Large Graph Mining

Christos Faloutsos
CMU

Thank you!

- Hillol Kargupta

Outline

- **Problem definition / Motivation**
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; fraud detection
- Conclusions

Motivation

Data mining: ~ find patterns (rules, outliers)

- Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

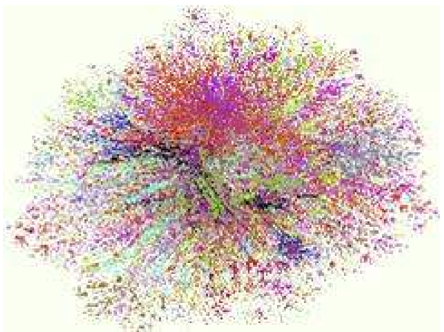
TOOLS

- Problem#4: Who is the ‘master-mind’?
- Problem#5: Fraud detection

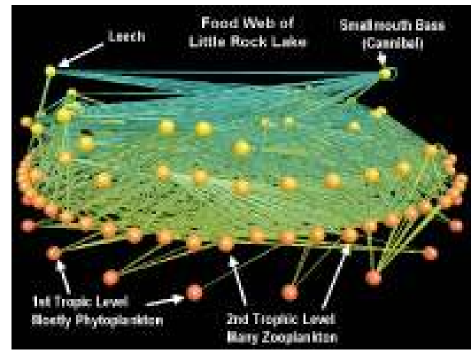
Problem#1: Joint work with

Dr. Deepayan Chakrabarti
(CMU/Yahoo R.L.)

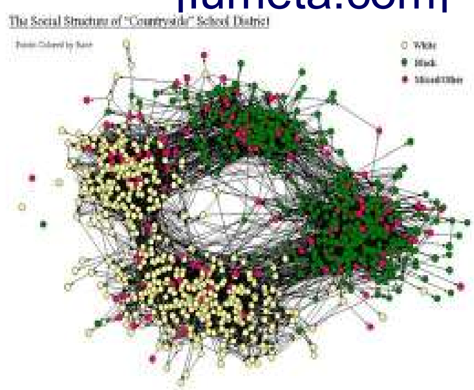
Graphs - why should we care?



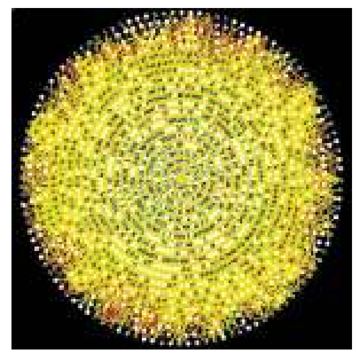
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



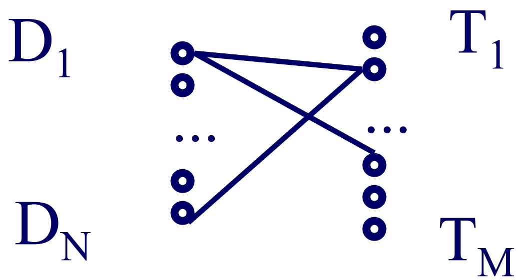
Friendship Network
[Moody '01]



Protein Interactions
[genomebiology.com]

Graphs - why should we care?

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



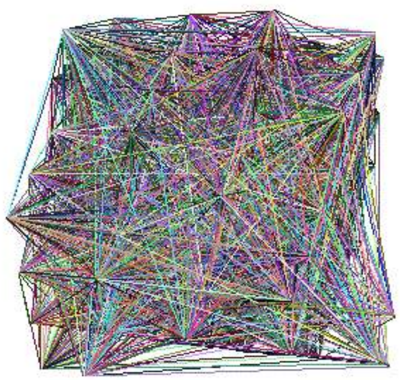
- web: hyper-text graph

- ... and more:

Graphs - why should we care?

- network of companies & board-of-directors members
- 'viral' marketing
- web-log ('blog') news propagation
- computer network security: email/IP traffic and anomaly detection
- ....

Problem #1 - network and graph mining



- How does the Internet look like?
- How does the web look like?
- What is 'normal'/'abnormal'?
- which patterns/laws hold?

Graph mining

- Are real graphs random?

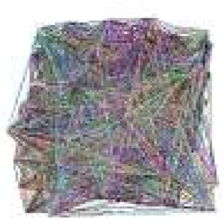
Laws and patterns

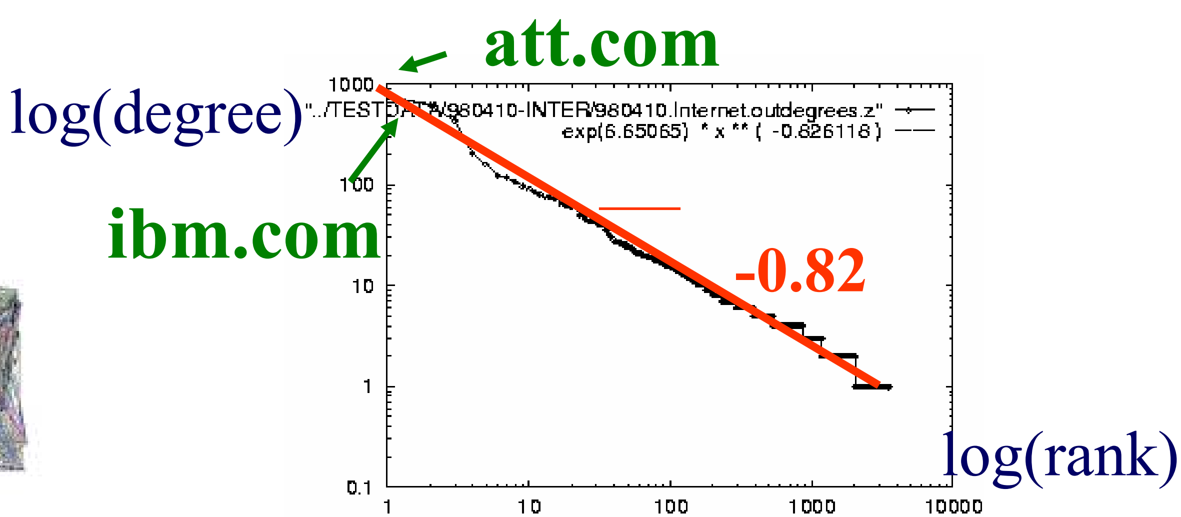
- Are real graphs random?
- A: NO!!
  - Diameter
  - in- and out- degree distributions
  - other (surprising) patterns

Solution#1

- Power law in the degree distribution [SIGCOMM99]

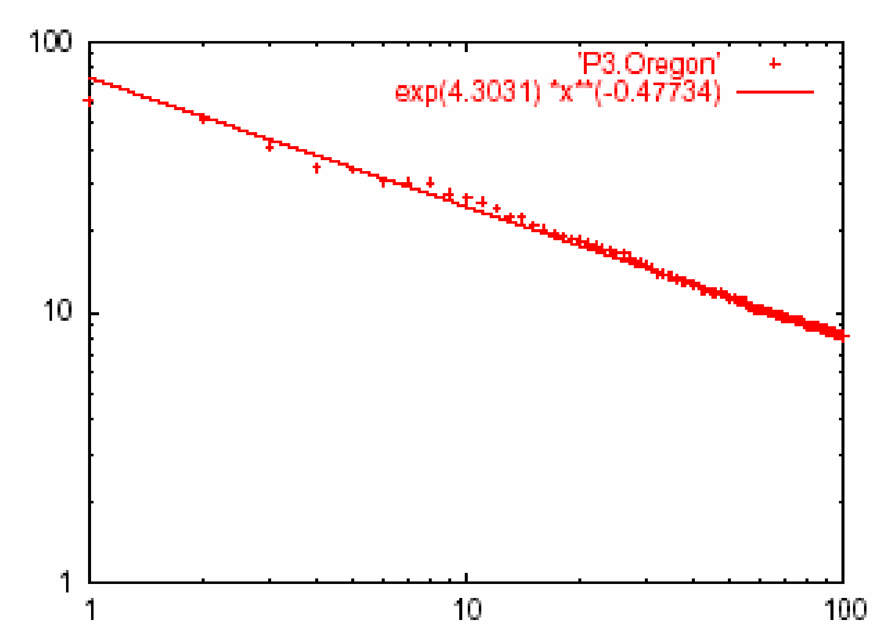
internet domains





Solution#1': Eigen Exponent *E*

Eigenvalue



Rank of decreasing eigenvalue

Exponent = slope

$E = -0.48$

May 2001

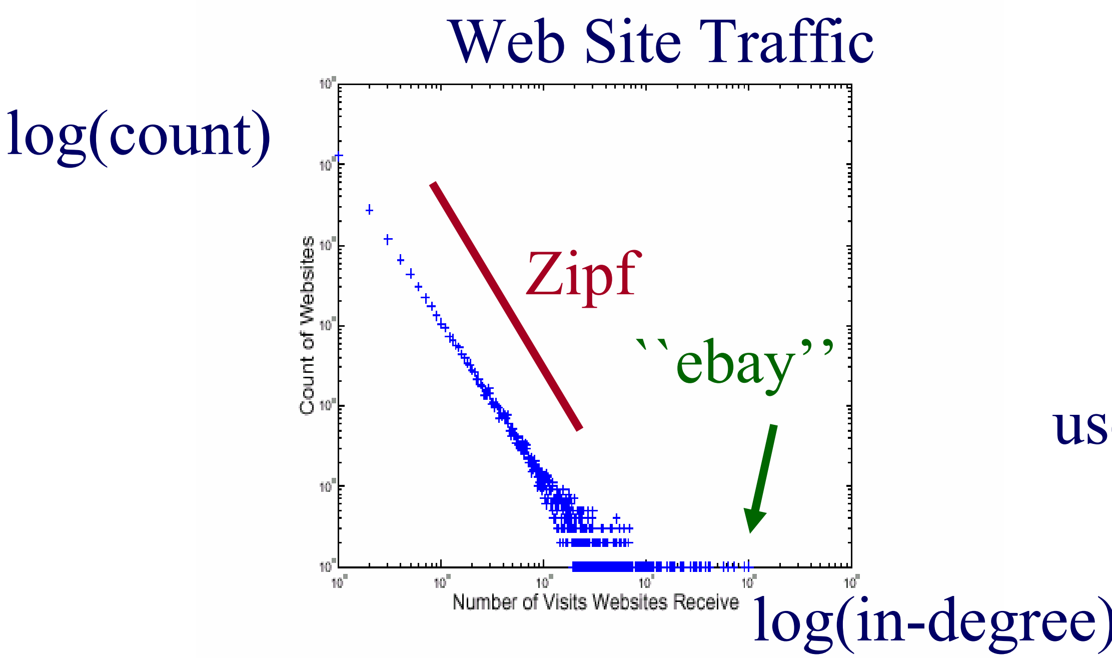
- A2: power law in the eigenvalues of the adjacency matrix

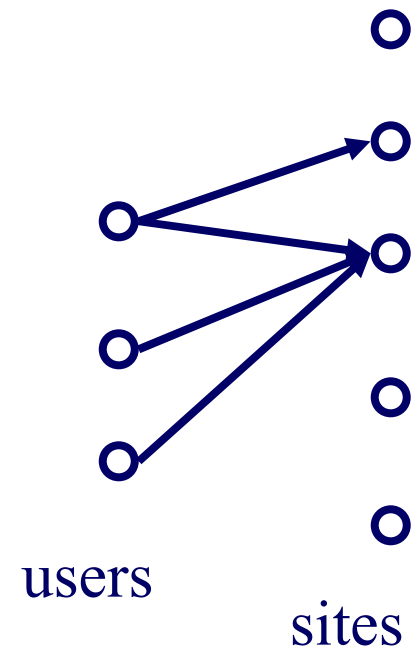
But:

How about graphs from other domains?

