Carnegie Mellon

15-826: Multimedia (Databases) and Data Mining

Lecture #26: Graph mining - patterns

Christos Faloutsos

Must-read Material – 1-of-2

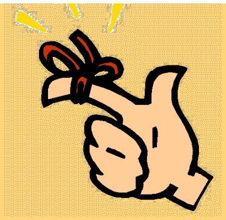
- [Graph minining textbook] Deepayan Chakrabarti and Christos Faloutsos *Graph Mining: Laws, Tools and Case Studies*, Springer, 2012 (internal evaluation copy)
  – Part I (patterns)

Must-read Material 2-of-2

- Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, On Power-Law Relationships of the Internet Topology, SIGCOMM 1999.
- R. Albert, H. Jeong, and A.-L. Barabasi, Diameter of the World Wide Web Nature, 401, 130-131 (1999).
- Reka Albert and Albert-Laszlo Barabasi Statistical mechanics of complex networks, Reviews of Modern Physics, 74, 47 (2002).
- Jure Leskovec, Jon Kleinberg, Christos Faloutsos Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, KDD 2005, Chicago, IL, USA

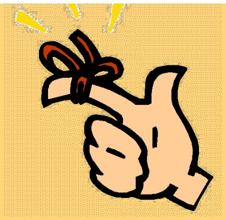
Problem

- Are real graphs random?



Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diameters
    - Skewed degree distribution
    - Shrinking diameters
  - Weighted
  - Time-evolving



Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diamete
    - Ske
  - e-evolving

- Many power laws – log-logistic
- Take logarithms

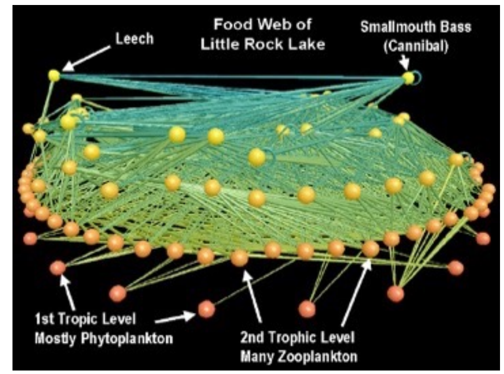
Main outline

- Introduction
- Indexing
- Mining
  - Graphs – patterns
  - Graphs – generators and tools
  - Association rules
  - …

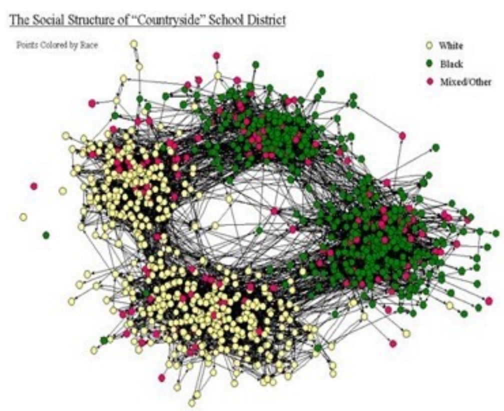
Outline

- Introduction – Motivation
- Problem: Patterns in graphs
- Problem#2: Scalability
- Conclusions

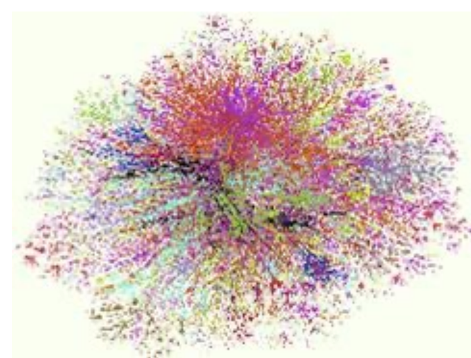
Graphs - why should we care?



Food Web
[Martinez ’91]



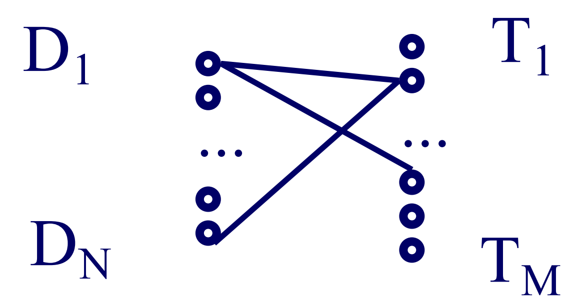
Friendship Network
[Moody ’01]



Internet Map
[lumeta.com]

Graphs - why should we care?

- IR: bi-partite graphs (doc-terms)



- web: hyper-text graph

- ... and more:

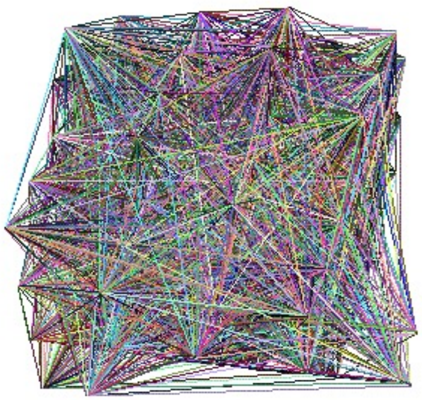
Graphs - why should we care?

- ‘viral’ marketing
- web-log (‘blog’) news propagation
- computer network security: email/IP traffic and anomaly detection
- ....

Outline

- Introduction – Motivation
- Problem: Patterns in graphs
  - Static graphs
  - Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

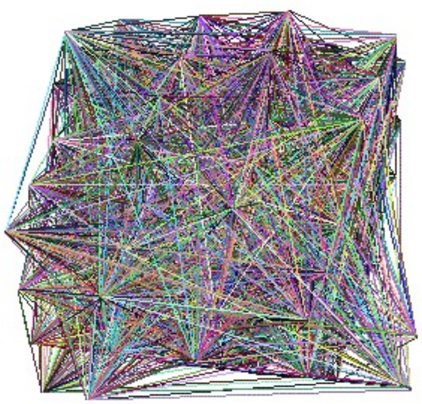
Problem #1 - network and graph mining



- What does the Internet look like?
- What does FaceBook look like?

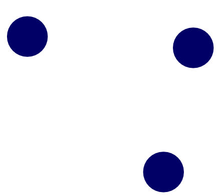
- What is 'normal' / 'abnormal' ?
- which patterns/laws hold?

Problem #1 - network and graph mining

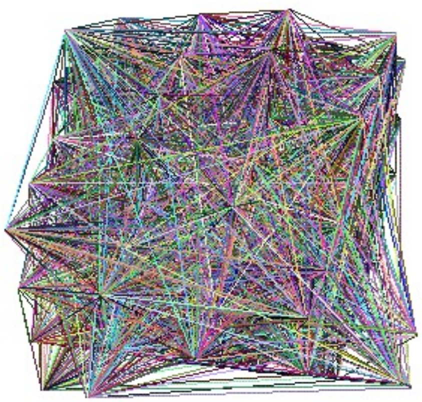


- What does the Internet look like?
- What does FaceBook look like?

- What is 'normal' / 'abnormal' ?
- which patterns/laws hold?
  - **anomalies** (rarities) <-> **patterns**



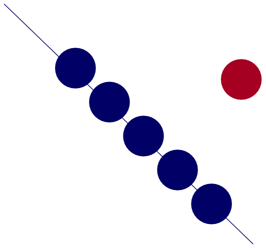
Problem #1 - network and graph mining



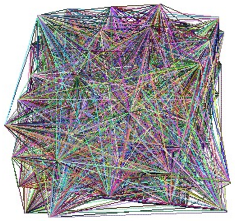
- What does the Internet look like?
- What does FaceBook look like?

- What is 'normal' / 'abnormal' ?
- which patterns/laws hold?

  - **anomalies** (rarities) <-> **patterns**
  - **Large** datasets reveal patterns/anomalies that may be invisible otherwise…



Graph mining



- Are real graphs random?

Laws and patterns

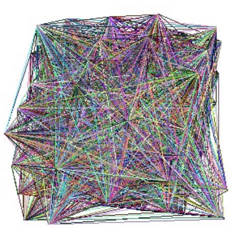
- Are real graphs random?
- A: NO!!
  - Diameter ('6 degrees', 'Kevin Bacon')
  - in- and out- degree distributions
  - other (surprising) patterns

- So, let's look at the data

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

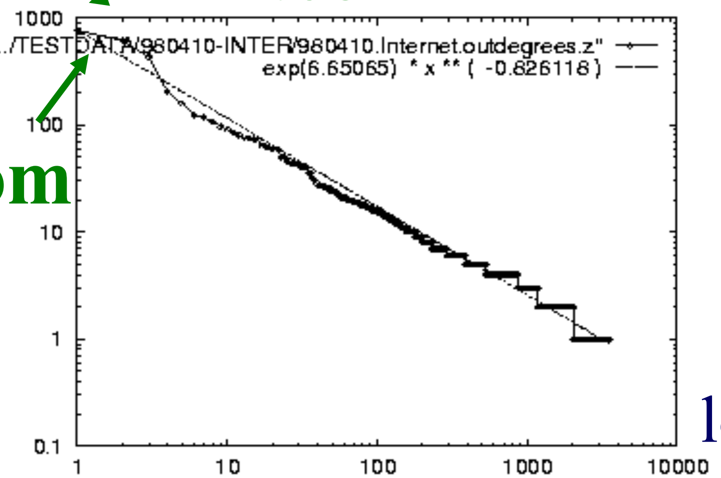
internet domains



att.com

log(degree)

ibm.com

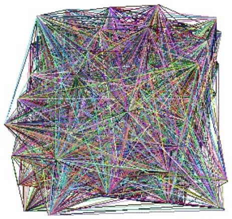


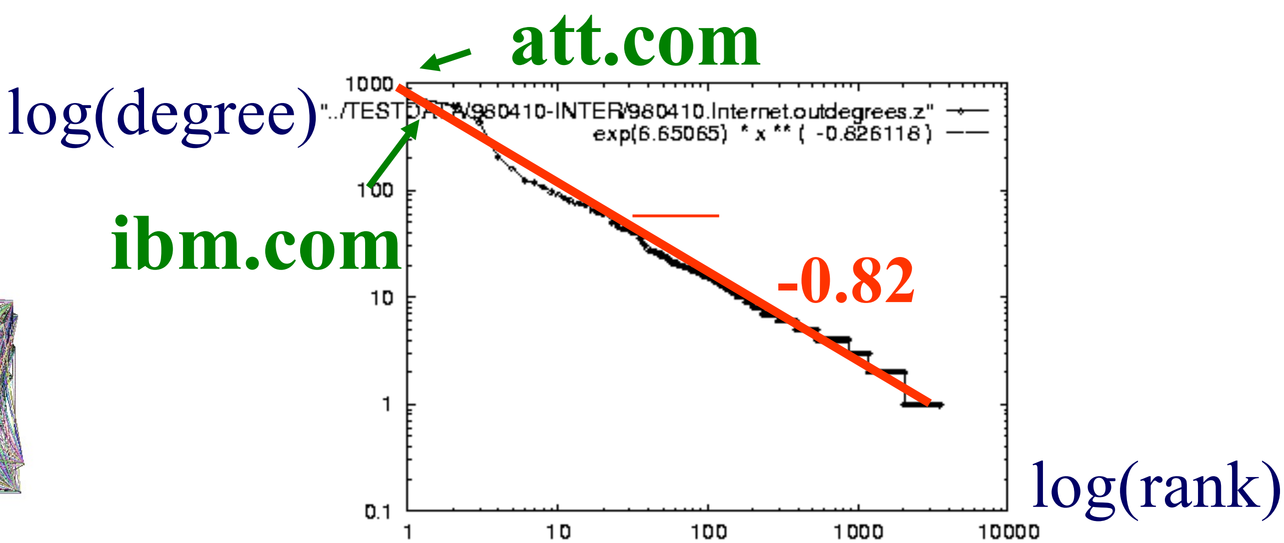
log(rank)

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains

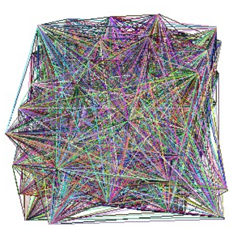


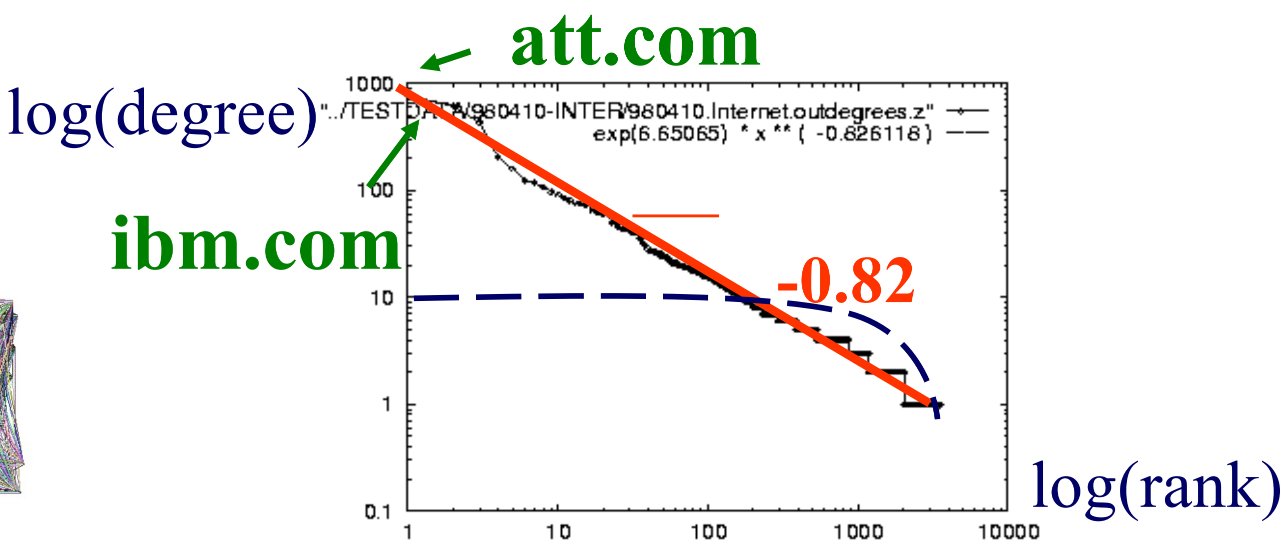


Solution# S.1

- Q: So what?

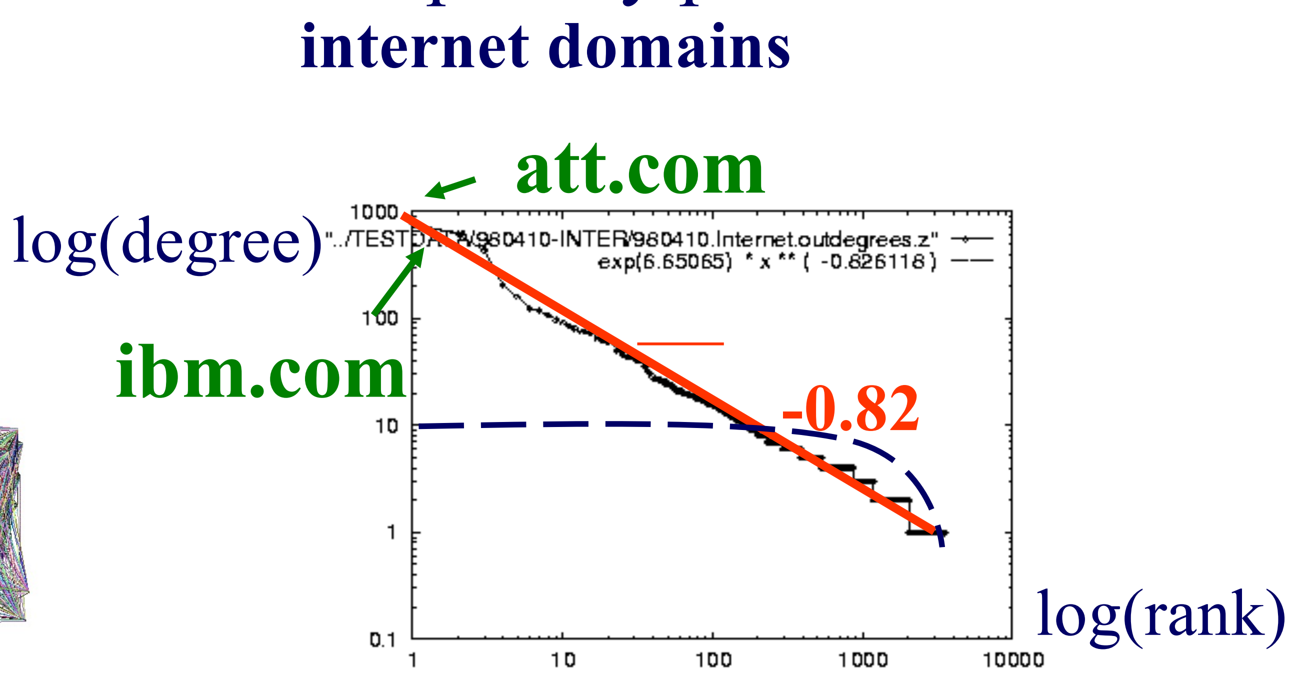
internet domains

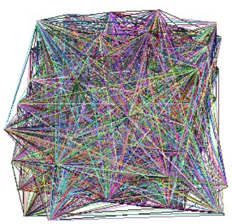




Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: = friends of friends (F.O.F.)



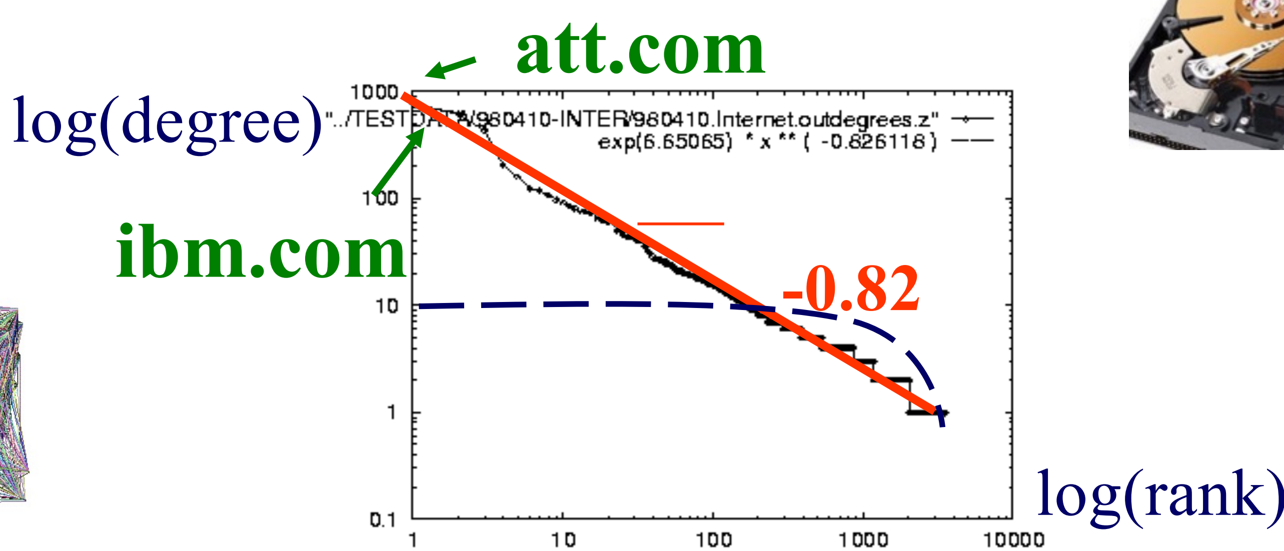


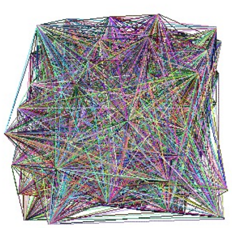
Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: 100^2 * N= 10 Trillion

= friends of friends (F.O.F.)

internet domains

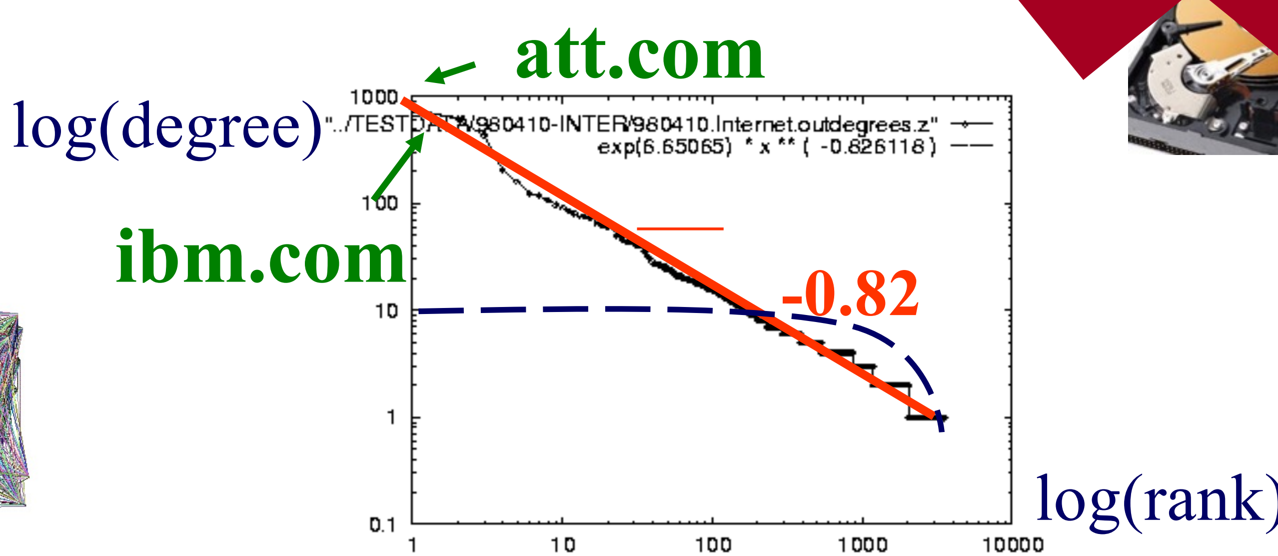


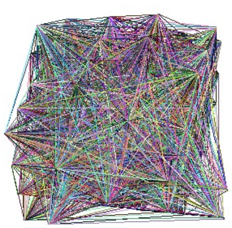
Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: 100^2 N Trillion

= friends of friends (F.O.F.)

internet domains





Carnegie Mellon

Gaussian trap

Solution# S.1

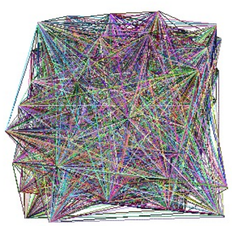
- Q: So what?
- A1: # of two-step-away pairs: O(d_max ^2) ~ 10M^2

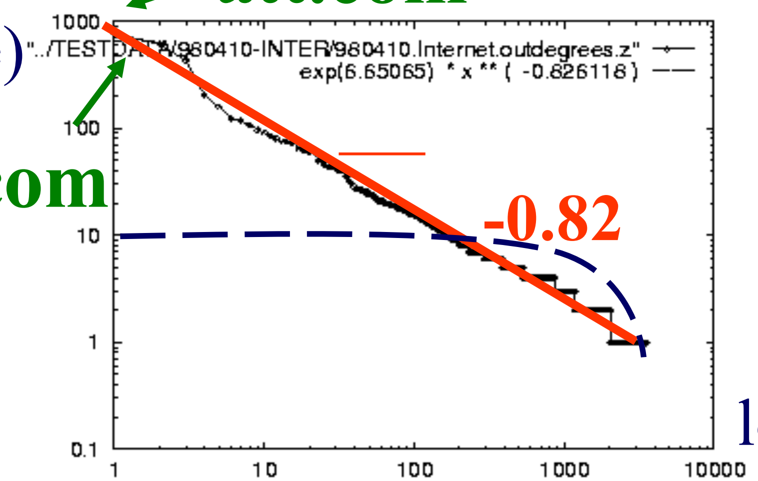
= friends of friends (F.O.F.)

⇩

~0.8PB ->
a data center(!)

internet domains





DCO @ CMU

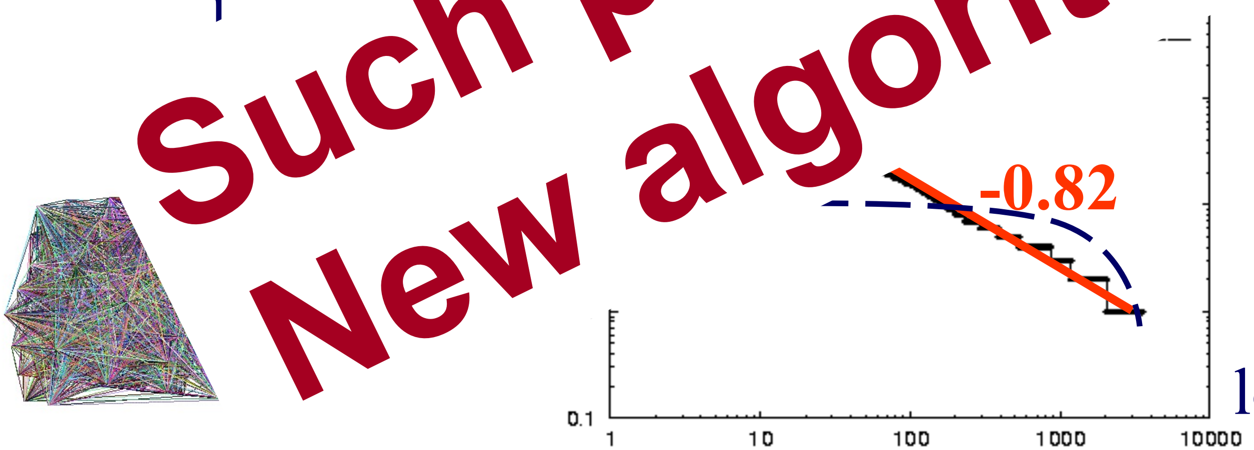
Gaussian trap

Solution# S.1

- Q: So what?
- A1: # of two-step-aw... ?) ~ 10M^2

inte...

Such patterns -> New algorithms

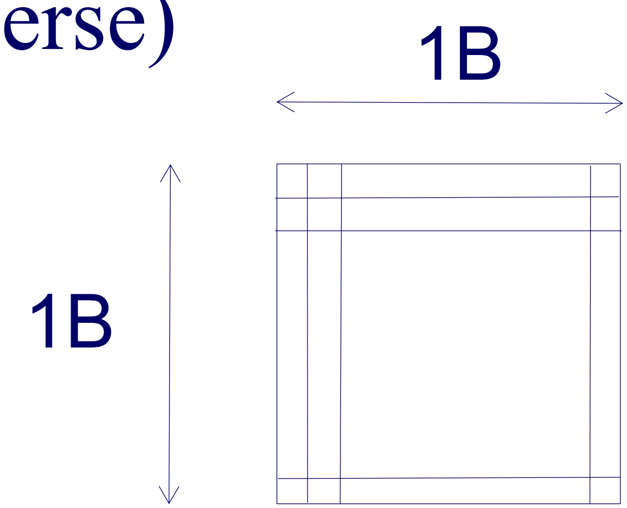


~0.8PB ->
a data center(!)

