Carnegie Mellon

# Mining Large Social Networks: Patterns and Anomalies

*Christos Faloutsos*

CMU

# Thank you

- The Department of Informatics
- Happy 20-th!

- Prof. Yannis Manolopoulos
- Prof. Kostas Tsichlas
- Mrs. Nina Daltsidou

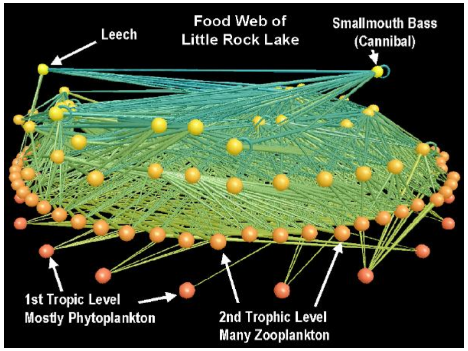
# International-caliber friends among AUTH alumni

- Prof. Evimaria Terzi (U. Boston)
- Prof. Kyriakos Mouratidis (SMU)
- Dr. Michalis Vlachos (IBM)
- …

# Outline

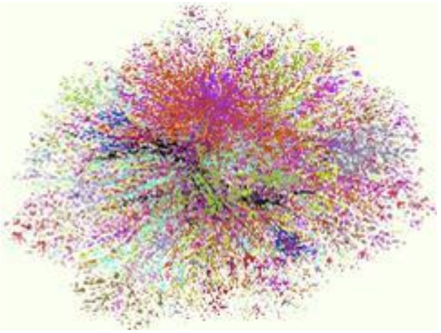
- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability
- Conclusions

# Graphs - why should we care?




Food Web
[Martinez '91]

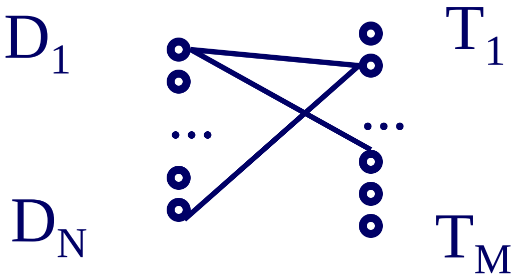
$10s of BILLIONS revenue
>500M users



Internet Map
[lumeta.com]

# Graphs - why should we care?

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




- web: hyper-text graph
- ... and more:

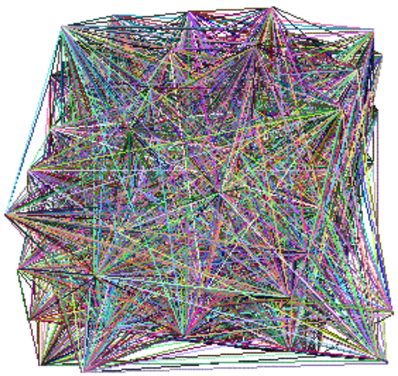
# Graphs - why should we care?

- web-log ('blog') news propagation
- computer network security: email/IP traffic and anomaly detection
- ....
- [subject-verb-object: → graph]
- Graph == relational table with 2 columns (src, dst)
- BIG DATA – big graphs

# Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
  - Static graphs
  - Weighted graphs
  - Time evolving graphs
- Problem#2: Tools
- Problem#3: 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?

# Graph mining

- Are real graphs random?

# Laws and patterns

- Are real graphs random?
- A: NO!!
  - Diameter
  - in- and out- degree distributions
  - other (surprising) patterns

- So, let's look at the data

# Solution# S.1

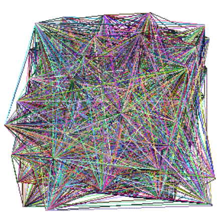
- Power law in the degree distribution [SIGCOMM99]

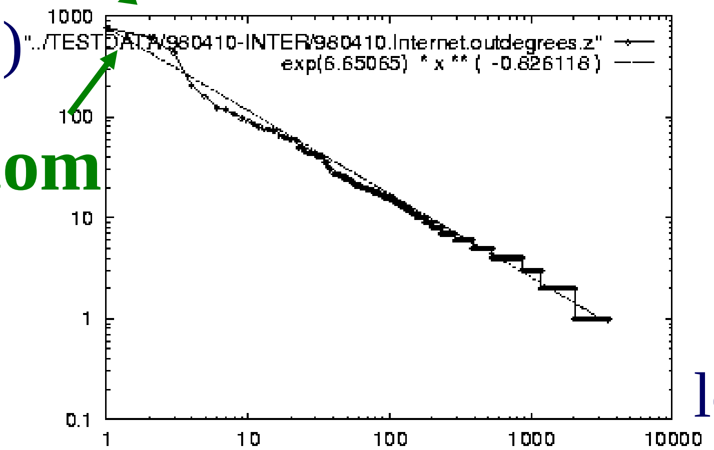
**internet domains**

**att.com**

log(degree)

**ibm.com**



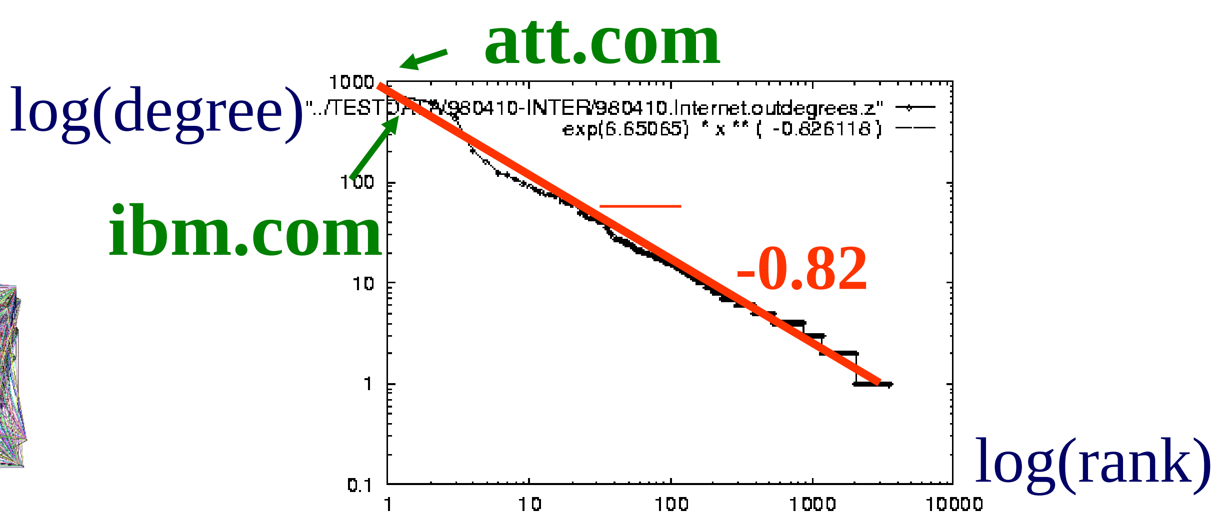


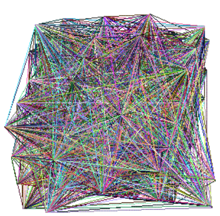

log(rank)

# Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

**internet domains**



# But:

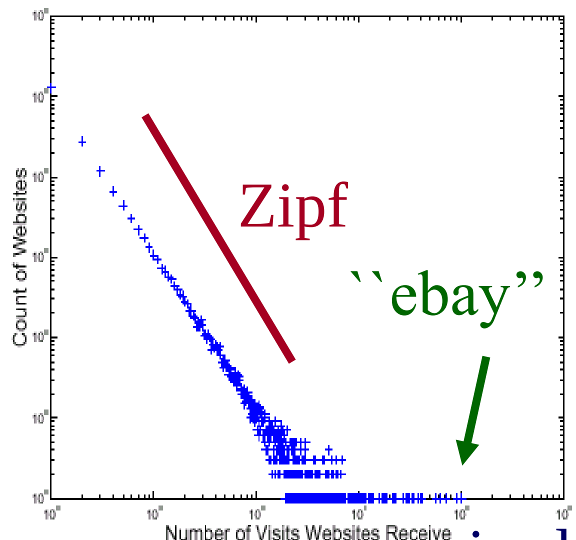
How about graphs from other domains?

