Graph Mining: Laws, Generators and Tools

Christos Faloutsos

CMU

Thanks

- Michael Mahoney

- Lek-Heng Lim
- Petros Drineas
- Gunnar Carlsson

Outline

- **Problem definition / Motivation**
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay 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

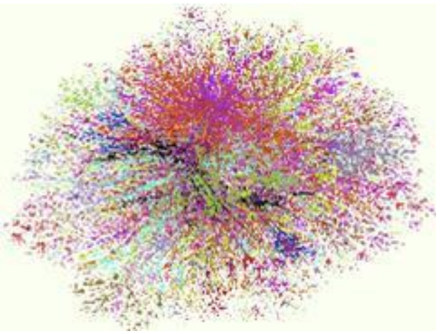
TOOLS

- Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

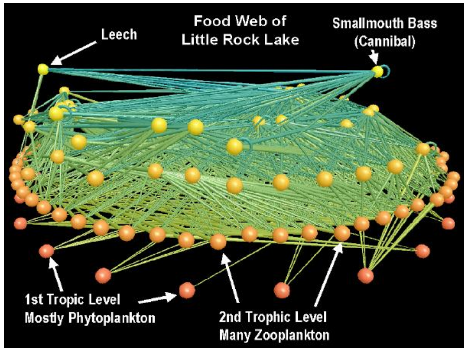
Problem#1: Joint work with

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

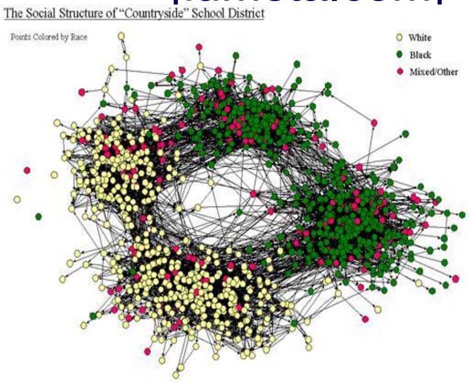
Graphs - why should we care?



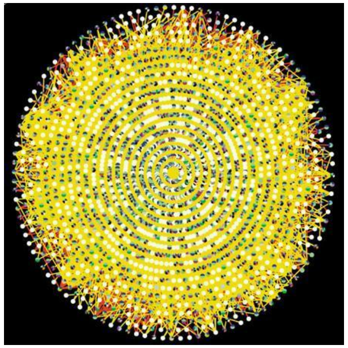
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



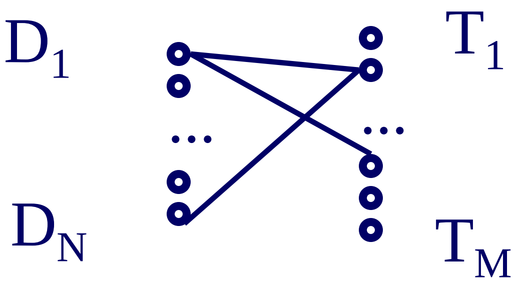
Friendship Network
[Moody '01]



Protein Interactions
[genomebiology.com]

Graphs - why should we care?

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



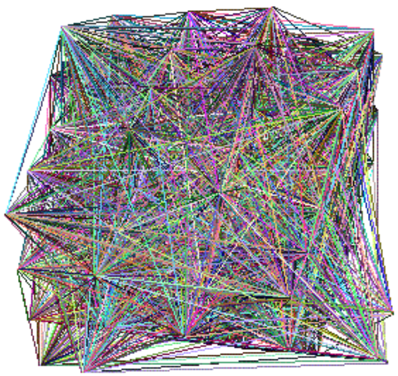
- web: hyper-text graph

- ... 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?

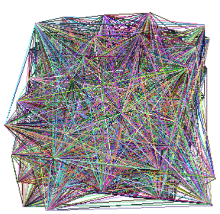
Laws and patterns

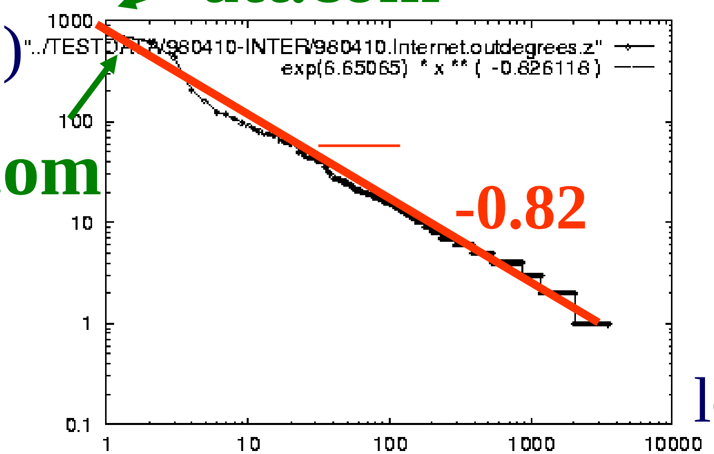
- Are real graphs random?
- A: NO!!
  - Diameter
  - in- and out- degree distributions
  - other (surprising) patterns

Solution#1

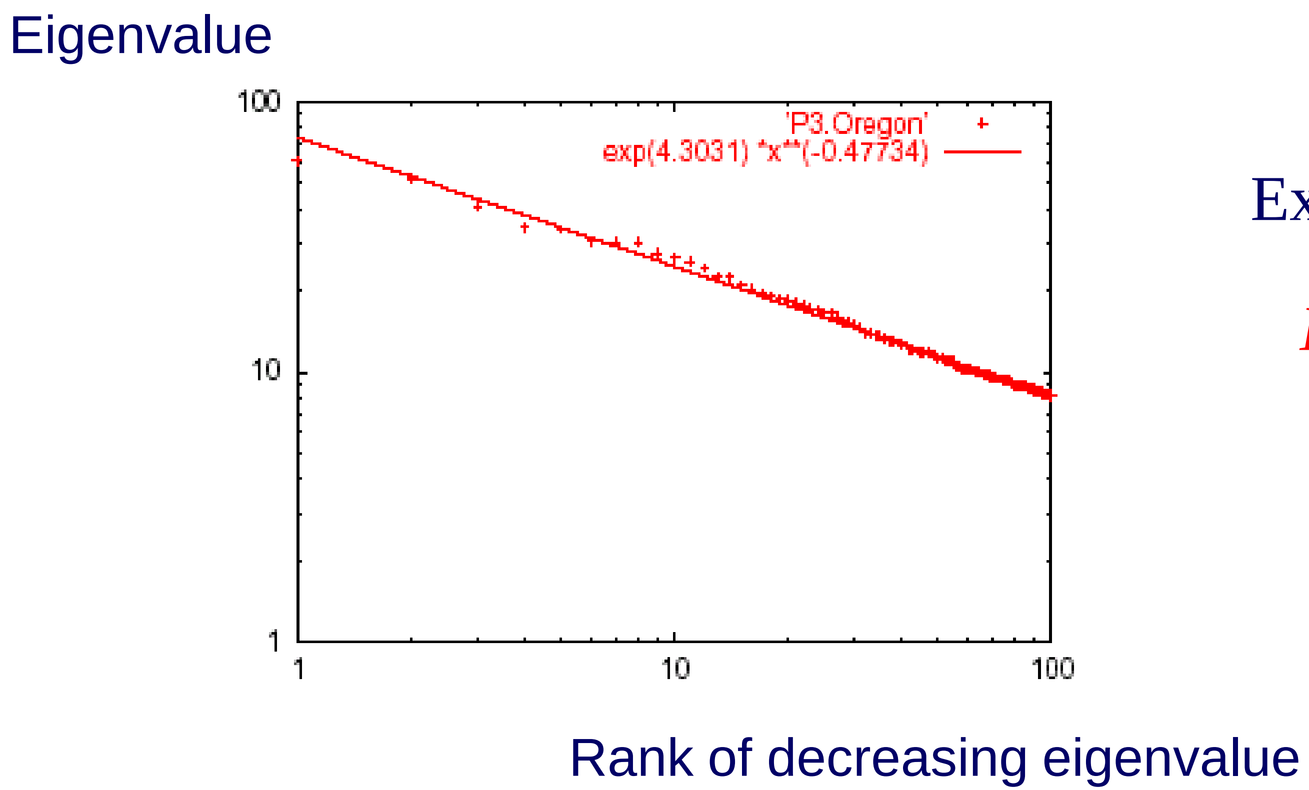
- Power law in the degree distribution [SIGCOMM99]