Observation – big-data:

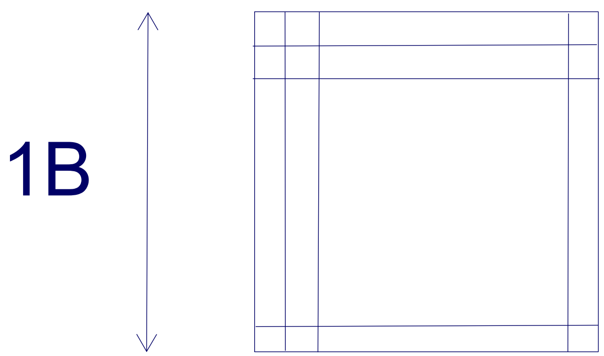
- O(N^2) algorithms are ~intractable - N=1B

- N^2 seconds = 31B years (>2x age of universe)



Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B
- N^2 seconds = ~~31B~~ 31M years
- 1,000 machines

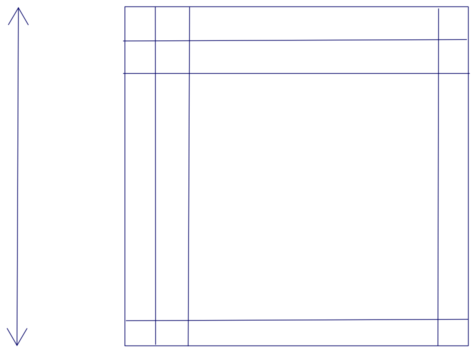


Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B

- N^2 seconds = ~~31B~~ 31K years
- 1M machines

1B



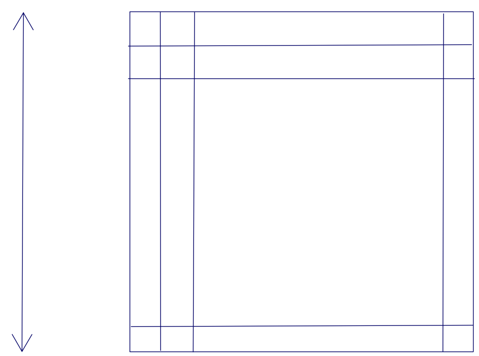
Observation – big-data:

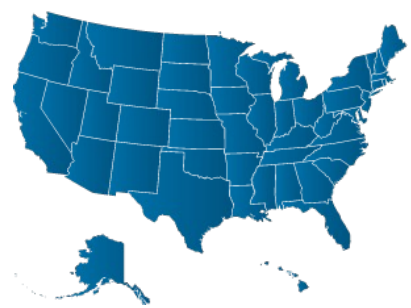
- O(N^2) algorithms are ~intractable - N=1B

3

- N^2 seconds = ~~31B~~ years
- 10B machines ~ $10Trillion

1B





Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B

And parallelism might not help

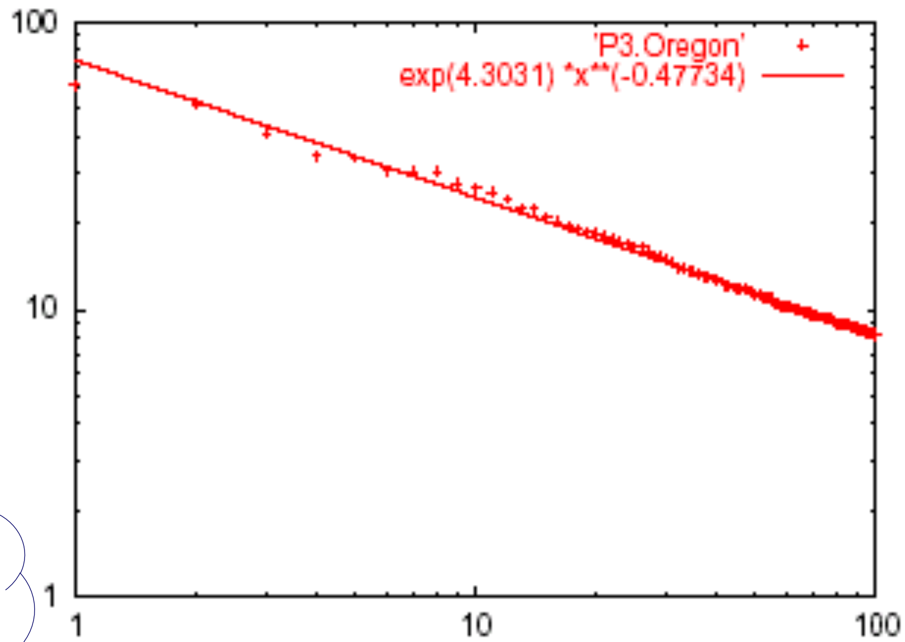
3

- N^2 seconds = 31B years
- 10B machines ~ $10Trillion

1B

Solution# S.2: Eigen Exponent *E*

Eigenvalue



Rank of decreasing eigenvalue

Exponent = slope

E = -0.48

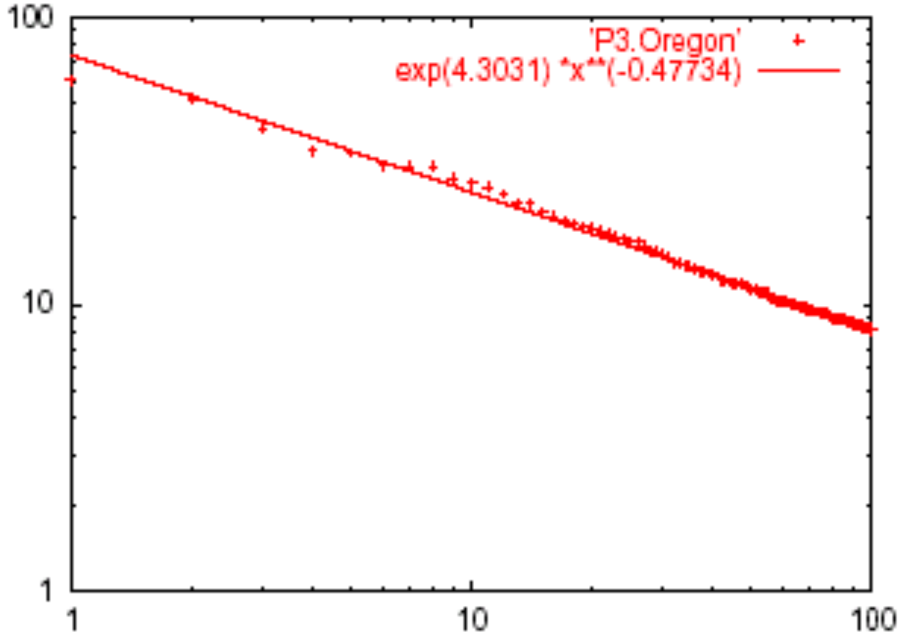
May 2001

$A\,x = \lambda\,x$

- A2: power law in the eigenvalues of the adjacency matrix

Solution# S.2: Eigen Exponent *E*

Eigenvalue

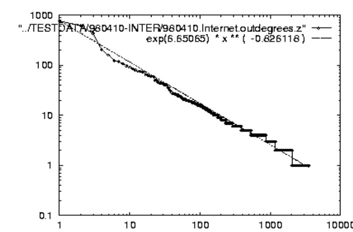


Rank of decreasing eigenvalue

Exponent = slope

$E = -0.48$

May 2001



- [Mihail, Papadimitriou '02]: slope is ½ of rank exponent

But:

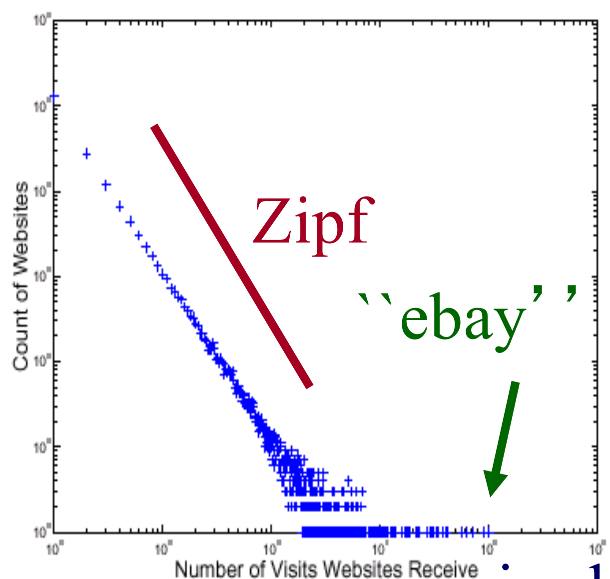
How about graphs from other domains?

More power laws:

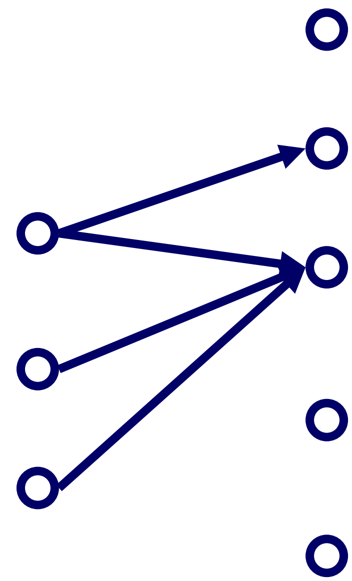
- web hit counts [w/ A. Montgomery]

Web Site Traffic

Count
(log scale)



in-degree (log scale)



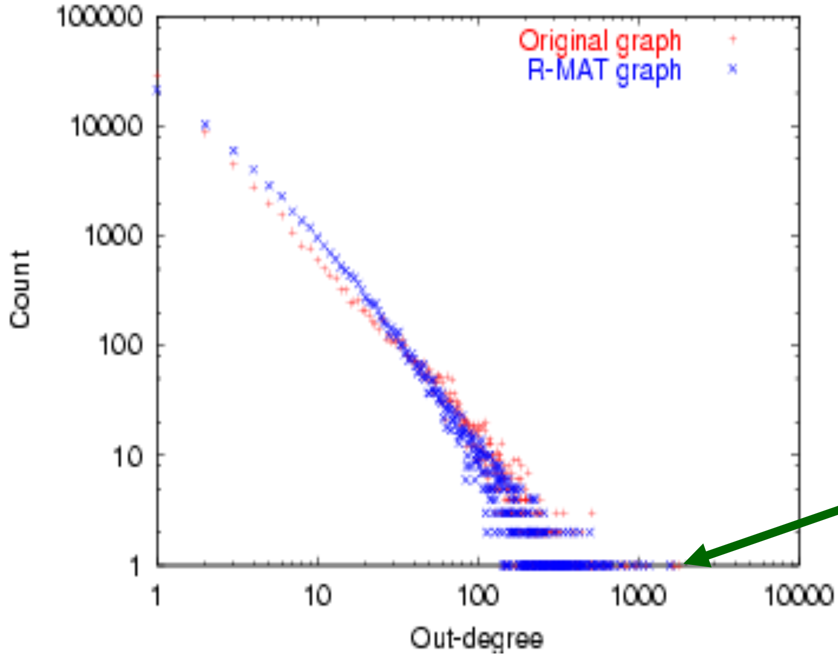
users

sites

epinions.com

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

count



trusts-2000-people user

(out) degree

And numerous more

- # of sexual contacts
- Income [Pareto] –' 80-20 distribution'
- Duration of downloads [Bestavros+]
- Duration of UNIX jobs ('mice and elephants')
- Size of files of a user
- …
- 'Black swans'

Carnegie Mellon

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

S.3 small diameters

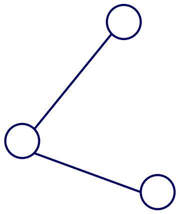
- Small diameter (~ constant!) –
  - six degrees of separation / 'Kevin Bacon'
  - small worlds [Watts and Strogatz]

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

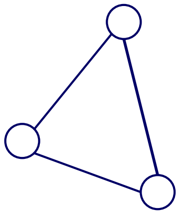
In textbook

Solution# S.4: Triangle ‘Laws’



- Real social networks have a lot of triangles

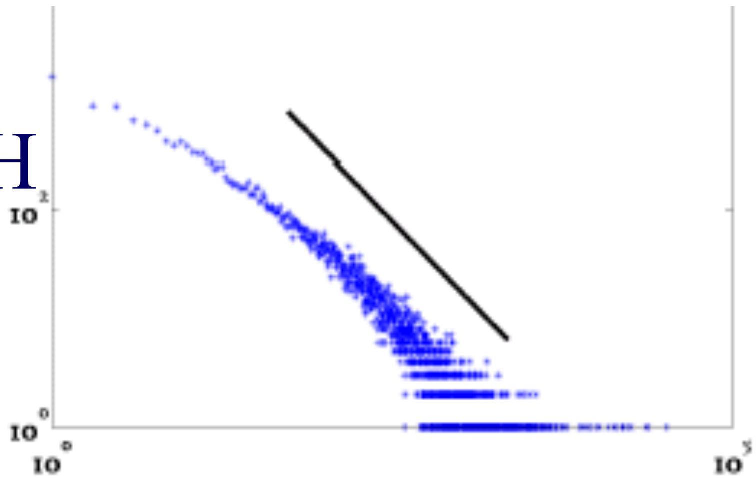
Solution# S.4: Triangle ‘Laws’



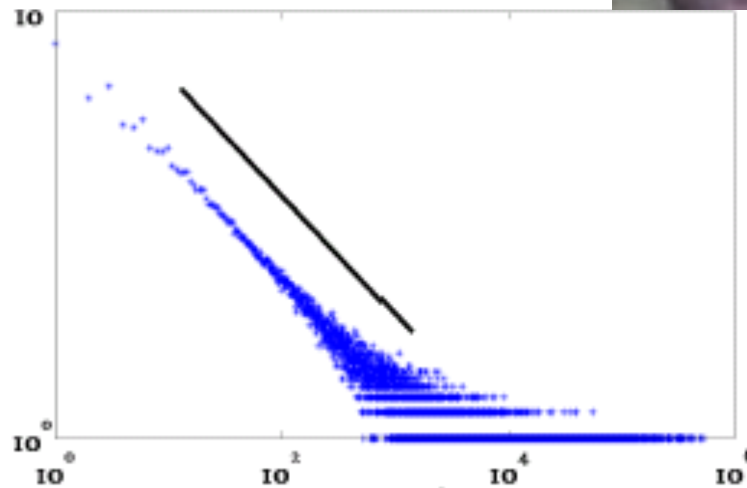
- Real social networks have a lot of triangles
  - Friends of friends are friends
- Any patterns?

Triangle Law: #S.4
[Tsourakakis ICDM 2008]

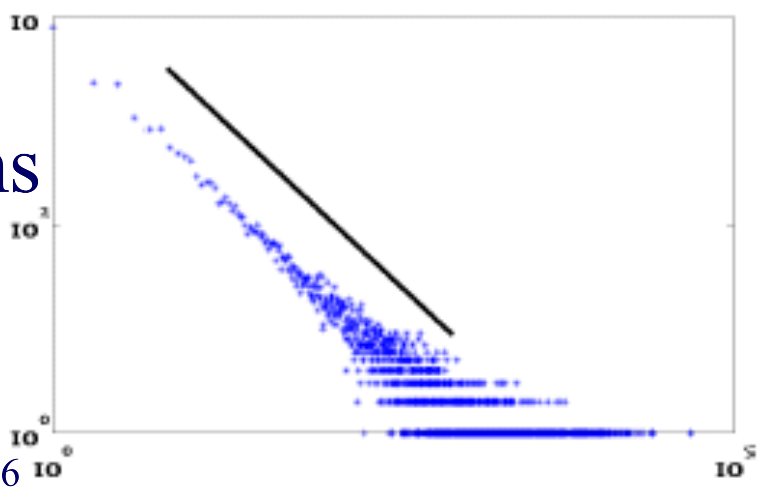
HEP-TH

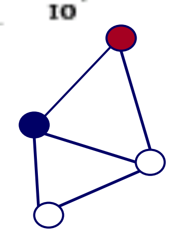


ASN



Epinions





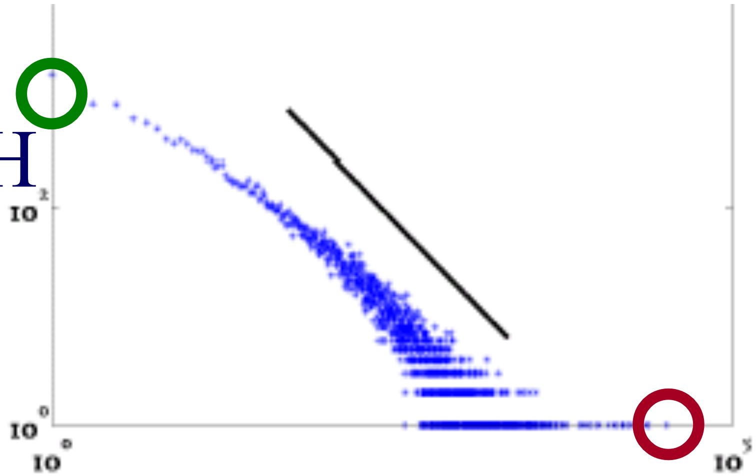
X-axis: # of participating triangles

Y: count (~ pdf)

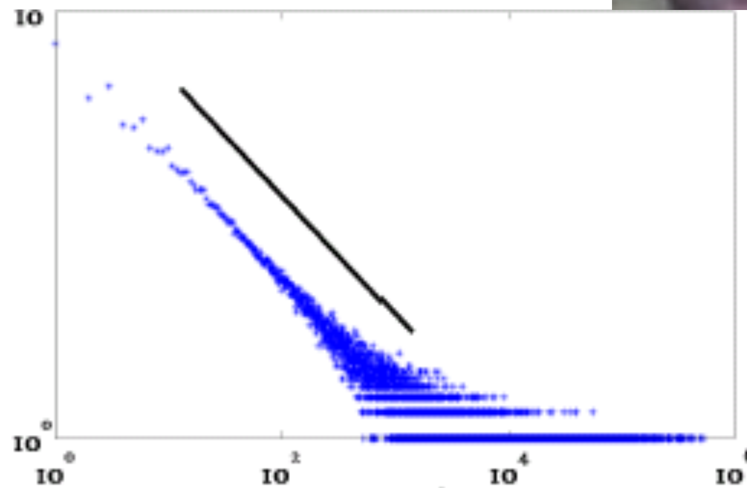
Carnegie Mellon

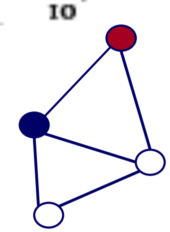
Triangle Law: #S.4
[Tsourakakis ICDM 2008]

HEP-TH

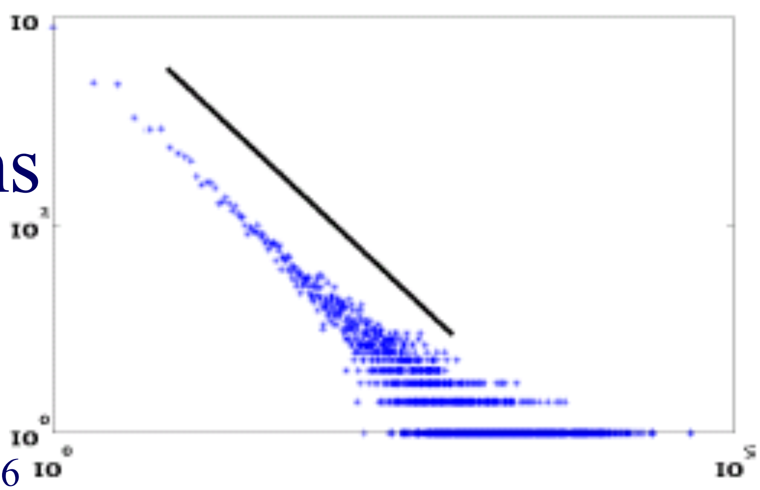


ASN





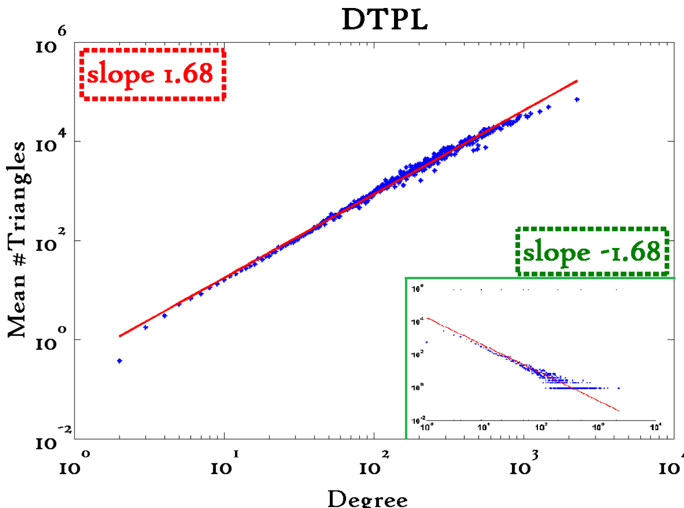
Epinions



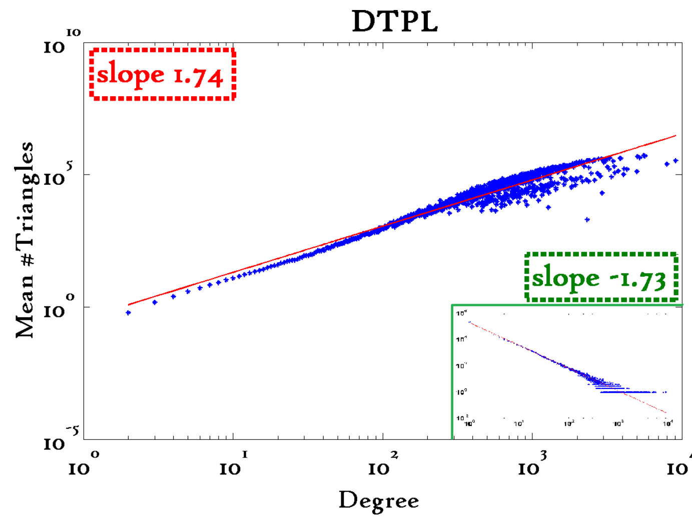
X-axis: # of participating triangles
Y: count (~ pdf)

Triangle Law: #S.5 [Tsourakakis ICDM 2008]

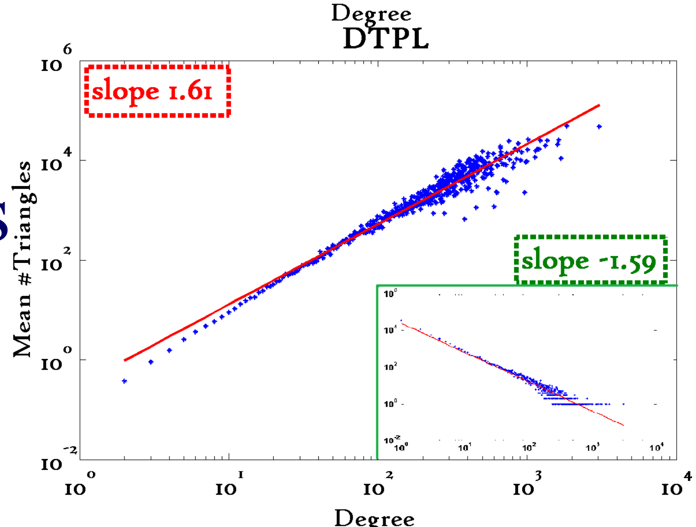
Reuters



SN



Epinions



X-axis: degree

Y-axis: mean # triangles

n friends -> ~$n^{1.6}$ triangles

details

Triangle Law: Computations [Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos)

Q: Can we do that quickly?

details

Triangle Law: Computations
[Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos)

Q: Can we do that quickly?

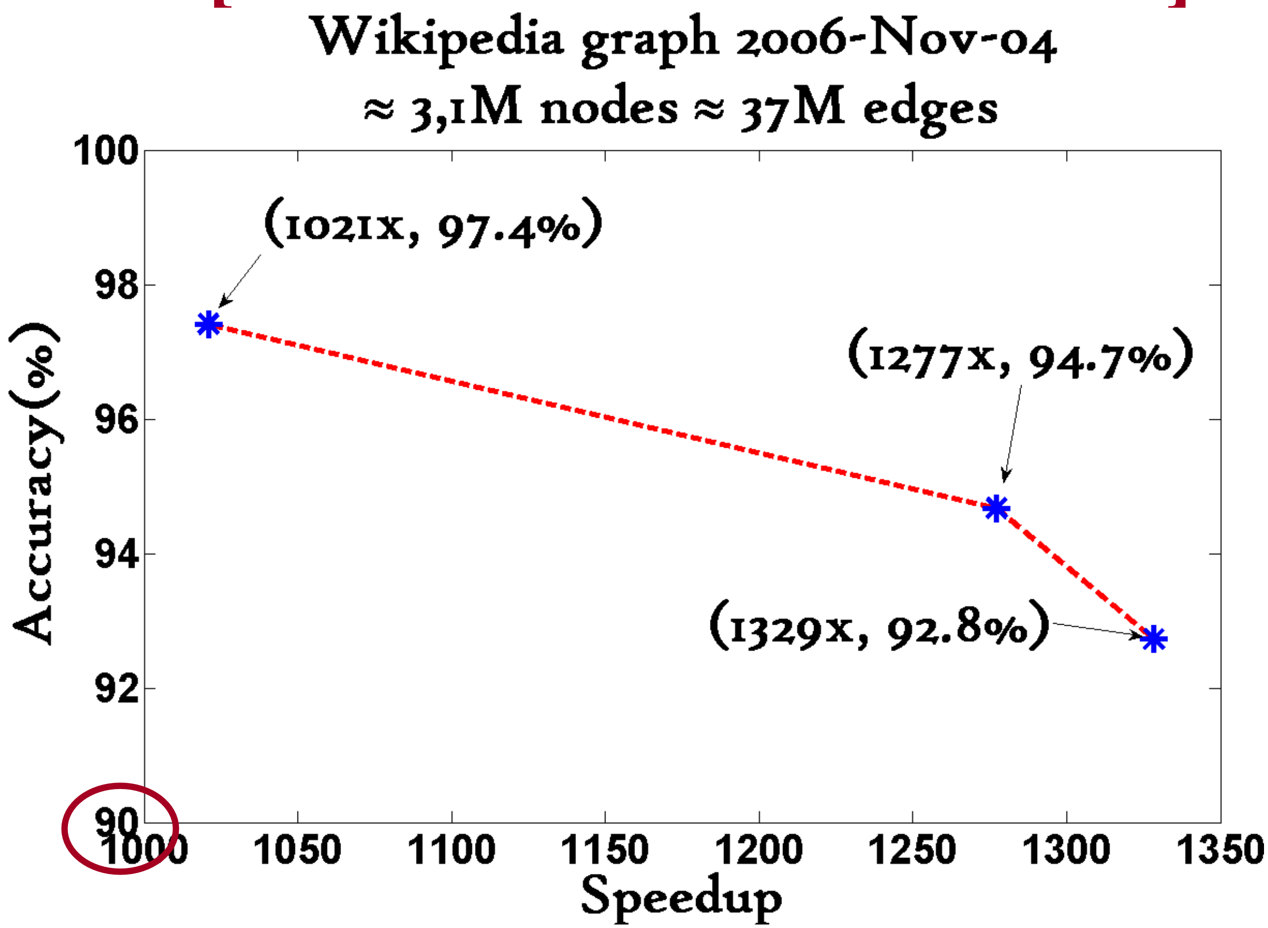
A: Yes!

#triangles = 1/6 Sum (λ_i^3)

(and, because of skewness (S2) ,
we only need the top few eigenvalues!

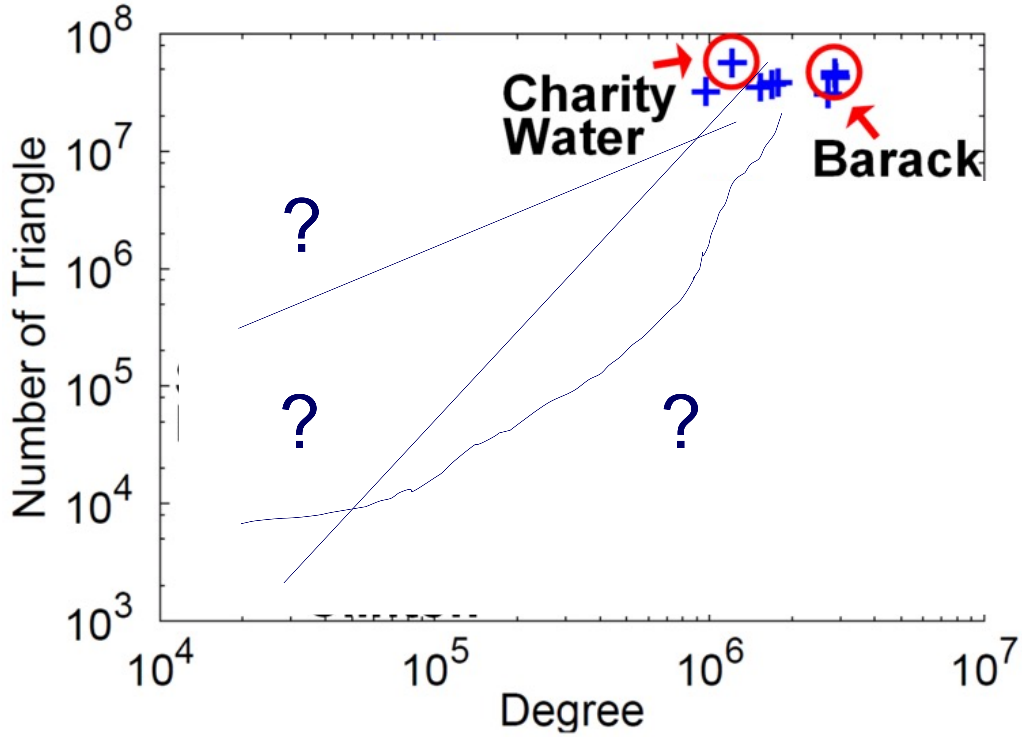
details

Triangle Law: Computations
[Tsourakakis ICDM 2008]



1000x+ speed-up, >90% accuracy

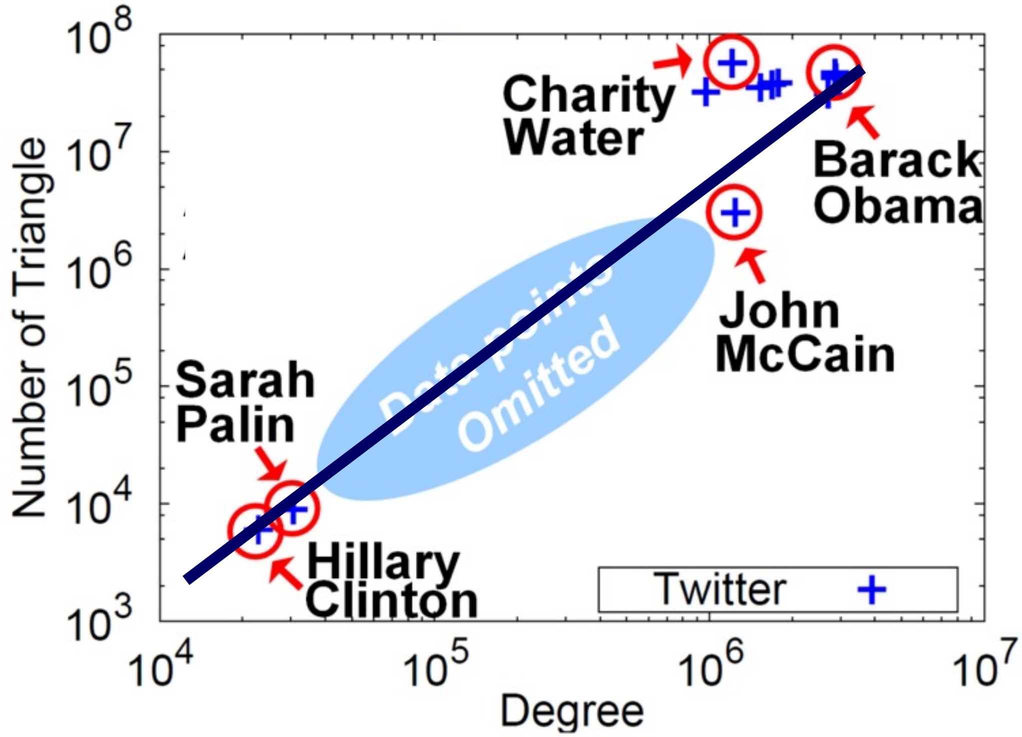
Triangle counting for large graphs?