More power laws:

- web hit counts [w/ A. Montgomery]

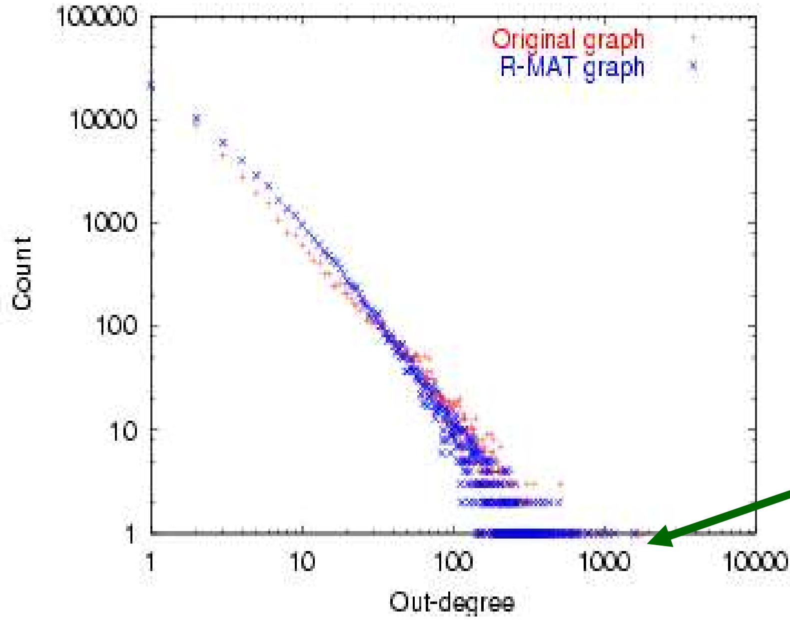




epinions.com

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

count



trusts-2000-people user

(out) degree

Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the 'master-mind'?
- Problem#5: Fraud detection

Problem#2: Time evolution

- with Jure Leskovec (CMU/MLD)

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

Evolution of the Diameter

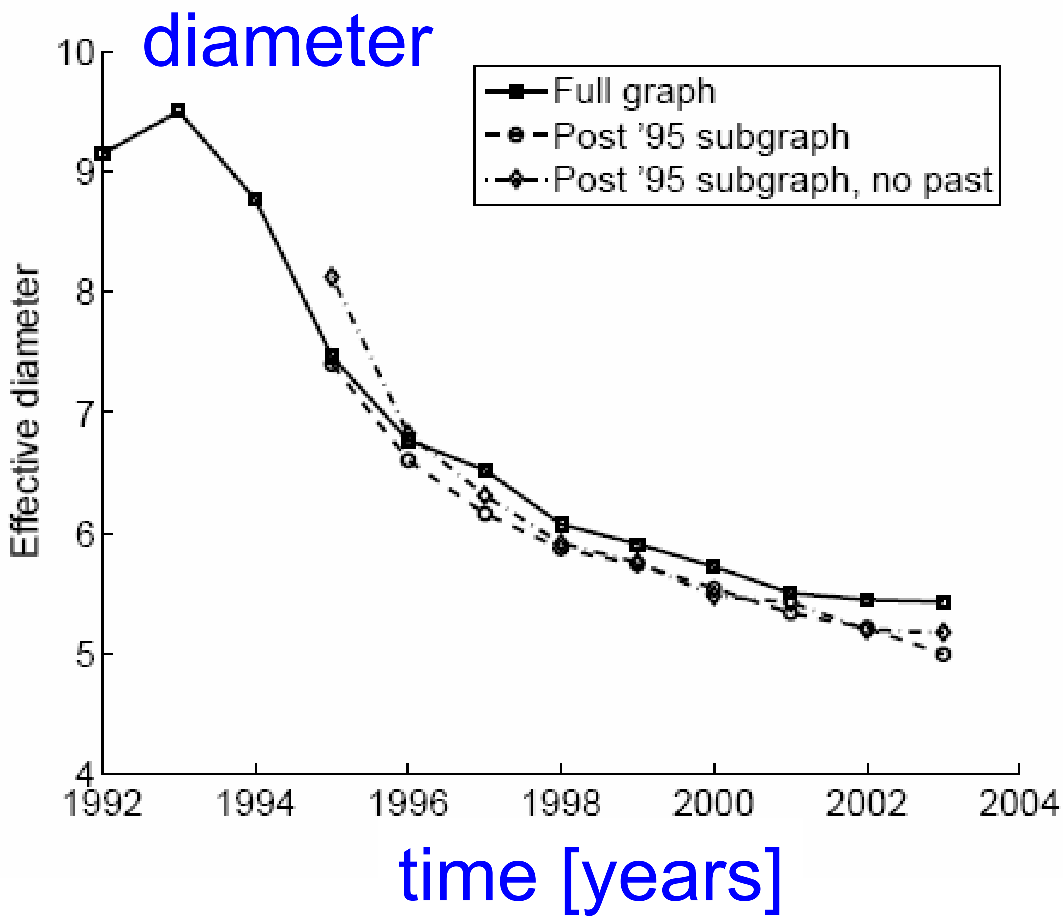
- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?

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

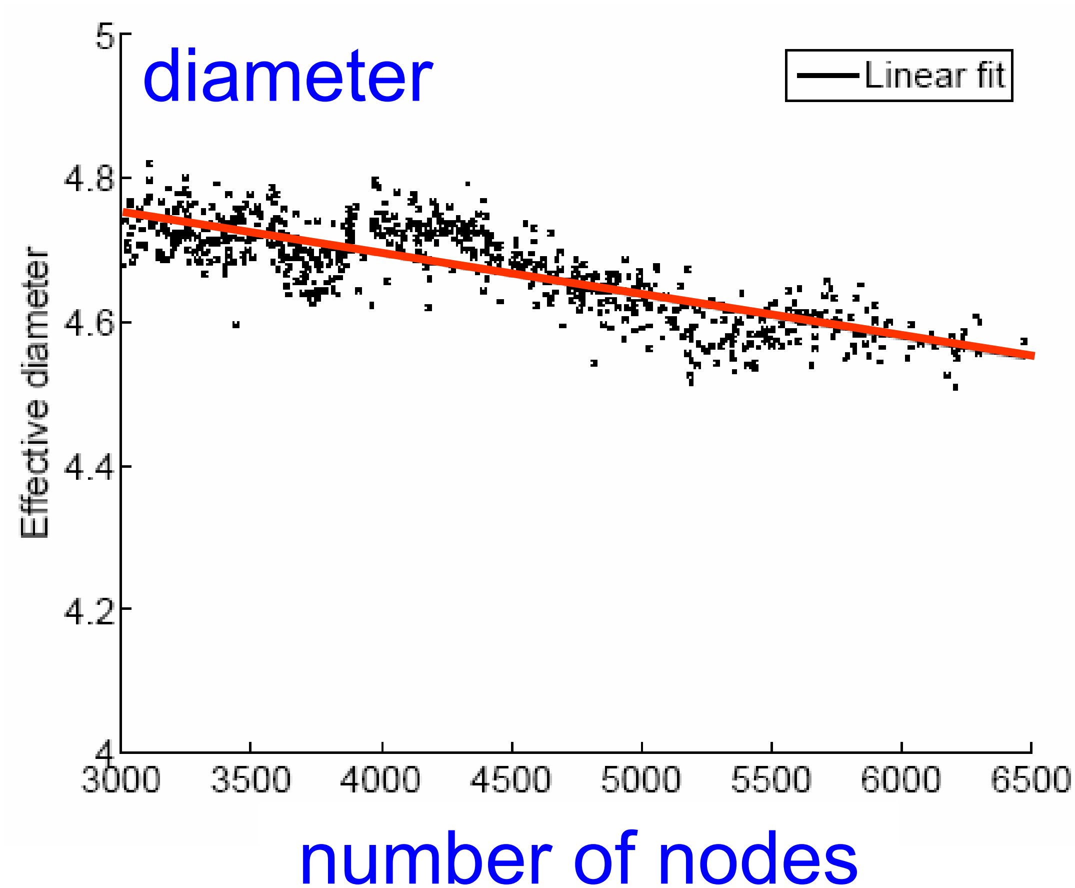
Diameter – ArXiv citation graph

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



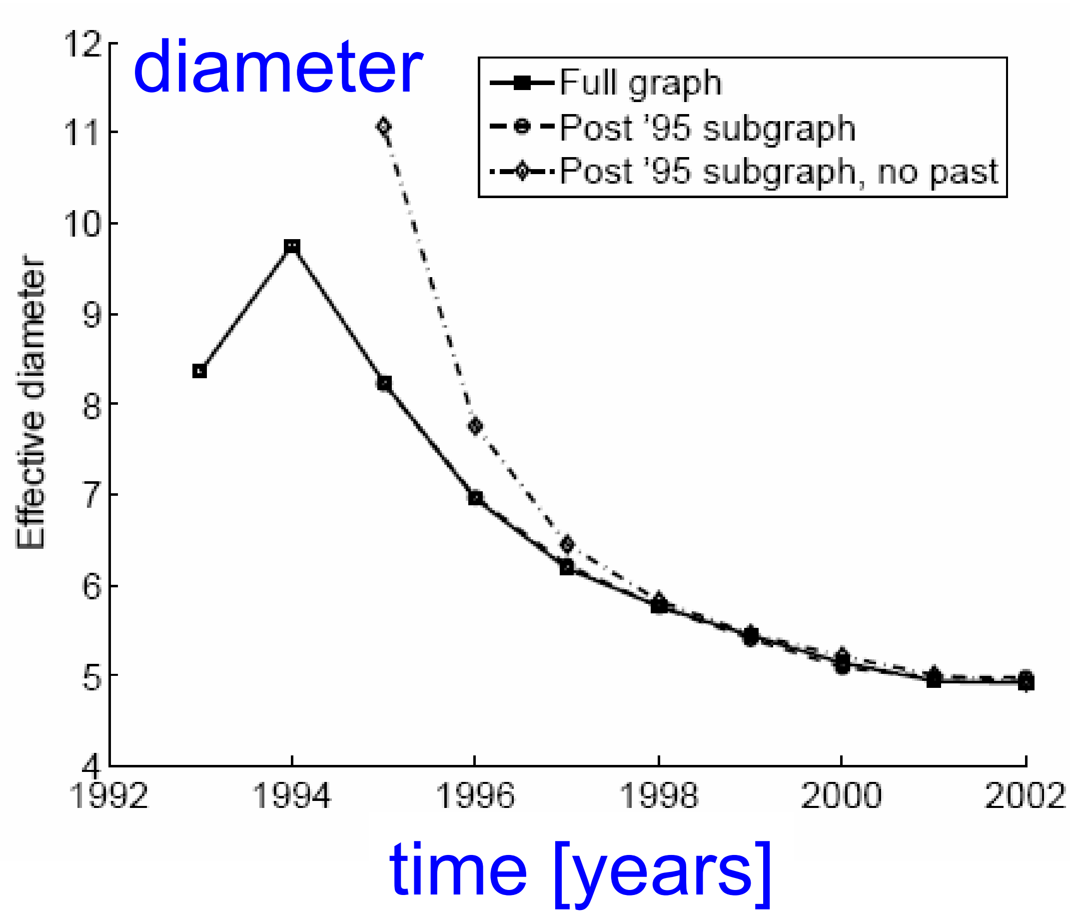
Diameter – "Autonomous Systems"

