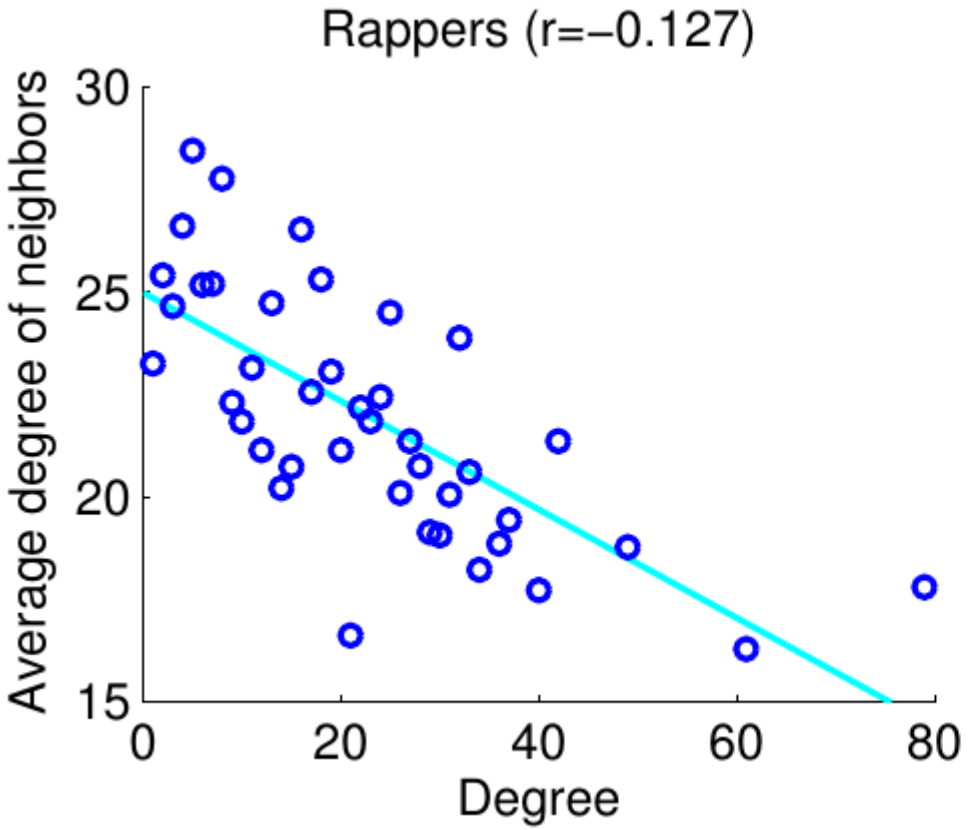
(b) Affiliation network

Assortativity

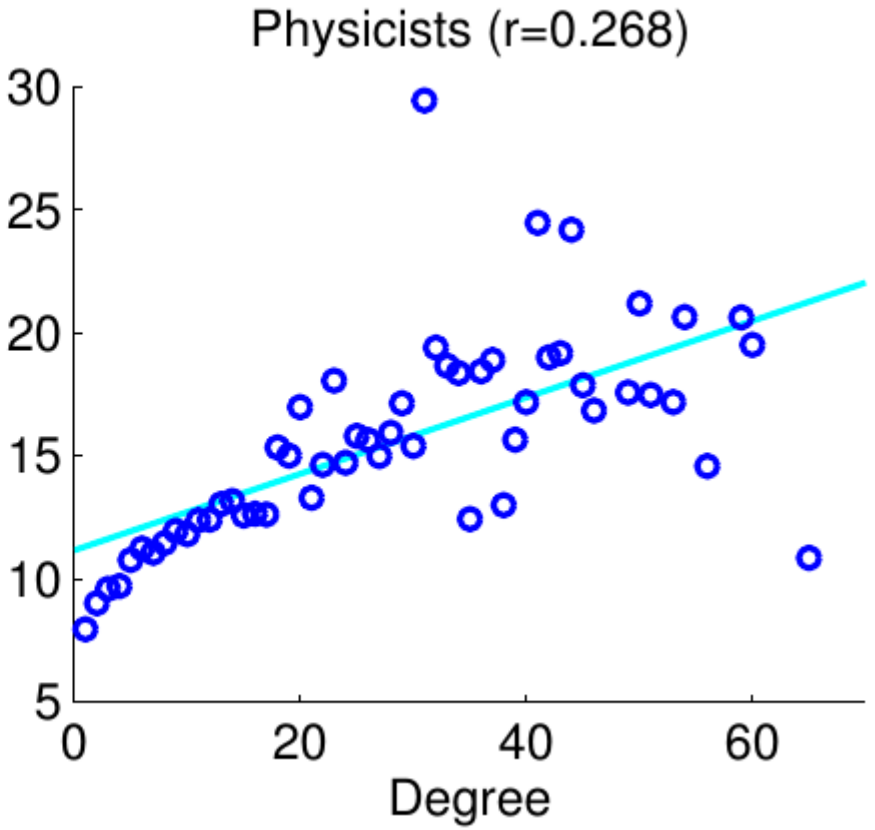
- “Similarly-linked people are together”
- Perfect assortativity if everyone is only connected to people of the same degree
- Perfect disassortativity if everyone is only connected to people of different degree

Assortativity

Disassortative



Assortative



https://blogs.aalto.fi/mining4meaning/2014/03/07/rappers-and-physicists/assortativity-2/

A simple way of understanding assortativity

- Draw two graphs
- Graph A: Disassortative
  - 5 nodes and 4 edges
  - Neighbors have very different degree
- Graph B: Assortative
  - 5 nodes and 4 edges
  - Neighbors have similar or equal degree

http://chato.cl/2015_data_analysis/exercise-answers/graph_models_exercise_01_answer.txt