internet domains





Solution#1': Eigen Exponent *E*



Exponent = slope

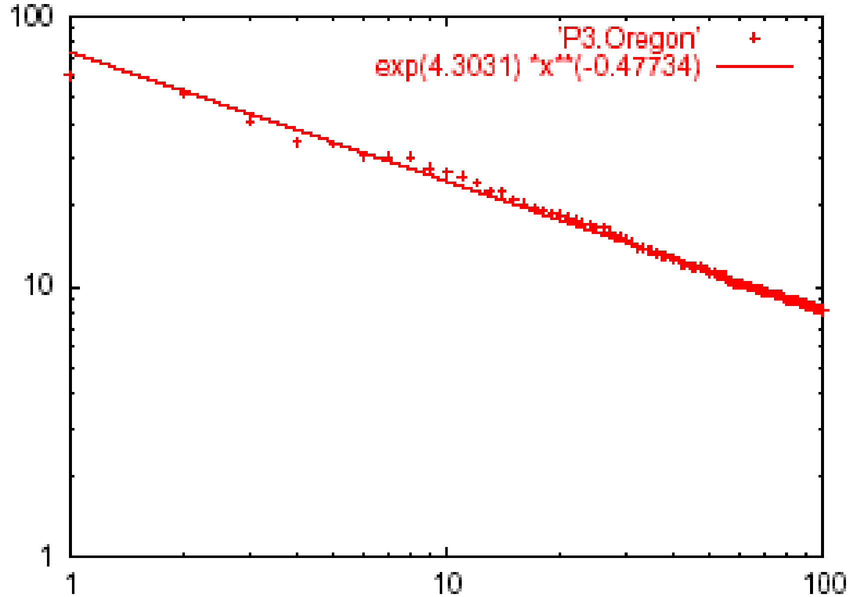
E = -0.48

May 2001

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

Solution#1’: Eigen Exponent *E*

Eigenvalue



Rank of decreasing eigenvalue

Exponent = slope

E = -0.48

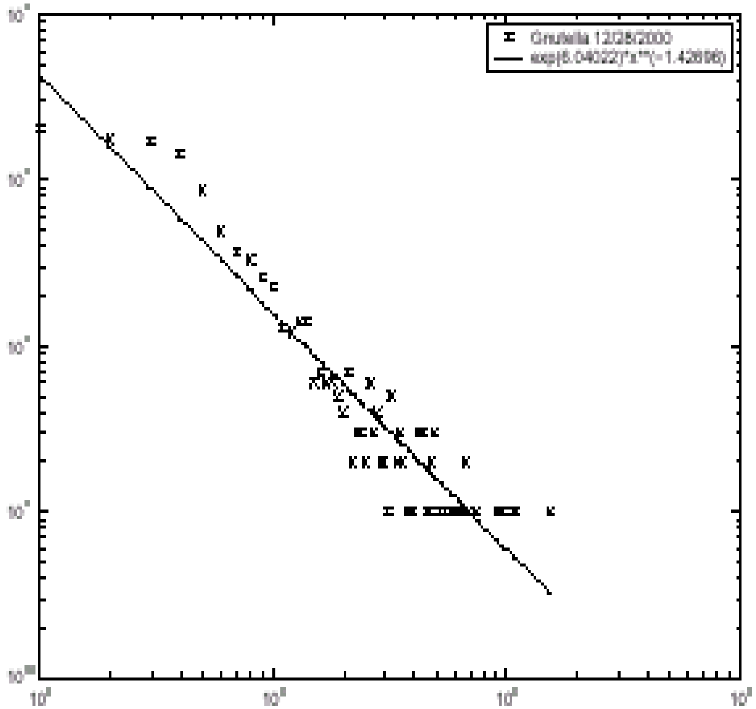
May 2001

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

But:

How about graphs from other domains?

The Peer-to-Peer Topology



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

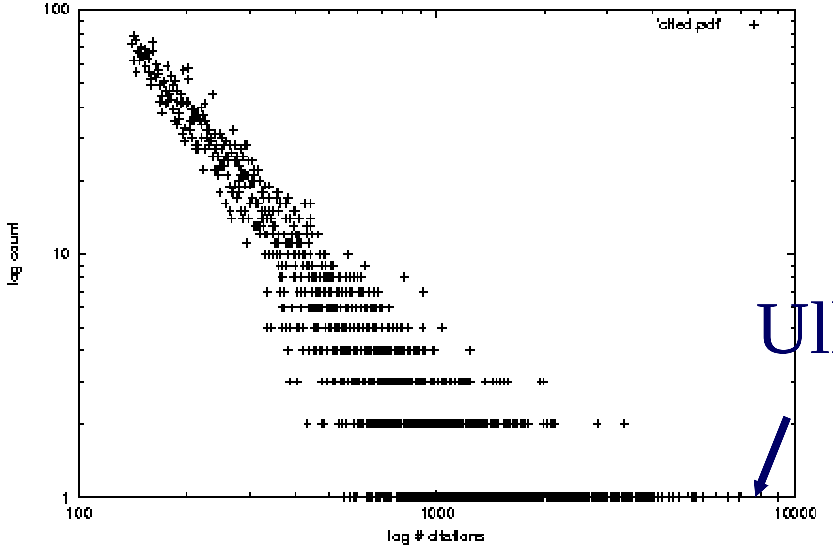
[Jovanovic+]

- Count versus degree
- Number of adjacent peers follows a power-law

More power laws:

citation counts: (*citeseer.nj.nec.com* 6/2001)

log(count)