Anomalous nodes in Twitter(~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD'11]

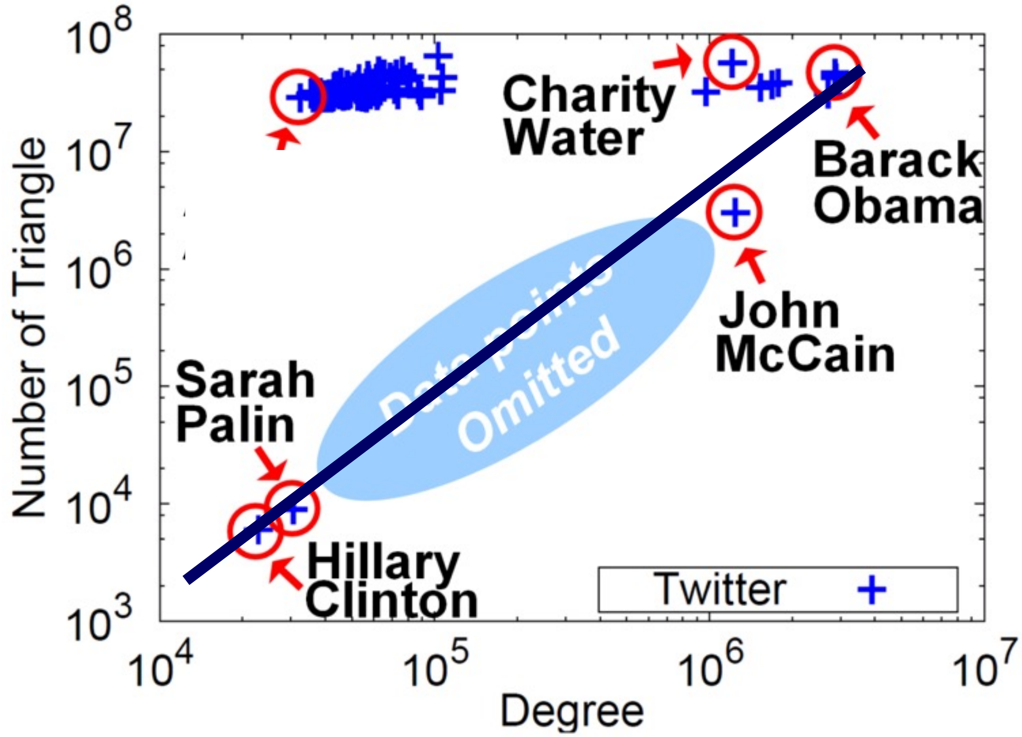
Triangle counting for large graphs?



Anomalous nodes in Twitter(~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD'11]

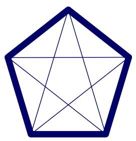
Triangle counting for large graphs?

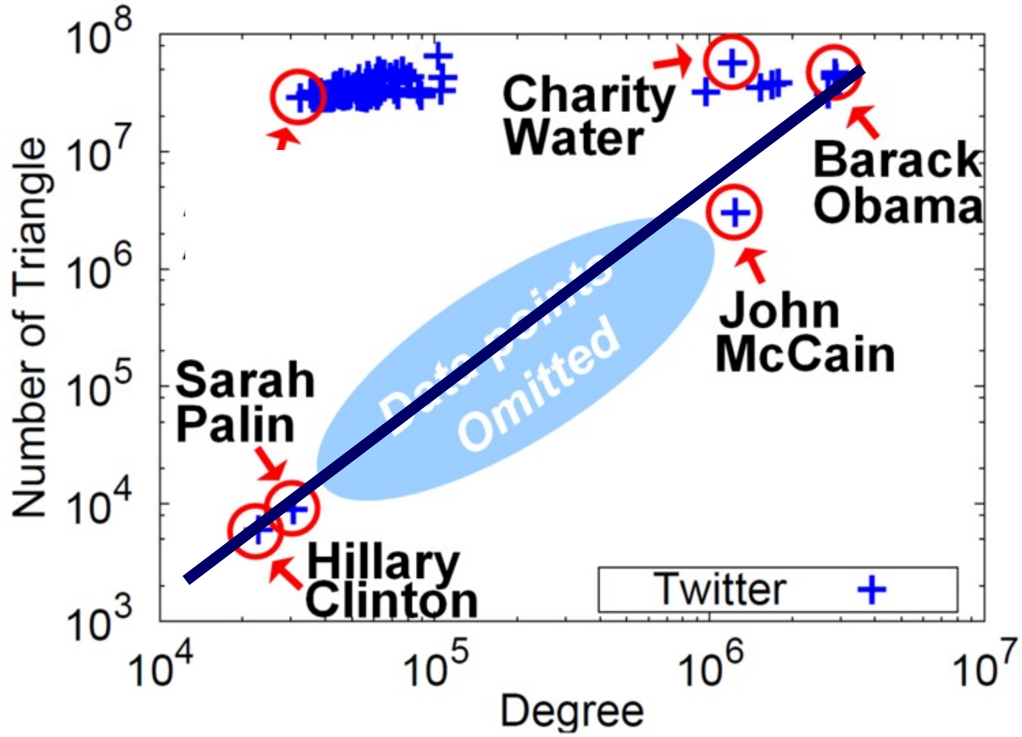


Anomalous nodes in Twitter(~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD'11]

Triangle counting for large graphs?

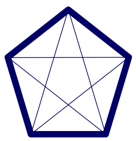


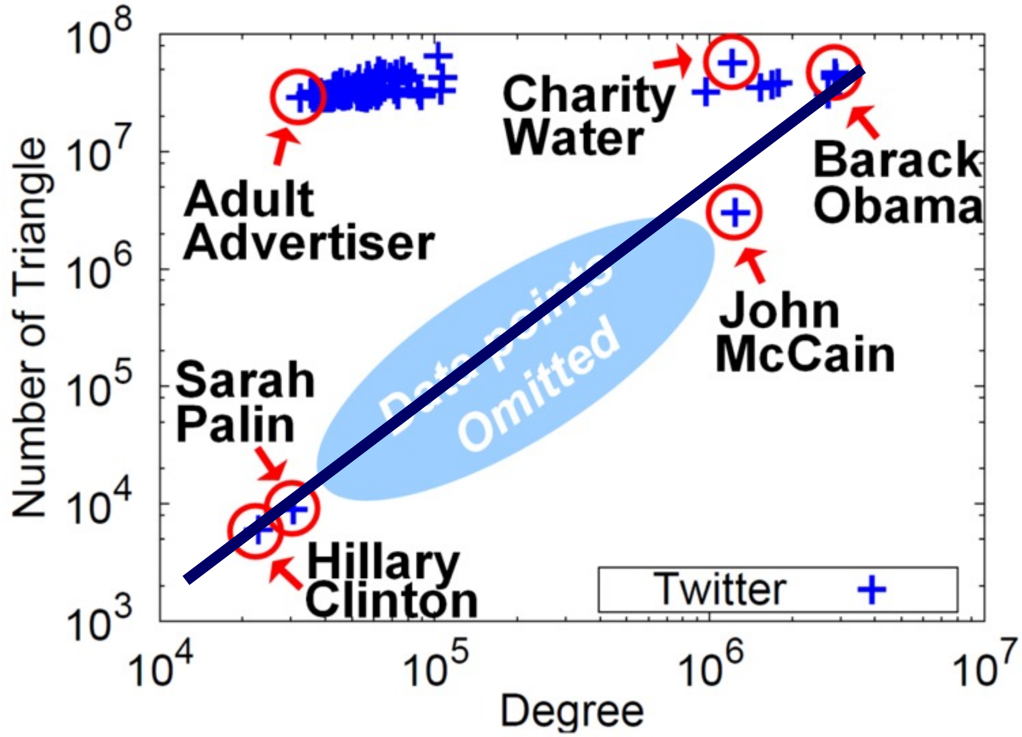


Anomalous nodes in Twitter(~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD'11]

Triangle counting for large graphs?





Anomalous nodes in Twitter(~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD'11]

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

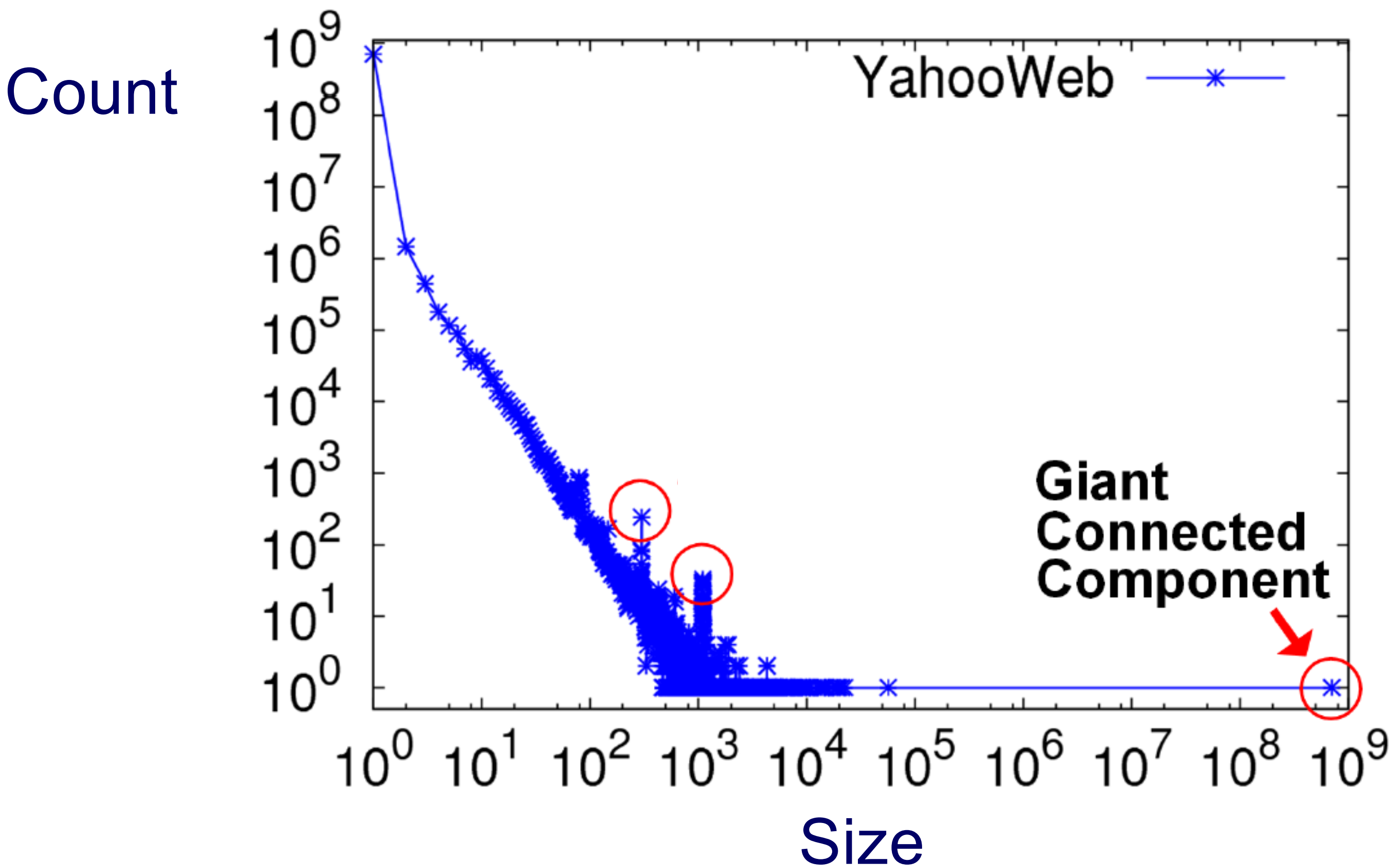
In textbook

Generalized Iterated Matrix Vector Multiplication (GIMV)

PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations.
U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos.
(ICDM) 2009, Miami, Florida, USA.
Best Application Paper (runner-up) and
10-yr highest impact award (2018)

S.6: NLCC

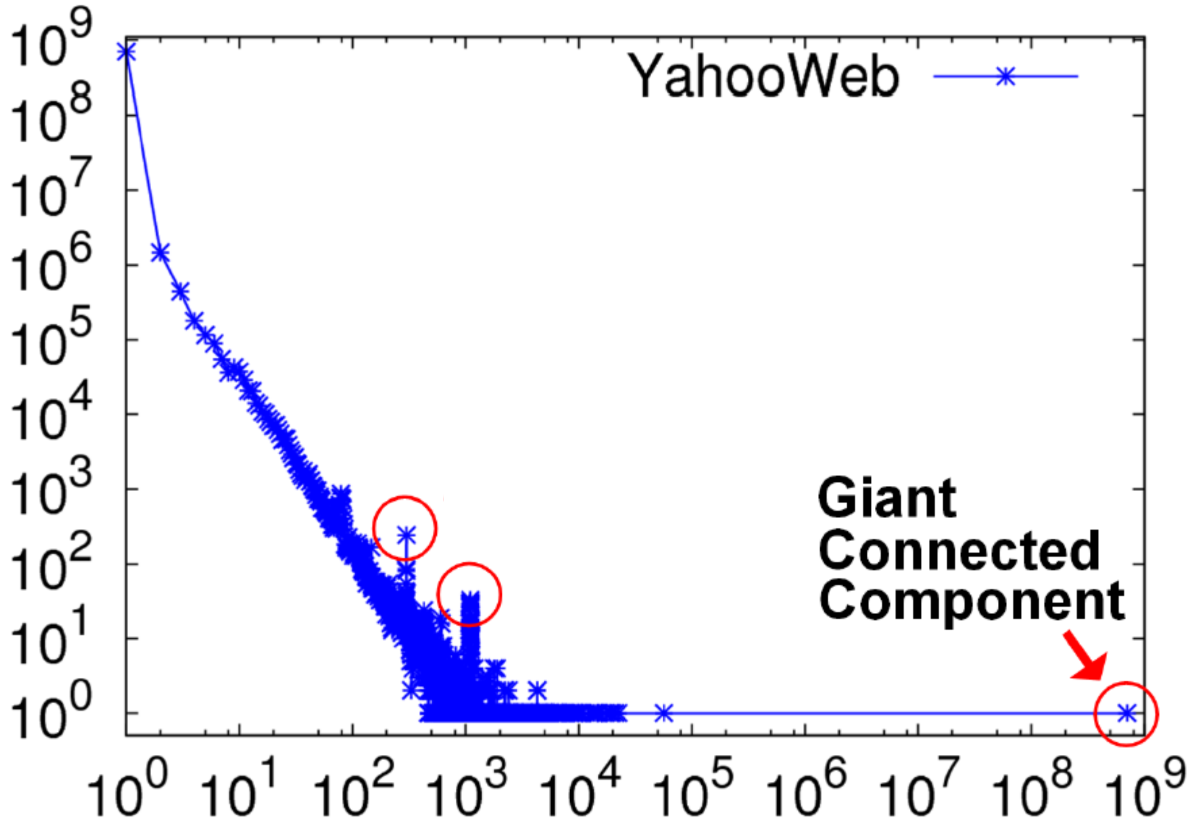
- Connected Components – 4 observations:



S.6: NLCC

- Connected Components

Count

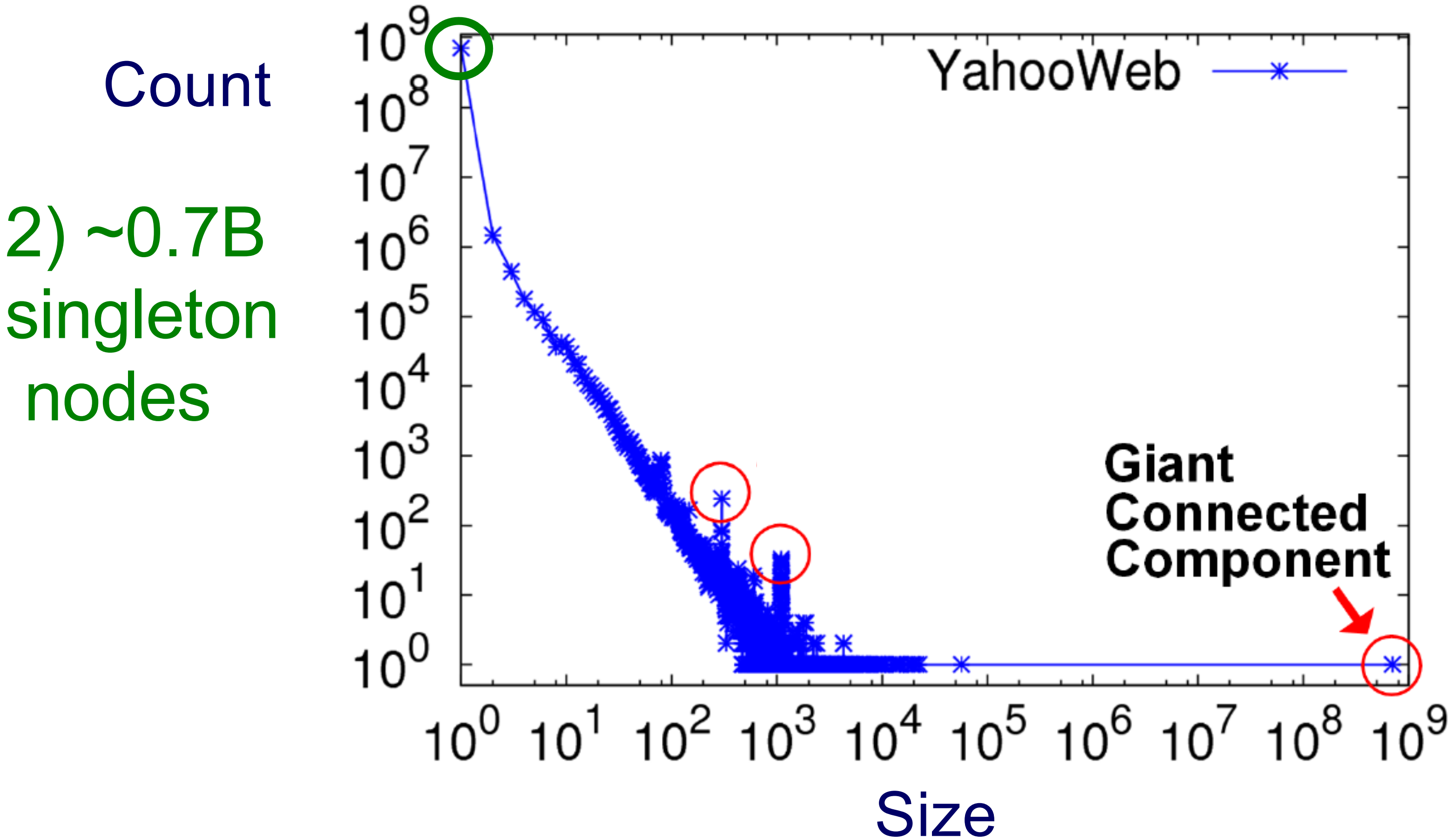


Size

1) 10K x larger than next

S.6: NLCC

- Connected Components

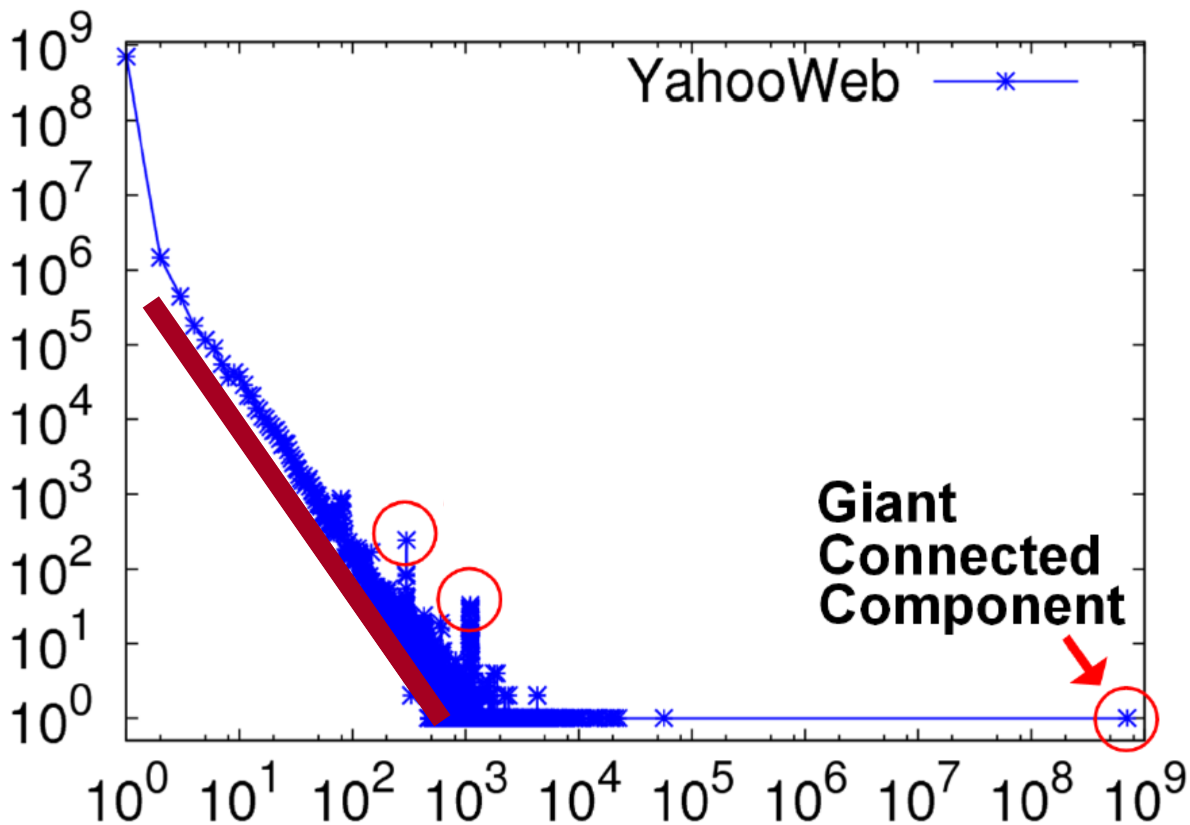


S.6: NLCC

- Connected Components

Count

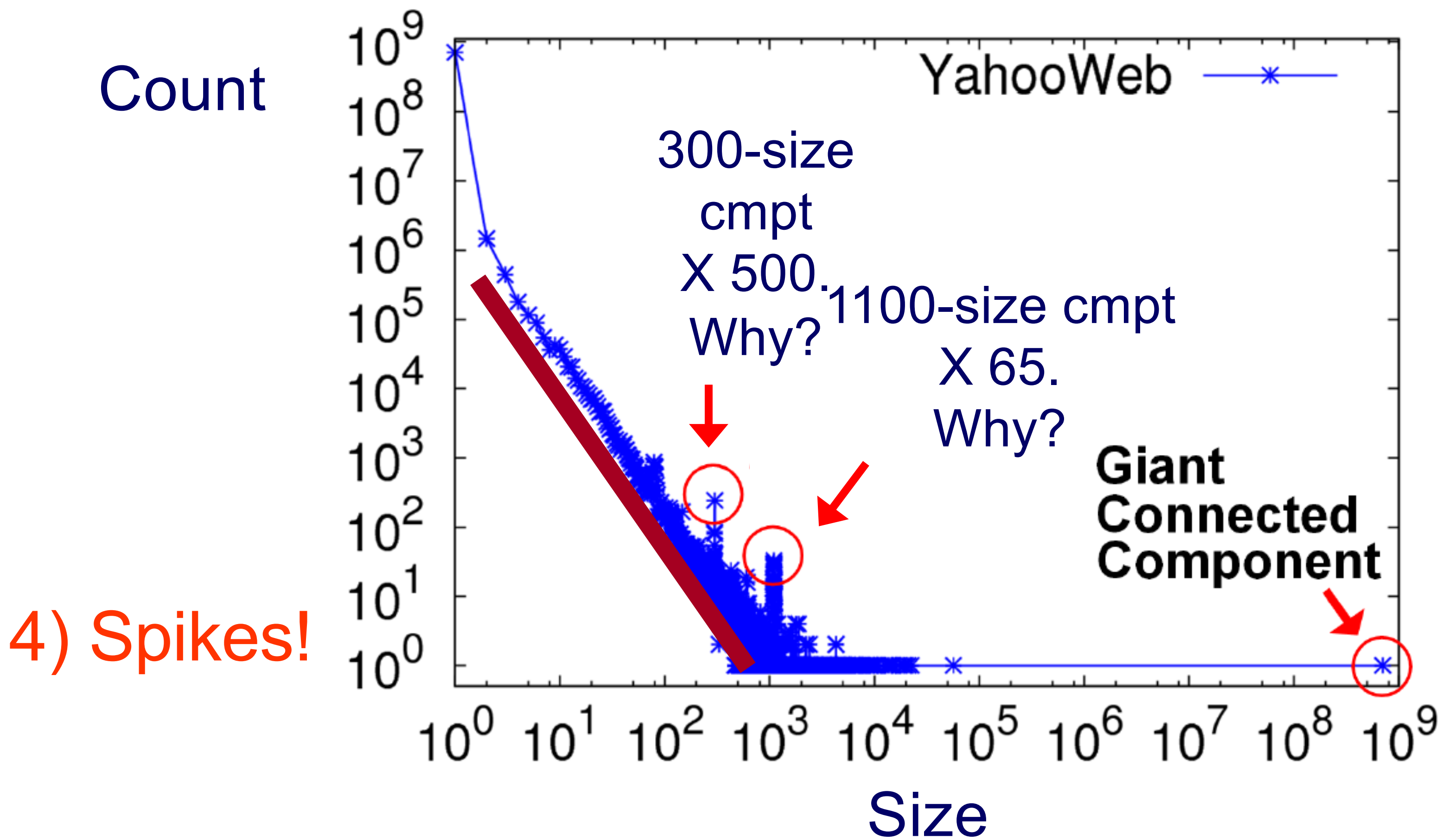
3) SLOPE!