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



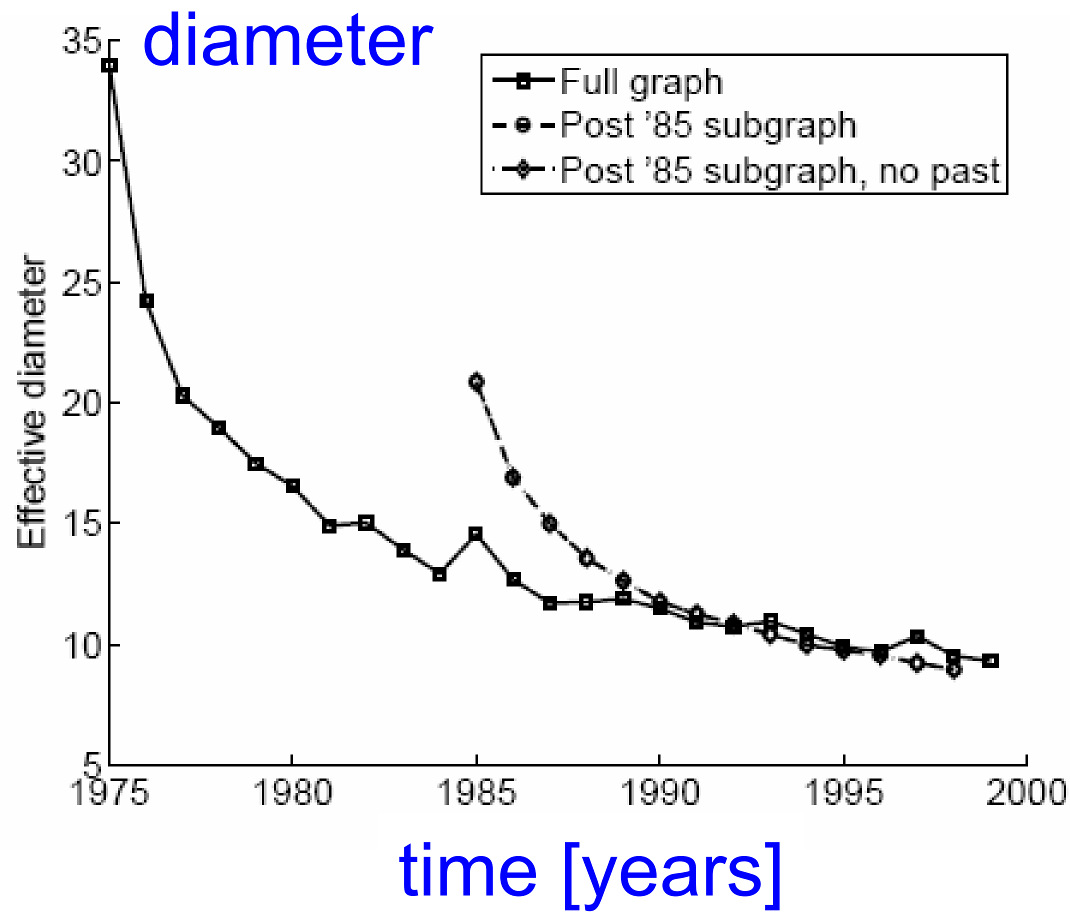
Diameter – "Affiliation Network"

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



Diameter – “Patents”

- Patent citation network
- 25 years of data



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

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

Temporal Evolution of the Graphs

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

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

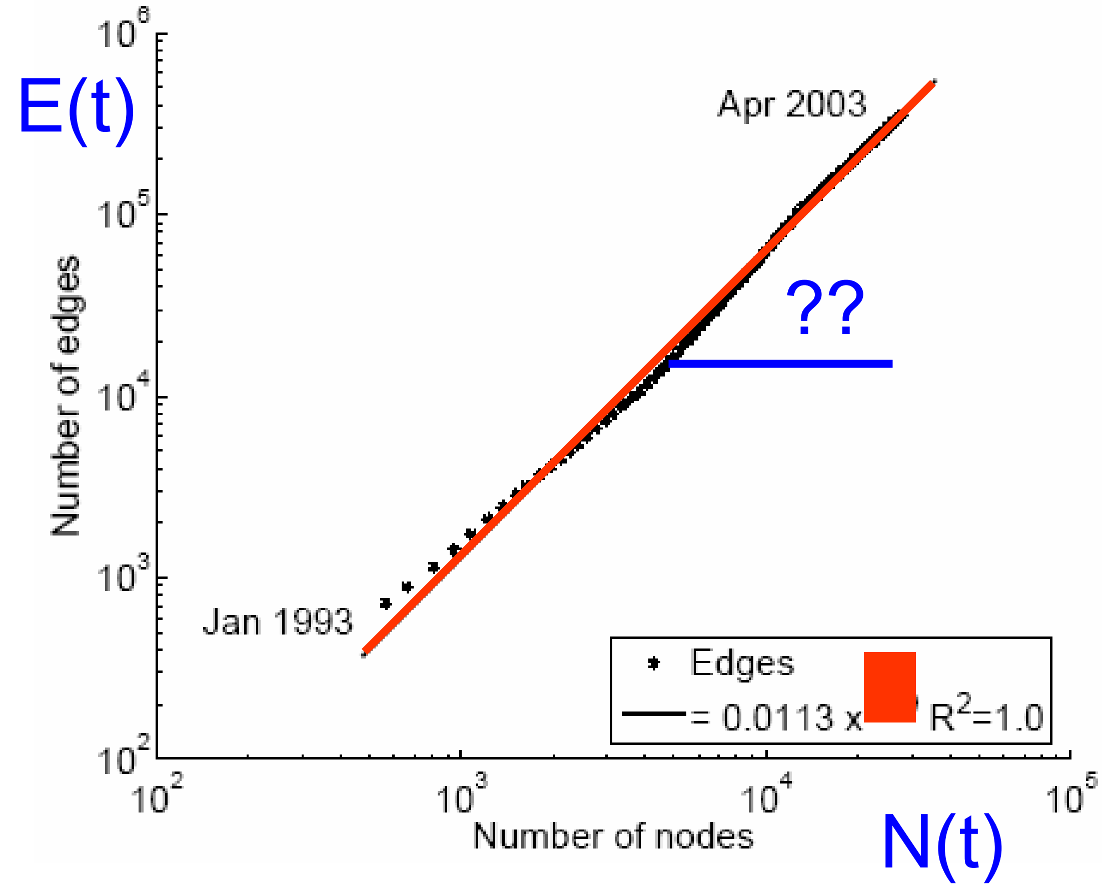
- Q: what is your guess for

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

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

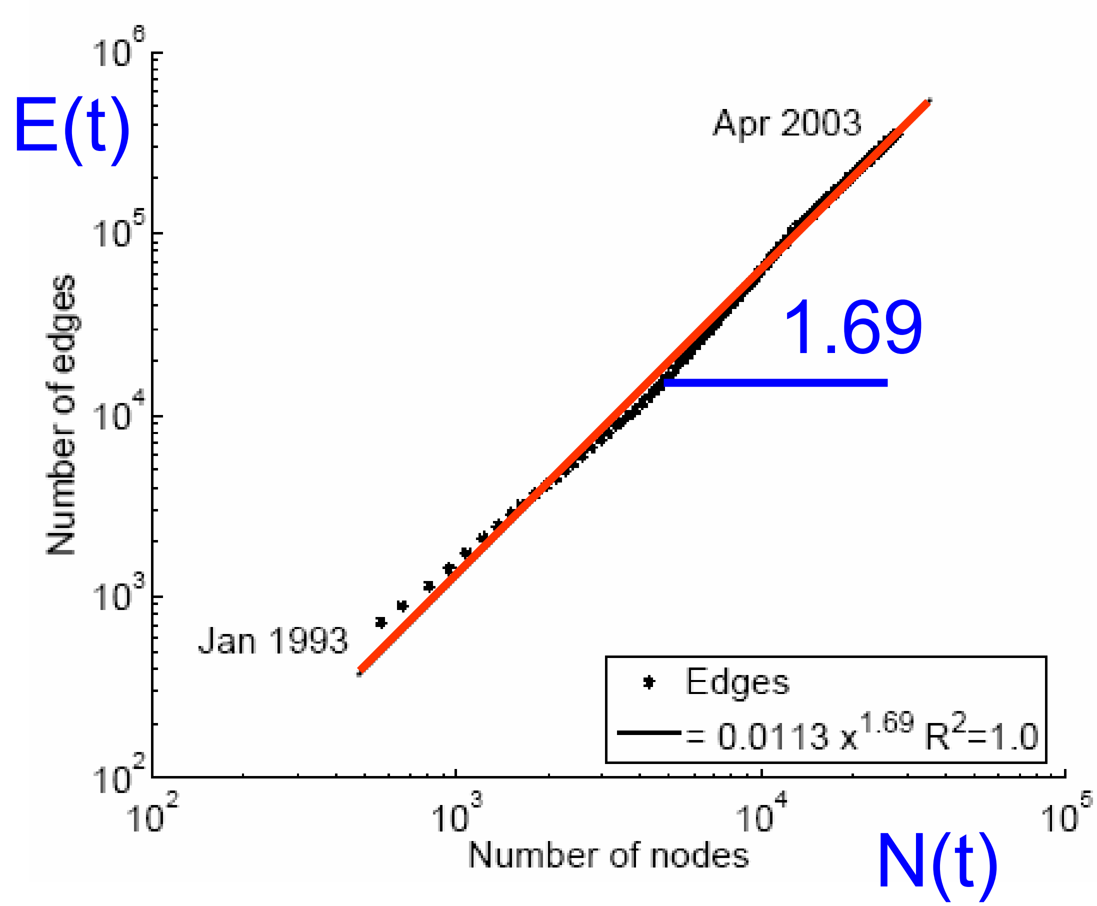
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