Ullman

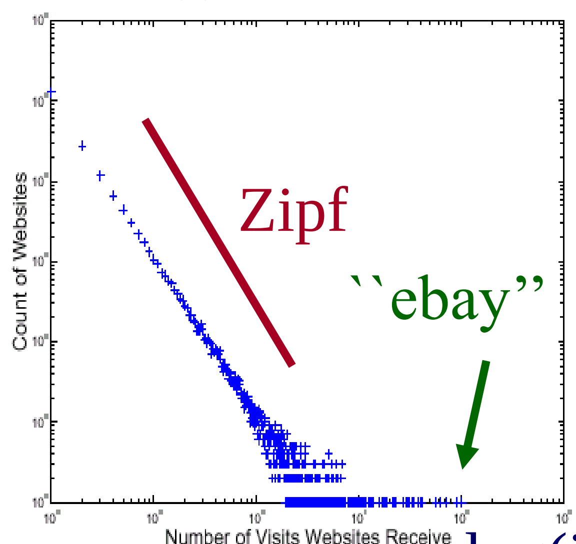
log(#citations)

More power laws:

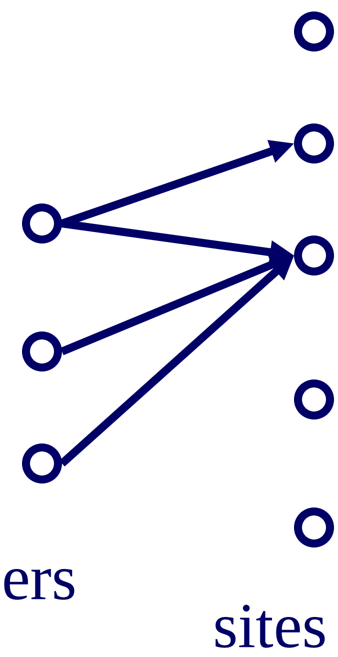
- web hit counts [w/ A. Montgomery]

Web Site Traffic

log(count)



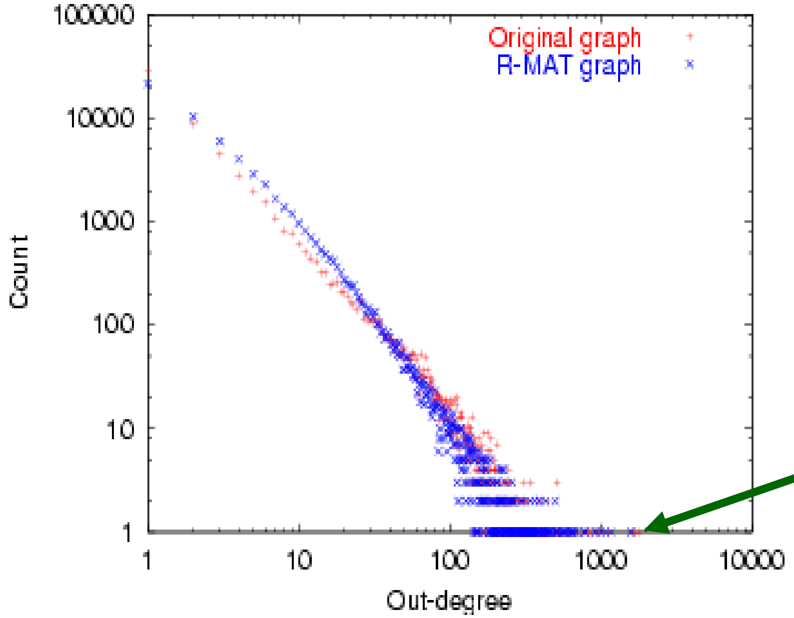
log(in-degree)



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

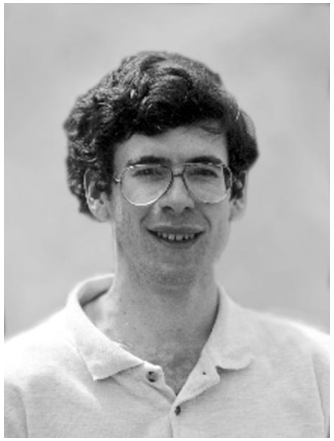
TOOLS

- Problem#4: Who is the ‘master-mind’?
- Problem#5: Track communities over time

Problem#2: Time evolution

- with Jure Leskovec (CMU/MLD)

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



Evolution of the Diameter

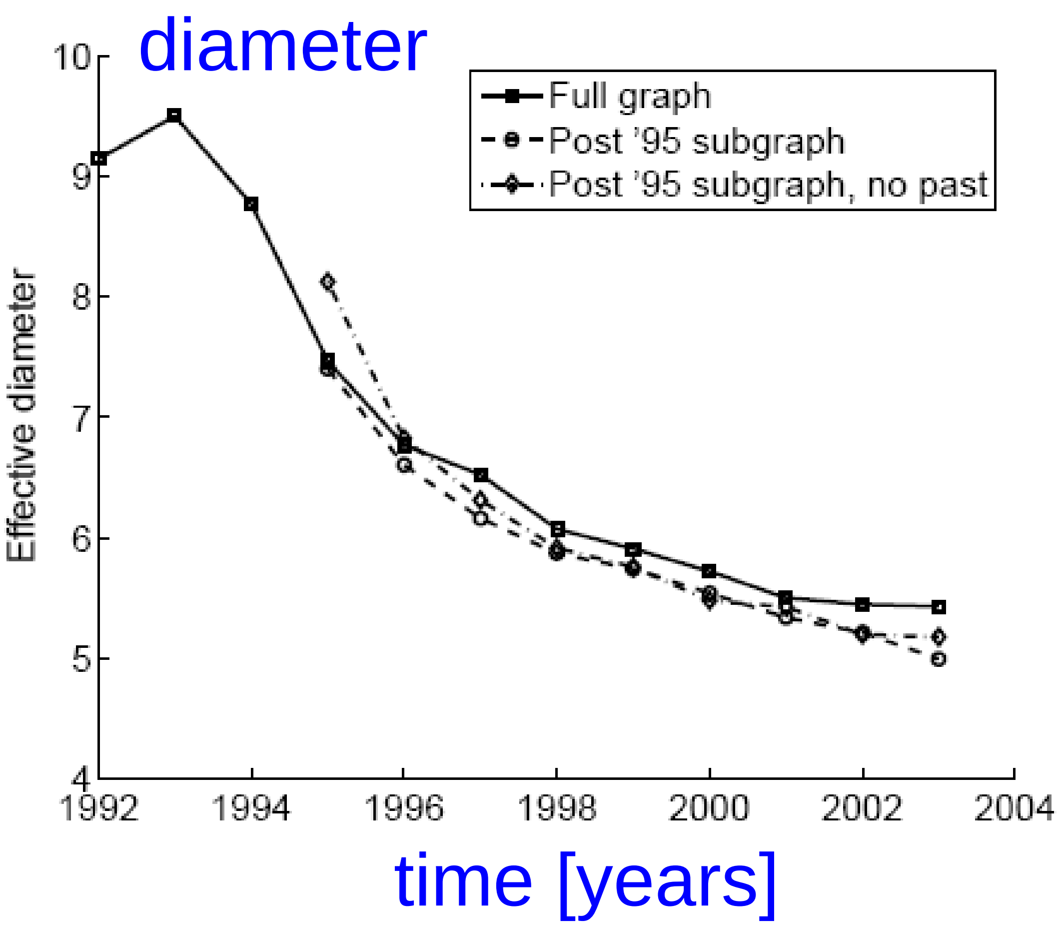
- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?