Size

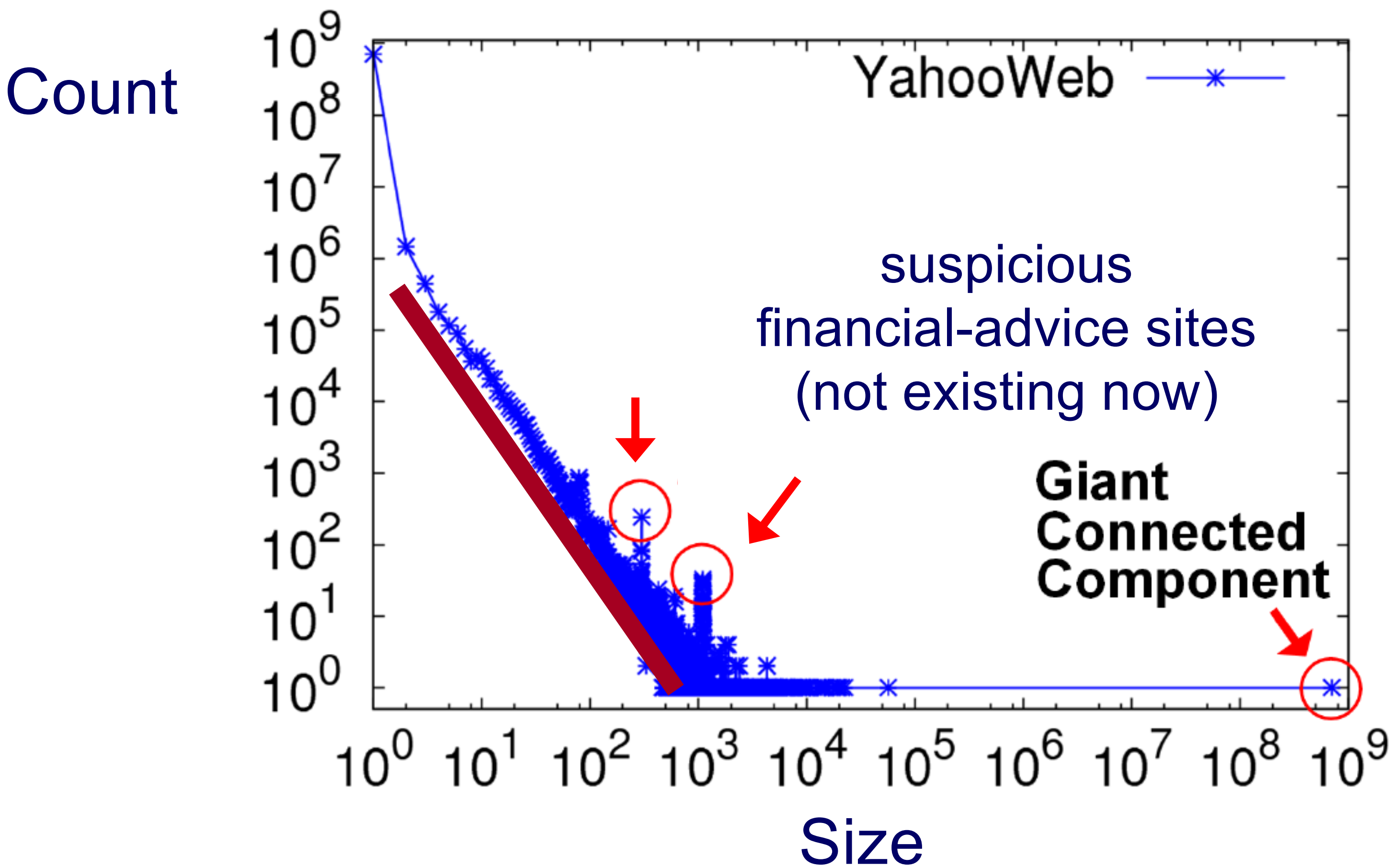
S.6: NLCC

- Connected Components



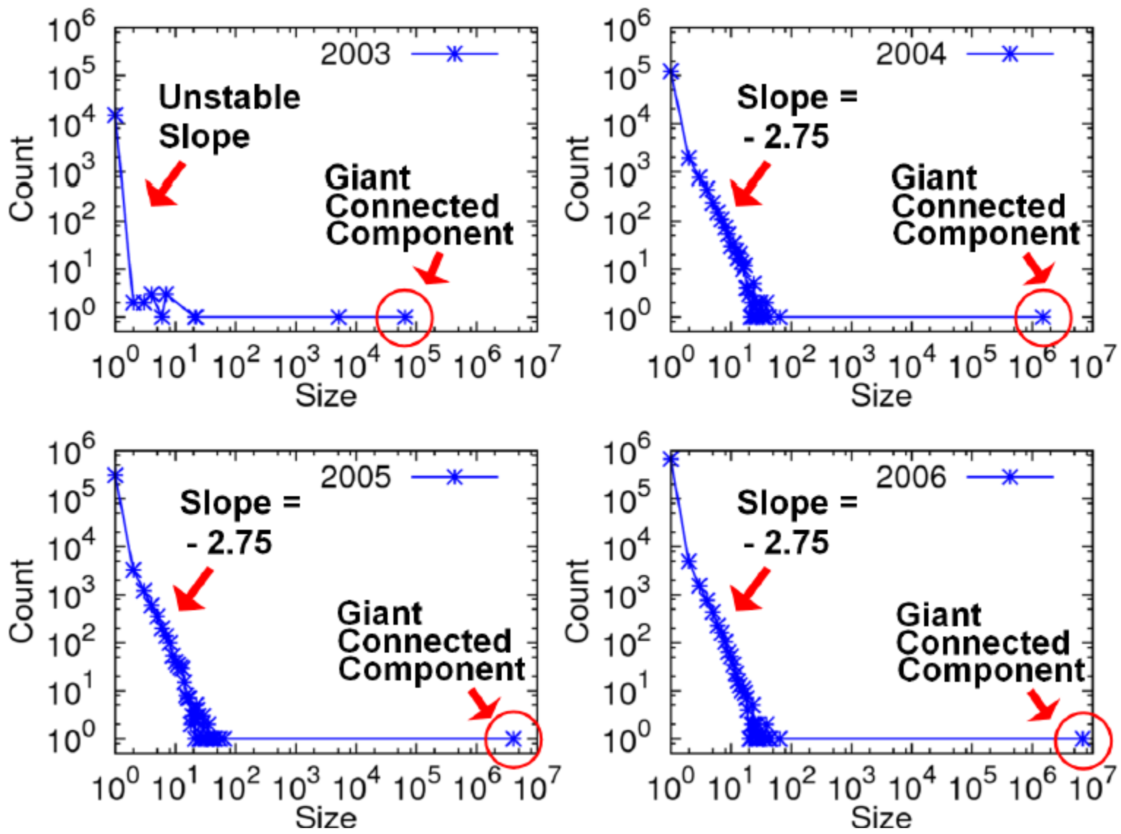
S.6: NLCC

- Connected Components



S.6: persists over time

- Connected Components over Time
- **LinkedIn: 7.5M nodes and 58M edges**



Stable tail slope after the gelling point

List of Static Patterns

- ✓ S.1 degree
- ✓ S.2 eigenvalues
- ✓ S.3 small diameter
- ✓ S.4/5 Triangle laws
- ✓ (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

EigenSpokes

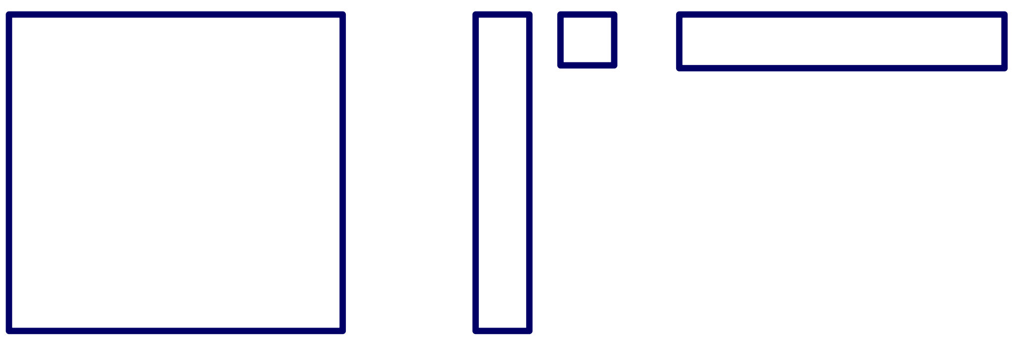
B. Aditya Prakash, Mukund Seshadri, Ashwin Sridharan, Sridhar Machiraju and Christos Faloutsos: *EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs*, PAKDD 2010, Hyderabad, India, 21-24 June 2010.

Useful for fraud detection!

EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

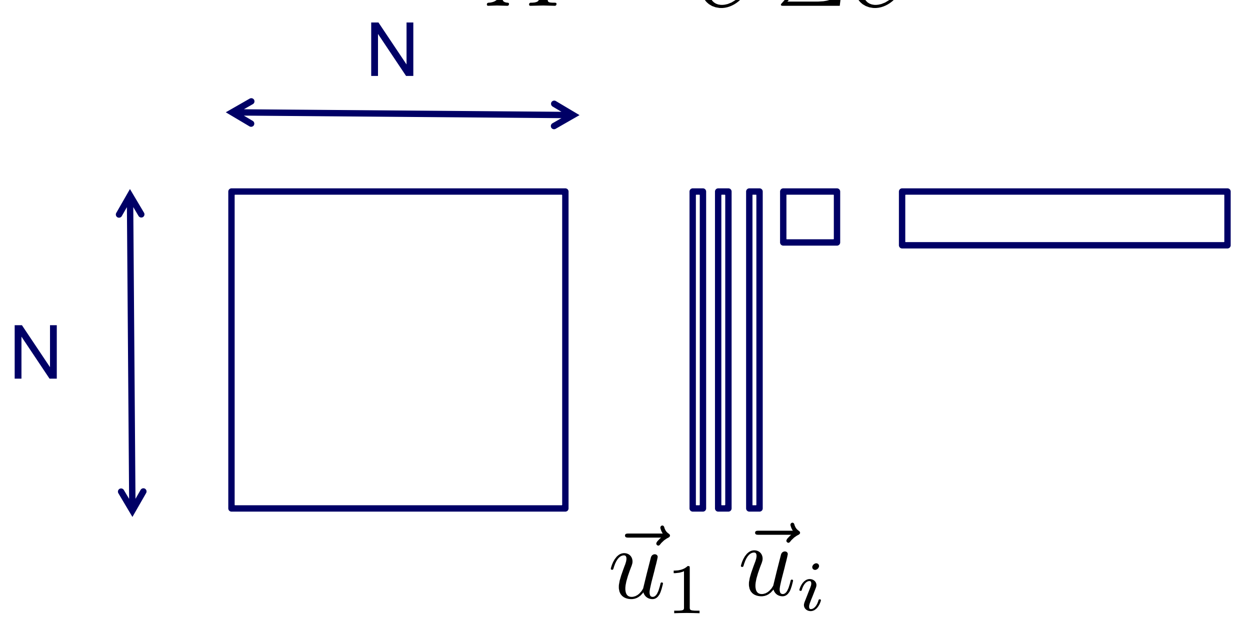
$$A = U \Sigma U^T$$



EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

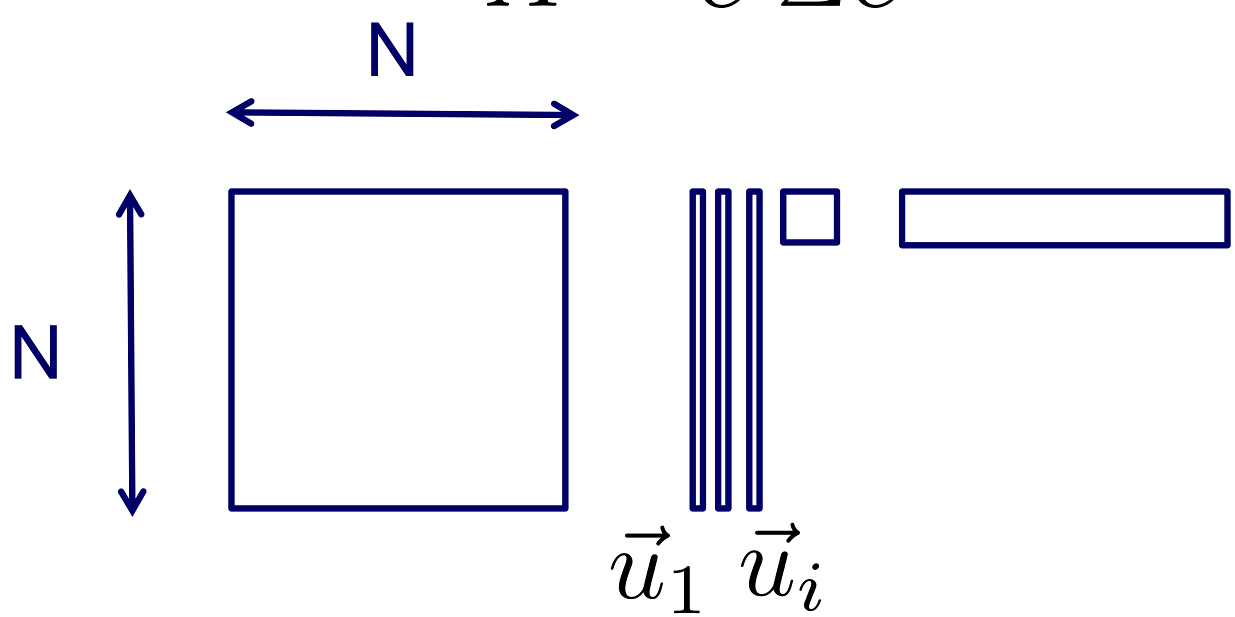
$$A = U \Sigma U^T$$

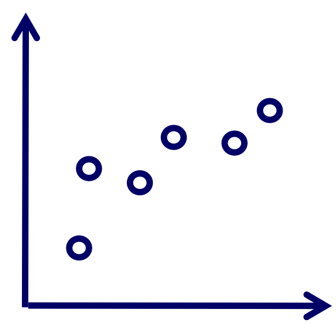


EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

$$A = U \Sigma U^T$$

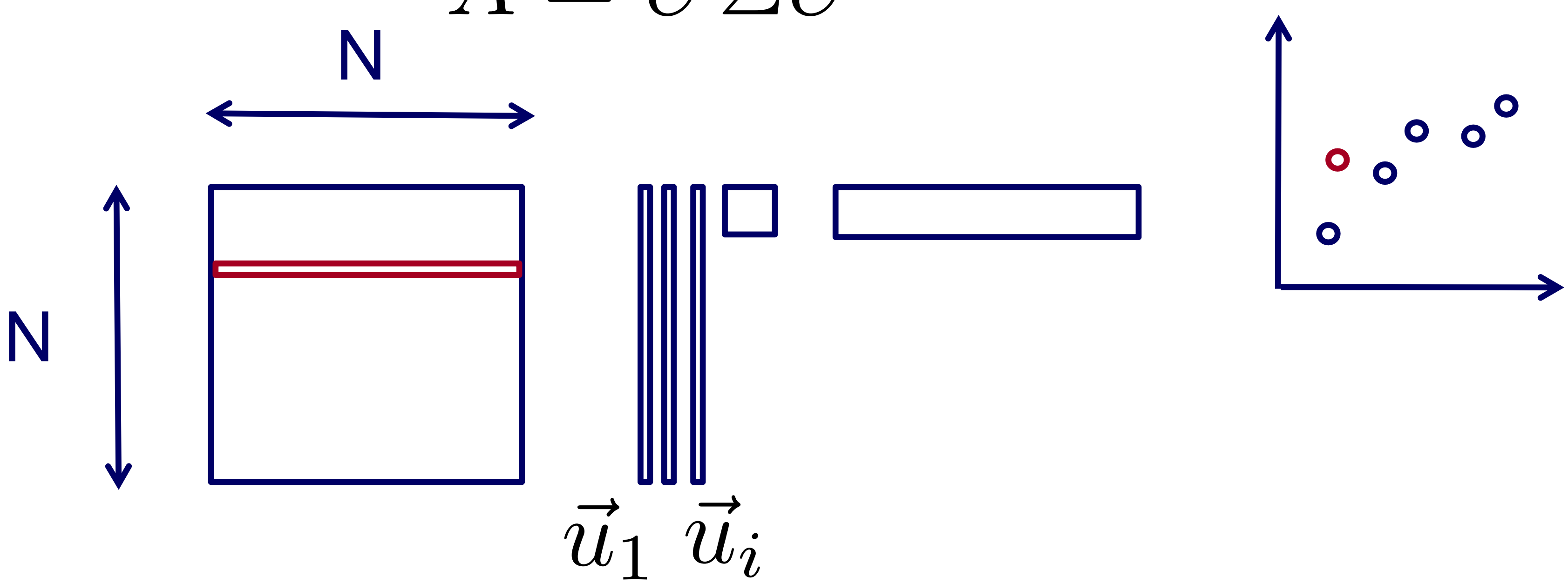




EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

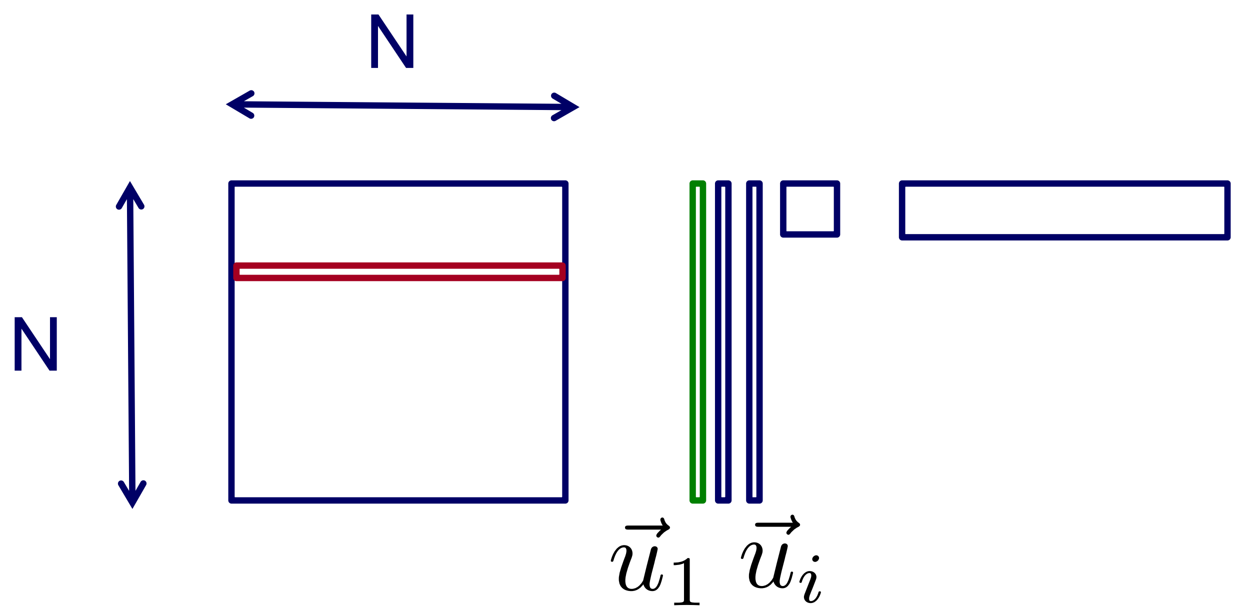
$$A = U \Sigma U^T$$

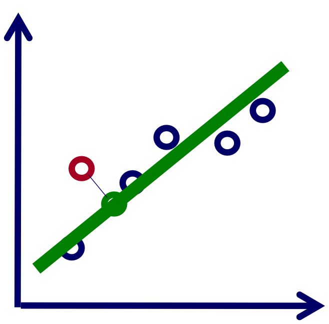


EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

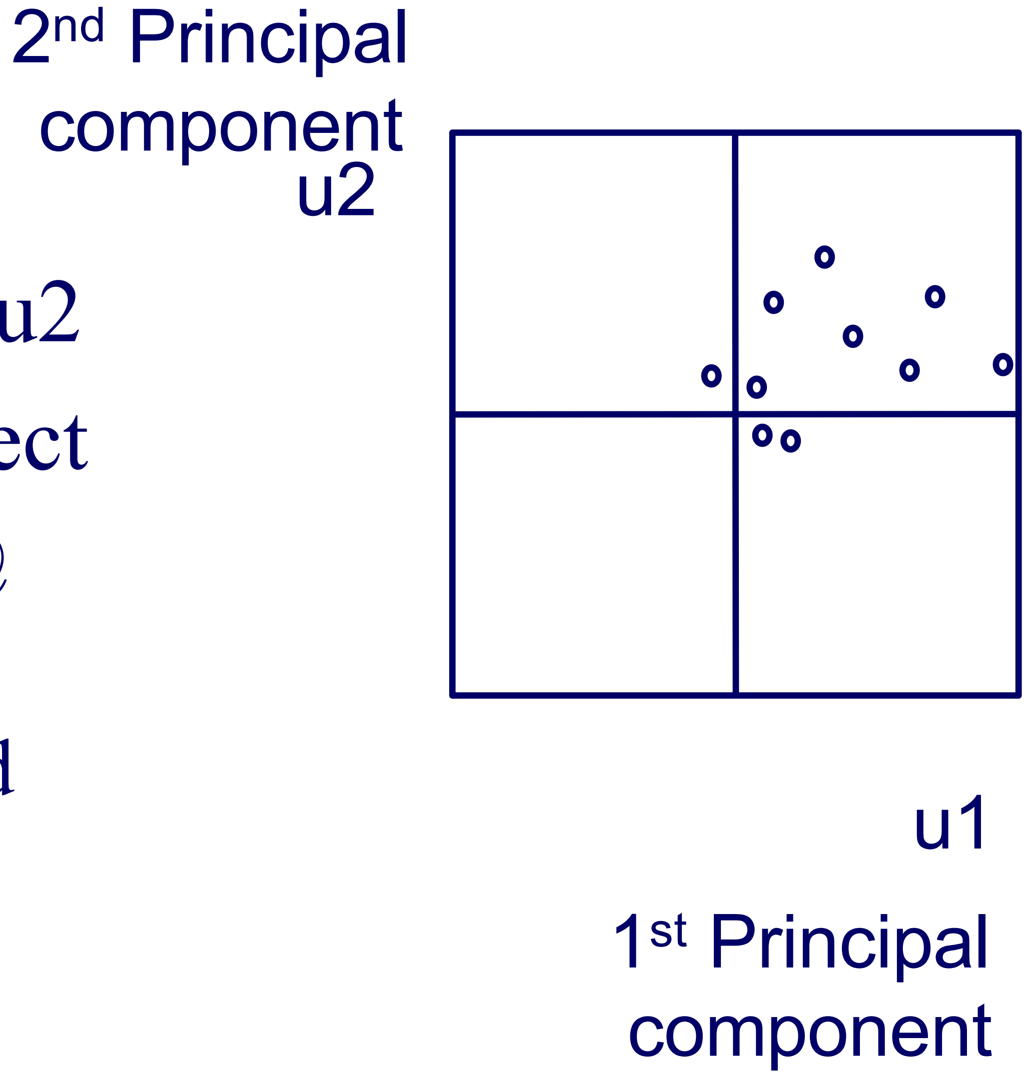
$$A = U \Sigma U^T$$





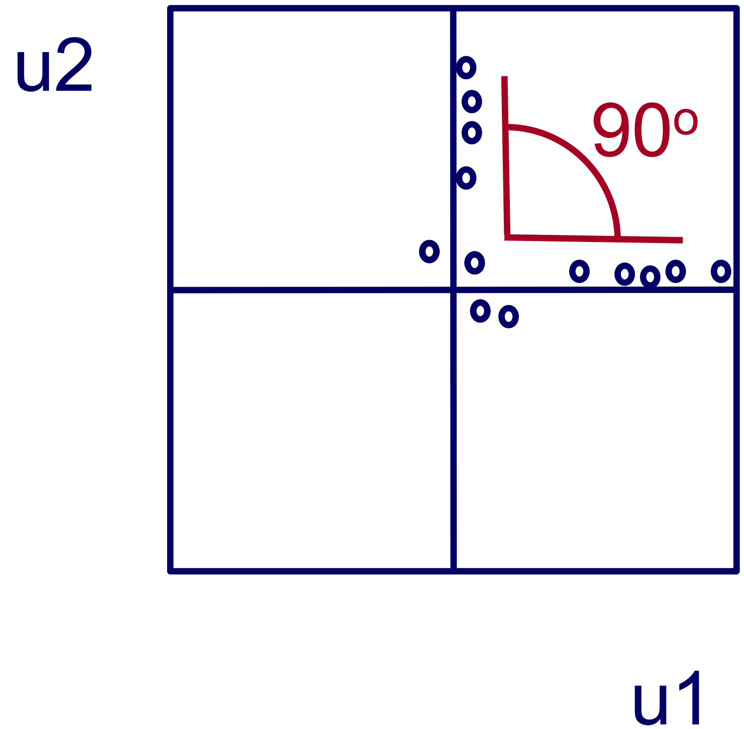
EigenSpokes

- EE plot:
- Scatter plot of scores of u1 vs u2
- One would expect
  - Many points @ origin
  - A few scattered ~randomly



EigenSpokes

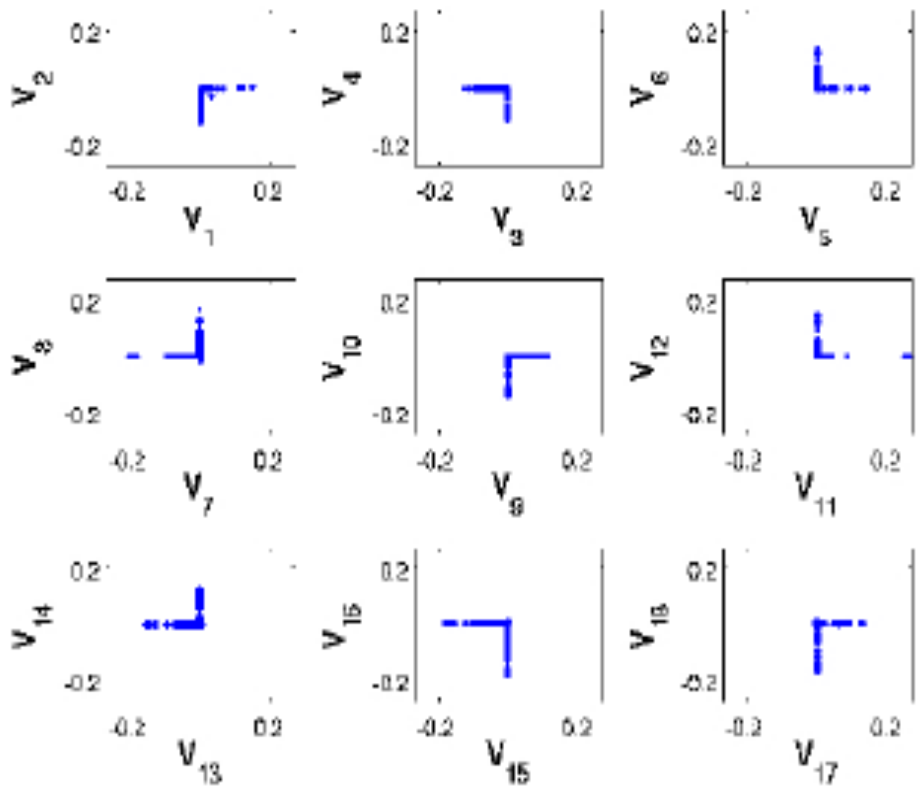
- EE plot:
- Scatter plot of scores of u1 vs u2
- One would expect
  - Many points @ origin
  - A fe[illegible]ttered ~r[illegible]mly



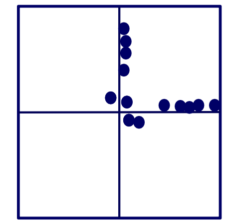
Carnegie Mellon

EigenSpokes - pervasiveness

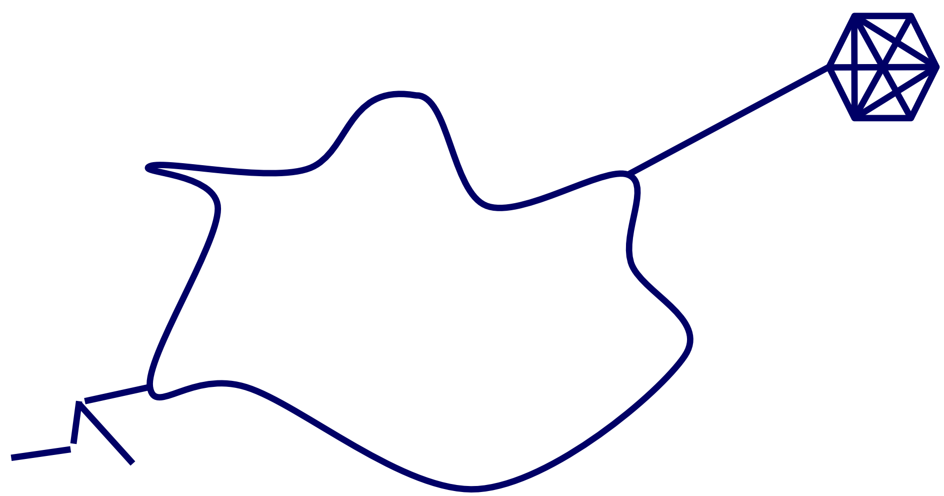
- Present in mobile social graph
  - across time and space
- Patent citation graph



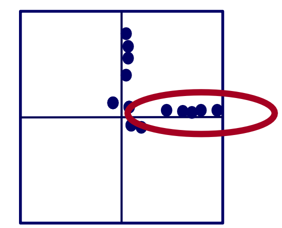
EigenSpokes - explanation



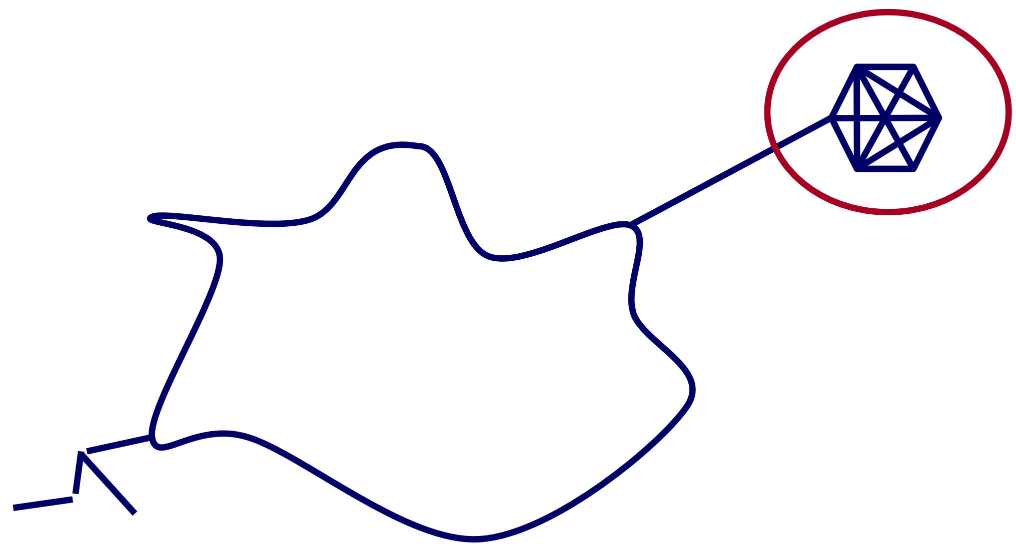
Near-cliques, or near-bipartite-cores, loosely connected