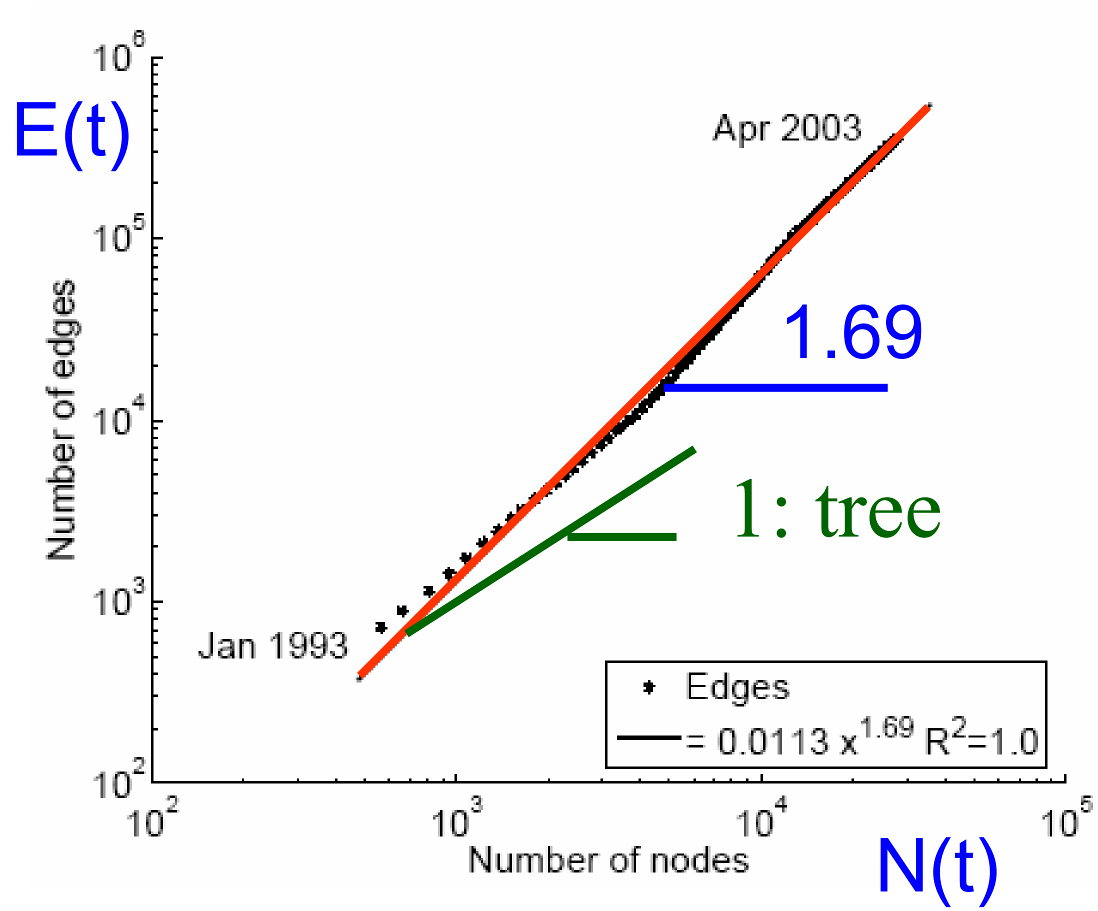
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



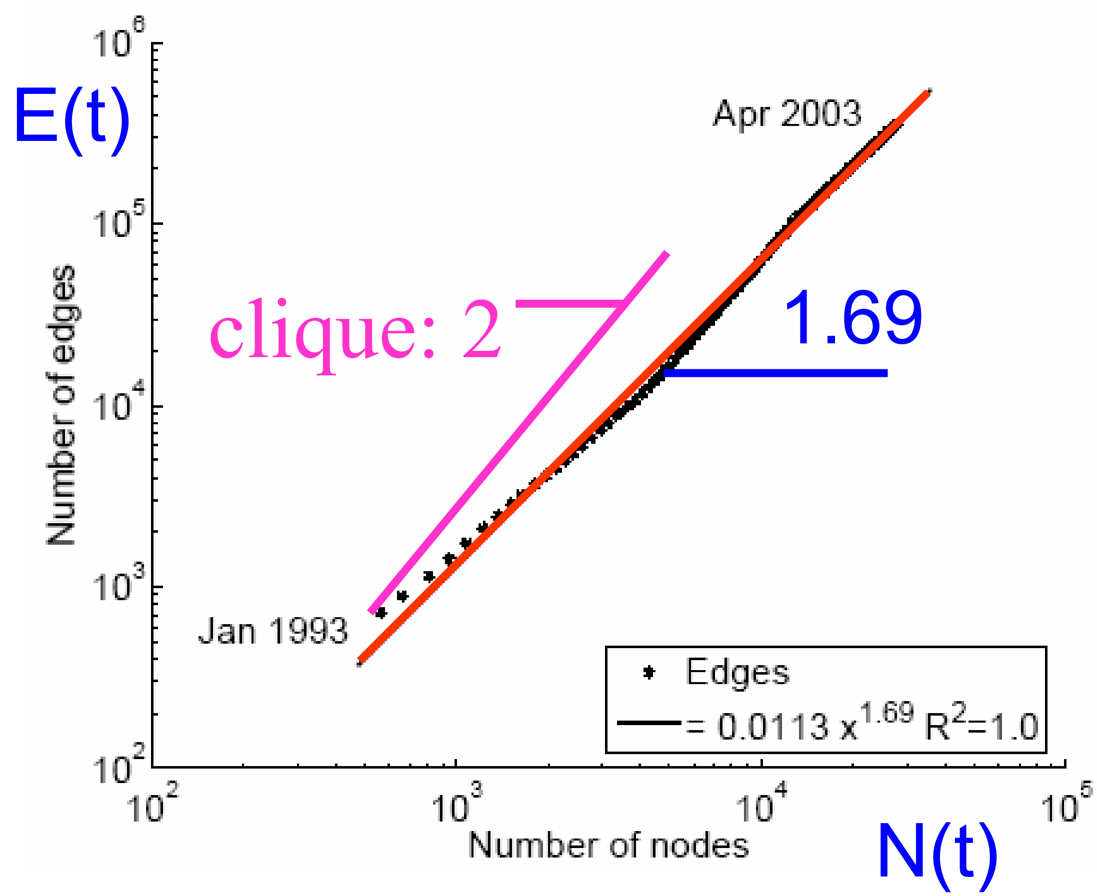
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



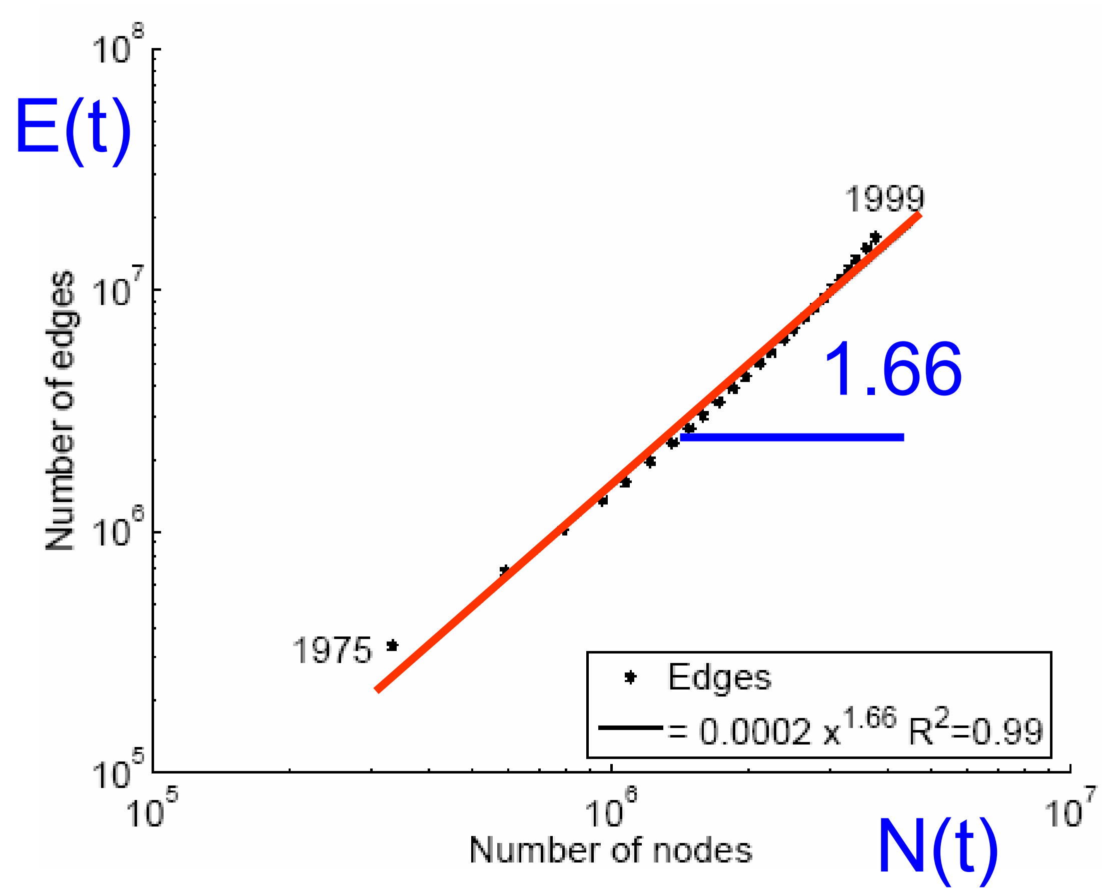
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