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

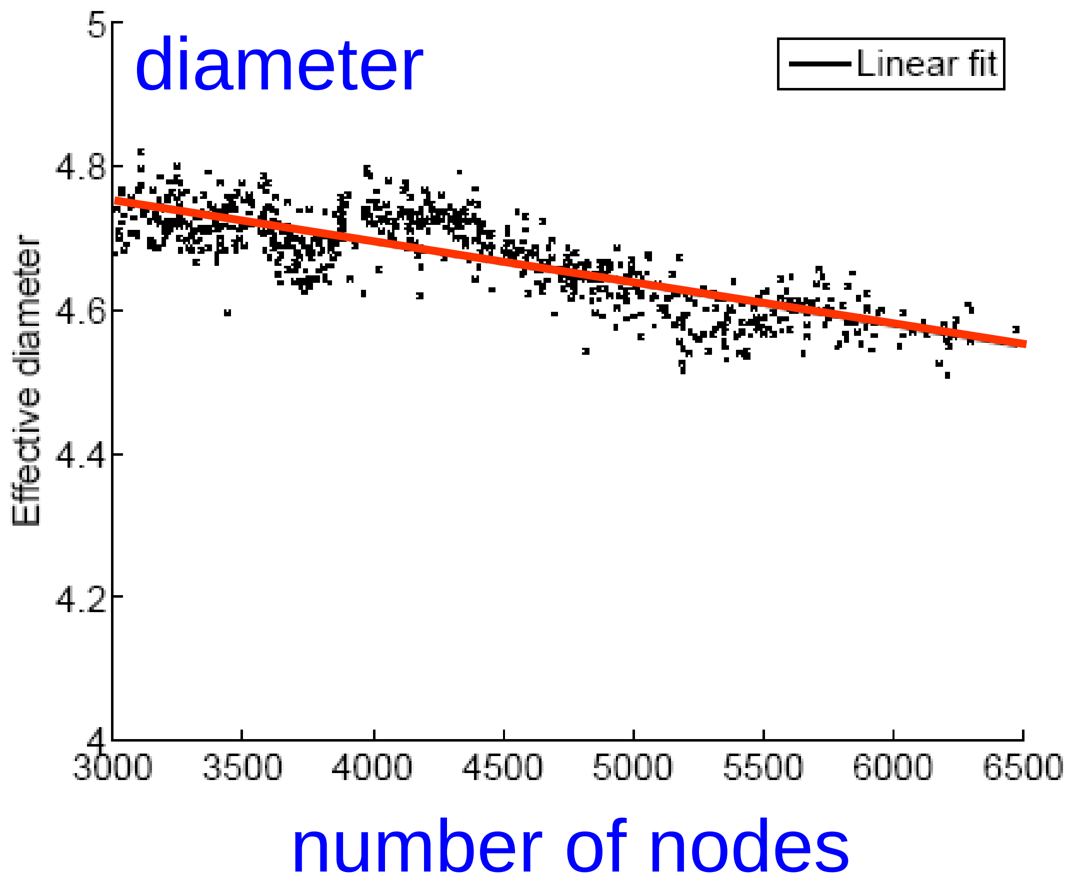
Diameter – ArXiv citation graph

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



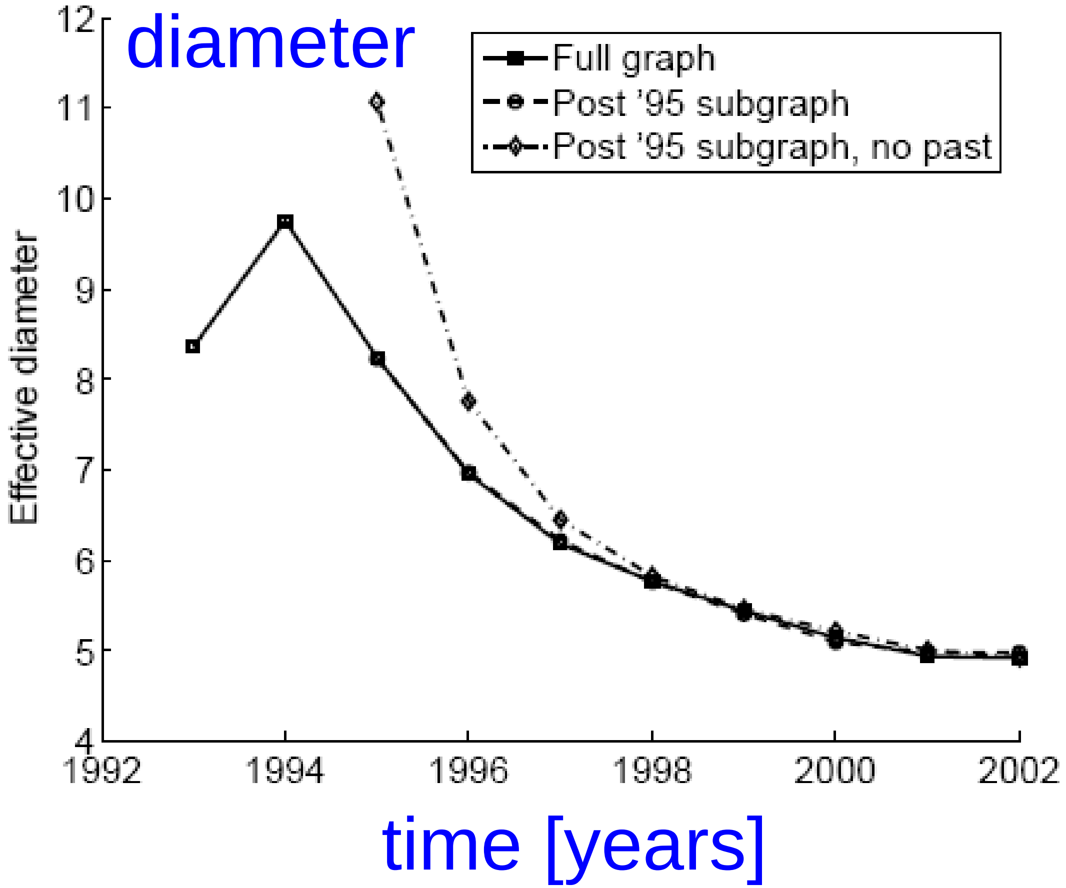
Diameter – "Autonomous Systems"