EigenSpokes - explanation

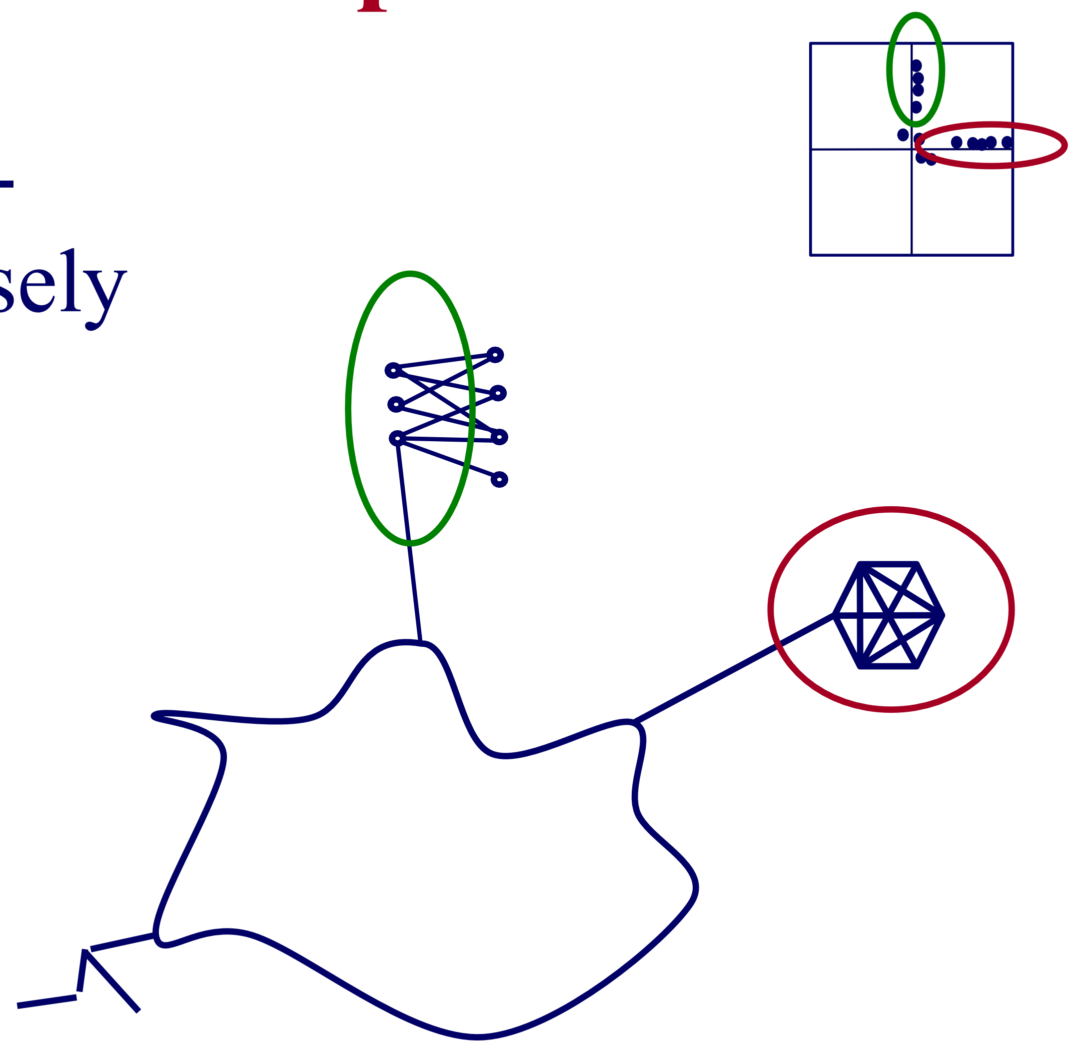


Near-cliques, or near-bipartite-cores, loosely connected

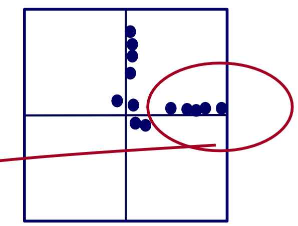


EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected

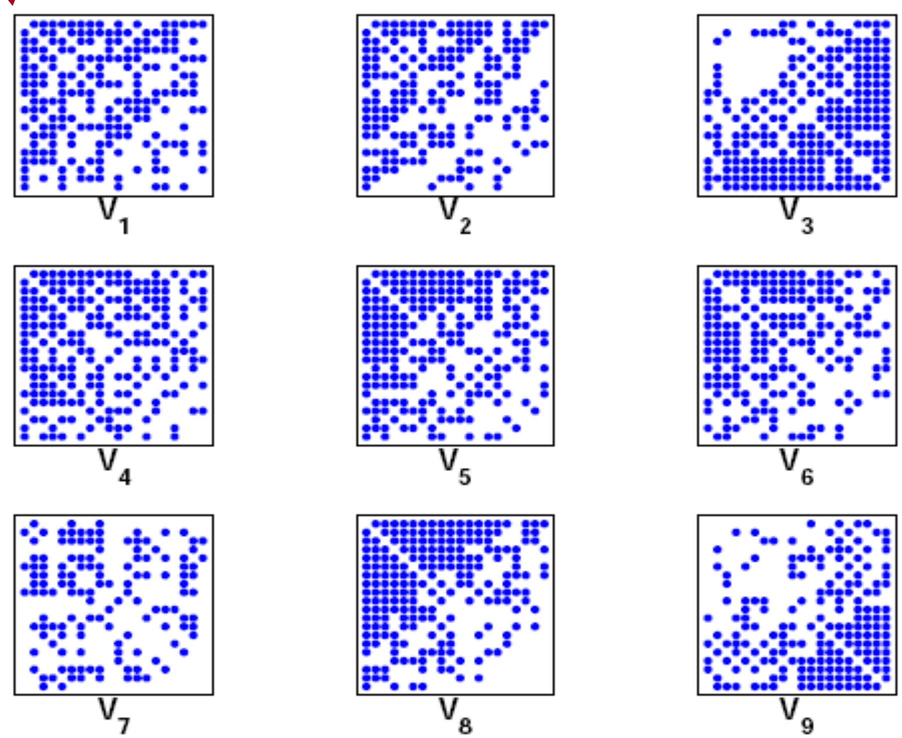


EigenSpokes - explanation



Near-cliques, or near-bipartite-cores, loosely connected

spy plot of top 20 nodes



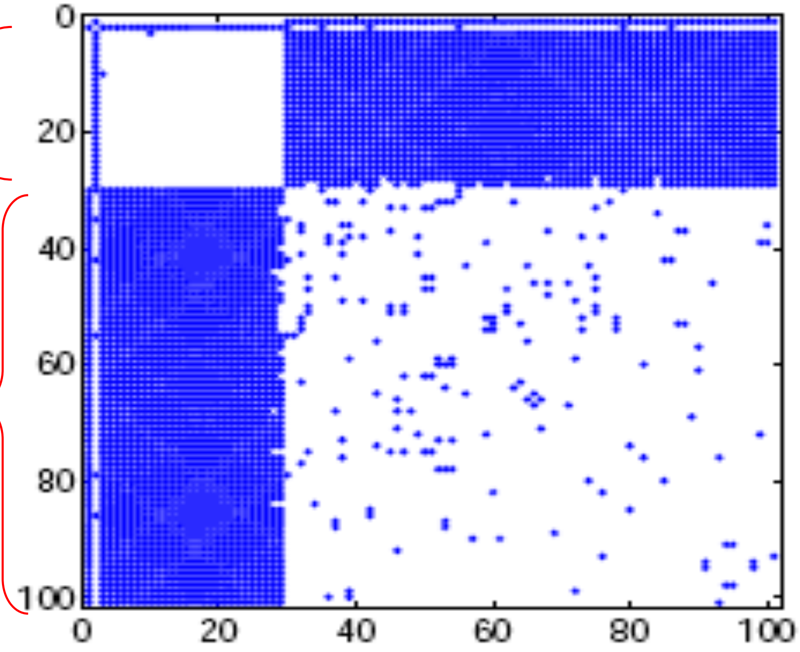
So what?

- Extract nodes with high *scores*
- high connectivity
- Good “communities”

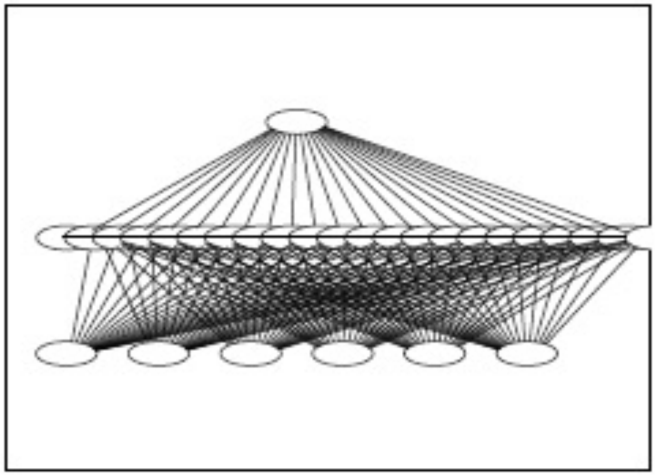
Bipartite Communities!

patents from
same inventor(s)

`cut-and-paste'
bibliography!



magnified bipartite community

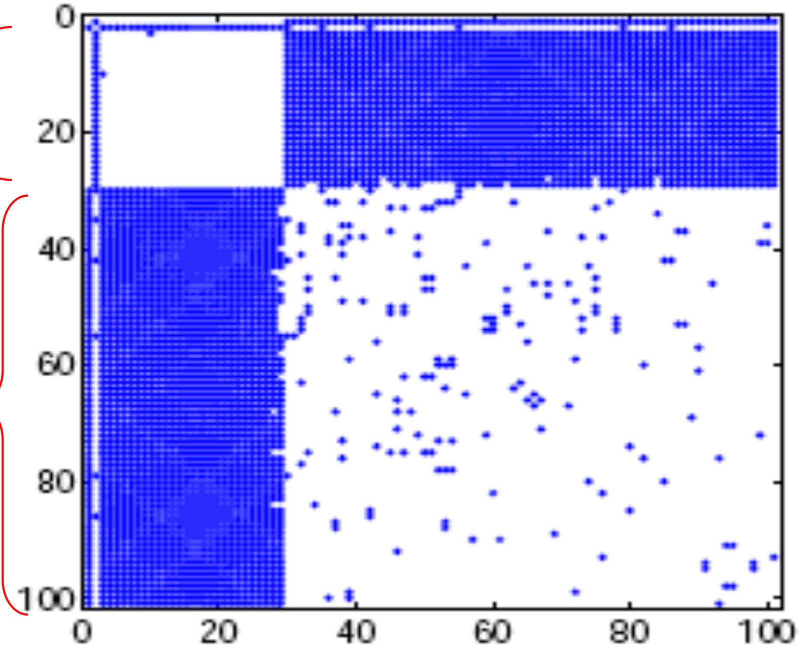


Useful for fraud detection!

Bipartite Communities!

IP – port scanners

victims



Useful for fraud detection!

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

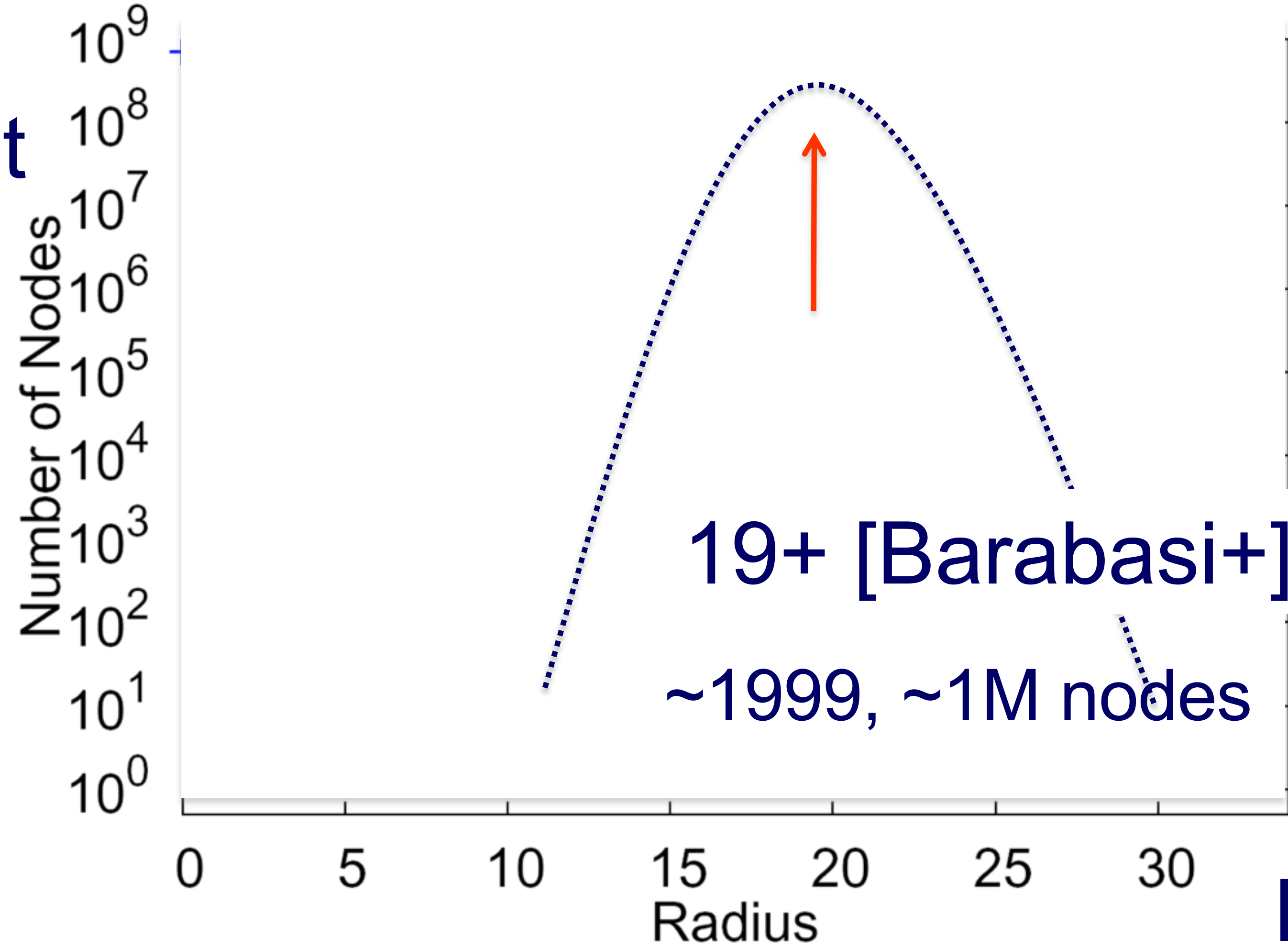
In textbook

Carnegie Mellon

HADI for diameter estimation

- *Radius Plots for Mining Tera-byte Scale Graphs* **U Kang**, Charalampos Tsourakakis, Ana Paula Appel, Christos Faloutsos, Jure Leskovec, SDM'10
- Naively: diameter needs **O(N**2)** space and up to O(N**3) time – **prohibitive** (N~1B)
- Our HADI: linear on E (~10B)
  - Near-linear scalability wrt # machines
  - Several optimizations -> 5x faster

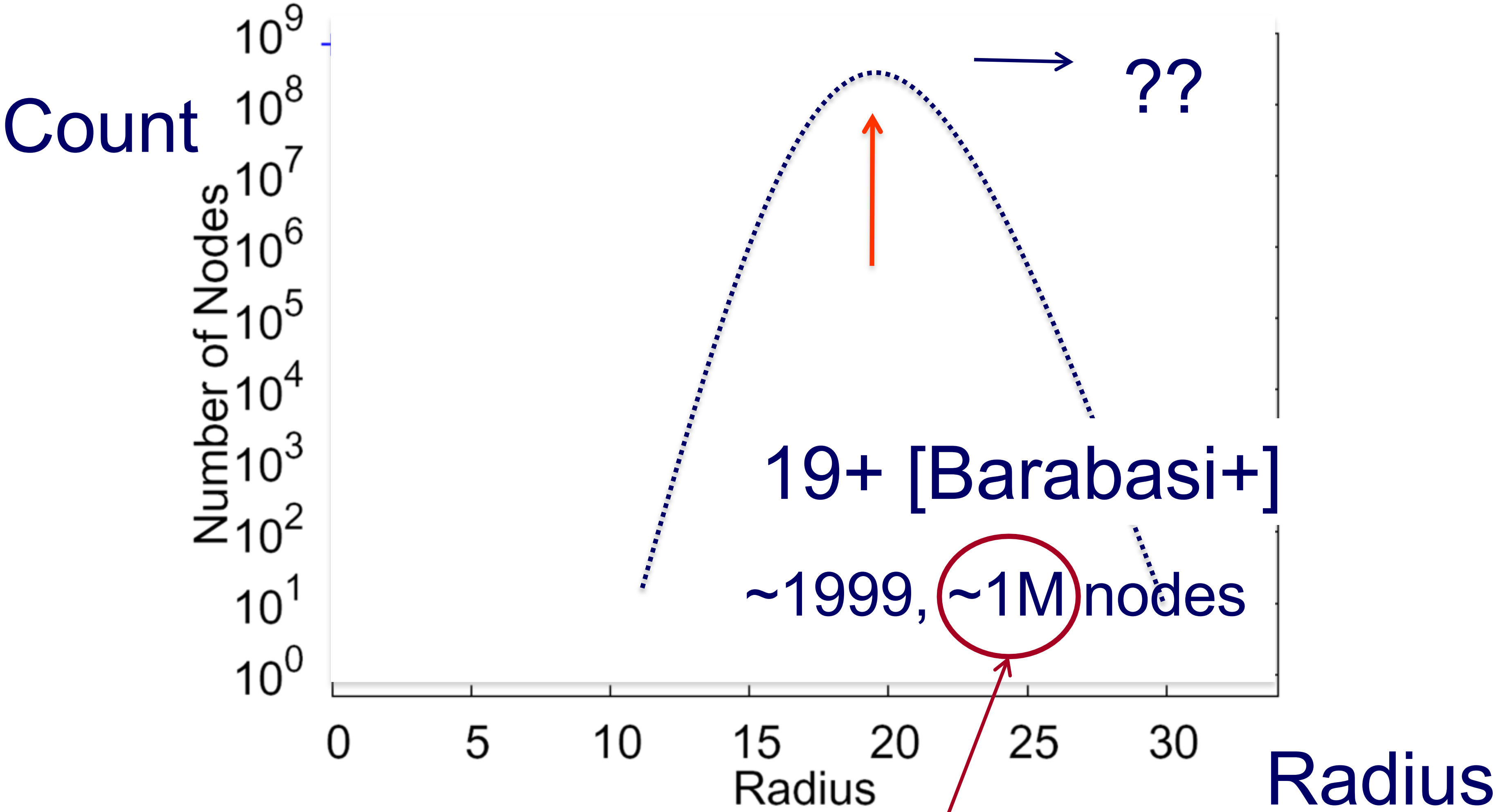
Count



19+ [Barabasi+]

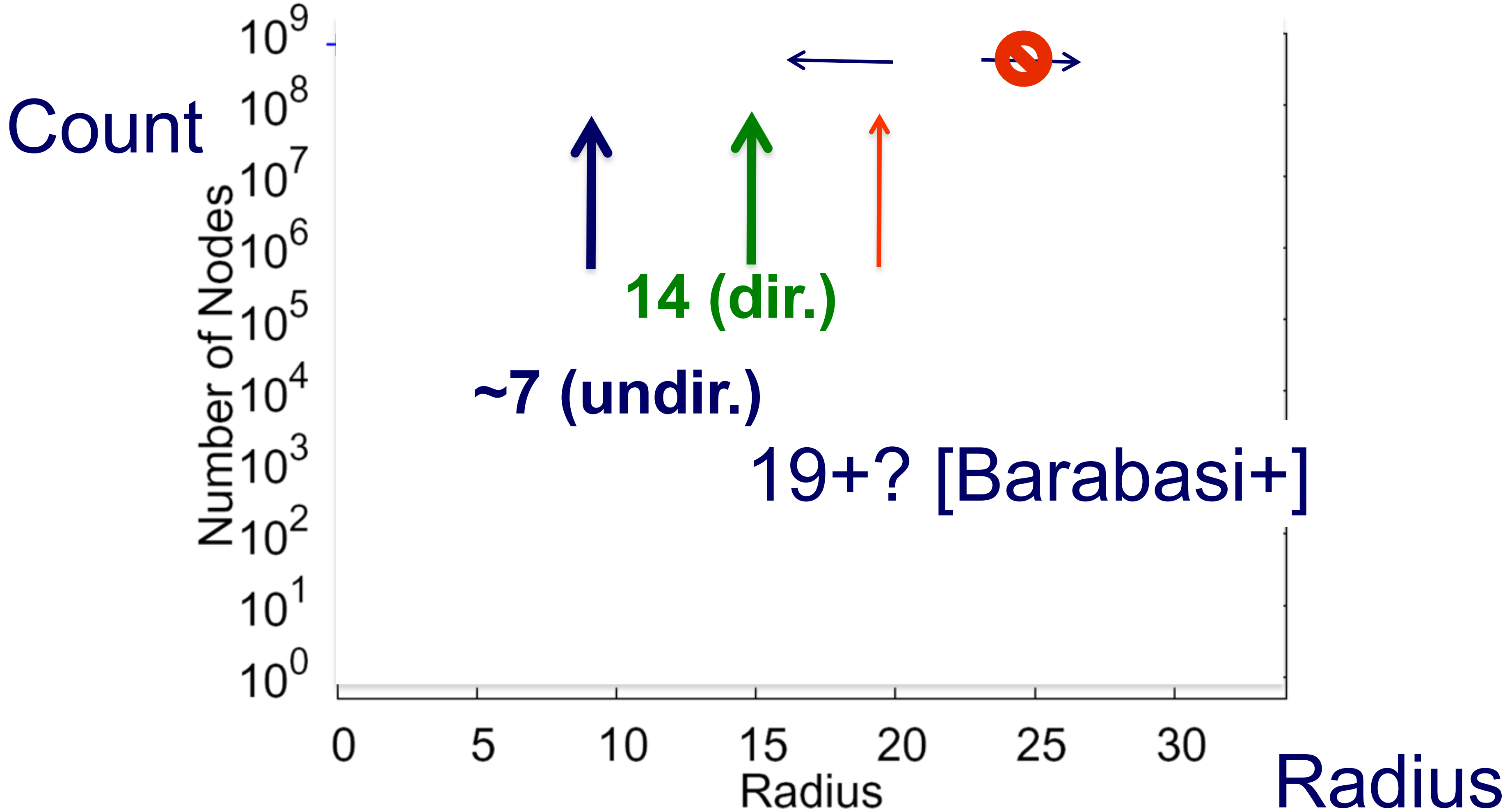
~1999, ~1M nodes

Radius



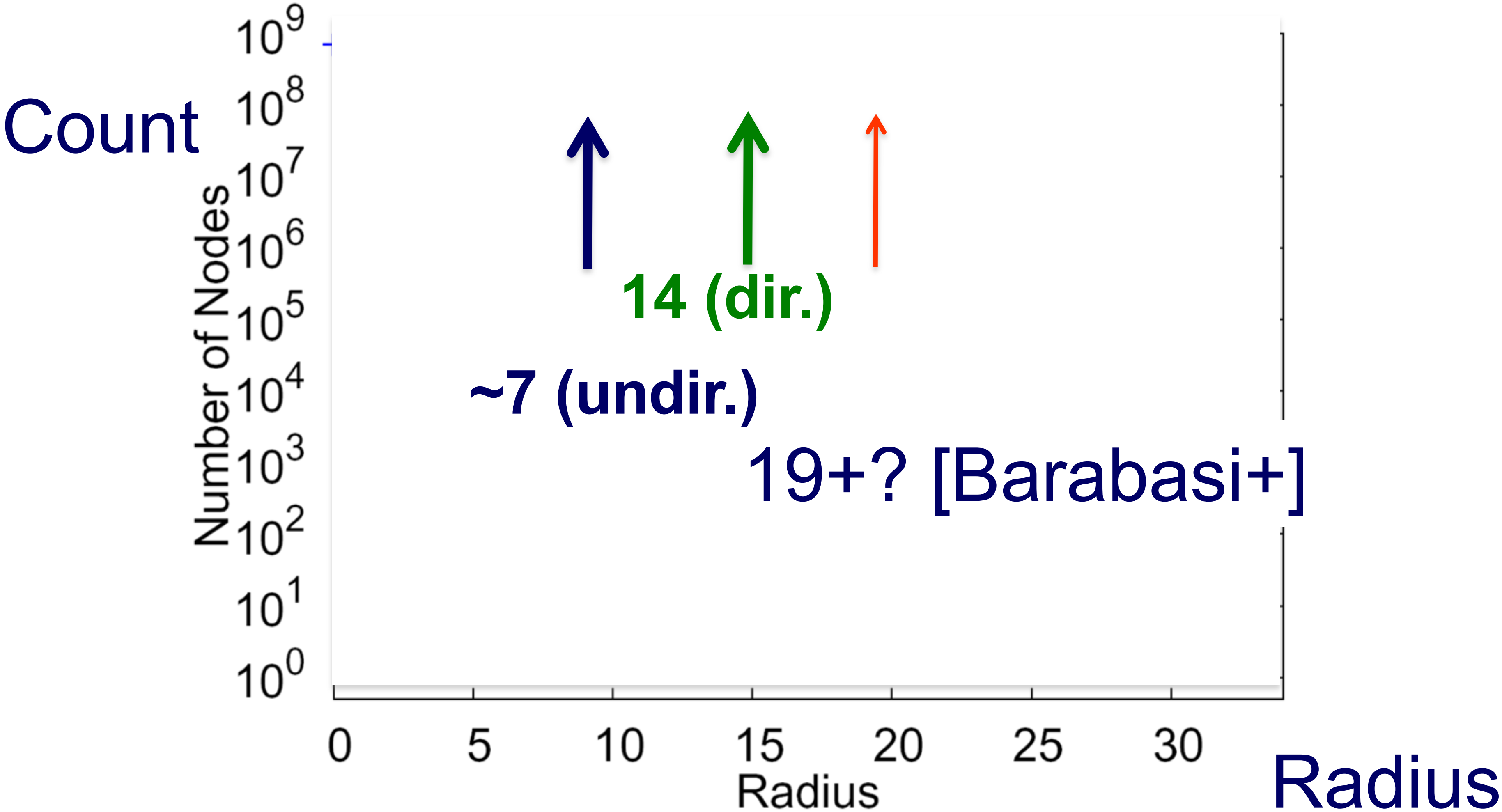
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.



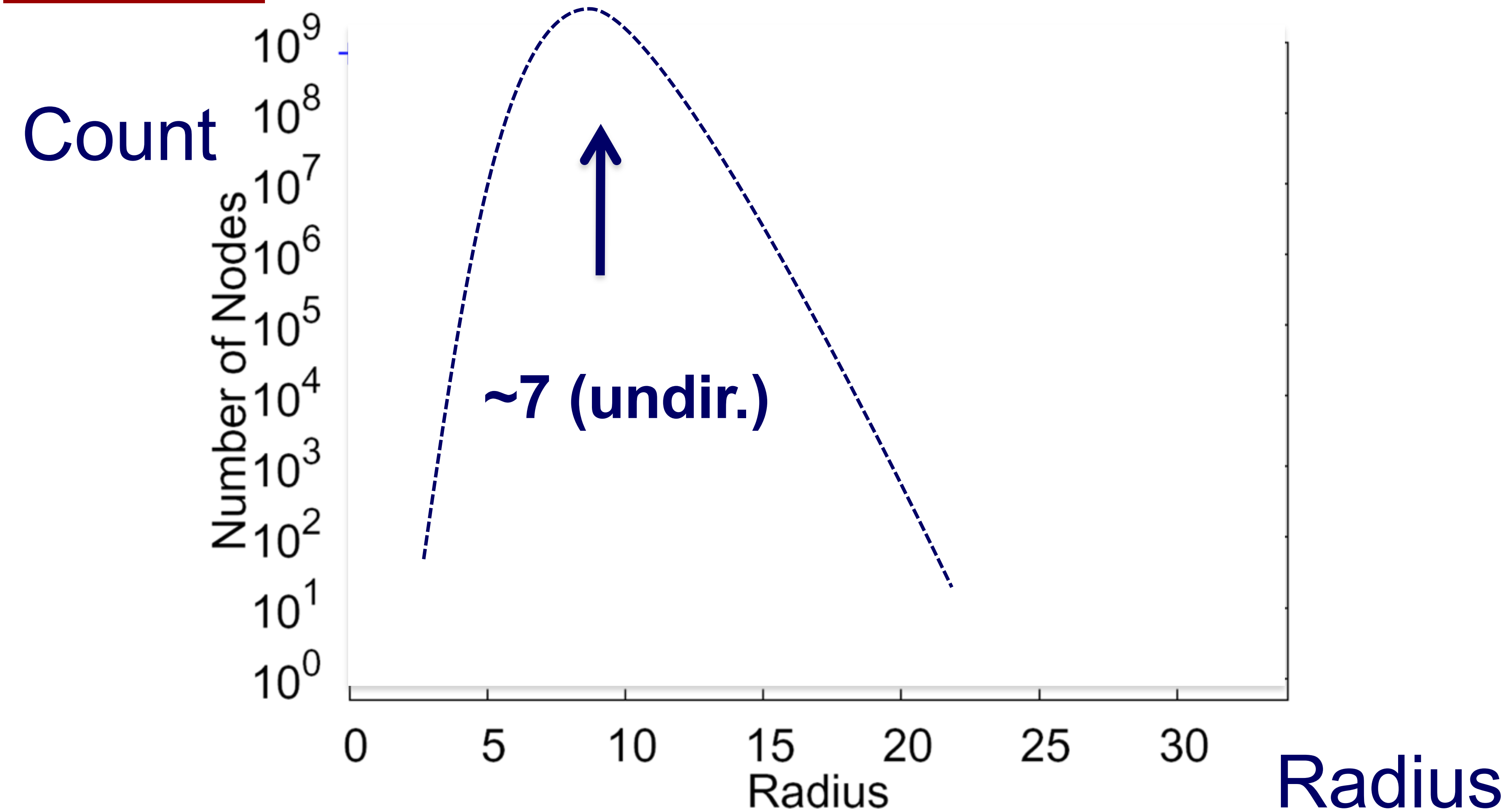
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.