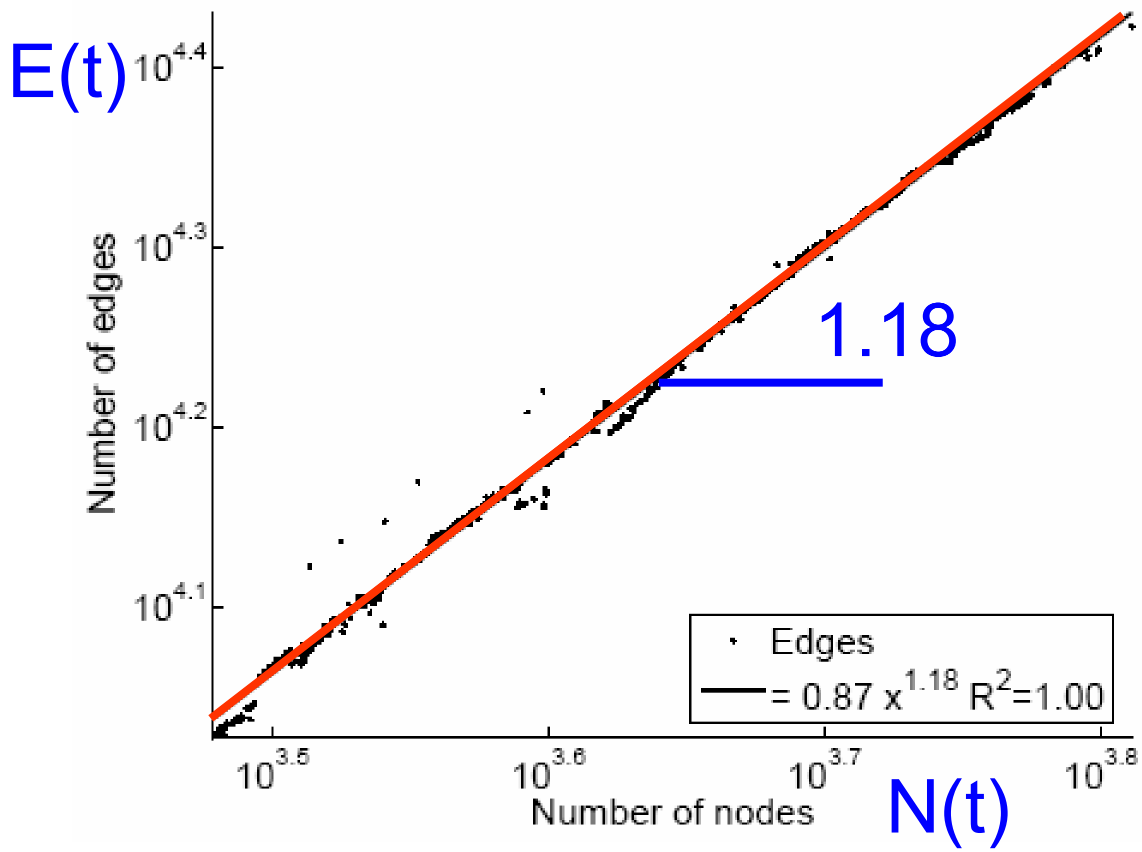
Densification – Patent Citations

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



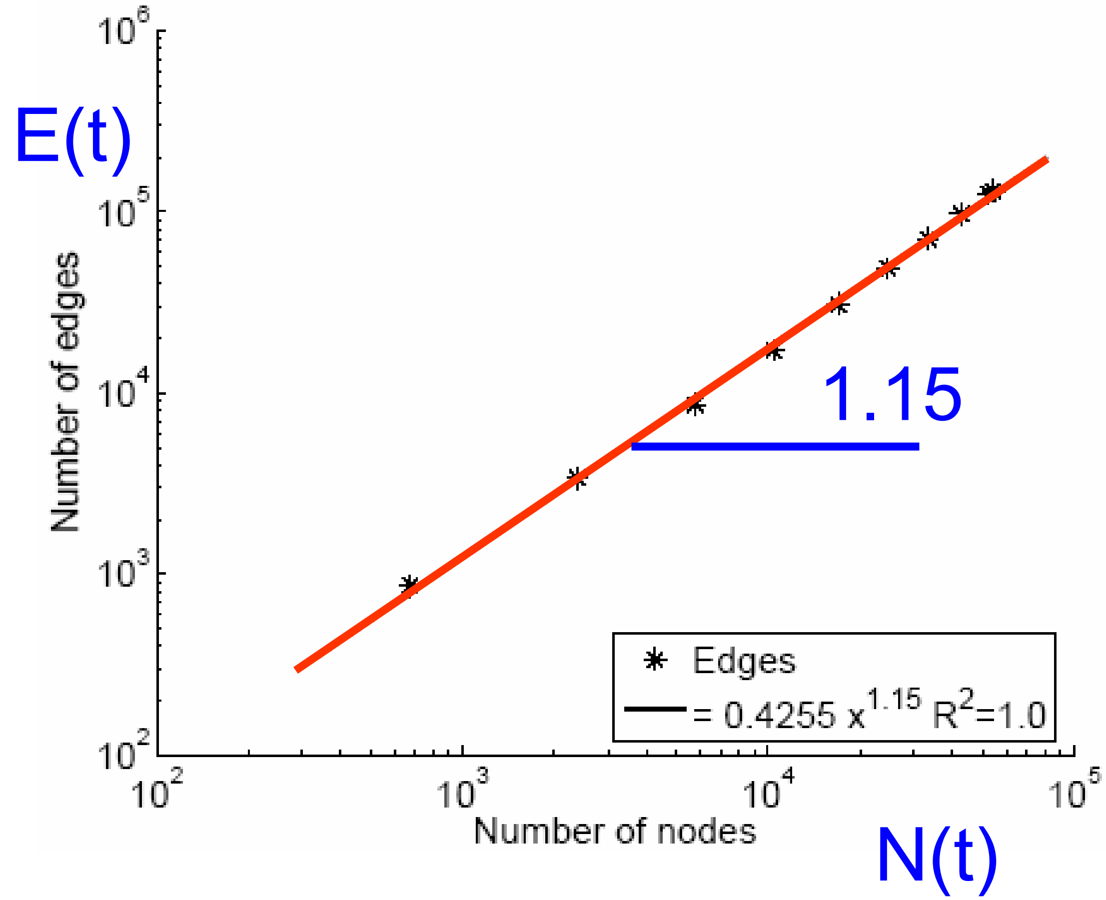
Densification – Autonomous Systems

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



Densification – Affiliation Network

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



Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the 'master-mind'?
- Problem#5: Fraud detection

Problem#3: Generation

- Given a growing graph with count of nodes N_1, N_2, ...
- Generate a realistic sequence of graphs that will obey all the patterns

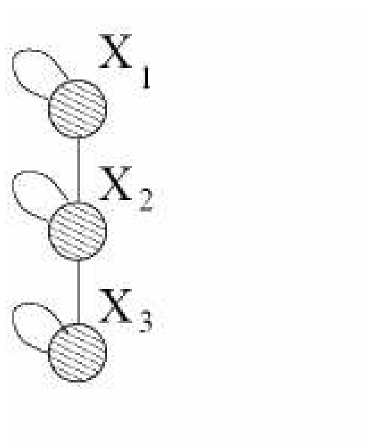
Problem Definition

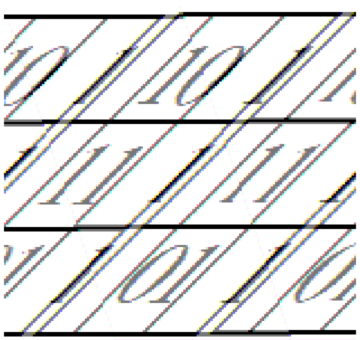
- Given a growing graph with count of nodes N_1, N_2, ...
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - Power Law Degree Distribution
    - Power Law eigenvalue and eigenvector distribution
    - Small Diameter
  - Dynamic Patterns
    - Growth Power Law
    - Shrinking/Stabilizing Diameters

Problem Definition

- Given a growing graph with count of nodes N_1, N_2, ...
- Generate a realistic sequence of graphs that will obey all the patterns
- **Idea: Self-similarity**
  - Leads to power laws
  - Communities within communities
  - …

Kronecker Product – a Graph



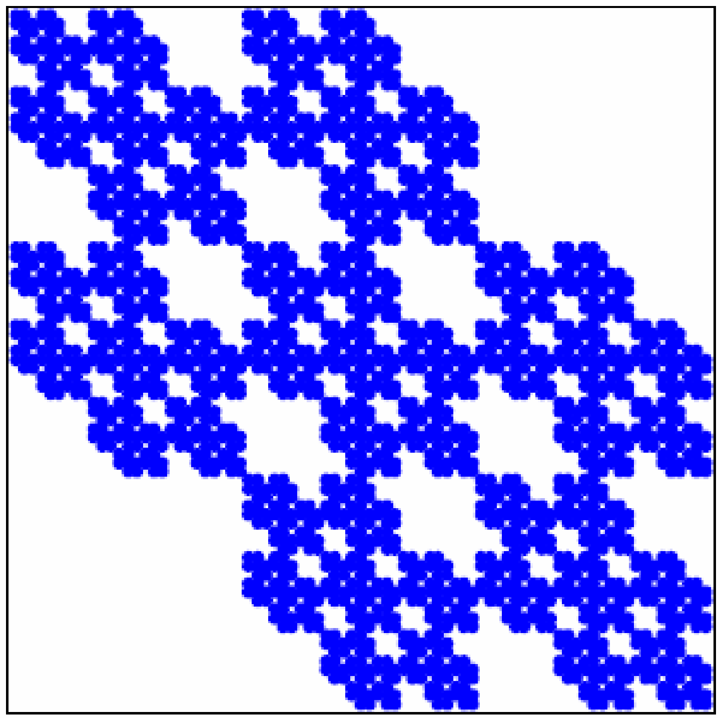


G_1

Adjacency matrix

Kronecker Product – a Graph

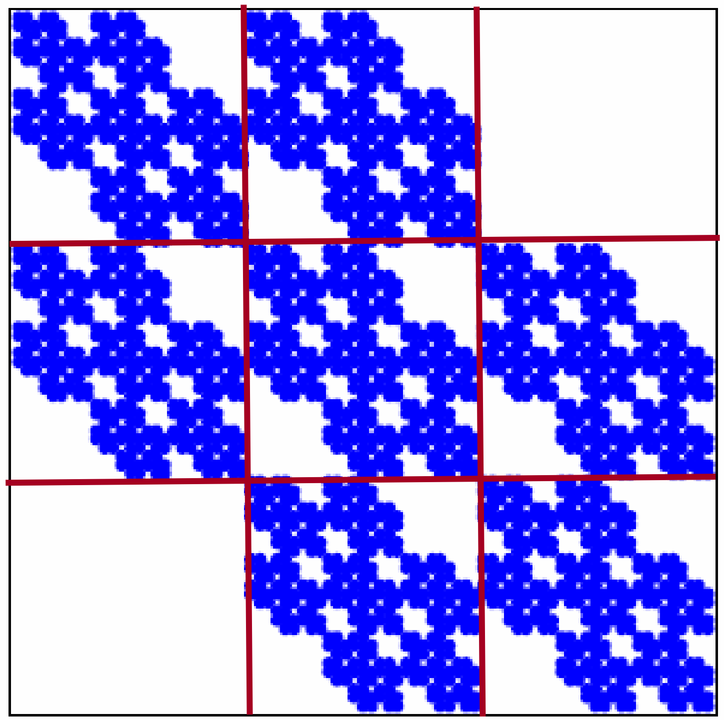
- Continuing multiplying with G_1 we obtain G_4 and so on …