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



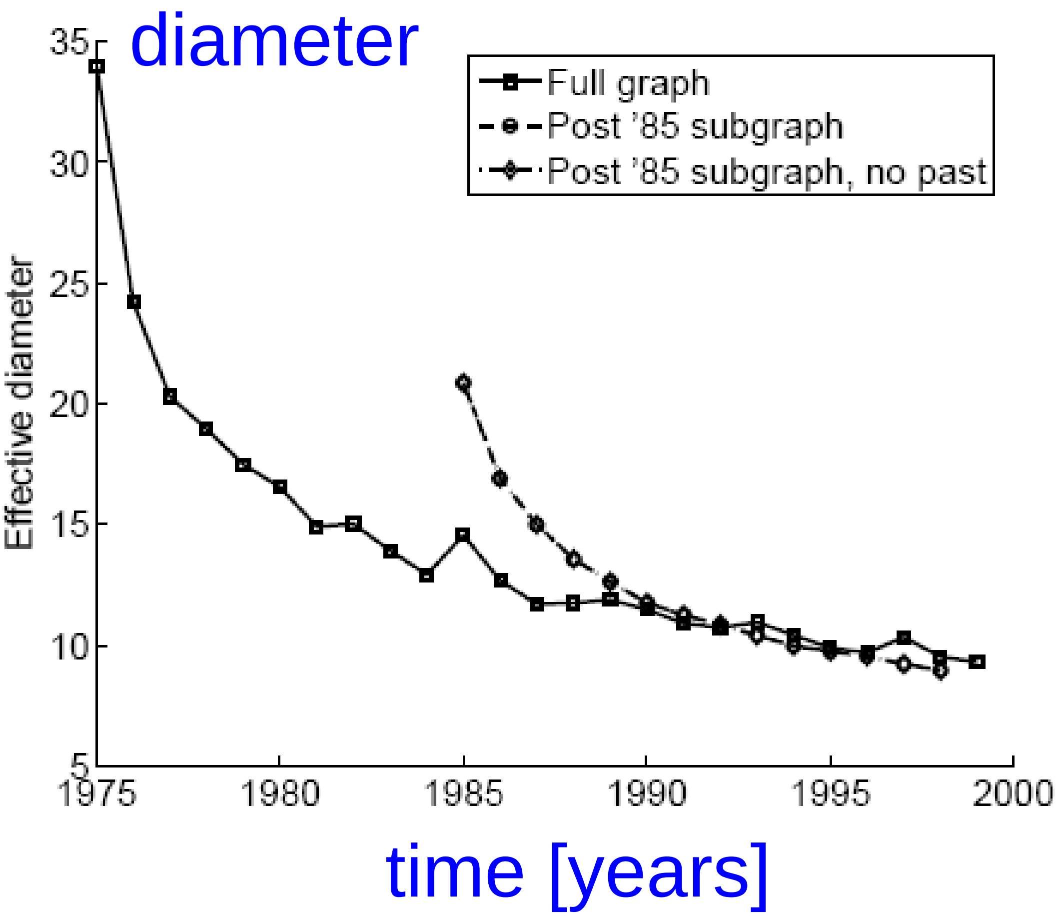
Diameter – "Affiliation Network"

- 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)