# More power laws:

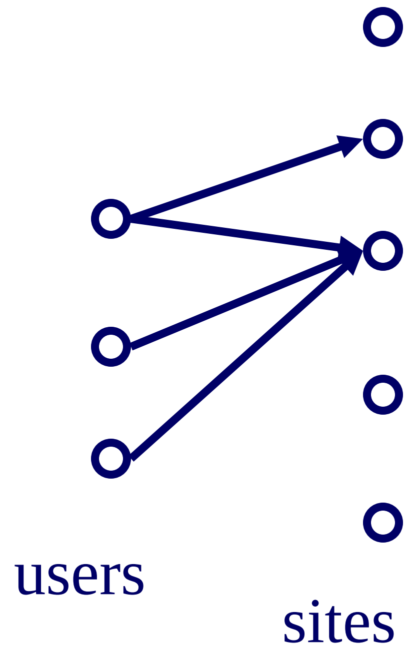
- web hit counts [w/ A. Montgomery]

Web Site Traffic

Count
(log scale)




in-degree (log scale)



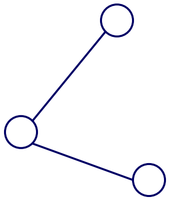
# And numerous more

- Who-trusts-whom (epinions.com)
- Income [Pareto] –'80-20 distribution'
- Duration of downloads [Bestavros+]
- Duration of UNIX jobs ('mice and elephants')
- Size of files of a user
- …
- 'Black swans'

# Outline

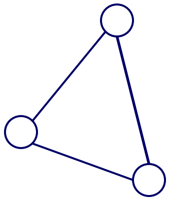
- Introduction – Motivation
- Problem#1: Patterns in graphs
  - Static graphs
    - degree, diameter, eigen,
    - Triangles
  - Time evolving graphs
- Problem#2: Tools

# Solution# S.3: Triangle 'Laws'



- Real social networks have a lot of triangles

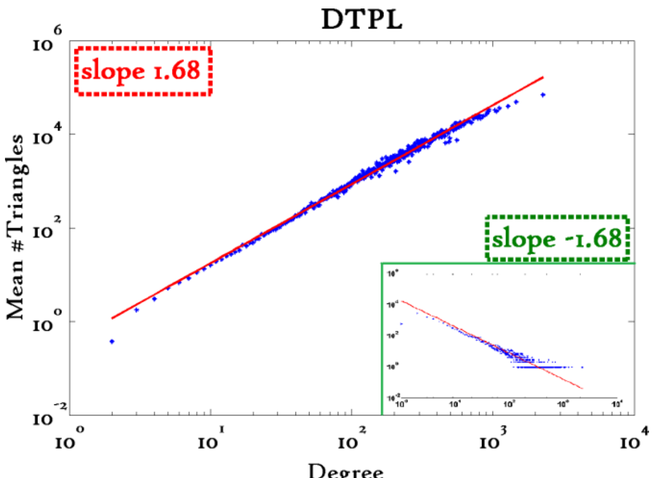
# Solution# S.3: Triangle 'Laws'



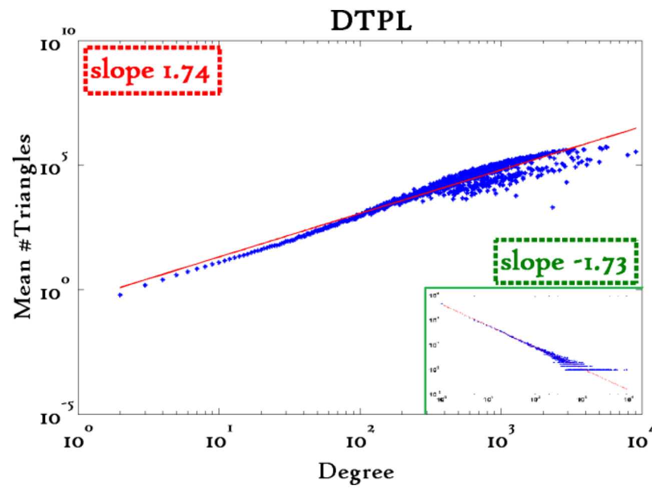
- Real social networks have a lot of triangles
  - Friends of friends are friends
- Any patterns?

# Triangle Law: #S.3
## [Tsourakakis ICDM 2008]

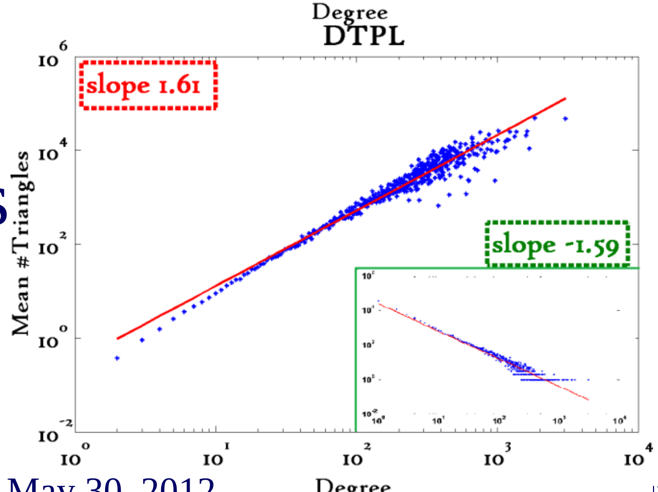
Reuters

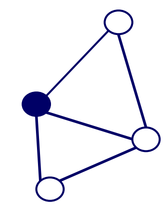



SN




Epinions



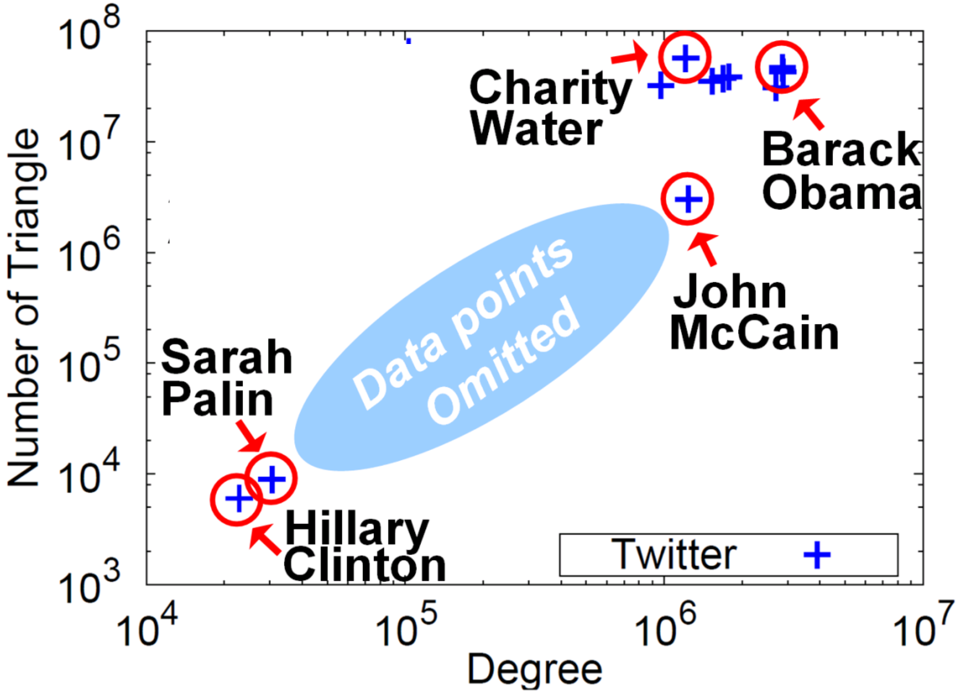



X-axis: degree

Y-axis: mean # triangles

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

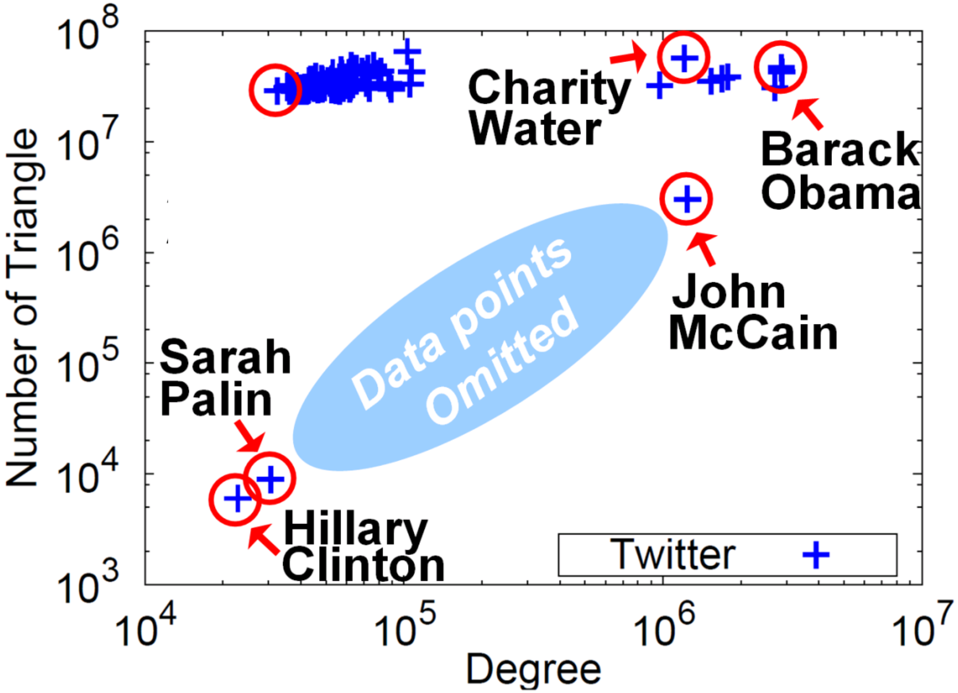
# Triangle counting for large graphs?