G_4 adjacency matrix

Kronecker Product – a Graph

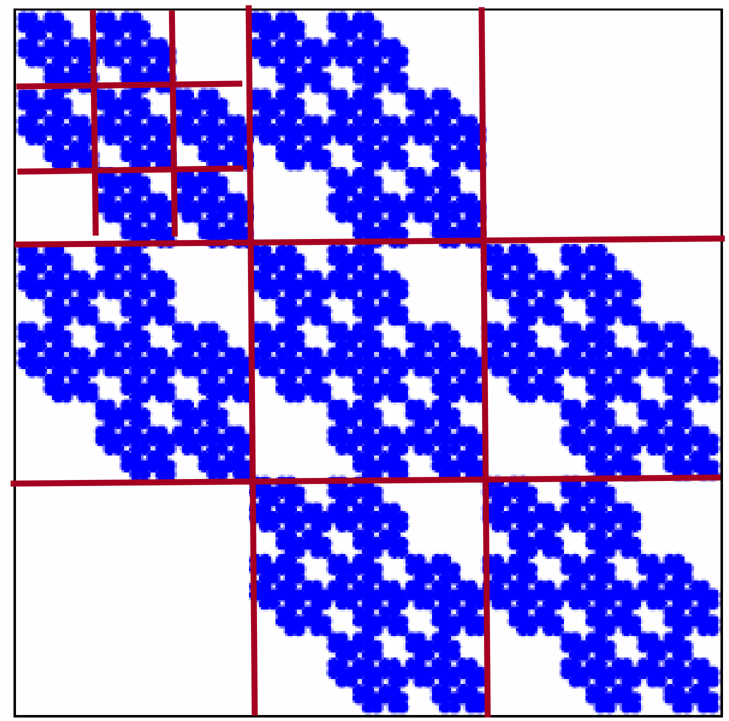
- Continuing multiplying with G_1 we obtain G_4 and so on …



G_4 adjacency matrix

Kronecker Product – a Graph

- Continuing multiplying with G_1 we obtain G_4 and so on …



G_4 adjacency matrix

Properties:

- We can PROVE that
  - Degree distribution is multinomial ~ power law
  - Diameter: constant
  - Eigenvalue distribution: multinomial
  - First eigenvector: multinomial
- See [Leskovec+, PKDD'05] for proofs

Problem Definition

- Given a growing graph with nodes N_1, N_2, ...
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - ✓ Power Law Degree Distribution
    - ✓ Power Law eigenvalue and eigenvector distribution
    - ✓ Small Diameter
  - Dynamic Patterns
    - ✓ Growth Power Law
    - ✓ Shrinking/Stabilizing Diameters
- First and only generator for which we can **prove** all these properties

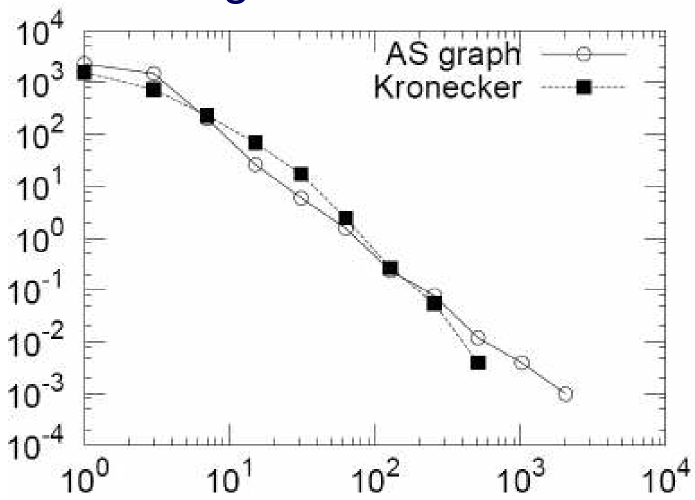
(Q: how to fit the parm's?)

A:

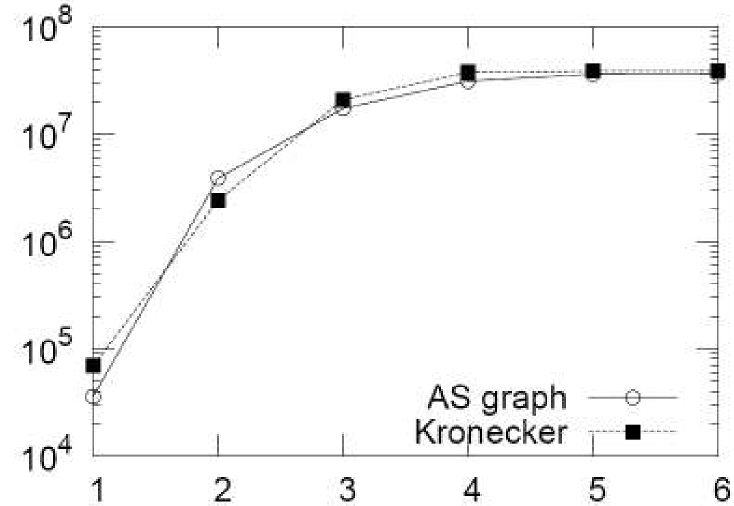
- Stochastic version of Kronecker graphs +
- Max likelihood +
- Metropolis sampling
- [Leskovec+, ICML'07]

Experiments on real AS graph

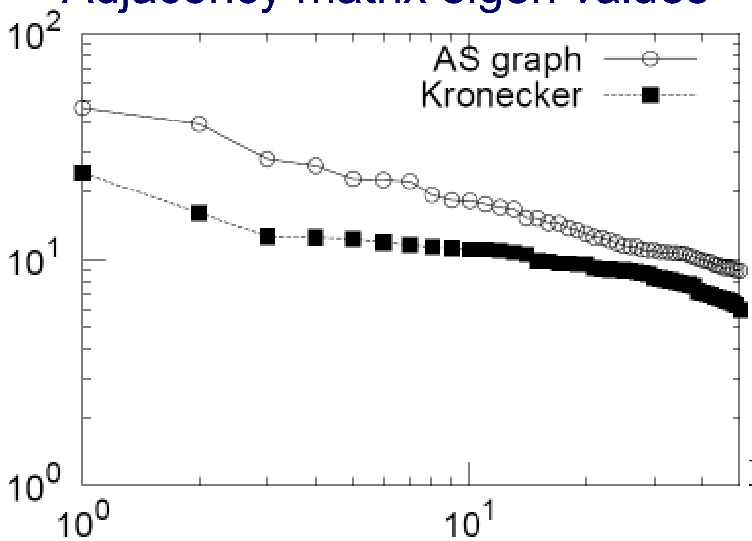
Degree distribution



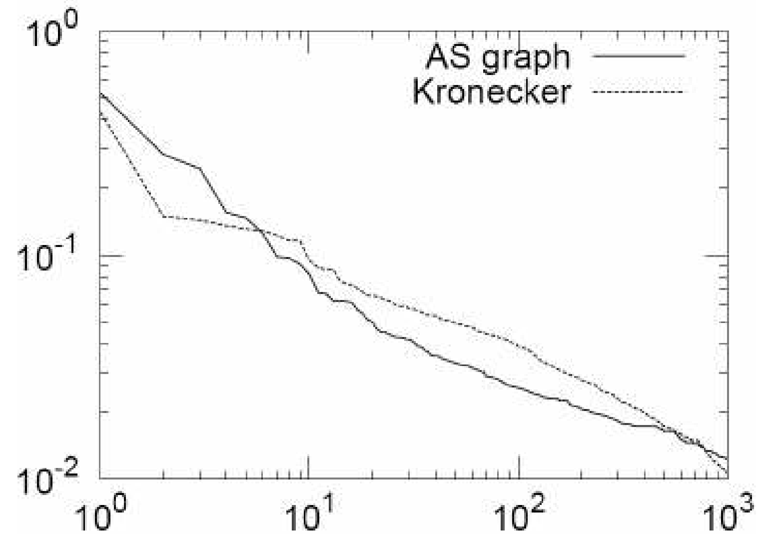
Hop plot



Adjacency matrix eigen values



Network value



Conclusions

- Kronecker graphs have:
  - All the static properties
    - ✓ Heavy tailed degree distributions
    - ✓ Small diameter
    - ✓ Multinomial eigenvalues and eigenvectors
  - All the temporal properties
    - ✓ Densification Power Law
    - ✓ Shrinking/Stabilizing Diameters
  - We can formally prove these results

Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- ✓ Problem#3: How to generate realistic graphs

TOOLS

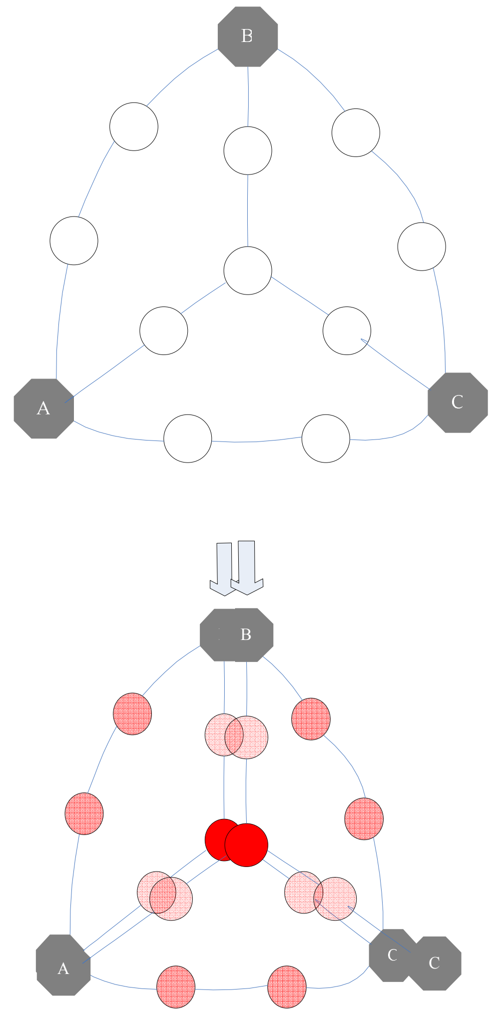
- Problem#4: Who is the 'master-mind'?
- Problem#5: Fraud detection

Problem#4: MasterMind – ‘CePS’

- w/ Hanghang Tong, KDD 2006
- htong <at> cs.cmu.edu

Center-Piece Subgraph(Ceps)