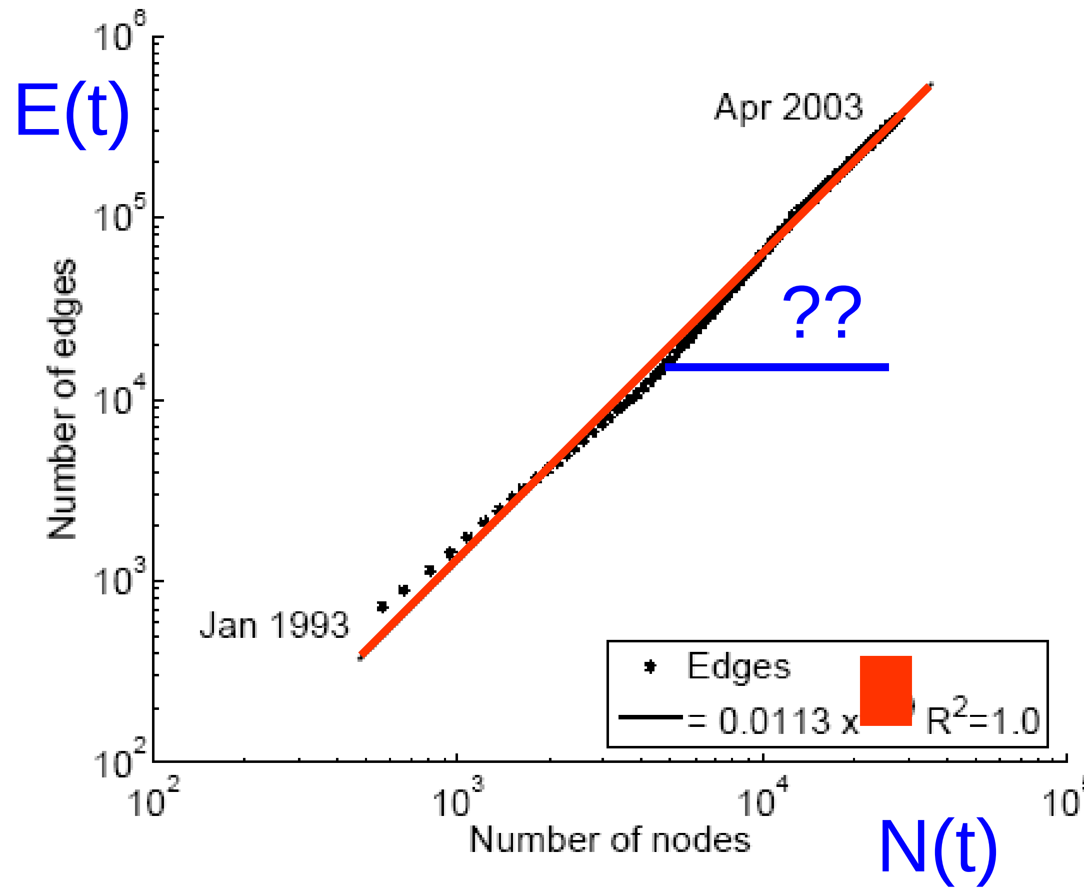
- Q: what is your guess for

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

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

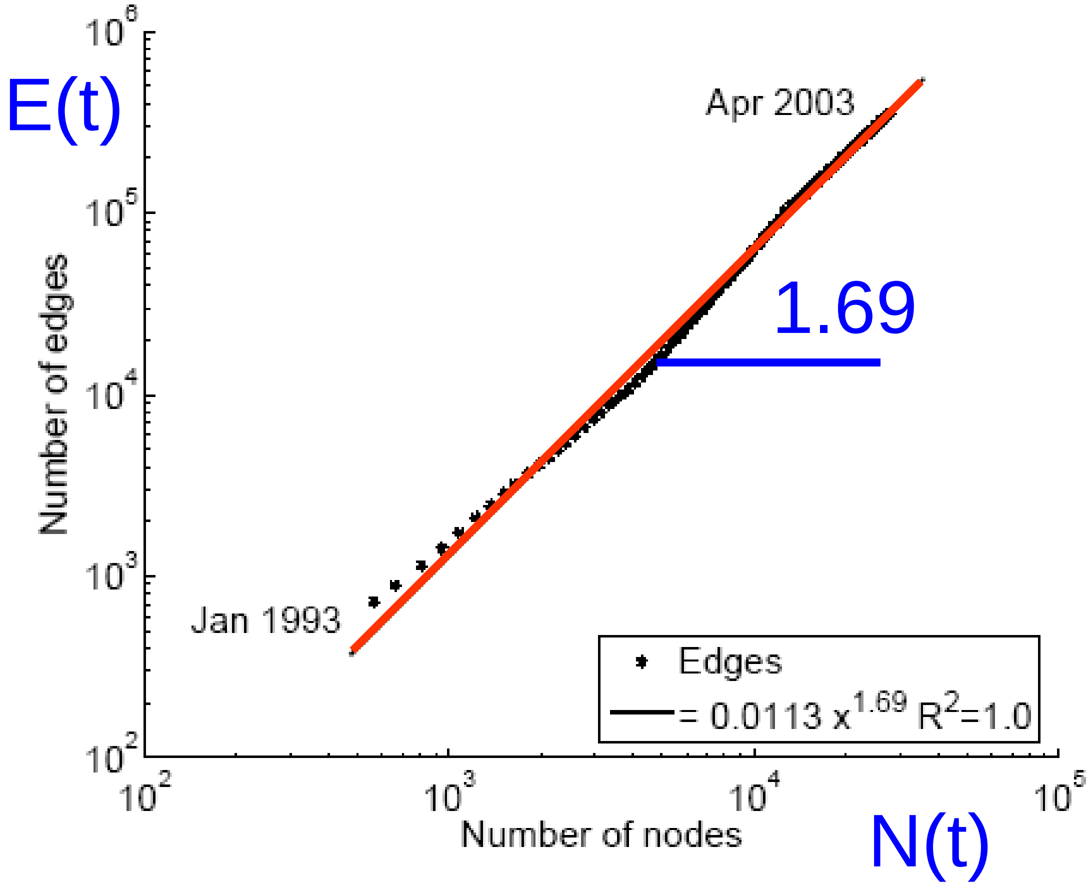
Densification – Physics Citations

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



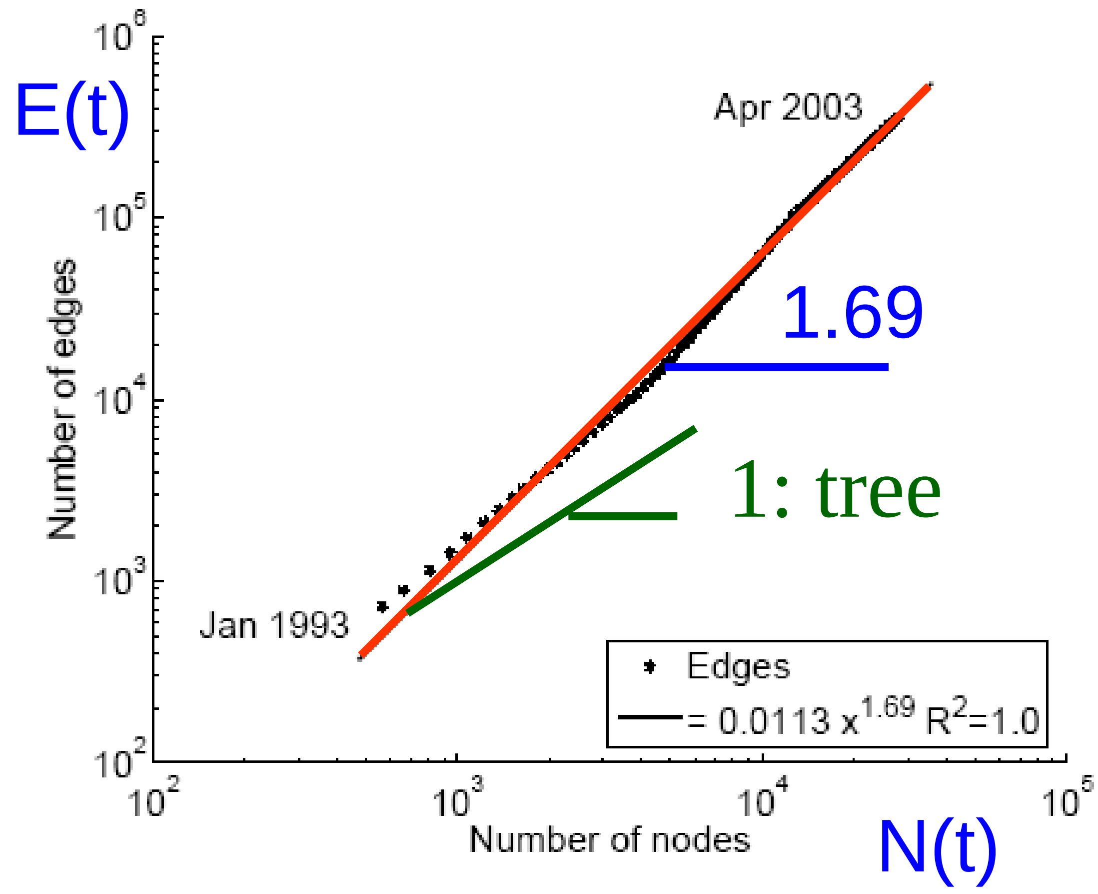
Densification – Physics Citations

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



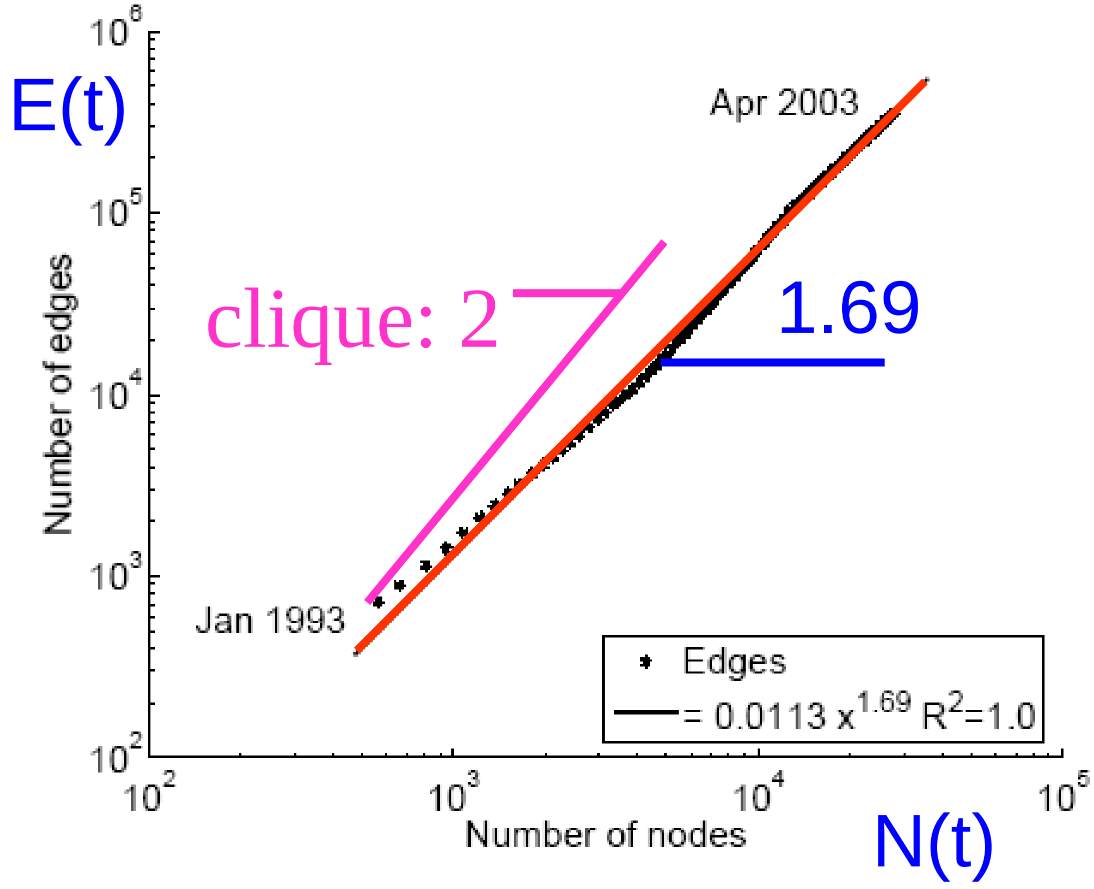
Densification – Physics Citations

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