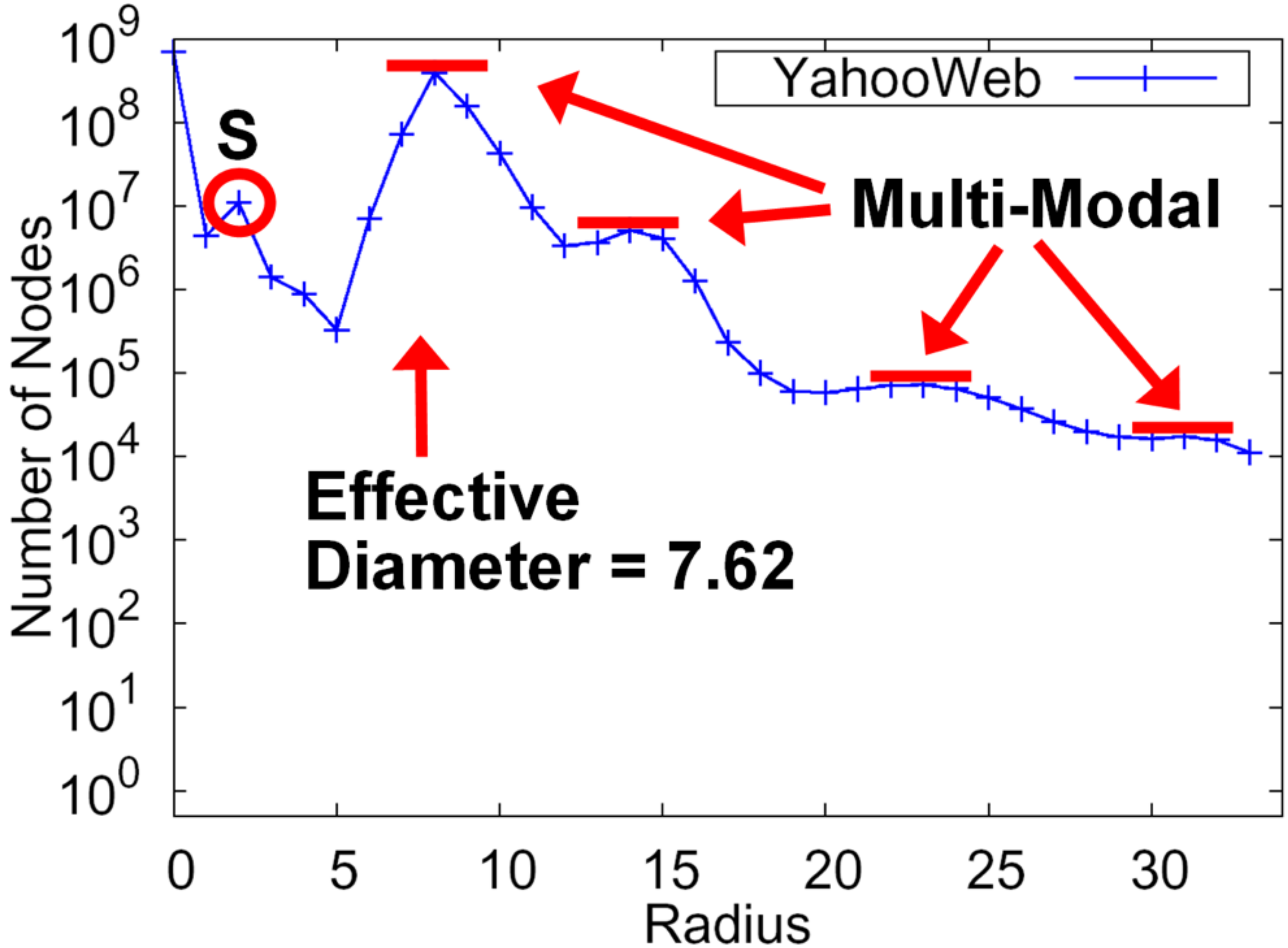


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- 7 degrees of separation (!)
- Diameter: shrunk



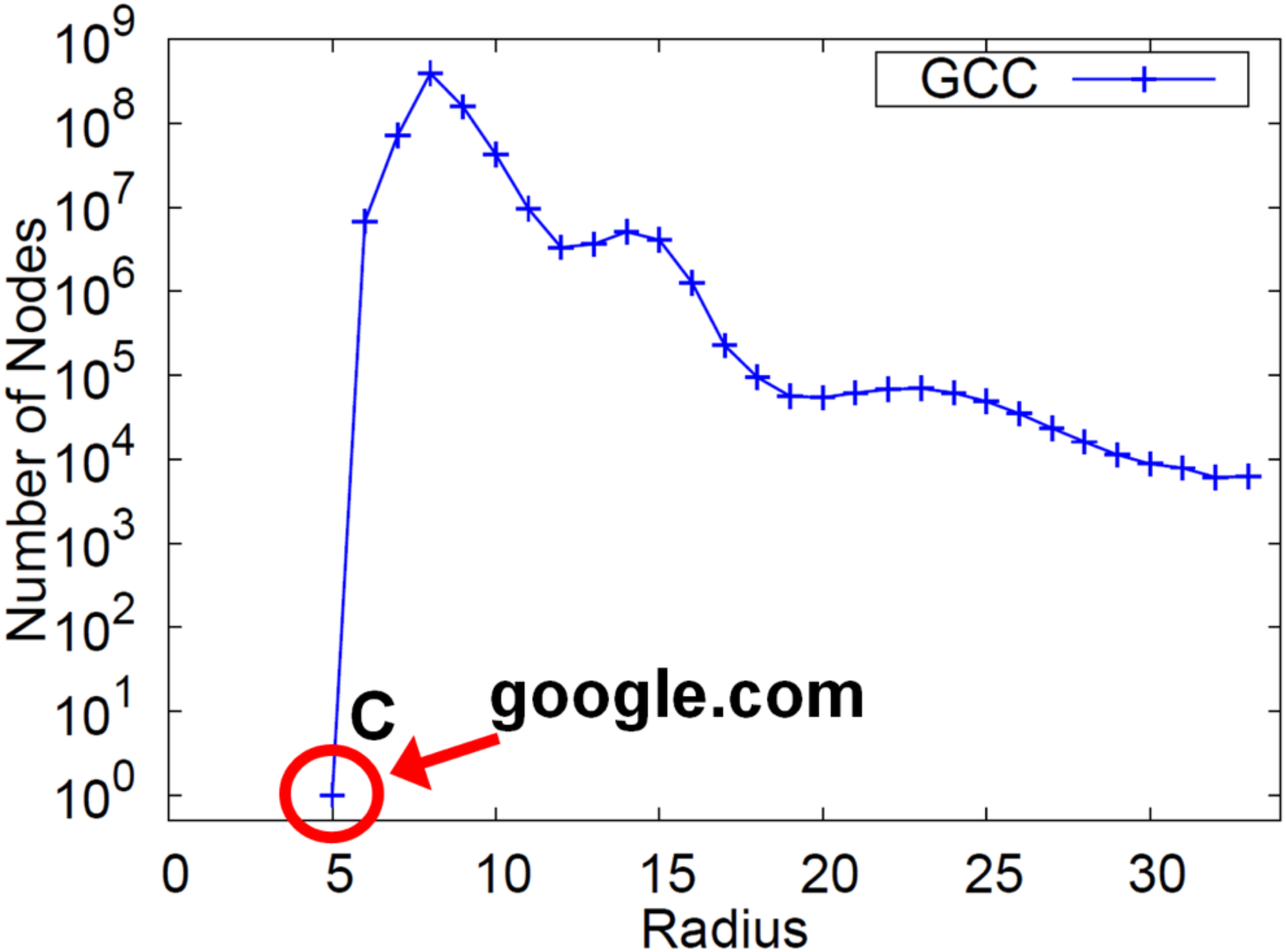
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
Q: Shape?



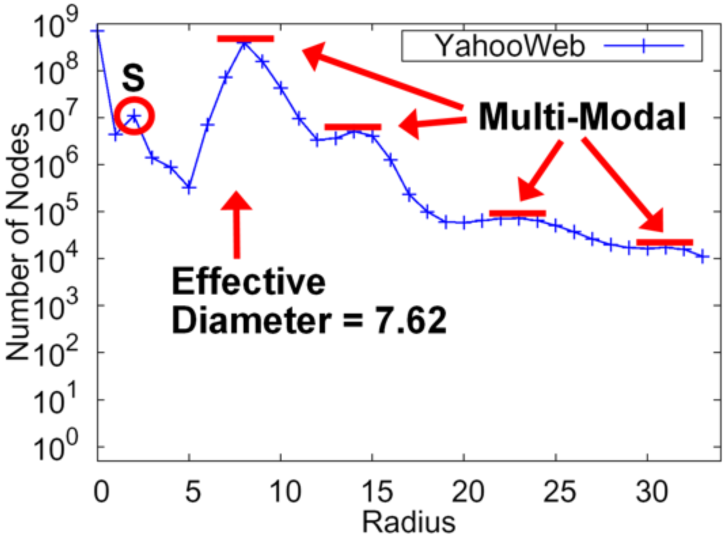
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality (?!)

Carnegie Mellon

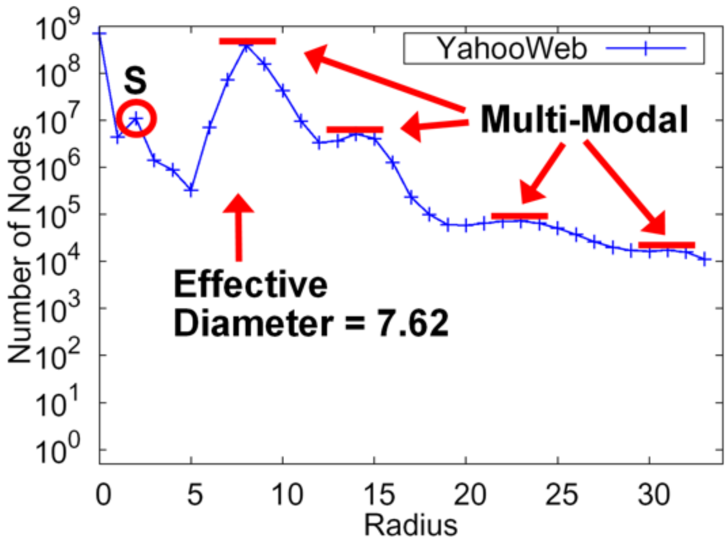


Radius Plot of **GCC** of YahooWeb.

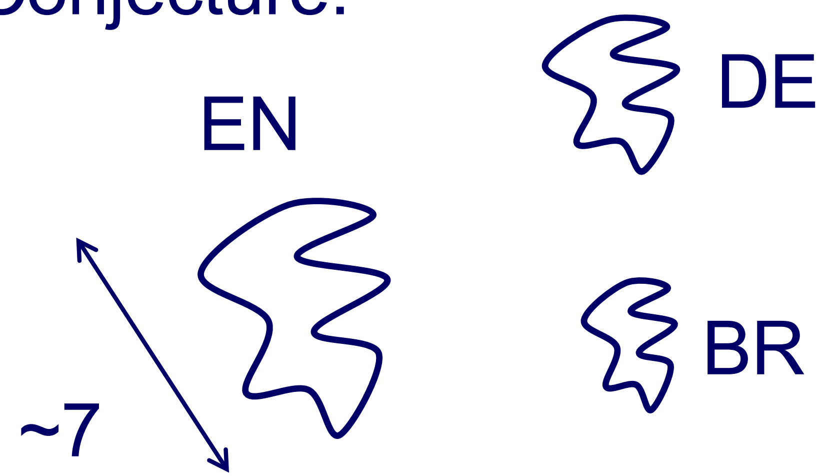


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .

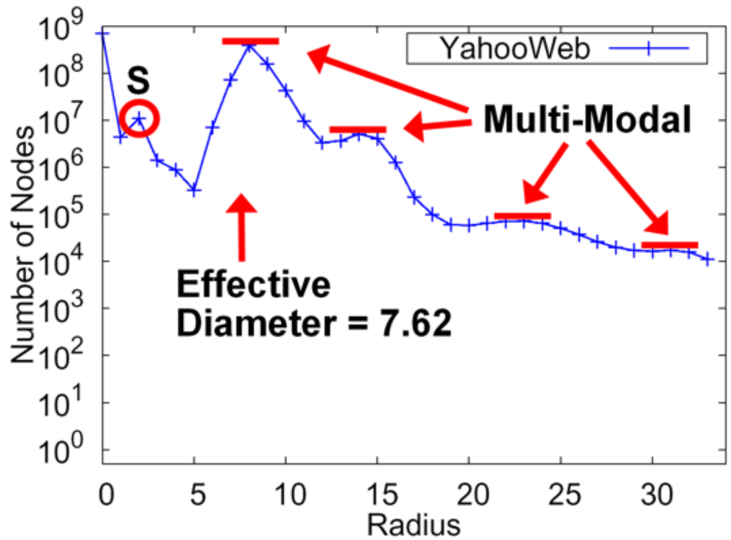


Conjecture:

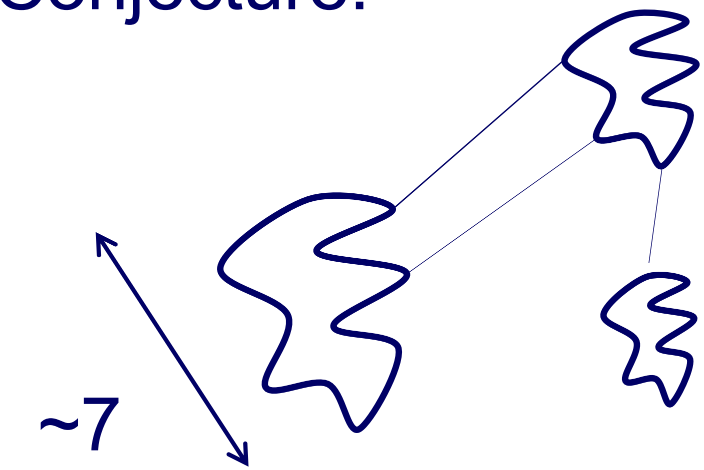


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .



Conjecture:



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot
- Other observations / patterns?

In textbook

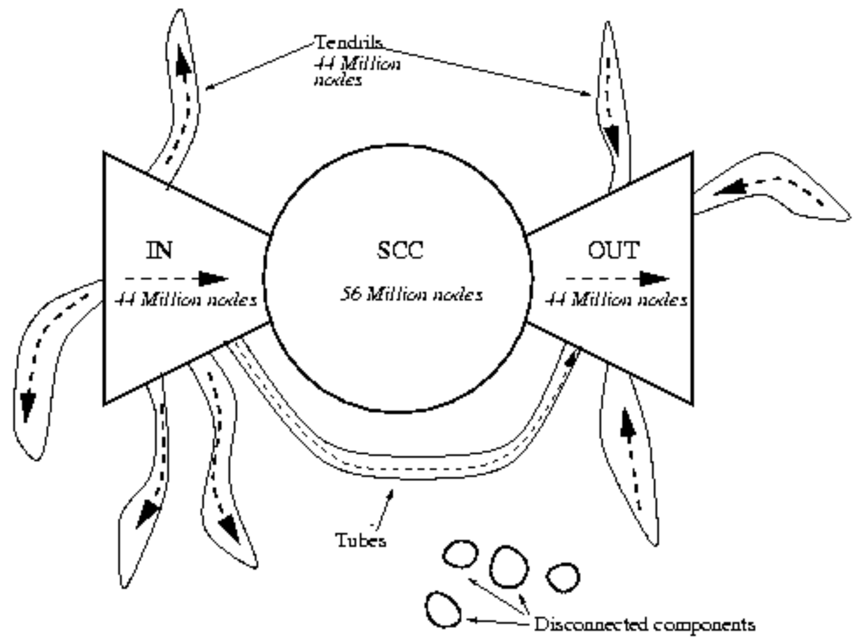
Any other 'laws'?

Yes!

- Small diameter (~ constant!) –
  - six degrees of separation / 'Kevin Bacon'
  - small worlds [Watts and Strogatz]

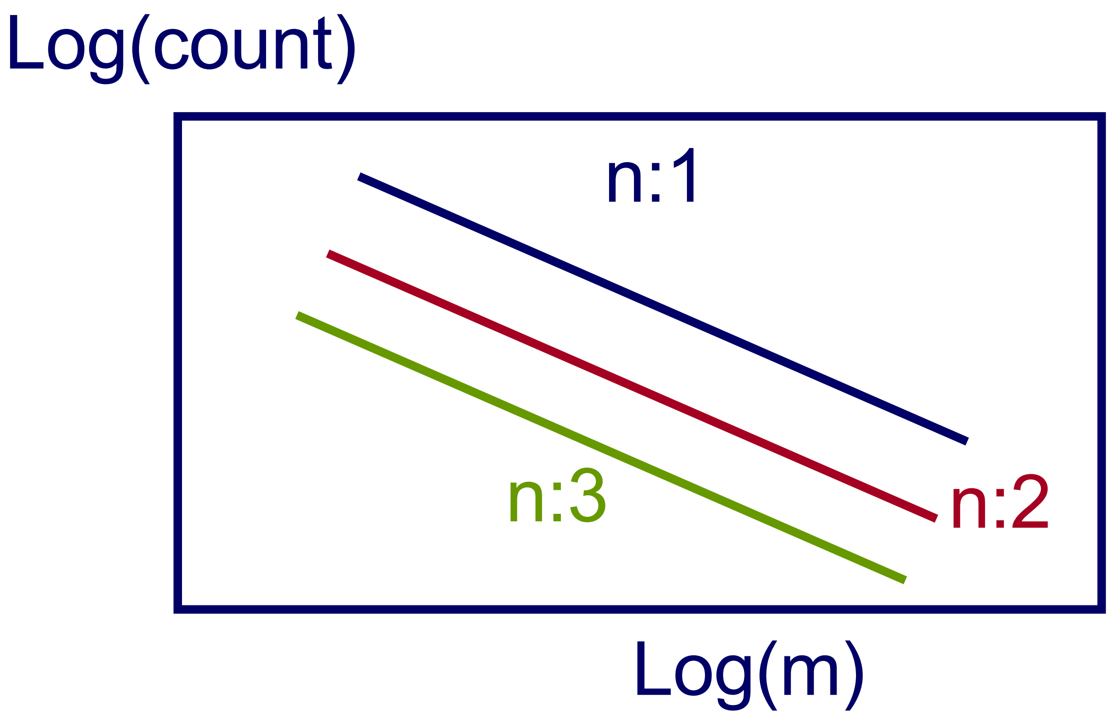
Any other 'laws'?

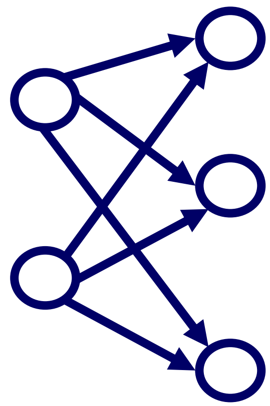
- Bow-tie, for the web [Kumar+ '99]
- IN, SCC, OUT, 'tendrils'
- disconnected components



Any other 'laws'?

- power-laws in communities (bi-partite cores) [Kumar+, '99]

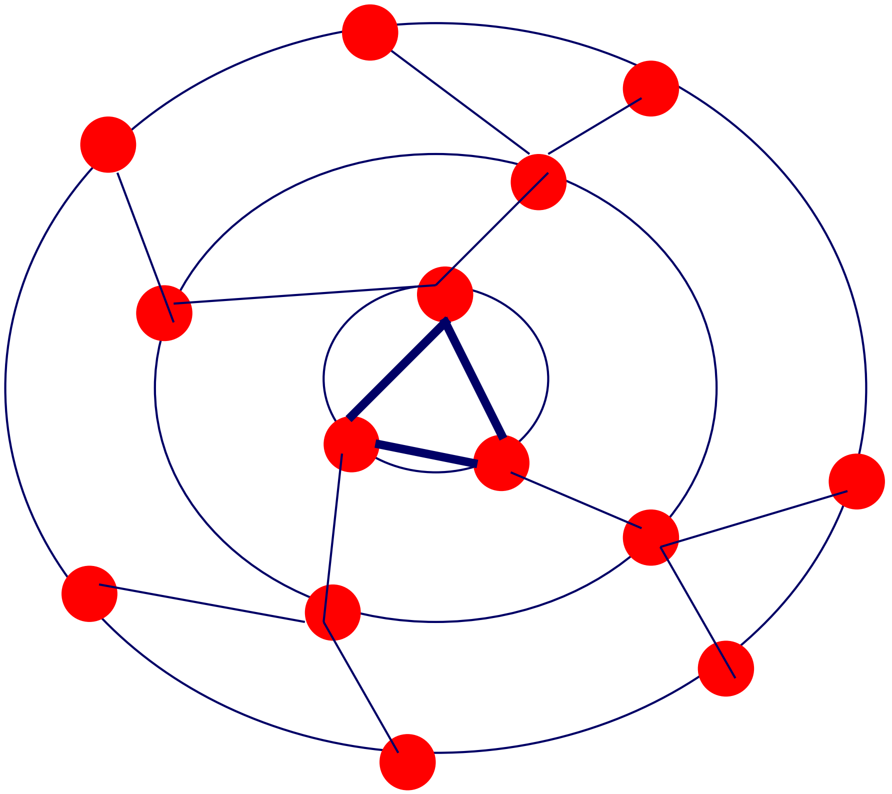




2:3 core
(m:n core)

Any other 'laws'?

- "Jellyfish" for Internet [Tauro+ '01]
- core: ~clique
- ~5 concentric layers
- many 1-degree nodes



Outline

- Introduction – Motivation
- Problem: Patterns in graphs
  - Static graphs
    - degree, diameter, eigen,
    - Triangles
  - ➡ Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

Observations on weighted graphs?

- A: yes - even more 'laws' !

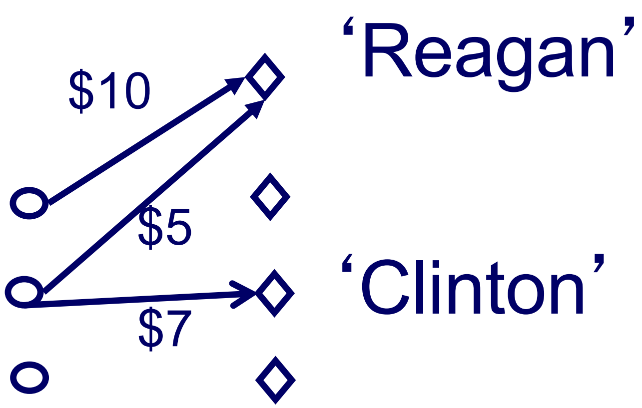
M. McGlohon, L. Akoglu, and C. Faloutsos
Weighted Graphs and Disconnected Components: Patterns and a Generator.
SIG-KDD 2008

Observation W.1: Fortification

Q: How do the weights of nodes relate to degree?

Observation W.1: Fortification

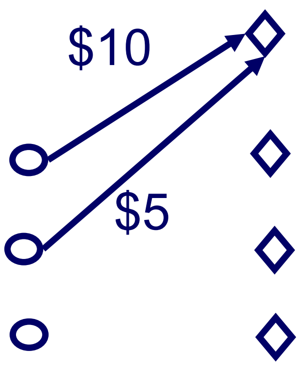
More donors, more $?



Observation W.1: fortification: Snapshot Power Law

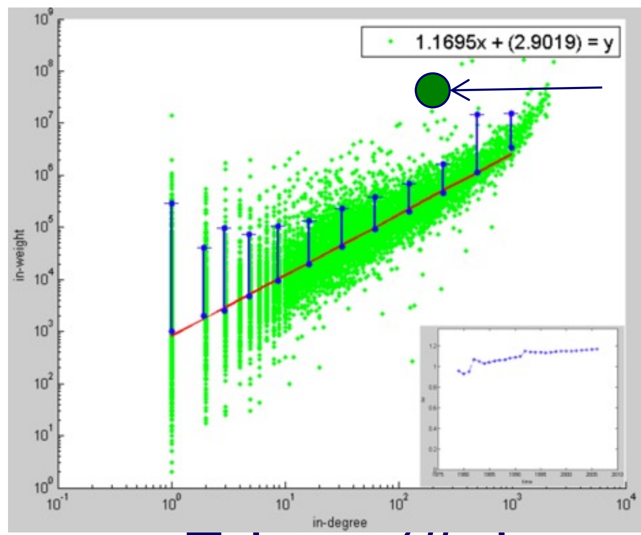
- Weight: super-linear on in-degree
- exponent ‘iw’ : 1.01 < iw < 1.26

More donors, even more $



In-weights ($)

Orgs-Candidates



e.g. John Kerry, $10M received, from 1K donors

Edges (# donors)

Outline

- Introduction – Motivation
- Problem: Patterns in graphs
  - Static graphs
  - Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

Problem: Time evolution

- with Jure Leskovec (CMU -> Stanford)

- and Jon Kleinberg (Cornell – sabb. @ CMU)

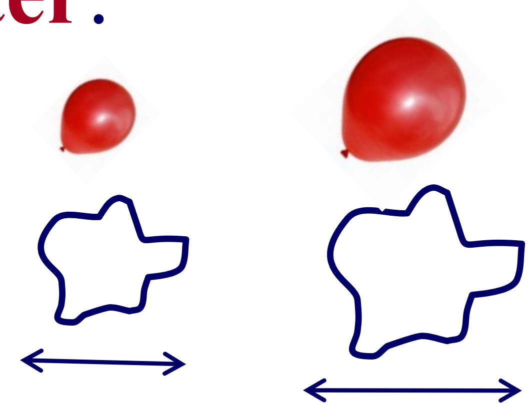
List of Dynamic Patterns

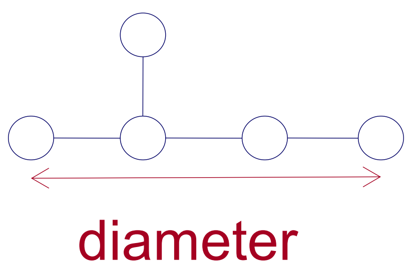
- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

D.1 Evolution of the Diameter

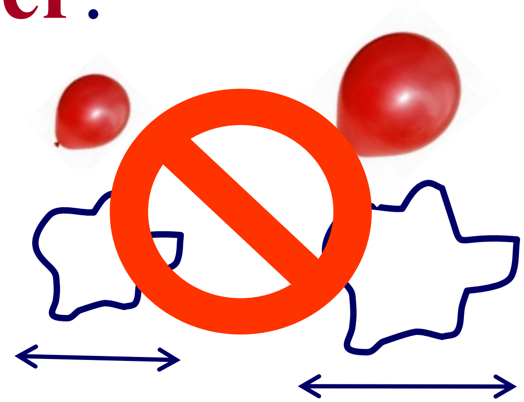
- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - [diameter ~ O($N^{1/3}$)]
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?