Anomalous nodes in Twitter(~ 3 billion edges)

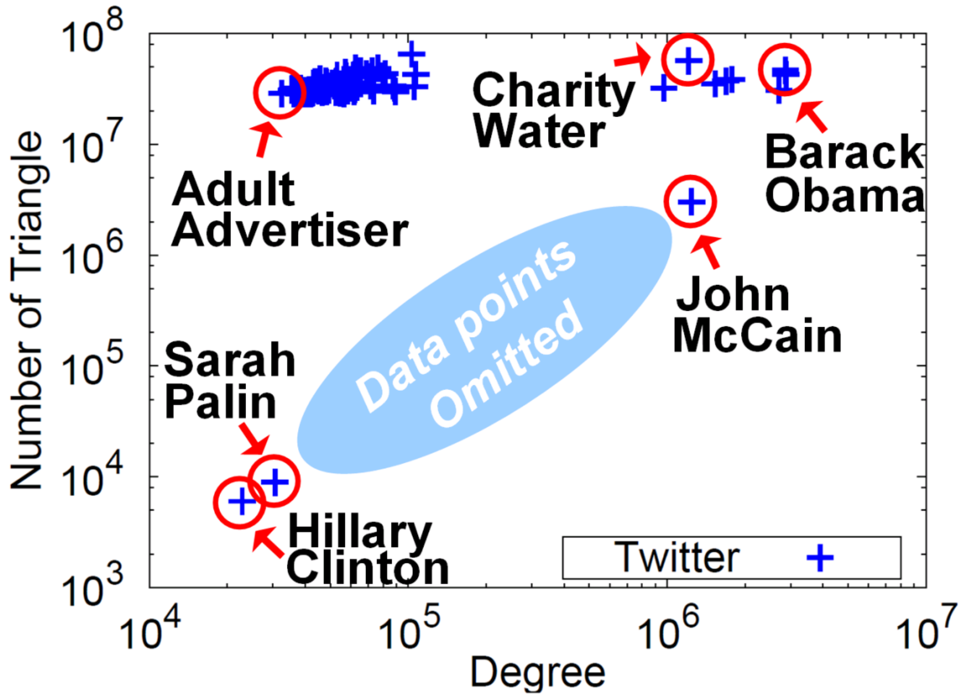
[U Kang, Brendan Meeder, +, PAKDD'11]

# 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]

# Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
  - Static graphs
  - Time evolving graphs
- Problem#2: Tools
- …

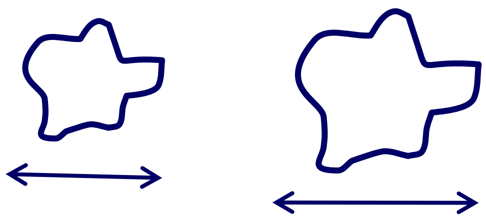
# Problem: Time evolution

- with Jure Leskovec (CMU -> Stanford)

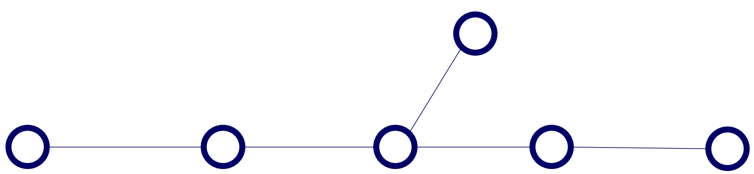
- and Jon Kleinberg (Cornell – sabb. @ CMU)

# T.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)

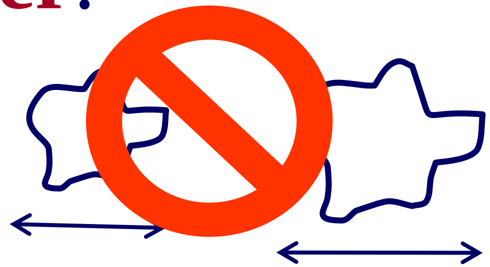


- What is happening in real data?

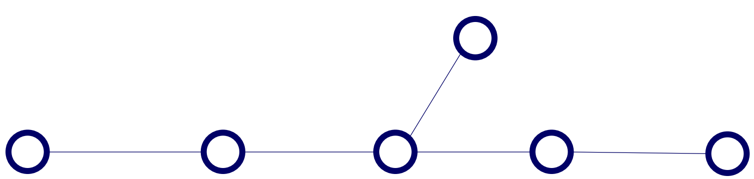


# T.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)

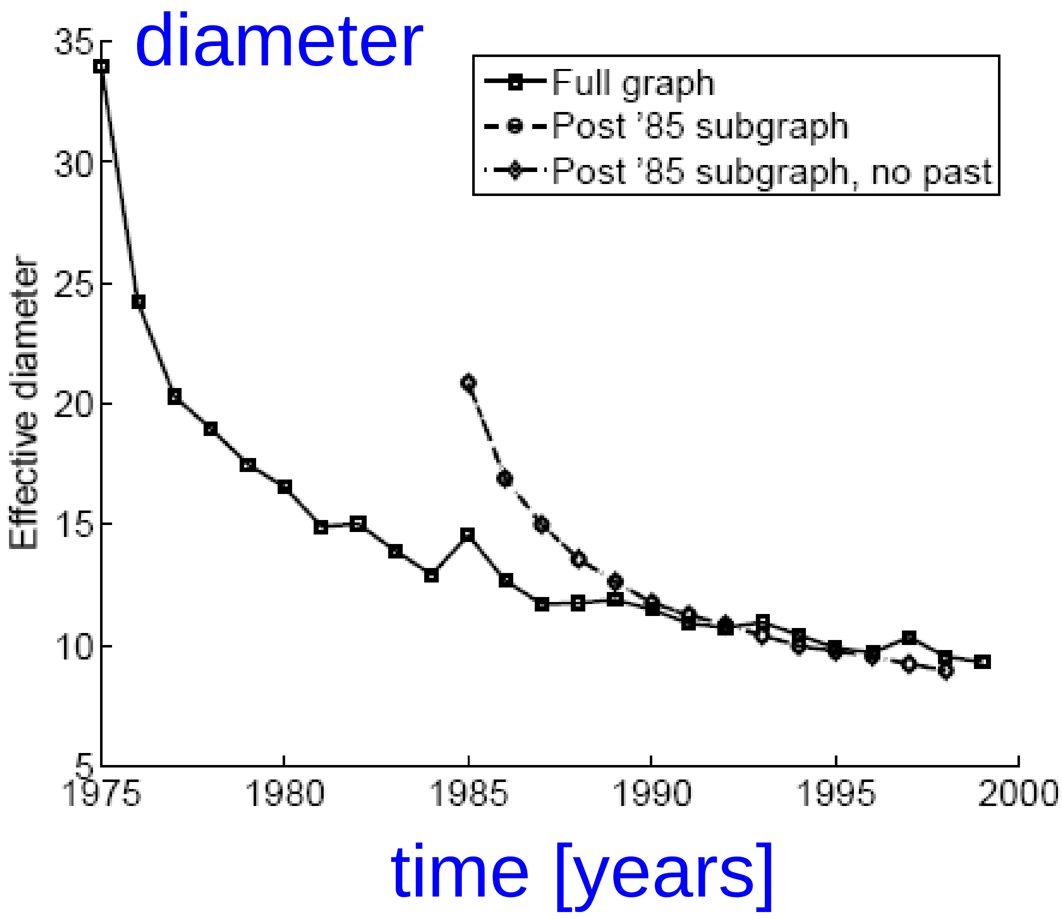


- What is happening in real data?
- Diameter **shrinks** over time



# T.1 Diameter – “Patents”

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

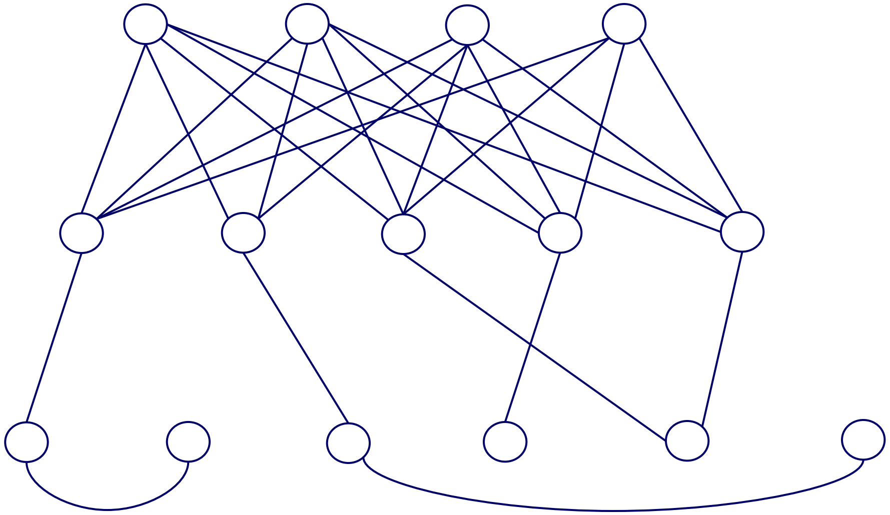


# Outline

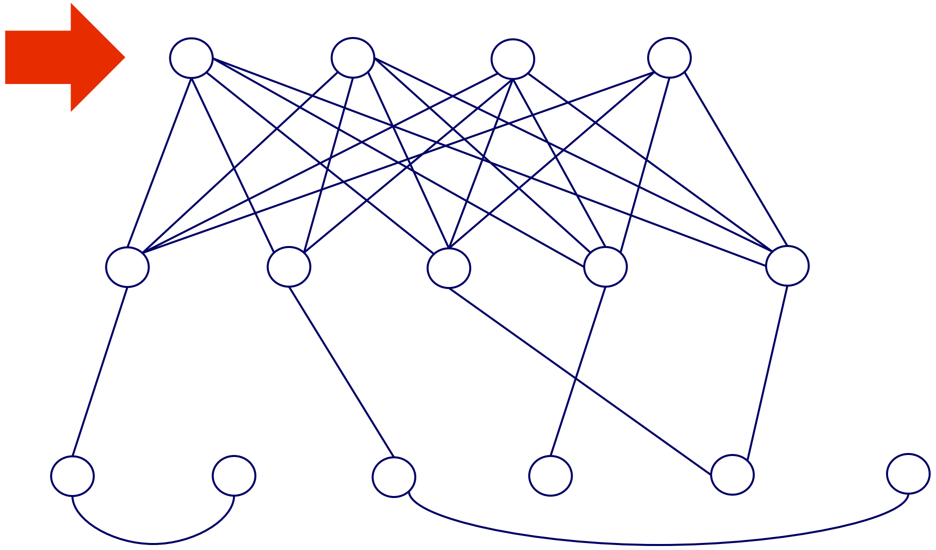
- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
  - Belief Propagation
- Problem#3: Scalability
- Conclusions