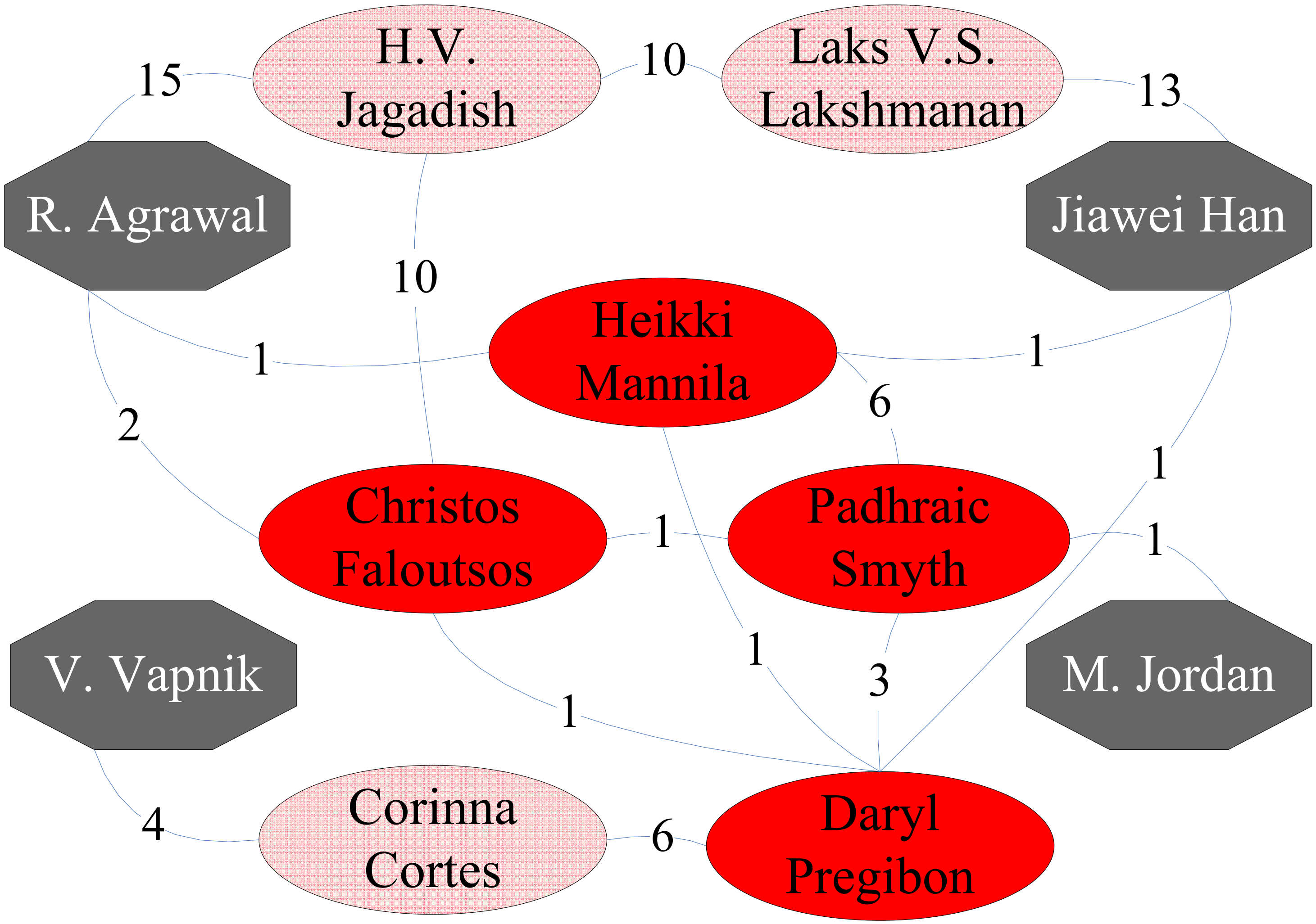
- **Given** Q query nodes
- **Find** Center-piece ($\leq b$)



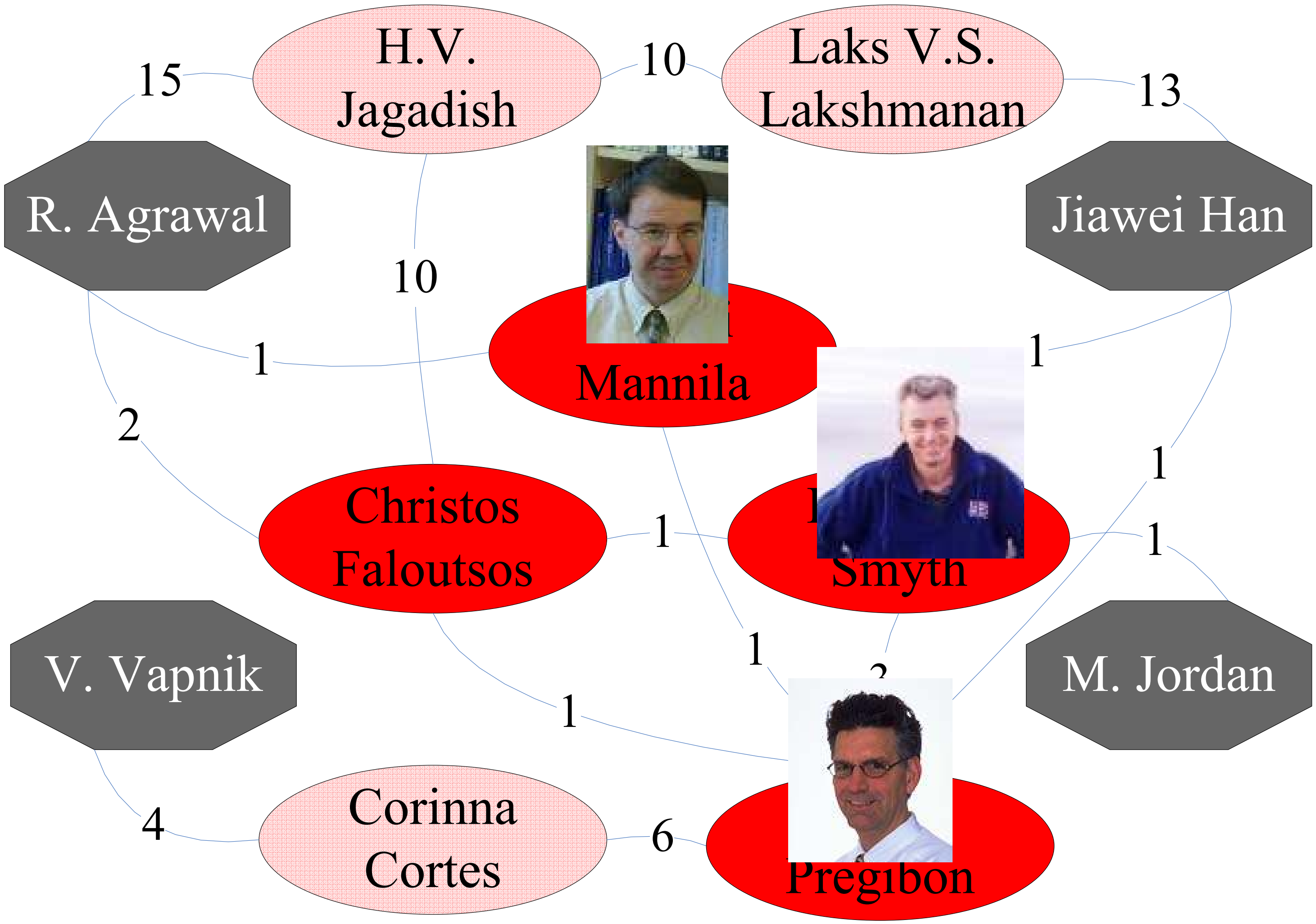
- App.
  - Social Networks
  - Law Inforcement, …
- Idea:
  - Proximity -> random walk with restarts

Case Study: AND query

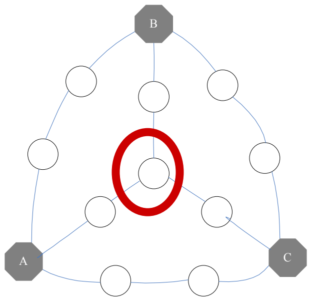
Case Study: AND query



Case Study: AND query



Conclusions



- Q1:How to measure the importance?
- A1: RWR+K_SoftAnd
- Q2:How to do it efficiently?
- A2:Graph Partition (Fast CePS)
  - ~90% quality
  - 150x speedup (ICDM'06)

Motivation

Data mining: ~ find patterns (rules, outliers)

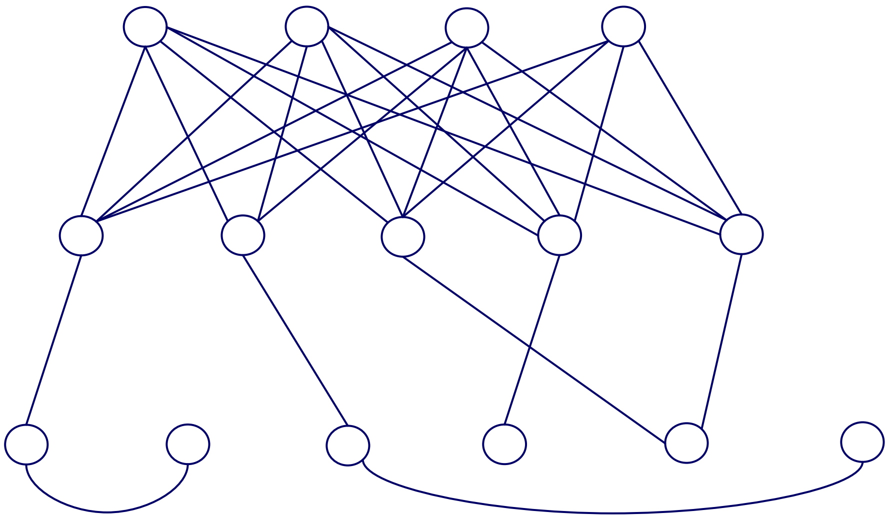
- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- ✓ Problem#3: How to generate realistic graphs