D.1 Evolution of the Diameter

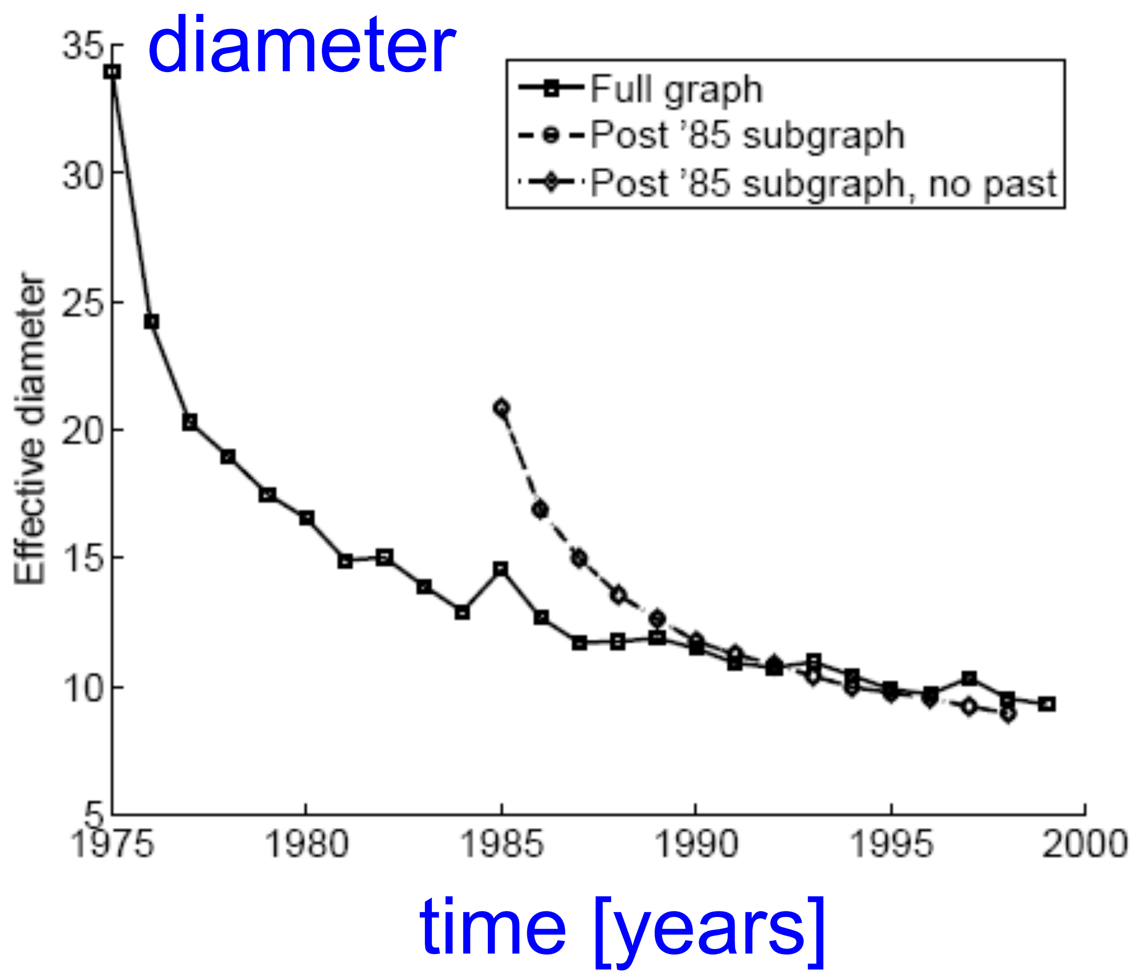
- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - [diameter ~ $O(N^{1/3})$]
  - diameter ~ $O(\log N)$
  - diameter ~ $O(\log \log N)$



- What is happening in real data?
- Diameter **shrinks** over time

D.1 Diameter – "Patents"

- Patent citation network
- 25 years of data
- @1999
  - – 2.9 M nodes
  - – 16.5 M edges



List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

D.2 Temporal Evolution of the Graphs

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

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

- Q: what is your guess for

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

D.2 Temporal Evolution of the Graphs

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

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

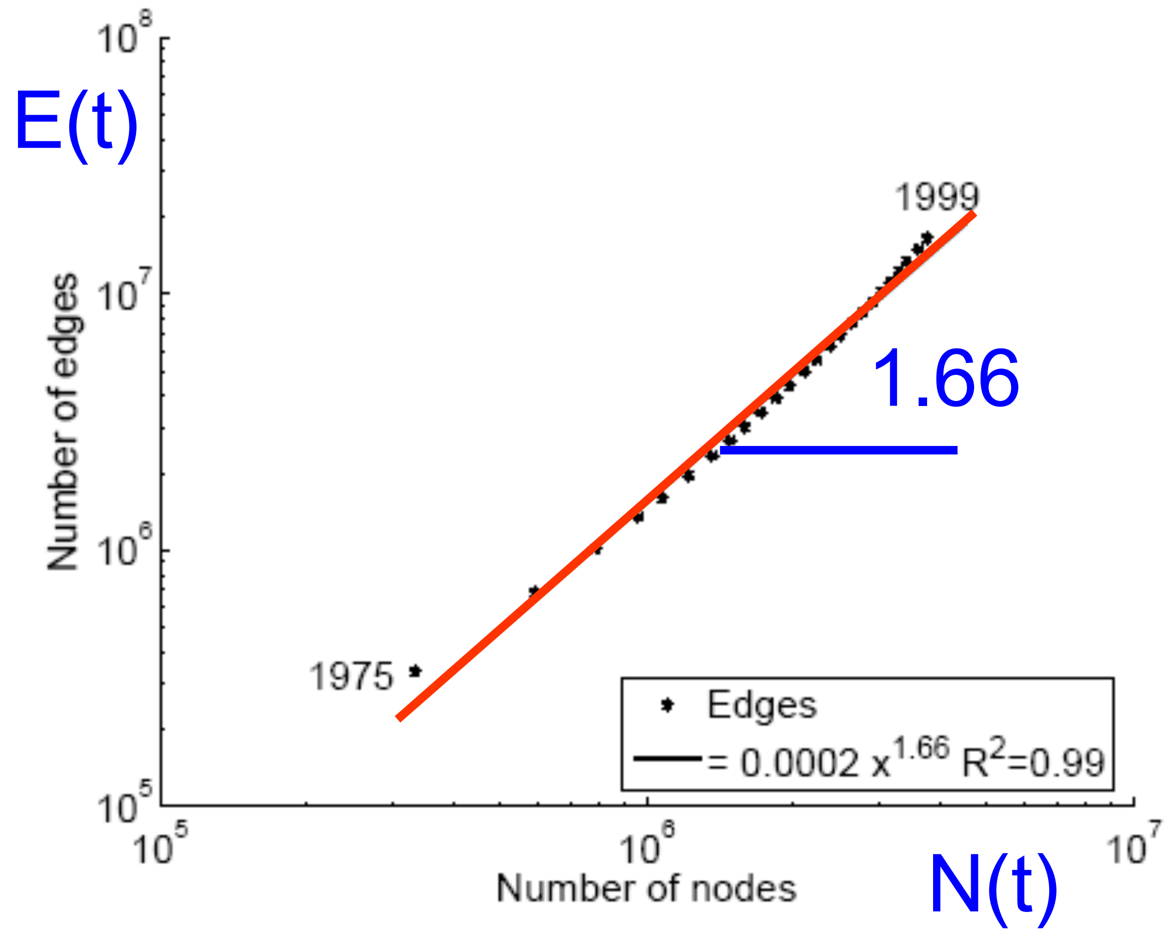
- Q: what is your guess for

  $$E(t+1) \cancel{=?\ 2} * E(t)$$

- A: over-doubled!
  - But obeying the ``Densification Power Law''

D.2 Densification – Patent Citations

- Citations among patents granted
- @1999
  - 2.9 M nodes
  - 16.5 M edges
- Each year is a datapoint



List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

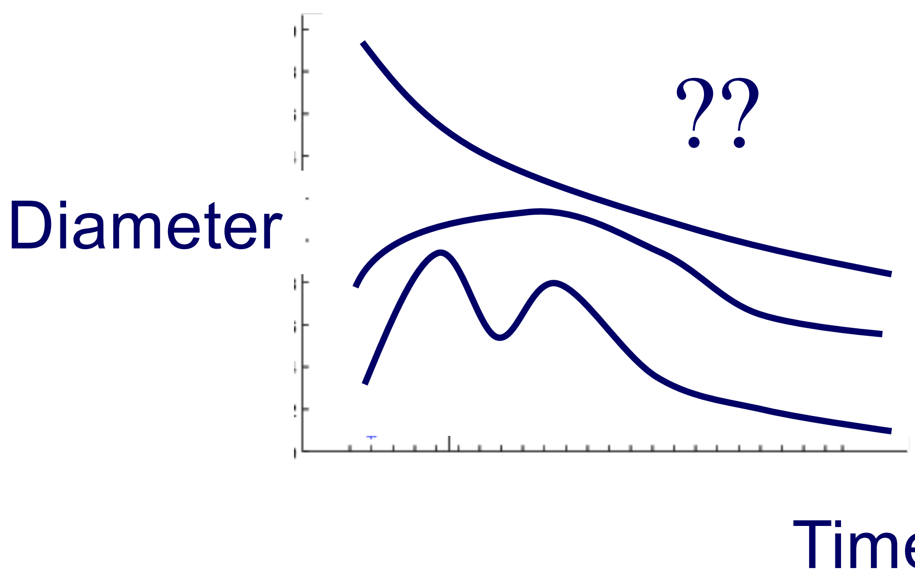
More on Time-evolving graphs

M. McGlohon, L. Akoglu, and C. Faloutsos
Weighted Graphs and Disconnected Components: Patterns and a Generator.
SIG-KDD 2008

D.3 Gelling Point

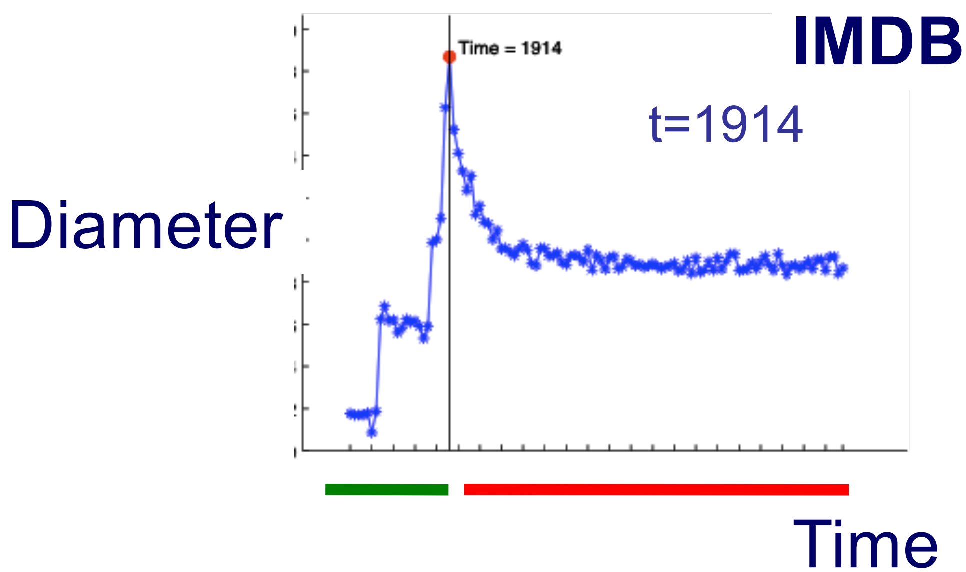
- Diameter, over time

IMDB



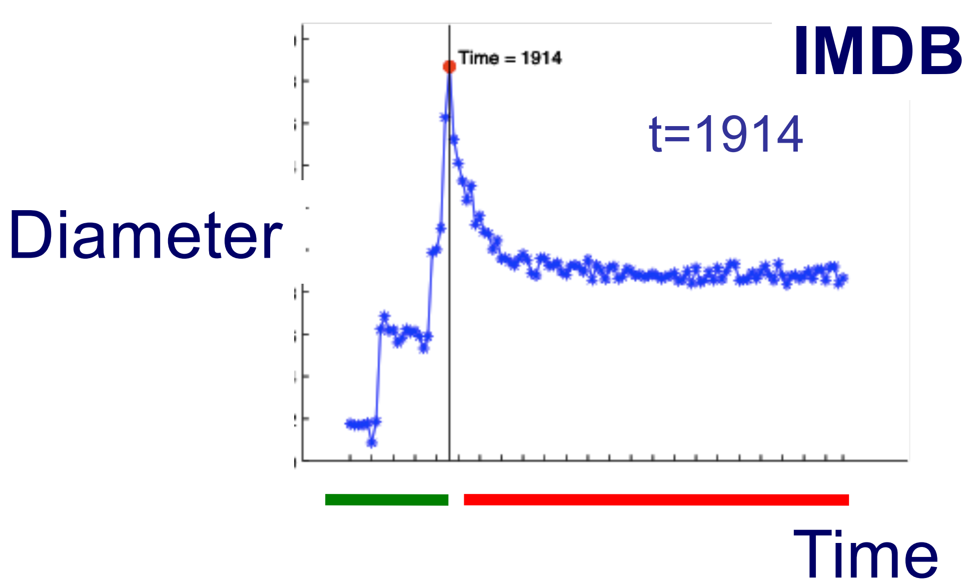
D.3 Gelling Point

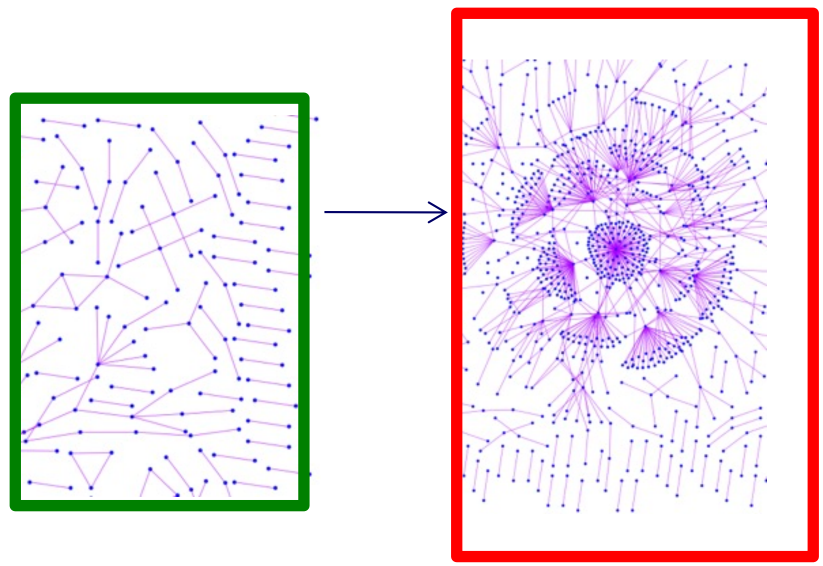
- Most real graphs display a gelling point
- After gelling point, they exhibit typical behavior. This is marked by a spike in diameter.



D.3 Gelling Point

- Most real graphs display a gelling point
- After gelling point, they exhibit typical behavior. This is marked by a spike in diameter.





List of Dynamic Patterns

- ✓ D.1 diameter
- ✓ D.2 densification
- ✓ D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

Observation D.4: NLCC behavior

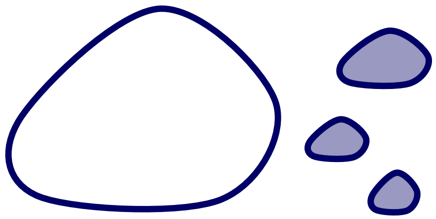
Q: How do NLCC's emerge and join with the GCC?

(``NLCC'' = non-largest conn. components)

– Do they continue to grow in size?

– or do they shrink?

– or stabilize?



Observation D.4: NLCC behavior

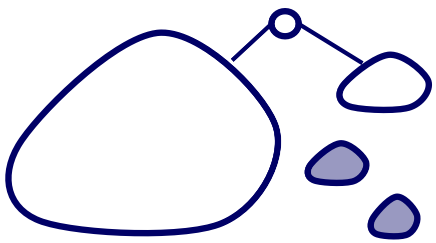
Q: How do NLCC's emerge and join with the GCC?

(``NLCC'' = non-largest conn. components)

– Do they continue to grow in size?

– or do they <u>shrink</u>?

– or stabilize?



Observation D.4: NLCC behavior

Q: How do NLCC's emerge and join with the GCC?

(``NLCC'' = non-largest conn. components)

YES – Do they continue to grow in size?

YES – or do they shrink?

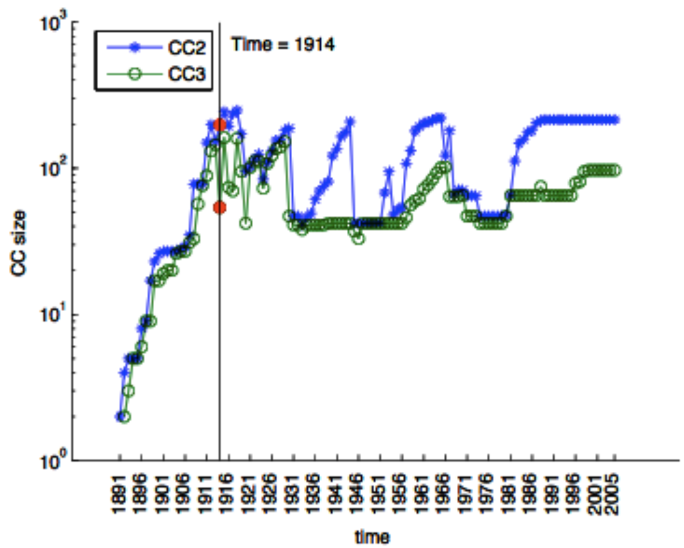
YES – or stabilize?

Observation D.4: NLCC behavior

- After the gelling point, the GCC takes off, but NLCC' s remain ~constant (actually, **oscillate**).

IMDB

CC size



Time-stamp

List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- ~~D.5 Eigenvalue over time~~
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

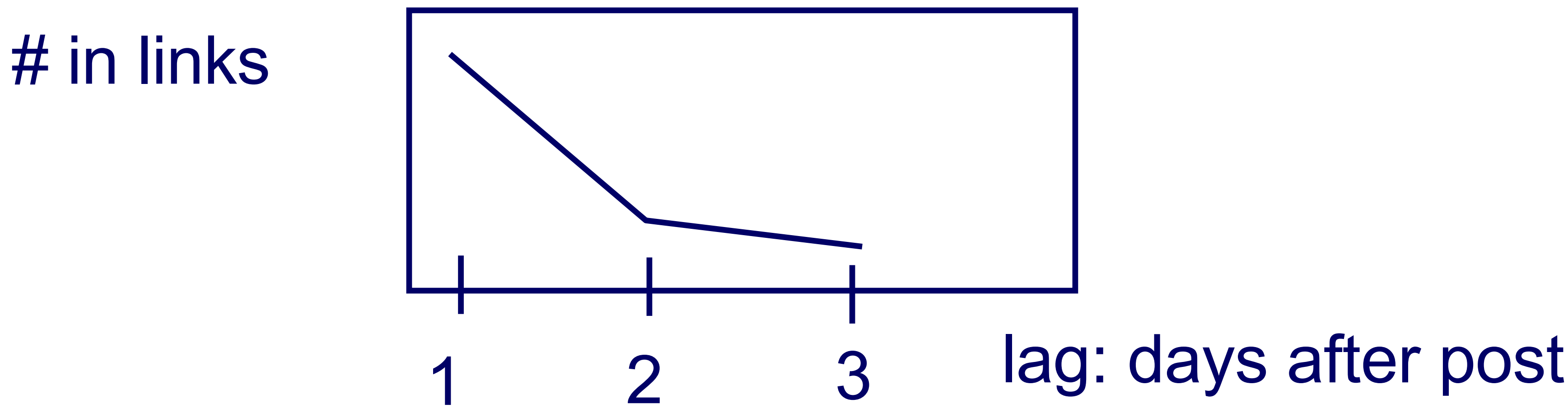
Timing for Blogs

Cascading Behavior in Large Blog Graphs: Patterns and a model

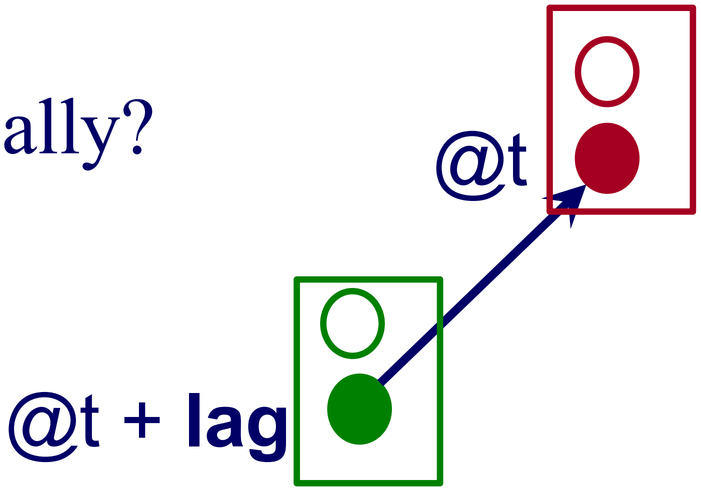
Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst

SDM'07

D.6 : popularity over time

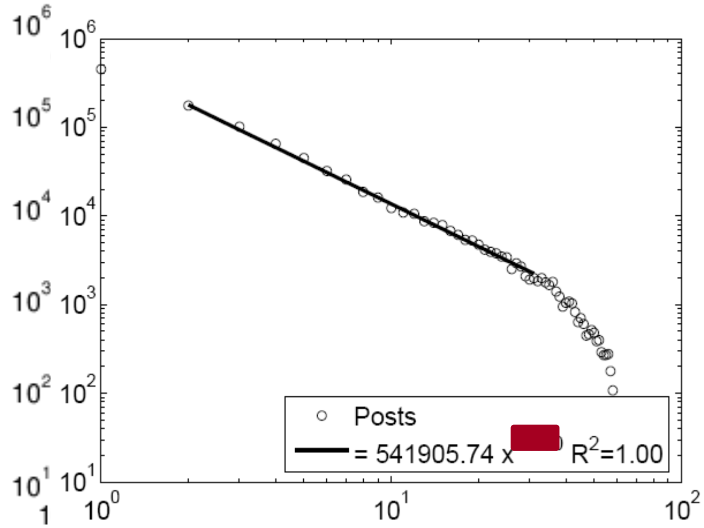


Post popularity drops-off – exponentially?



D.6 : popularity over time

in links (**log**)

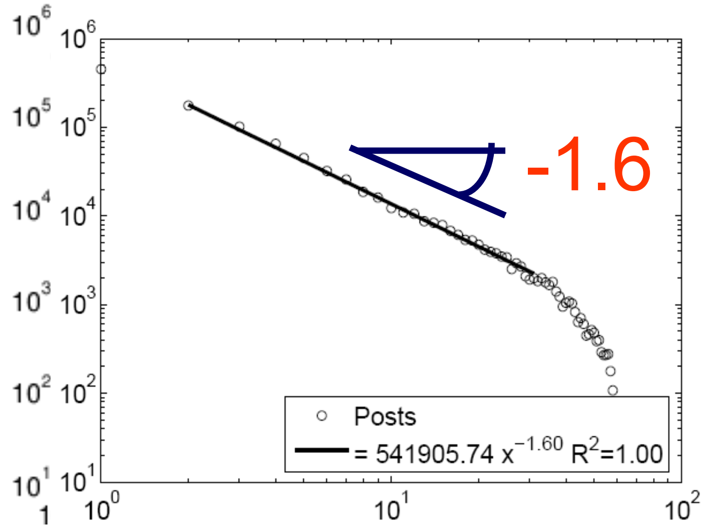


days after post (**log**)

Post popularity drops-off – exponentially?
POWER LAW!
Exponent?

D.6 : popularity over time

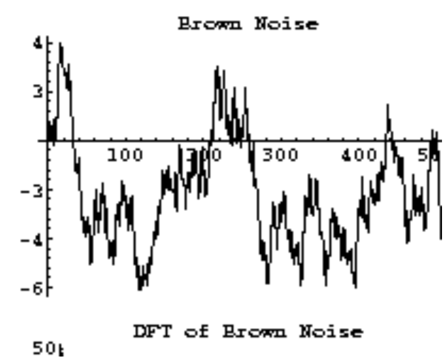
in links
(log)



days after post
(log)

Post popularity drops-off – exponentially?
POWER LAW!
Exponent? -1.6

- close to -1.5: Barabasi's stack model
- and like the zero-crossings of a random walk



Carnegie Mellon

-1.5 slope

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein. *Nature* **437,** 1251 (2005) . [PDF]

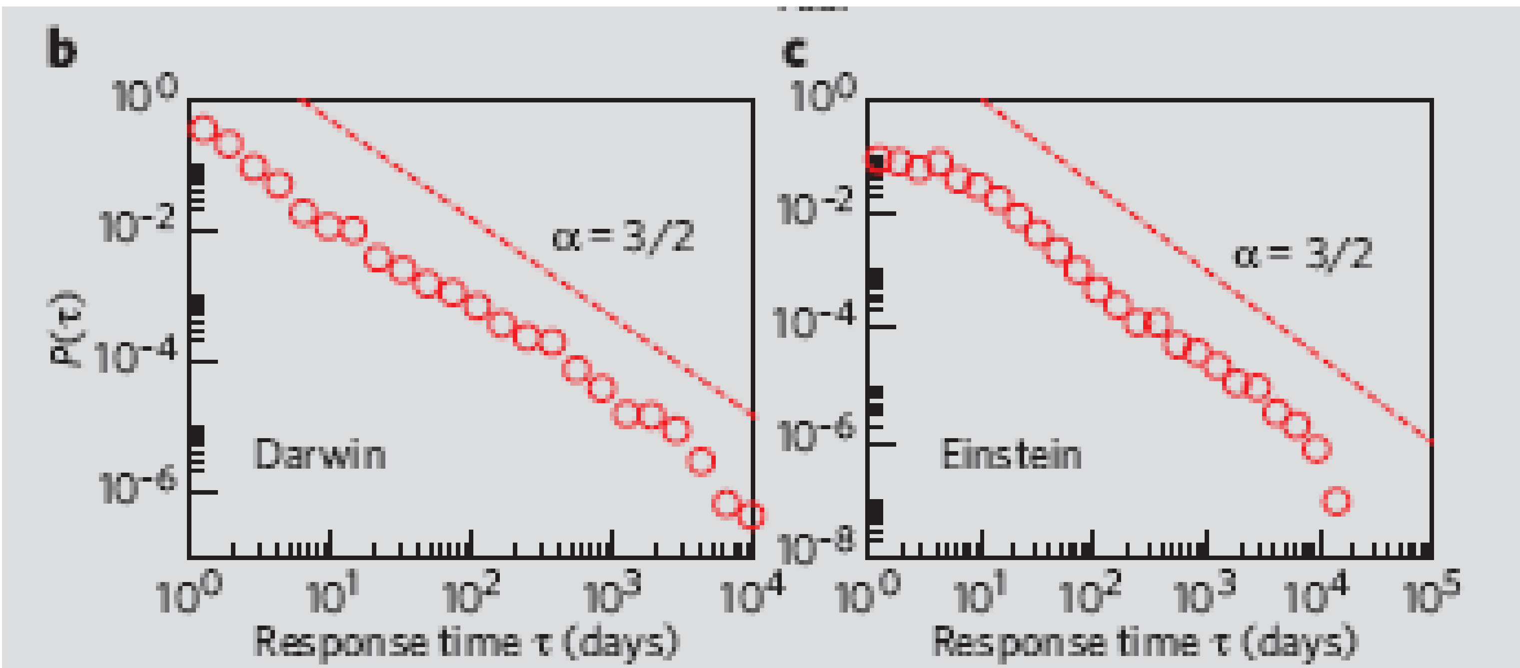


Figure 1 | The correspondence patterns of Darwin and Einstein.

List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- ~~D.5 Eigenvalue over time~~
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

Carnegie Mellon

D.7: duration of phonecalls

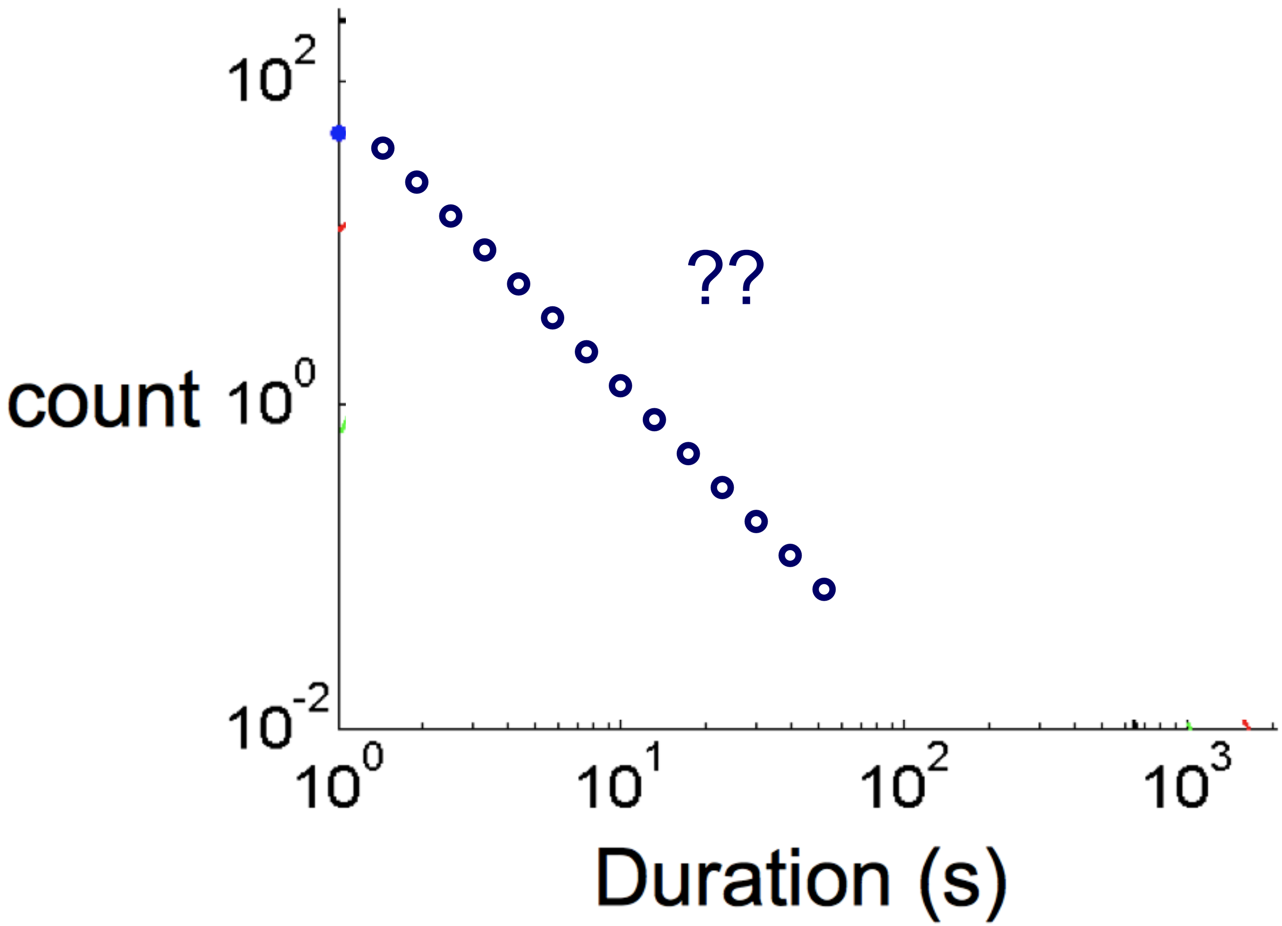
Surprising Patterns for the Call Duration Distribution of Mobile Phone Users

Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, Antonio A. F. Loureiro

PKDD 2010

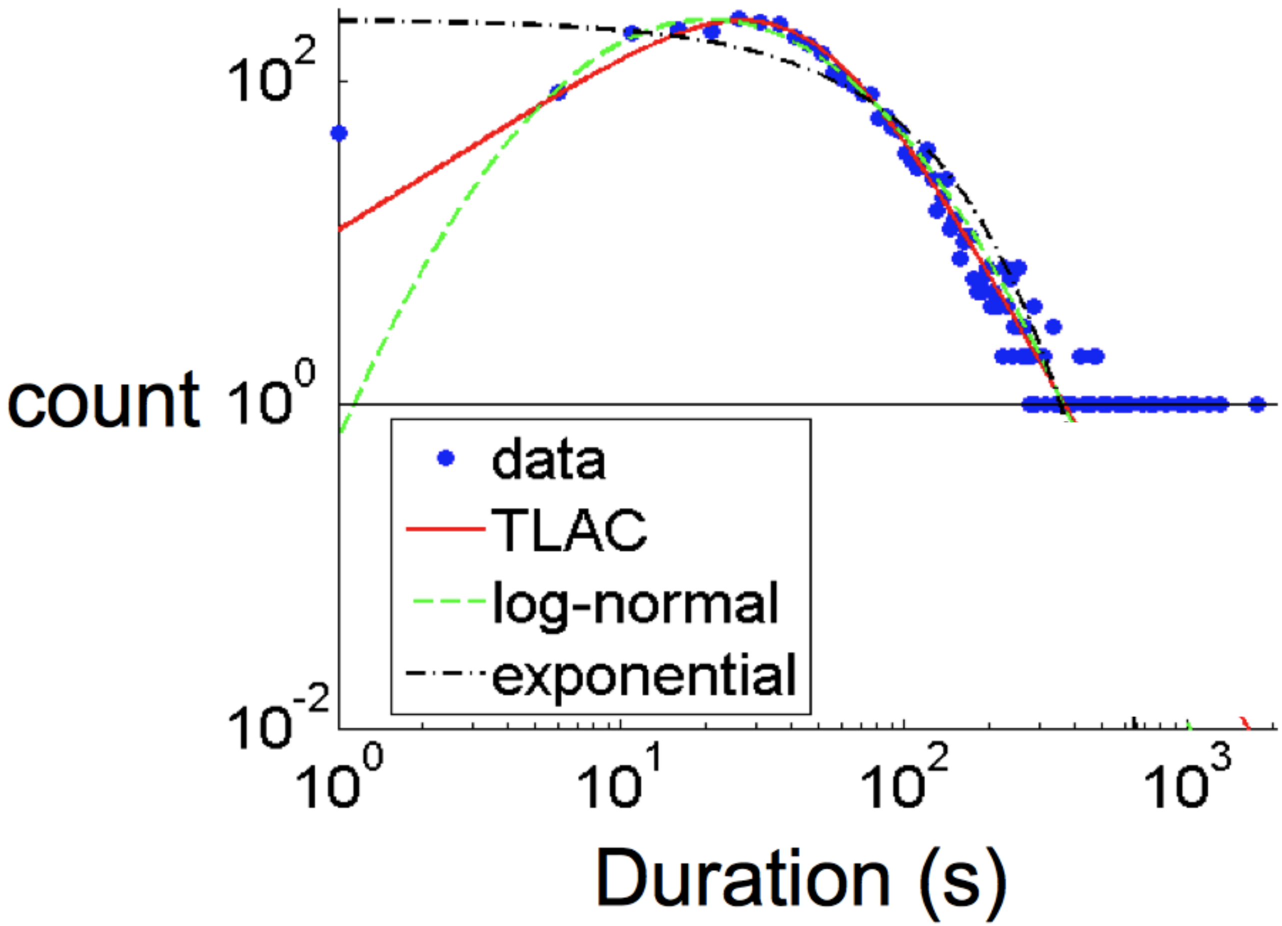
Carnegie Mellon

Probably, power law (?)