# E-bay Fraud detection

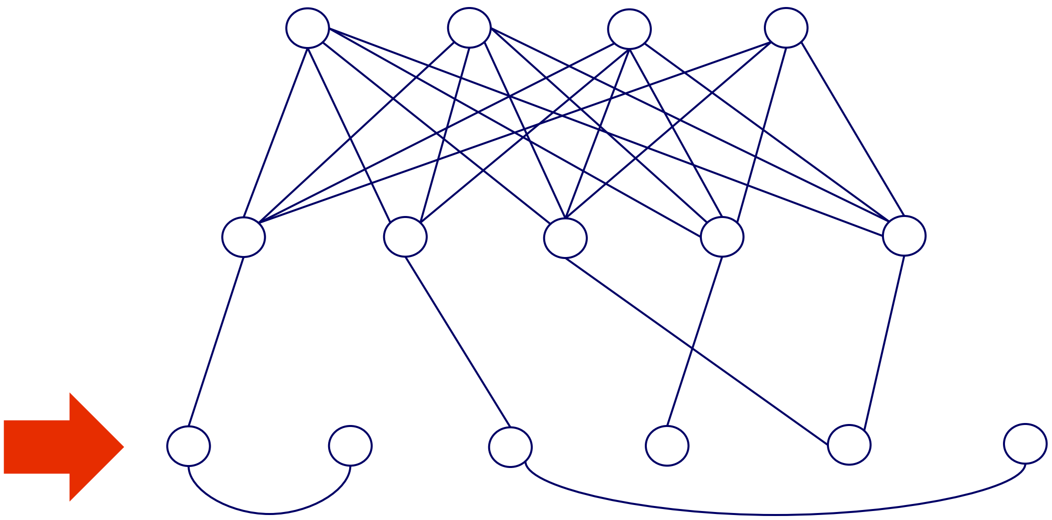
w/ Polo Chau &
Shashank Pandit, CMU
[www'07]



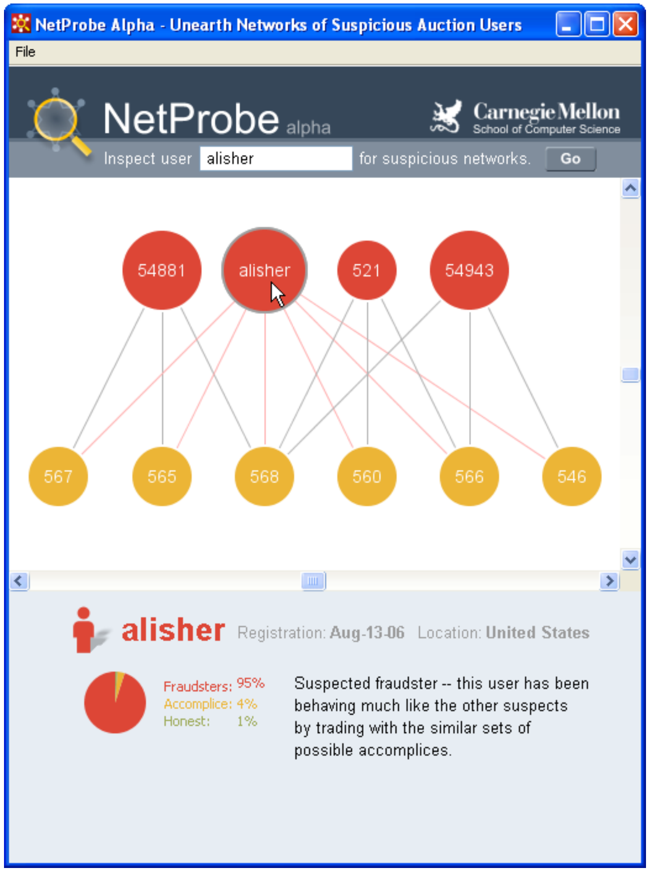
# E-bay Fraud detection

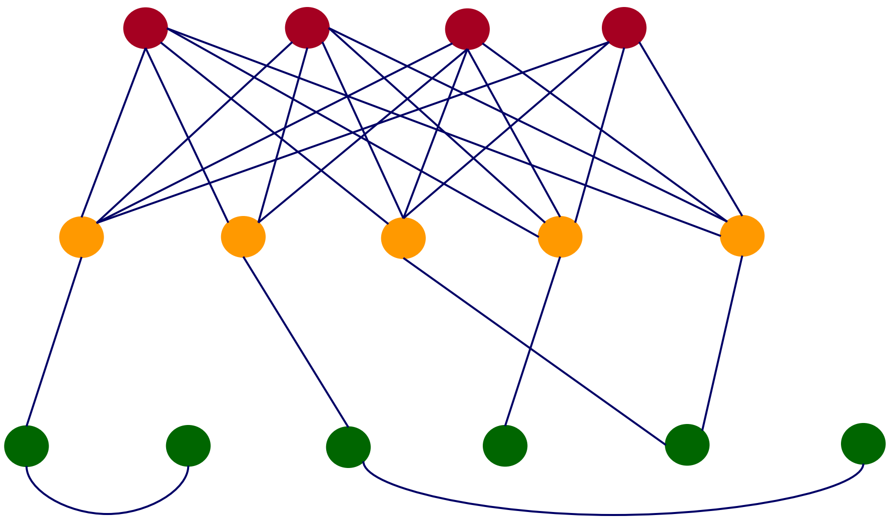


# E-bay Fraud detection



# E-bay Fraud detection - NetProbe



# Popular press

USA TODAY

The Washington Post

Los Angeles Times

And less desirable attention:

- E-mail from ‘Belgium police’ (‘copy of your code?’)

# Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability -PEGASUS
- Conclusions

# Scalability

- Google: > 450,000 processors in clusters of ~2000 processors each [Barroso, Dean, Hölzle, *“Web Search for a Planet: The Google Cluster Architecture”* IEEE Micro 2003]
- Yahoo: 5Pb of data [Fayyad, KDD’07]
- Problem: machine failures, on a daily basis
- How to parallelize data mining tasks, then?
- A: map/reduce – hadoop (open-source clone) http://hadoop.apache.org/