TOOLS

- ✓ Problem#4: Who is the 'master-mind'?
- Problem#5: Fraud detection

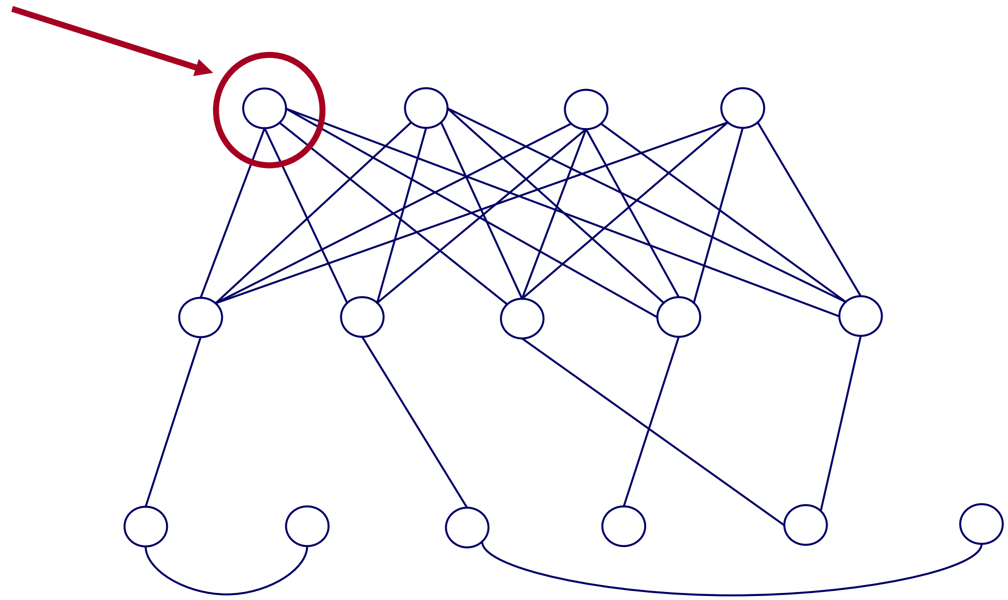
E-bay Fraud detection

w/ Polo Chau &
Shashank Pandit, CMU



E-bay Fraud detection

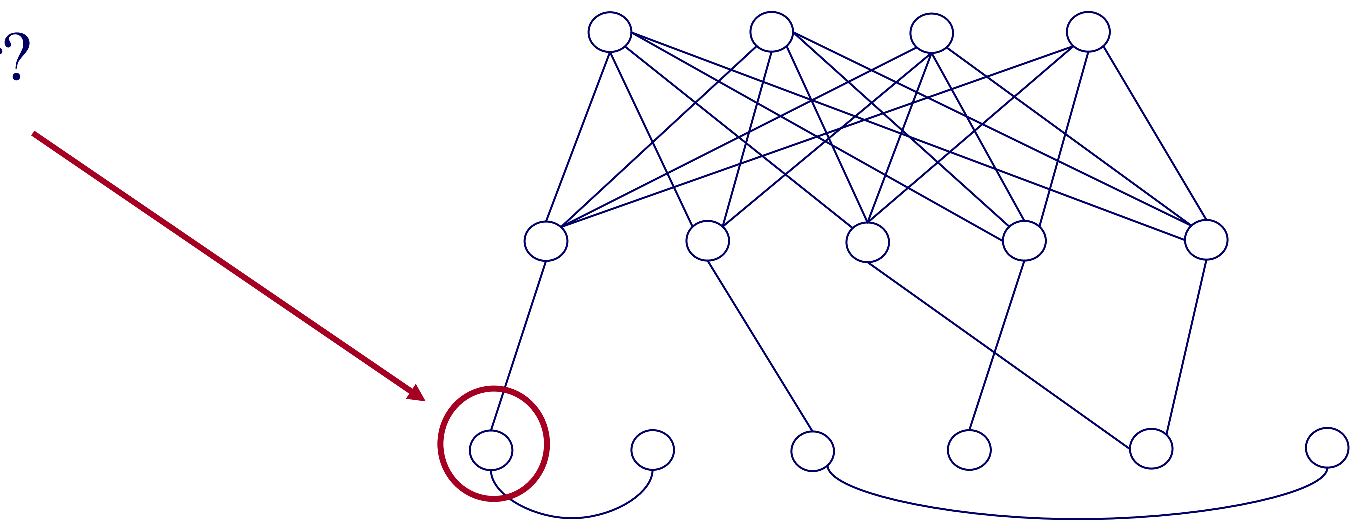
- lines: positive feedbacks
- would you buy from him/her?



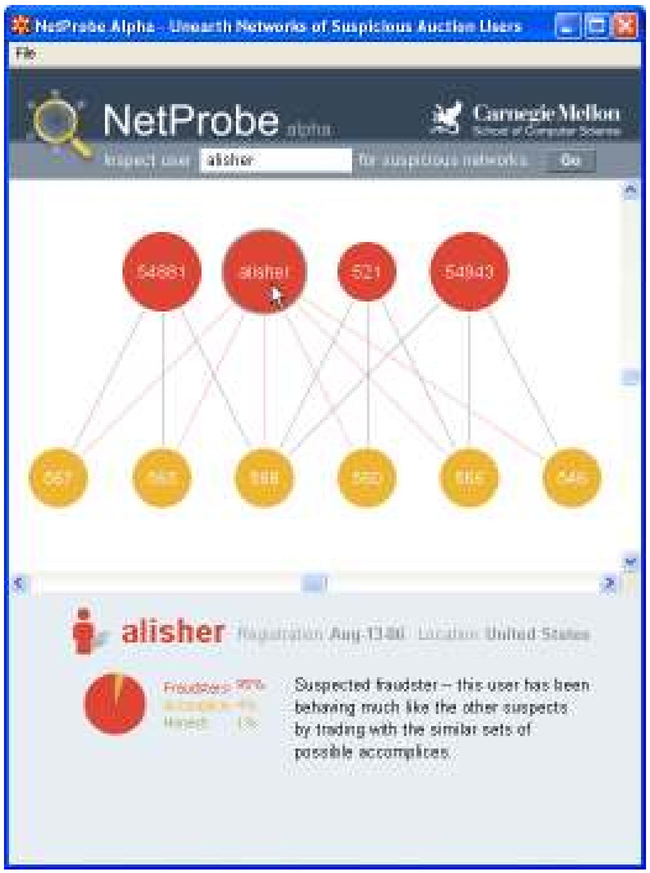
E-bay Fraud detection

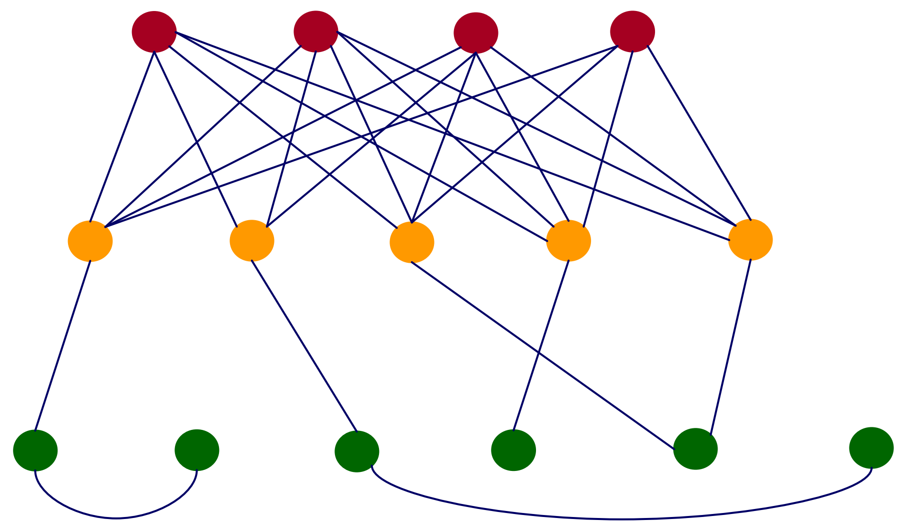
- lines: positive feedbacks
- would you buy from him/her?

- or him/her?



E-bay Fraud detection - NetProbe



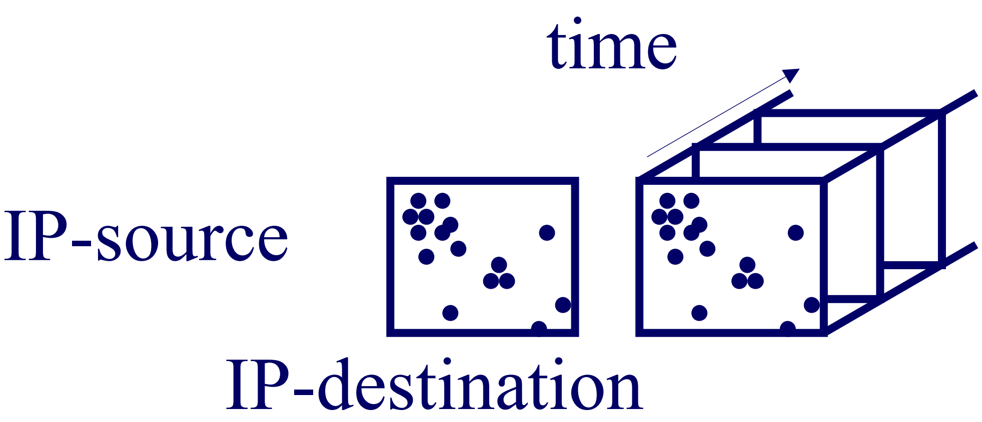


OVERALL CONCLUSIONS

- Graphs pose a wealth of fascinating problems
- self-similarity and power laws work, when textbook methods fail!
- New patterns (shrinking diameter!)
- New generator: Kronecker

Promising directions

- Reaching out
  - Sociology, epidemiology; physics, ++…
  - Computer networks, security, intrusion det.
  - Num. analysis (tensors)



Promising directions – cont'd

- Scaling up, to Gb/Tb/Pb
  - Storage Systems
  - Parallelism (hadoop/map-reduce)

E.g.: self-* system @ CMU

- >200 nodes
- 40 racks of computing equipment
- 774kw of power.
- target: 1 PetaByte
- goal: self-correcting, self-securing, self-monitoring, self-...

DM for Tera- and Peta-bytes

Two-way street:

<- DM can use such infrastructures to find patterns

-> DM can help such infrastructures become self-healing, self-adjusting, 'self-*'

References

- Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan *Fast Random Walk with Restart and Its Applications* ICDM 2006, Hong Kong.
- Hanghang Tong, Christos Faloutsos *Center-Piece Subgraphs: Problem Definition and Fast Solutions,* KDD 2006, Philadelphia, PA

References

- Jure Leskovec, Jon Kleinberg and Christos Faloutsos *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations* KDD 2005, Chicago, IL. ("Best Research Paper" award).
- Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication* (ECML/PKDD 2005), Porto, Portugal, 2005.

References

- Jure Leskovec and Christos Faloutsos, *Scalable Modeling of Real Graphs using Kronecker Multiplication*, ICML 2007, Corvallis, OR, USA
- Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang and Christos Faloutsos *NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks* WWW 2007, Banff, Alberta, Canada, May 8-12, 2007.
- Jimeng Sun, Dacheng Tao, Christos Faloutsos *Beyond Streams and Graphs: Dynamic Tensor Analysis,* KDD 2006, Philadelphia, PA

References

- Jimeng Sun, Yinglian Xie, Hui Zhang, Christos Faloutsos. *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, SDM, Minneapolis, Minnesota, Apr 2007. [pdf]

THANK YOU!

Contact info:

www. cs.cmu.edu /~christos

(w/ papers, datasets, code, etc)