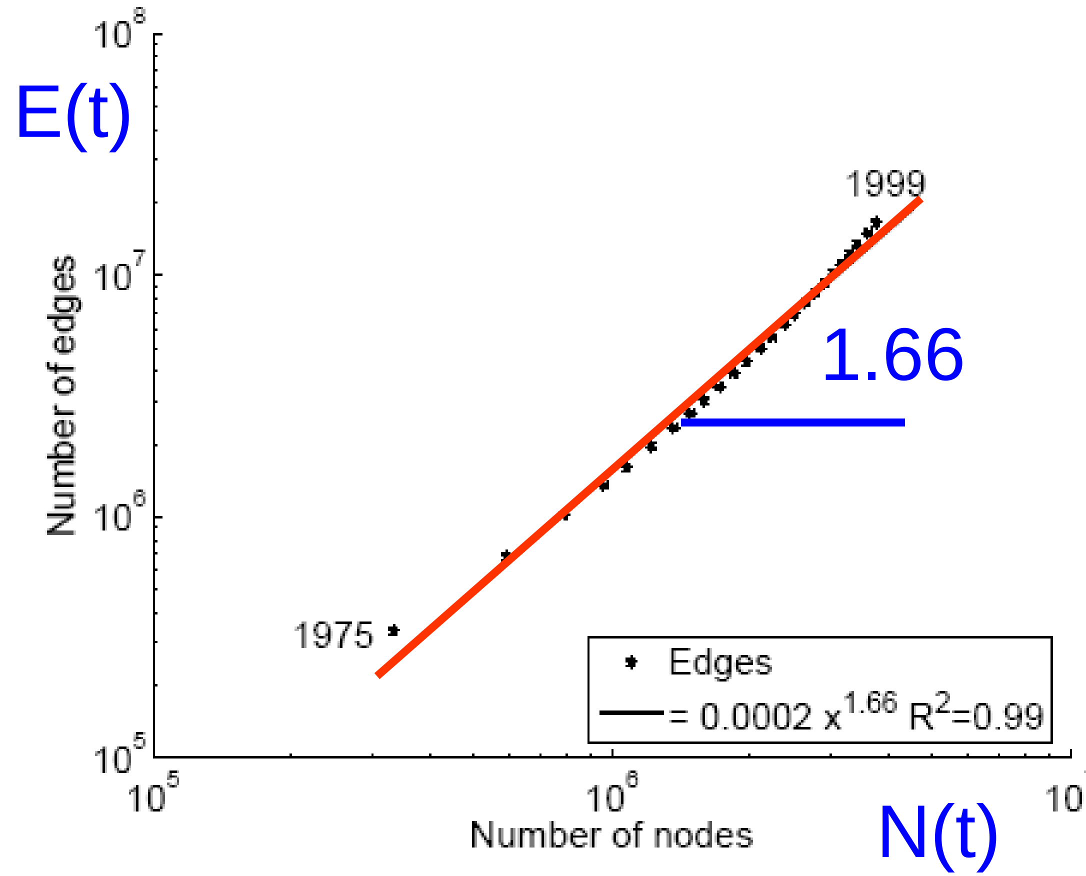
Densification – Physics Citations

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



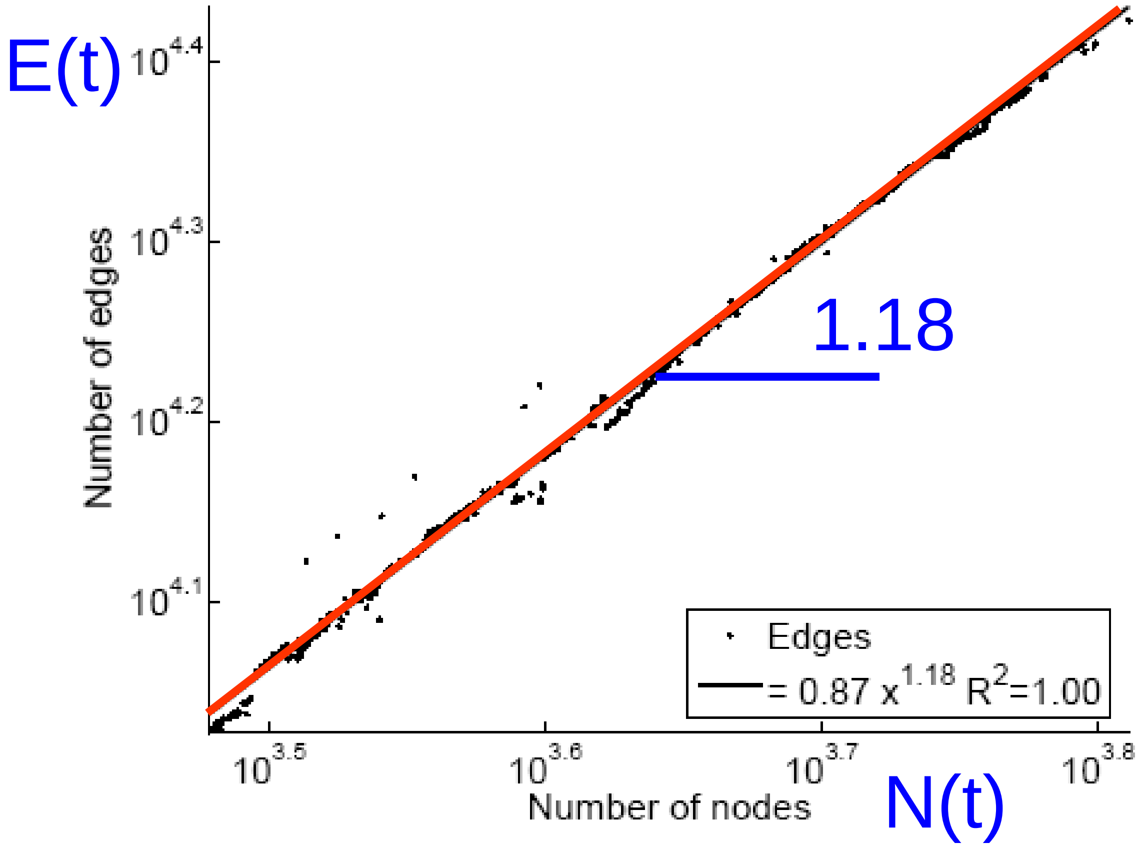
Densification – Patent Citations

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



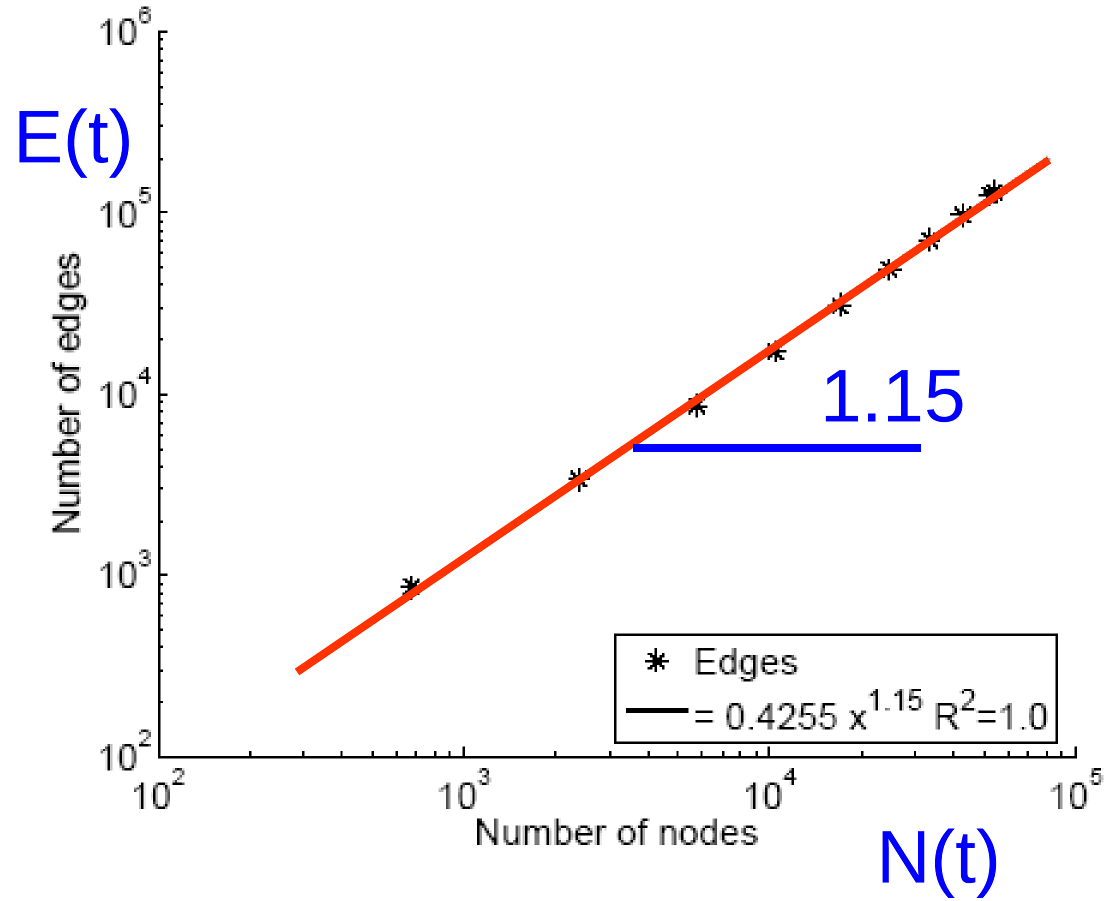
Densification – Autonomous Systems

- 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: Track communities over time

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

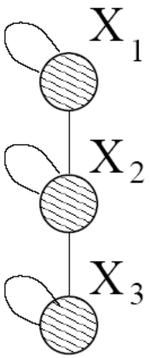