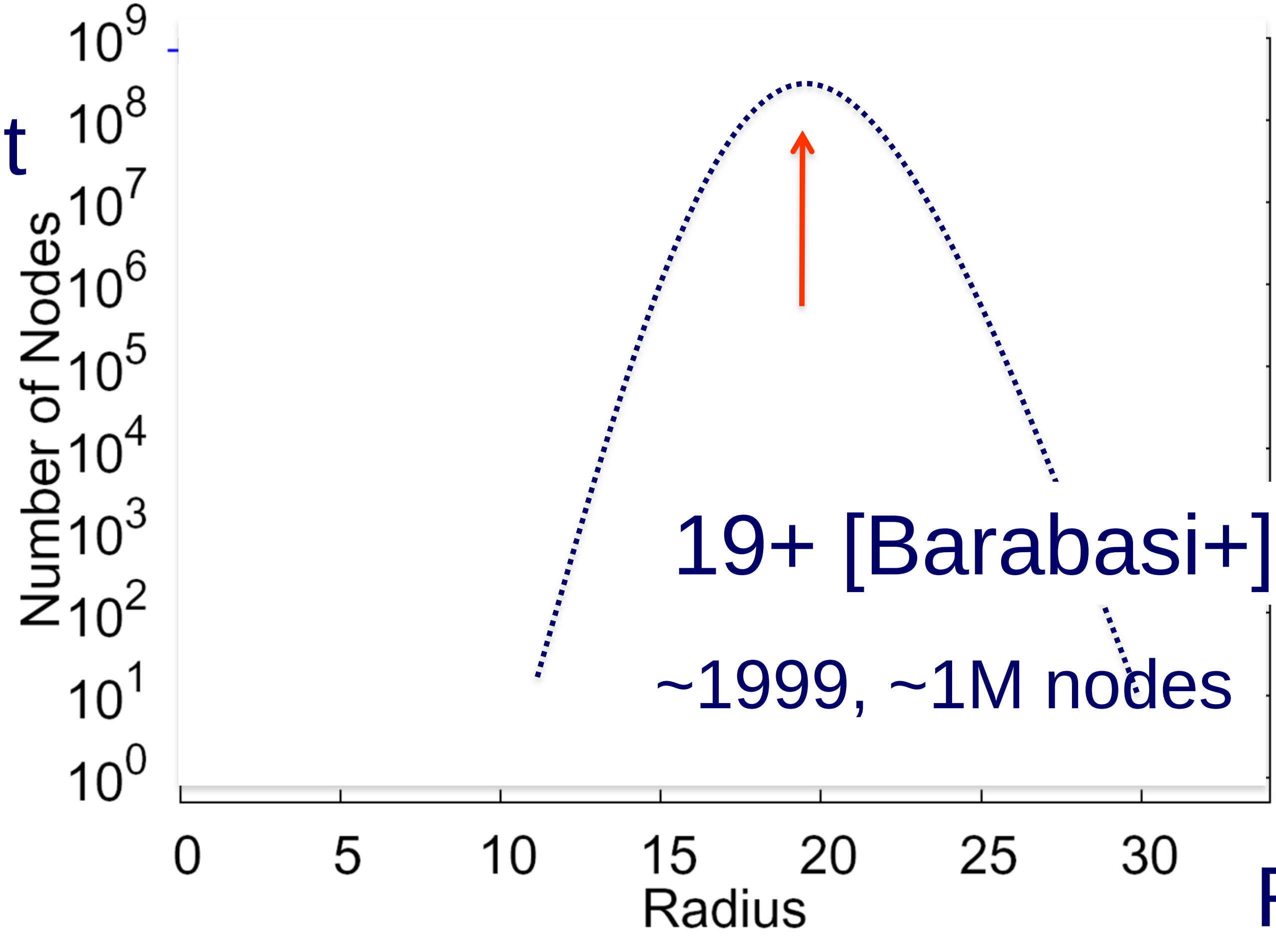
# Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability –PEGASUS
    - Radius plot
- Conclusions