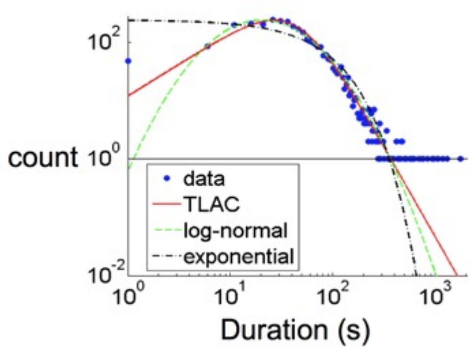
Carnegie Mellon

No Power Law!



'TLaC: Lazy Contractor'

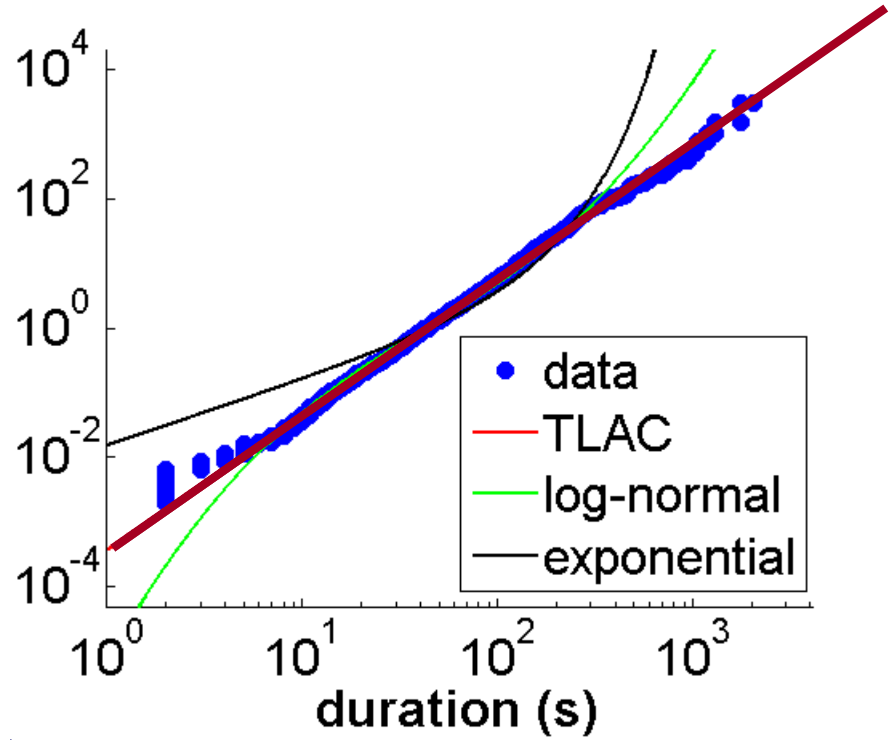
- The longer a task (phonecall) has taken,
- The even longer it will take



Odds ratio=

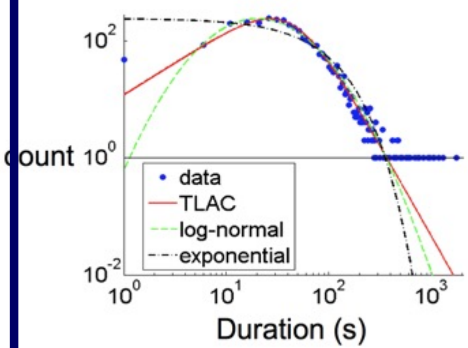
Casualties(<x):
Survivors(>=x)

== power law



Log-logistic distribution

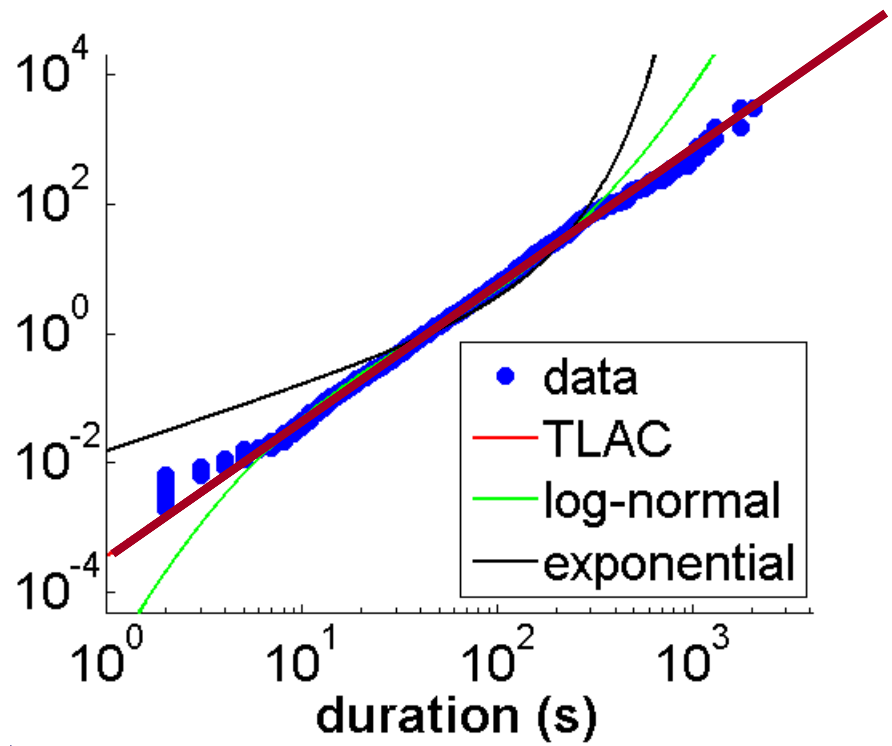
- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)



Odds ratio=

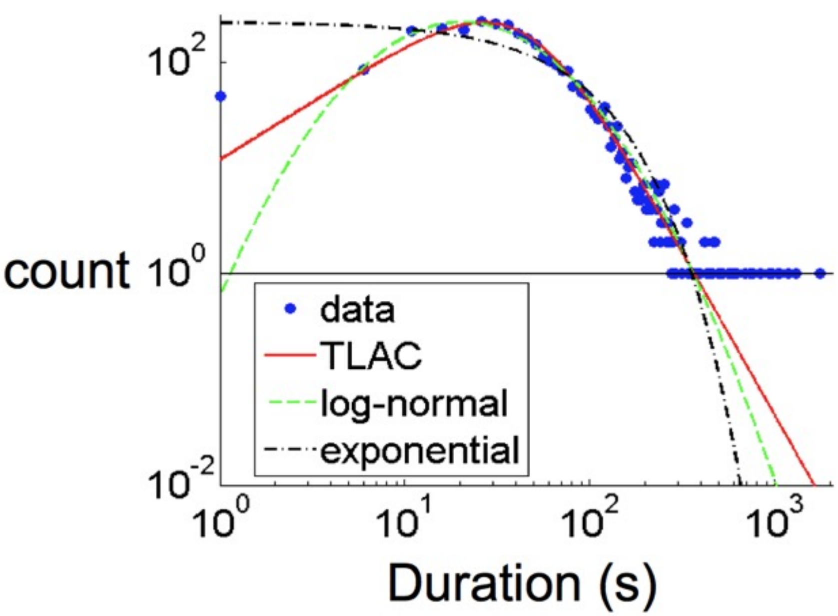
Casualties(<x):
Survivors(>=x)

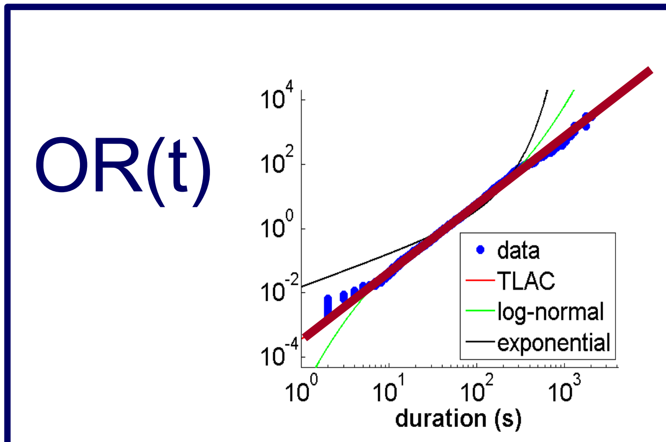
== power law



Log-logistic distribution

- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)



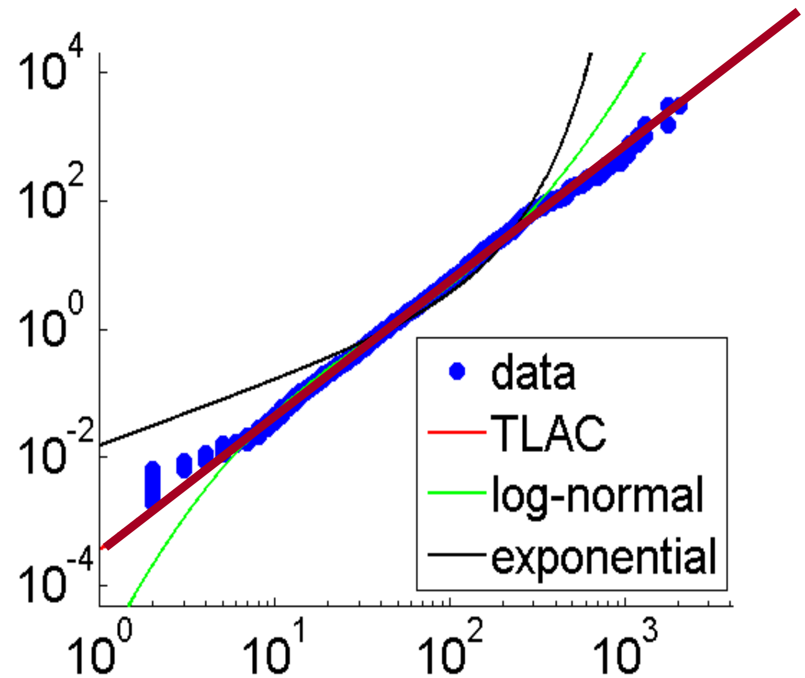


- PDF looks like hyperbola;
- and, if clipped, like power-law

Log-logistic distribution

- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: $\boxed{\log[OR(t)] = \beta + \rho * \log(t)}$

OR(t)

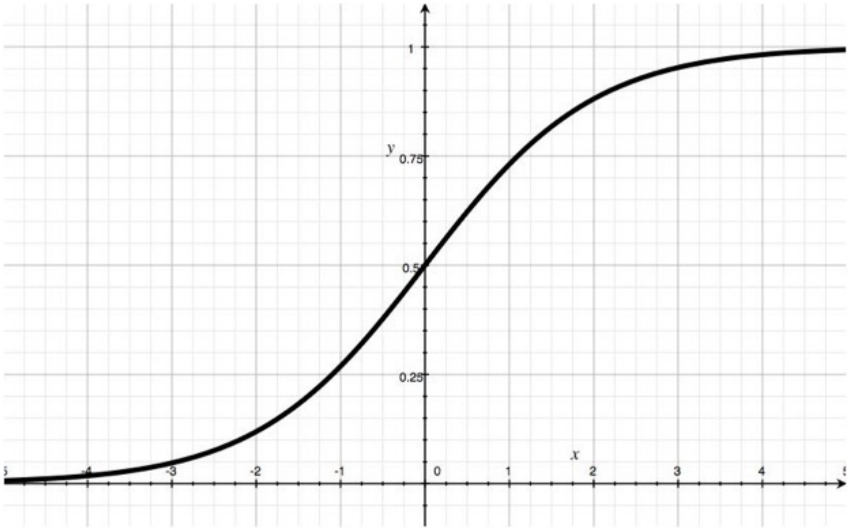


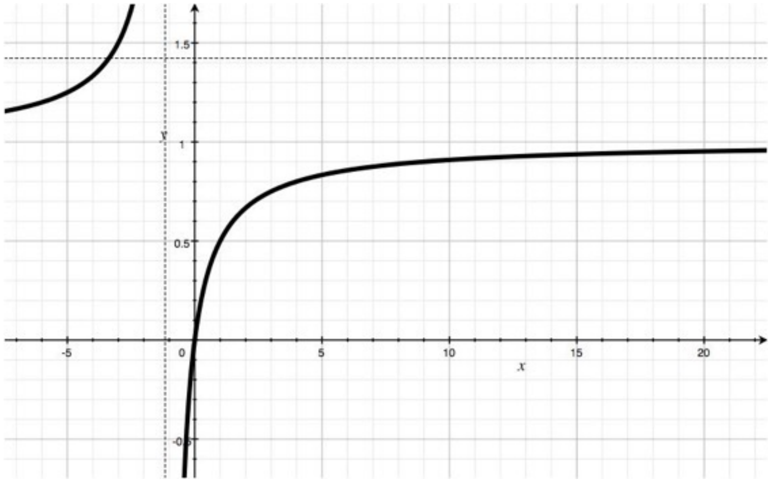
Duration (t)

Log-logistic distribution

- Logistic distribution: CDF -> sigmoid
- LOG-Logistic distribution:

x-> ln(x)



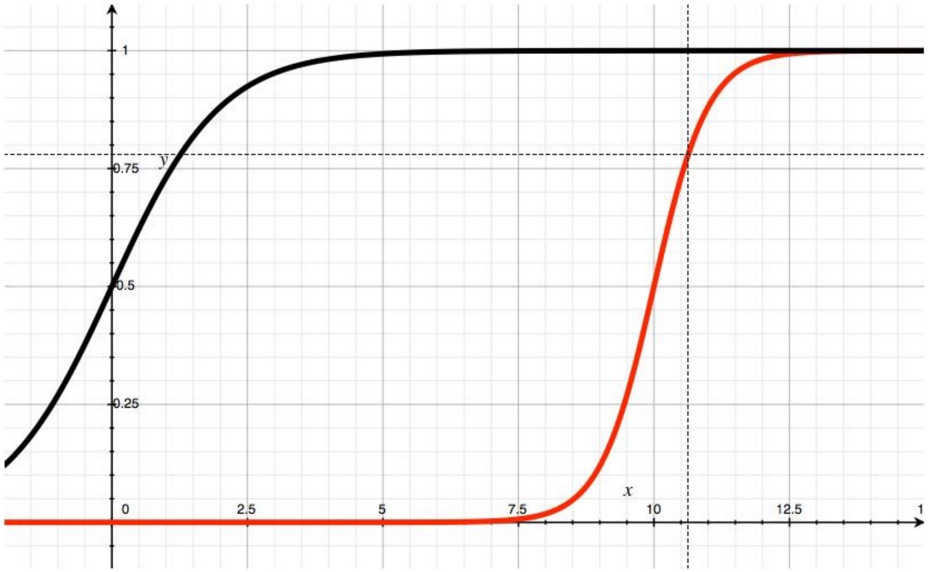


CDF(x) = 1/(1+exp(-x))

CDF(x) = 1/(1+1/x)

Log-logistic distribution

- Logistic distribution: CDF -> sigmoid



- LOG-Logistic distribution:

CDF(x) = 1/(1+exp(-(x-m)/s)) CDF(x) = 1/(1+exp(-(ln(x)-m)/s))

Log-logistic distribution

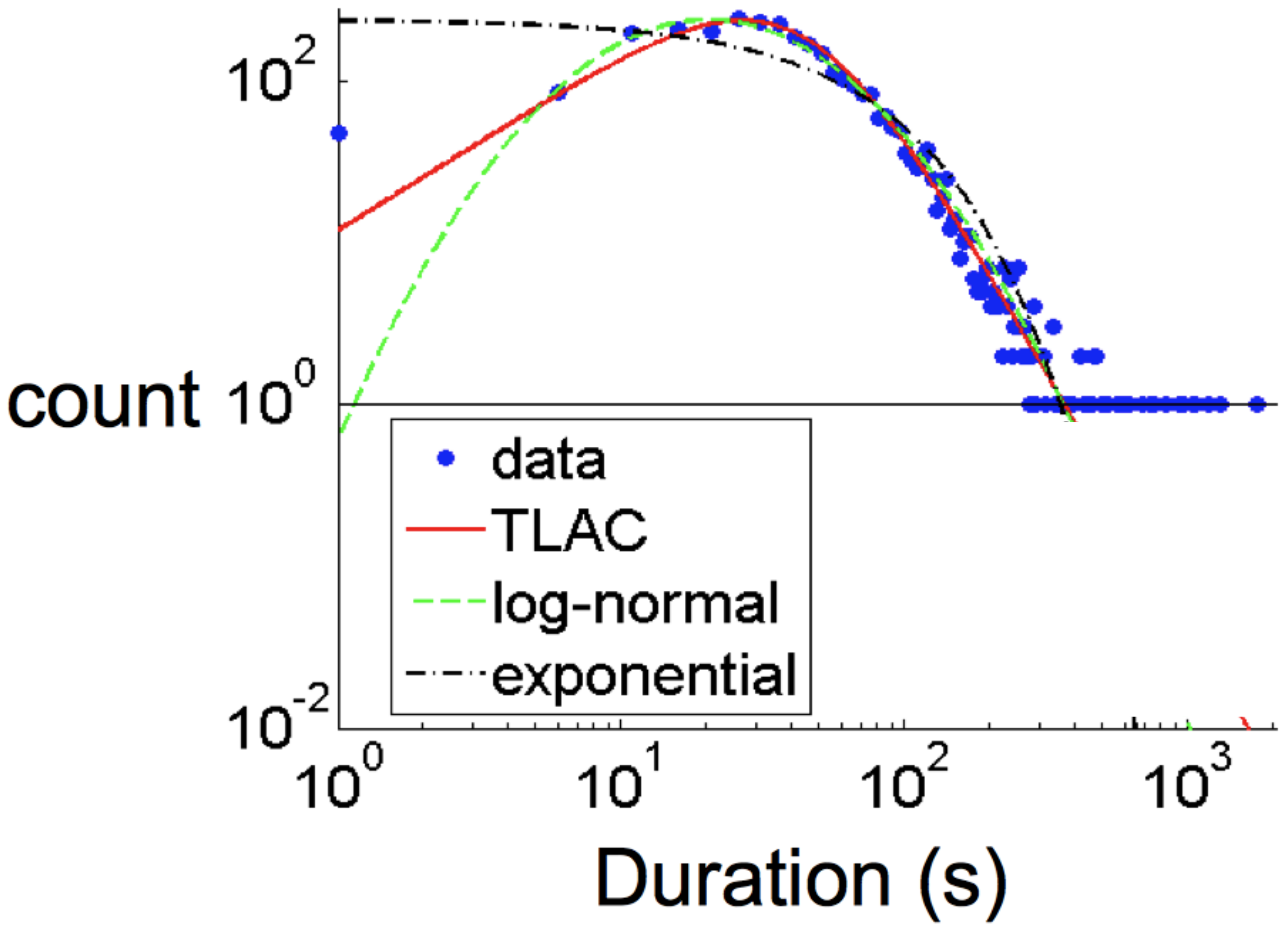
Nice 1 page description: section II of

Pravallika Devineni, Danai Koutra, Michalis Faloutsos, and Christos Faloutsos.

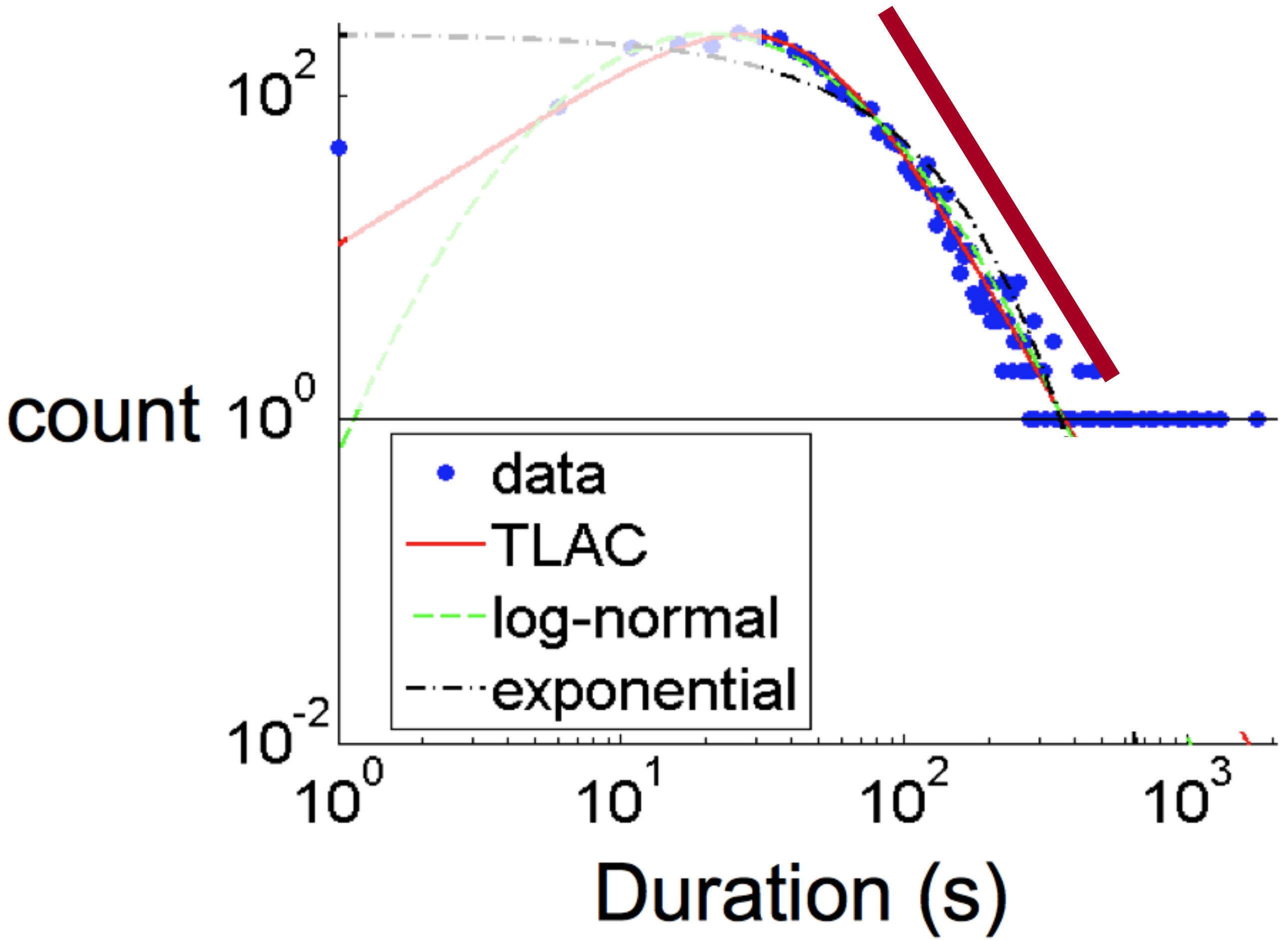
If walls could talk: Patterns and anomalies in Facebook wallposts.

ASONAM 2015, pp 367-374.

Log-logistic: ~ power law



Log-logistic: ~ power law

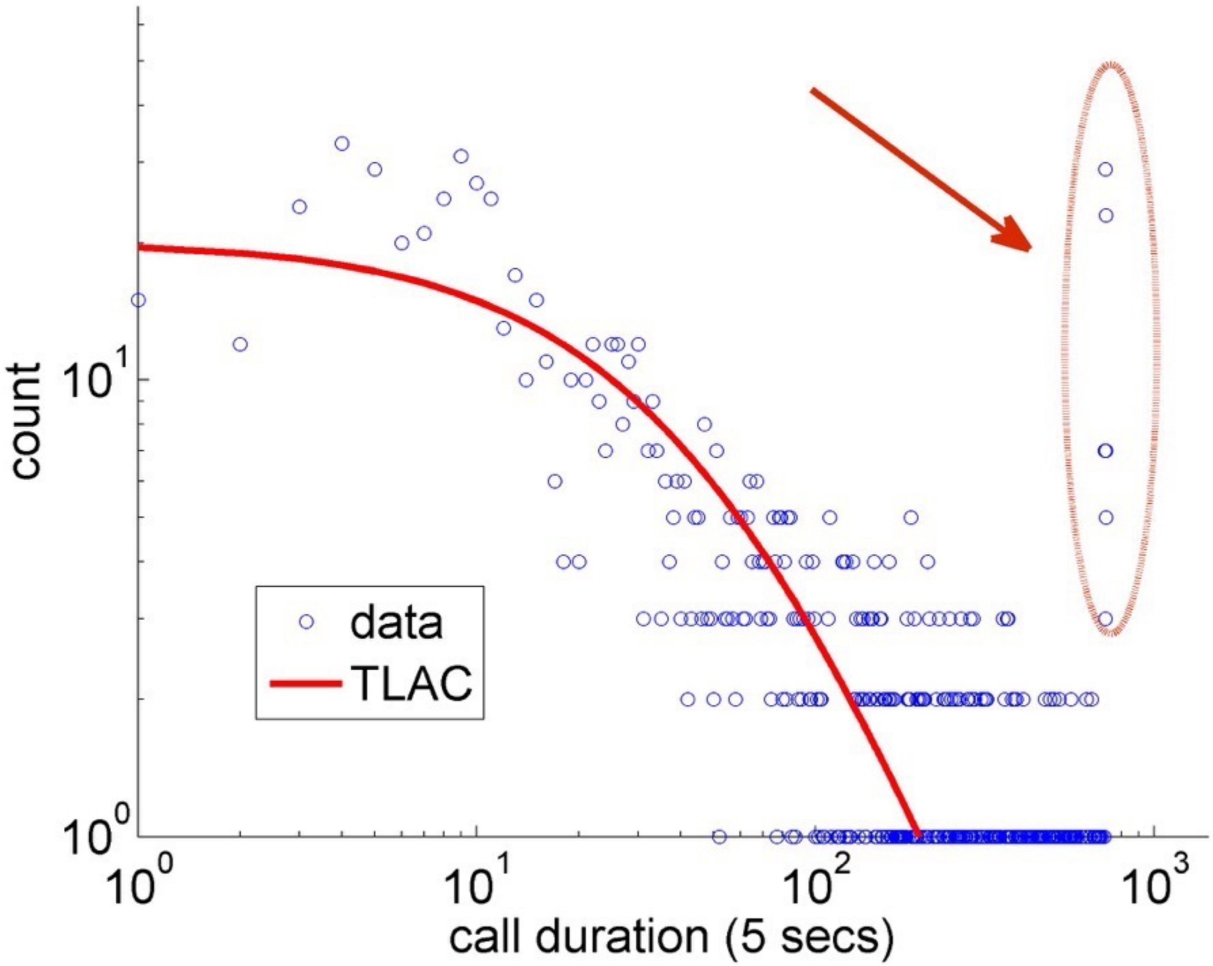


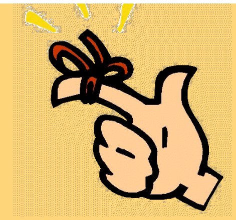
Data Description

- Data from a private mobile operator of a large city
  - 4 months of data
  - 3.1 million users
  - more than 1 billion phone records
- Over 96% of 'talkative' users obeyed a TLAC distribution ('talkative': >30 calls)

Carnegie Mellon

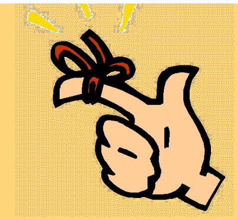
Outliers:





Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diameters
    - Skewed degree distribution
    - Shrinking diameters
  - Weighted
  - Time-evolving



Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diamete
    - Ske
  - e-evolving

- Many power laws – log-logistic
- Take logarithms