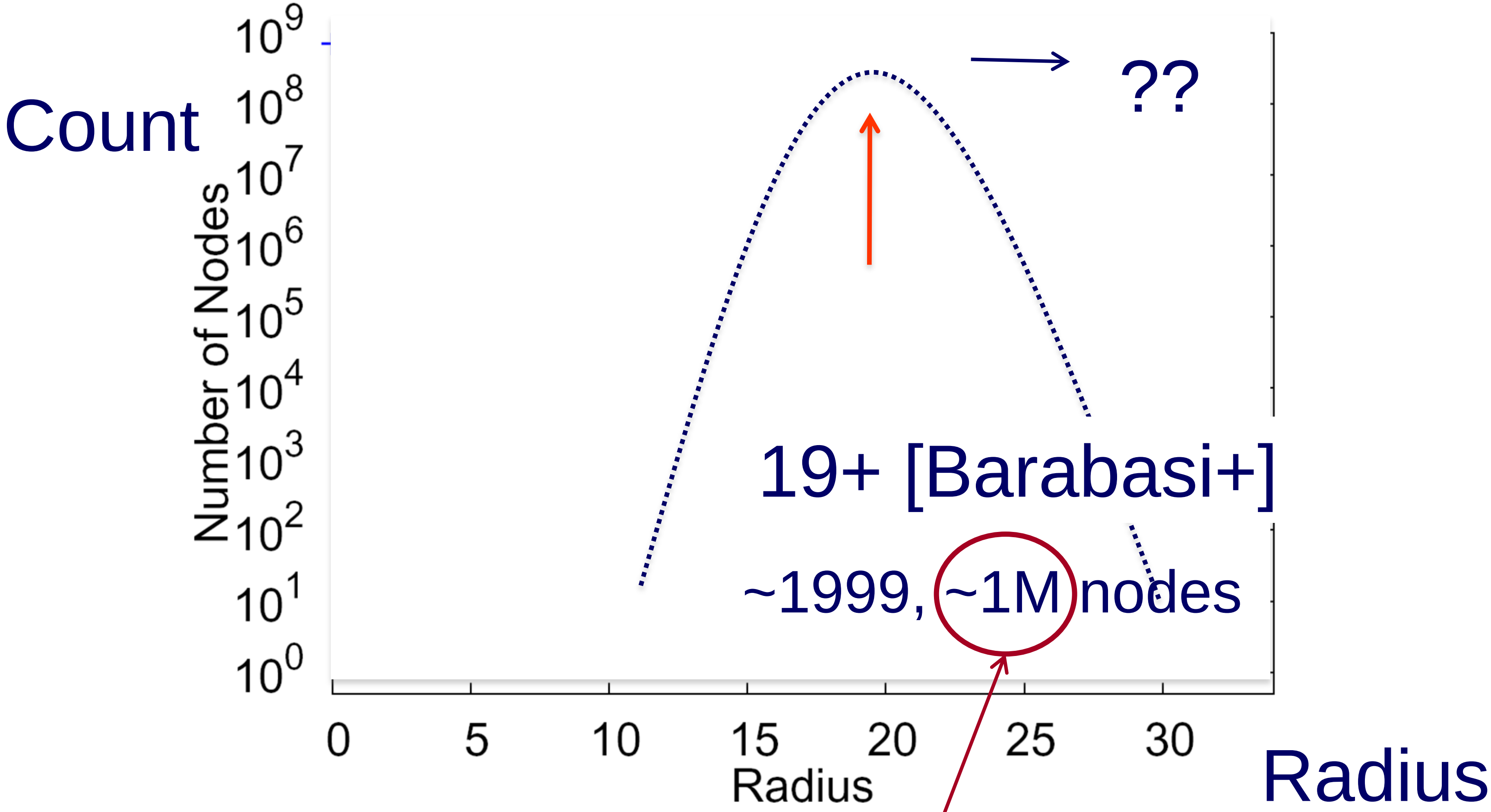
# HADI for diameter estimation

- *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

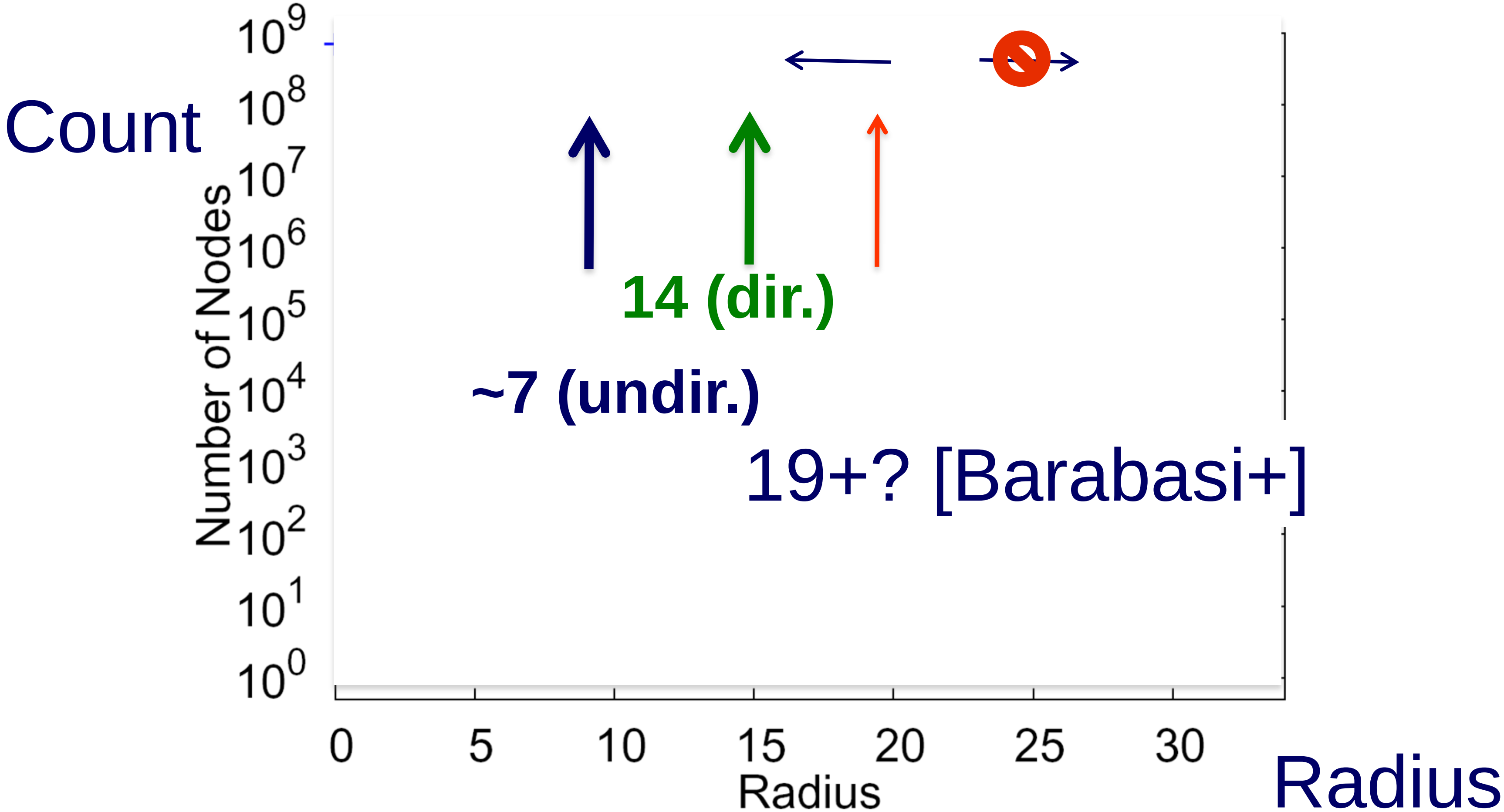



Radius



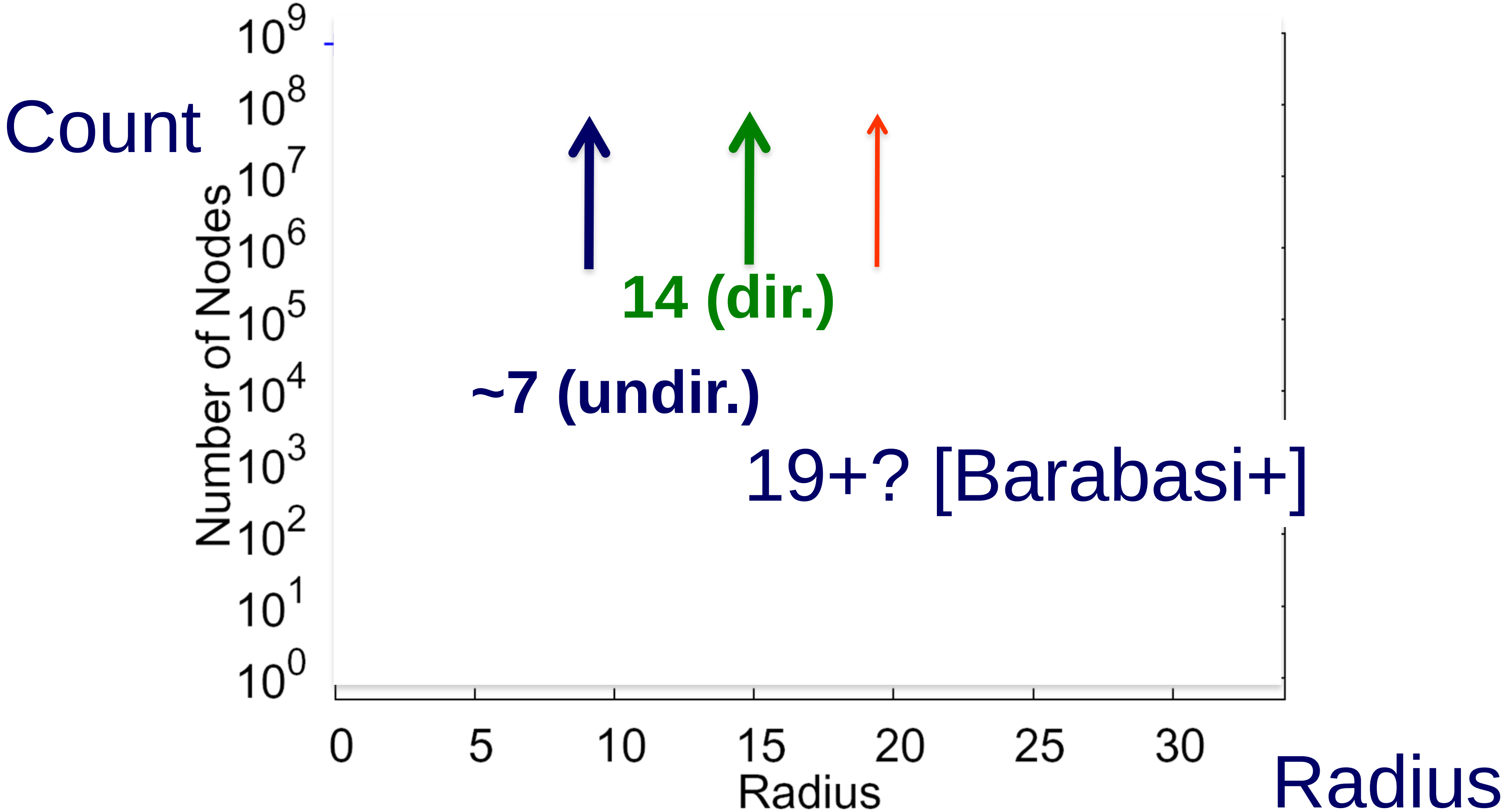

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.



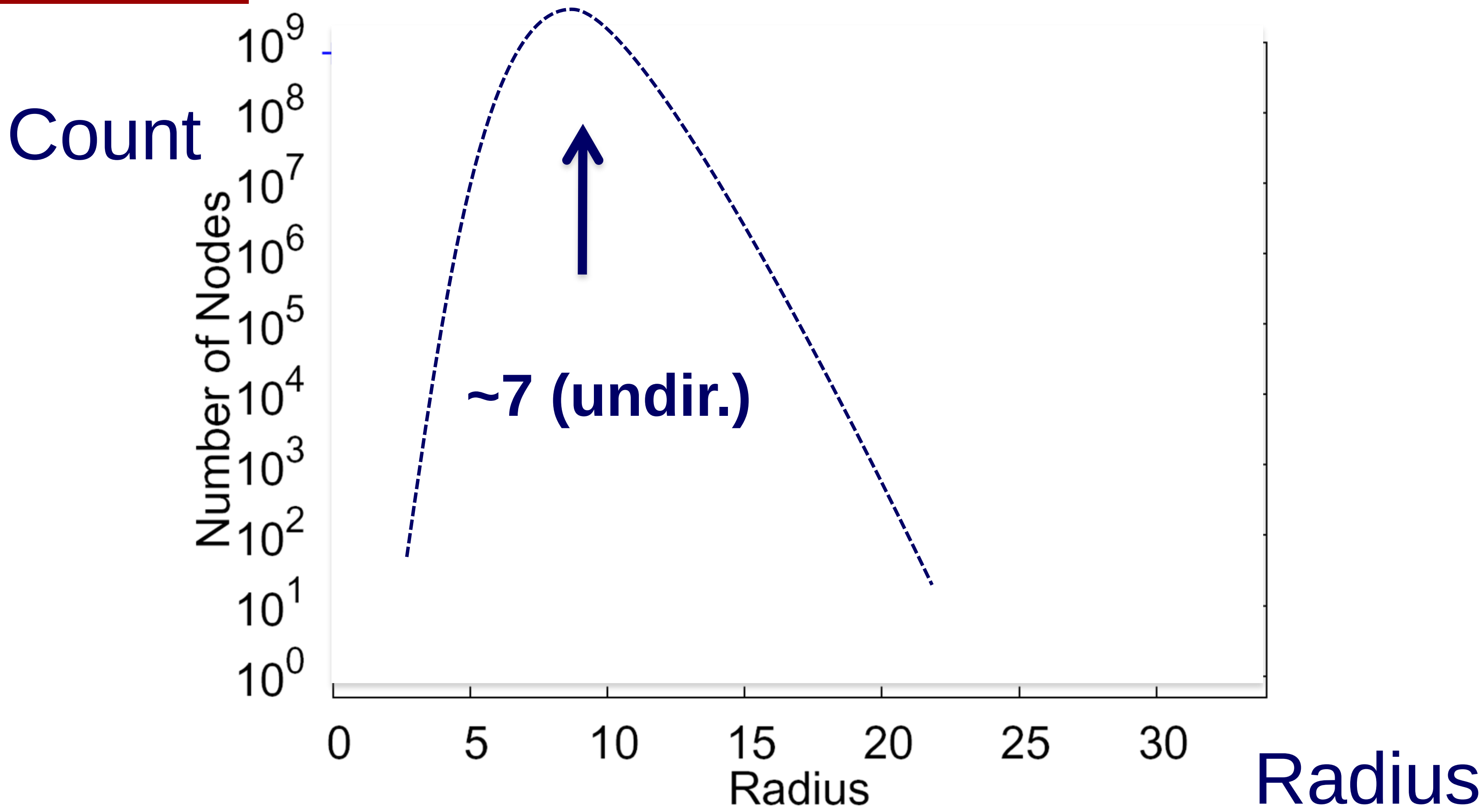

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.

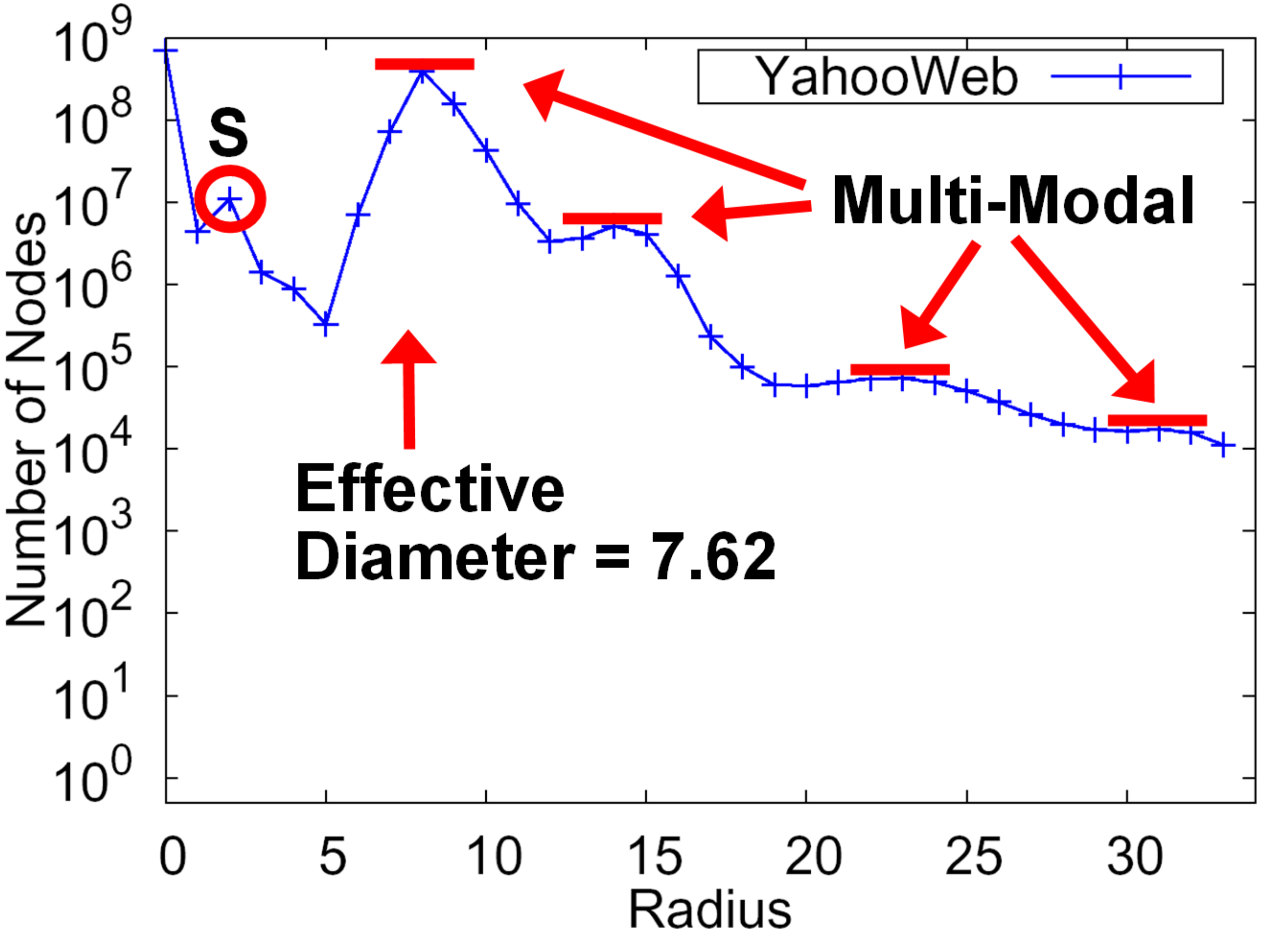



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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

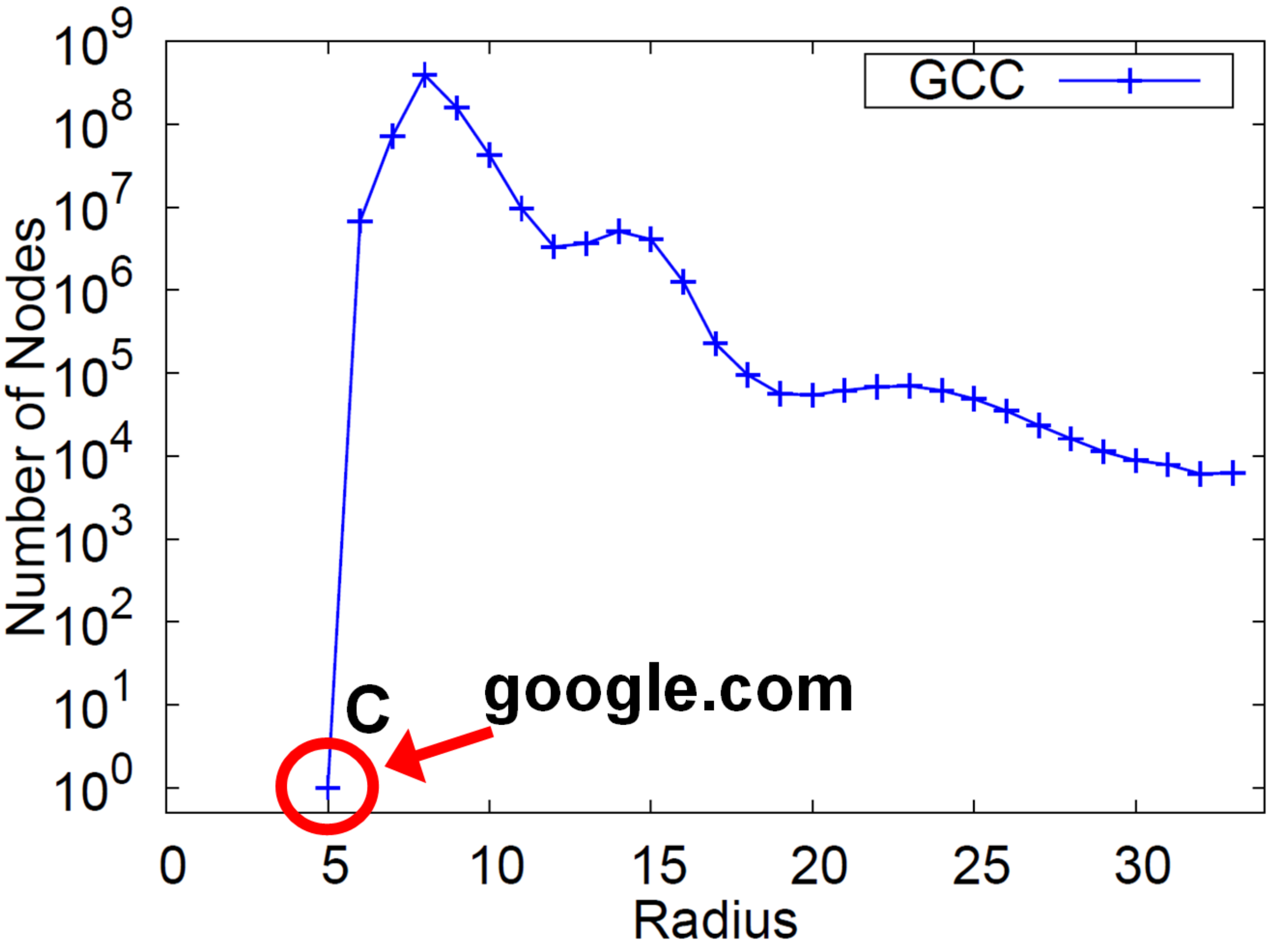



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
Q: Shape?

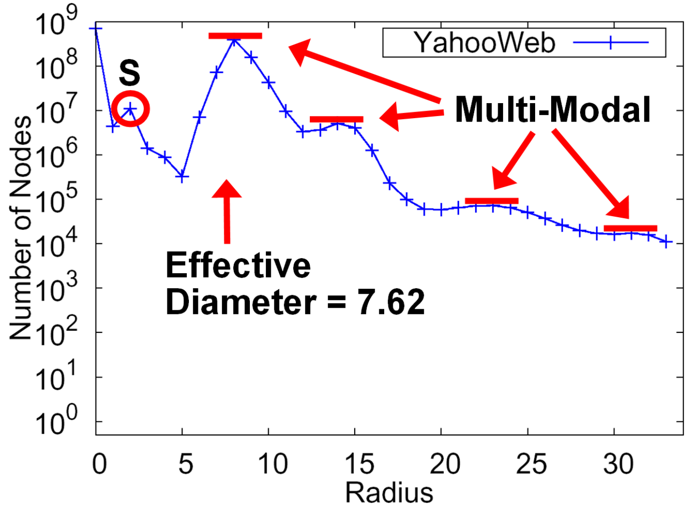



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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

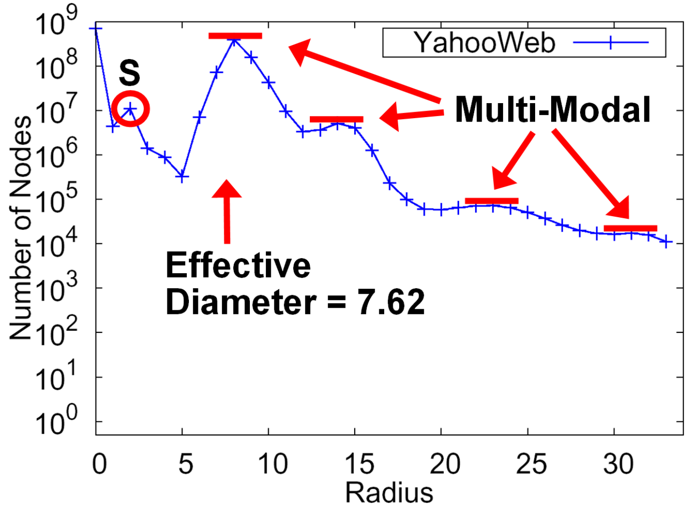


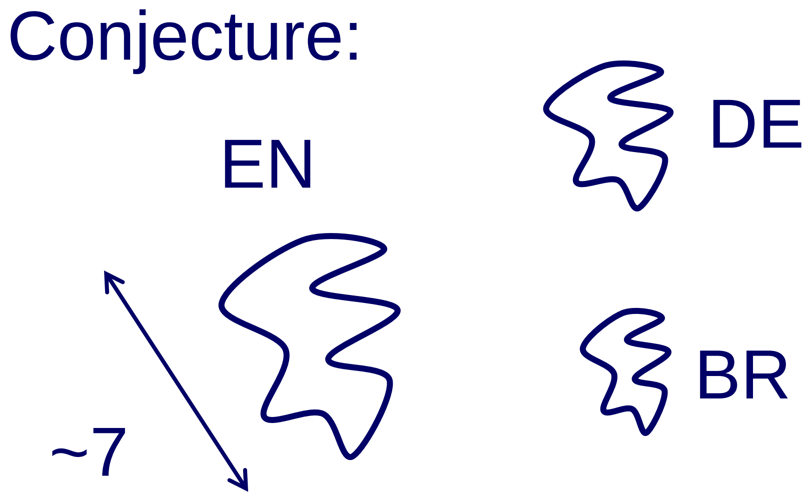

Radius Plot of **GCC** of YahooWeb.




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

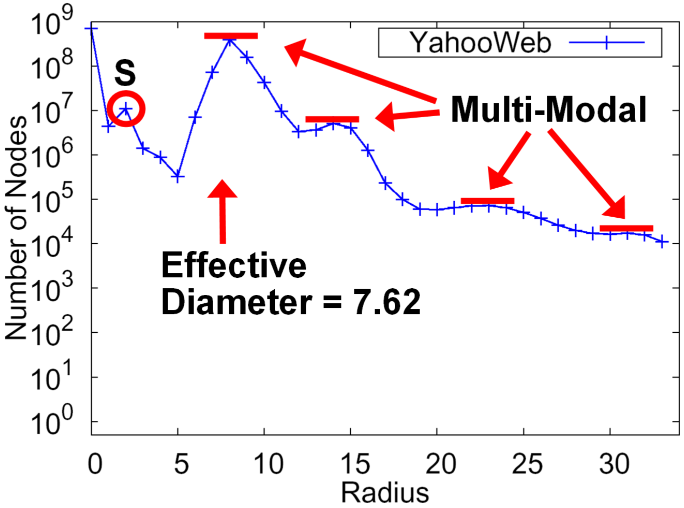
- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .



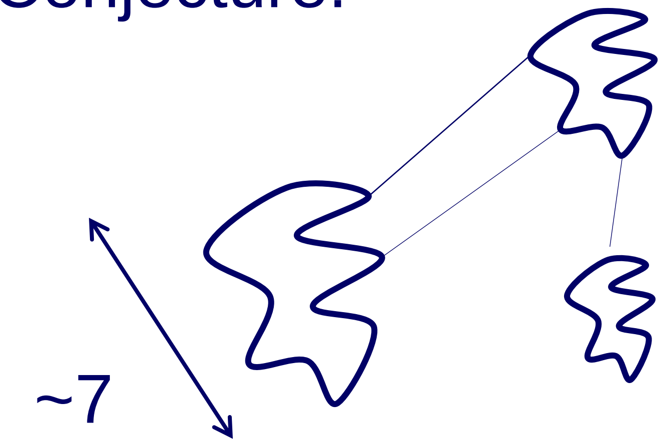




YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- 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 .

# Outline

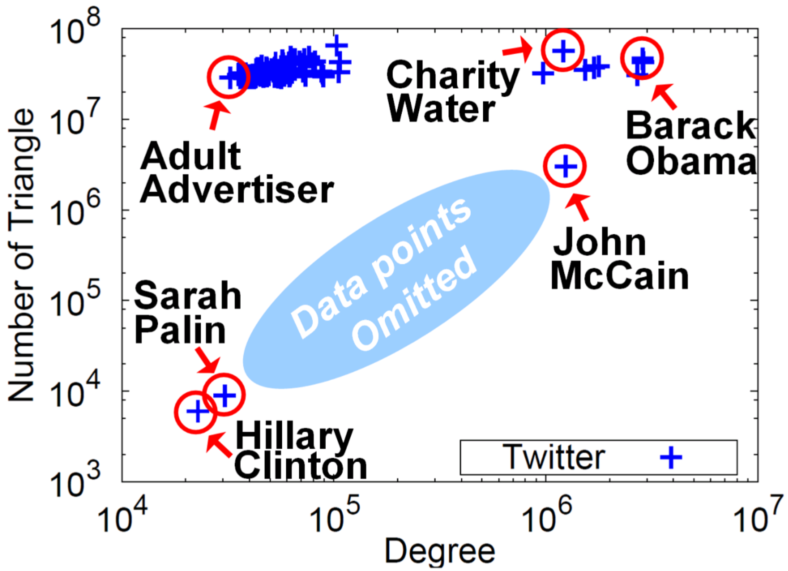
- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability
- Conclusions

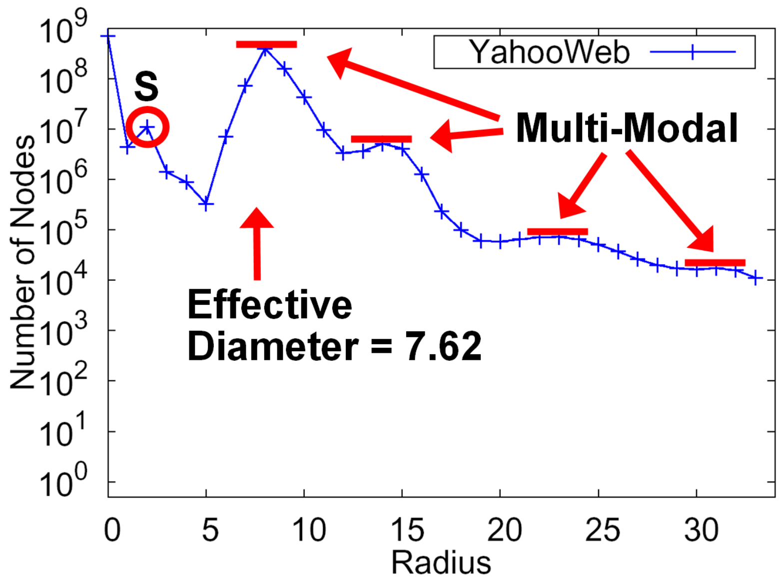
# OVERALL CONCLUSIONS – low level:

- Several new **patterns** (shrinking diameters, triangle-laws, etc)
- New **tools**:
  - Fraud detection (belief propagation)
- **Scalability**: PEGASUS / hadoop

# OVERALL CONCLUSIONS – medium-level

- **BIG DATA: Large** datasets reveal patterns/outliers that are invisible otherwise



# Project info

www.cs.cmu.edu/~pegasus

Chau, Polo

Koutra, Danae

Prakash, Aditya







Akoglu, Leman

Kang, U

McGlohon, Mary

Tong, Hanghang

Thanks to: NSF IIS-0705359, IIS-0534205, CTA-INARC; Yahoo (M45), LLNL, IBM, SPRINT, Google, INTEL, HP, iLab

# Thank you for the honor!

- Congratulations for 20-th anniversary

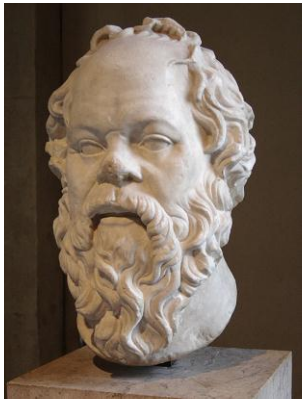
and…

# High-level conclusion: Collaborations

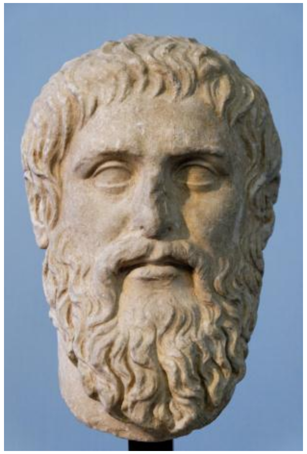
- Sociology + CS (triangles)
- Civil engineering + CS (sensor placement)
- fMRI/medical + graphs (medical db’s)
- …

# Never stop learning

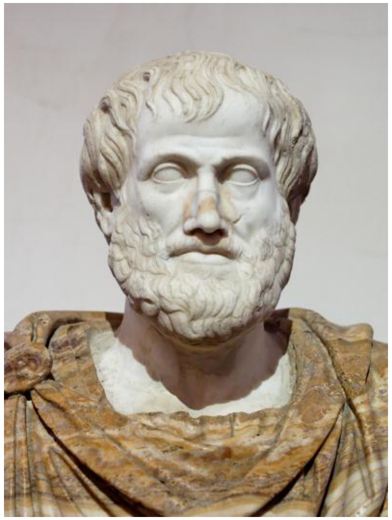
ΓΗΡΑΣΚΩ ΑΕΙ ΔΙΔΑΣΚΟΜΕΝΟΣ



Socrates



Plato



Aristotle