- Given a growing graph with count of nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - Power Law Degree Distribution
    - Power Law eigenvalue and eigenvector distribution
    - Small Diameter
  - Dynamic Patterns
    - Growth Power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

Problem Definition

- Given a growing graph with count of nodes $N_1, N_2, \ldots$
- 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



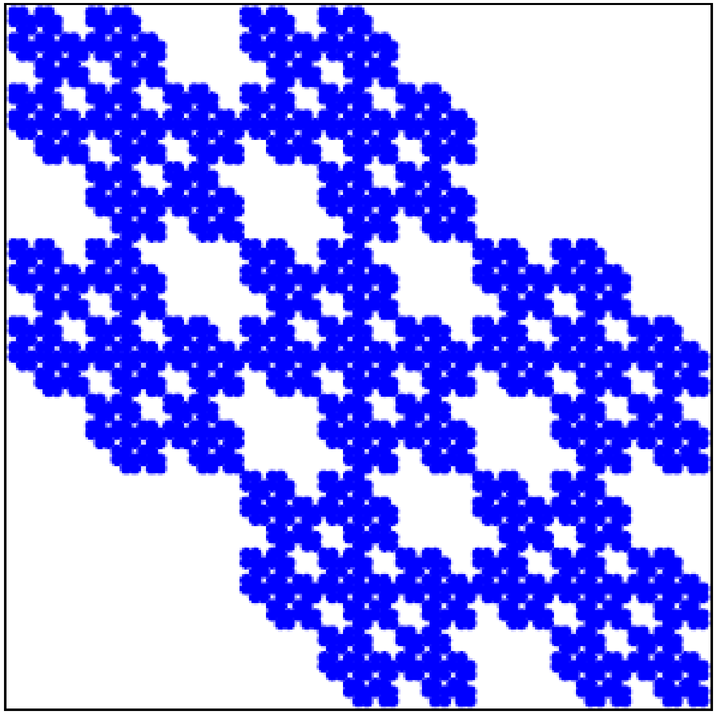
1	1	0
1	1	1
0	1	1

G_1

Adjacency matrix

Kronecker Product – a Graph

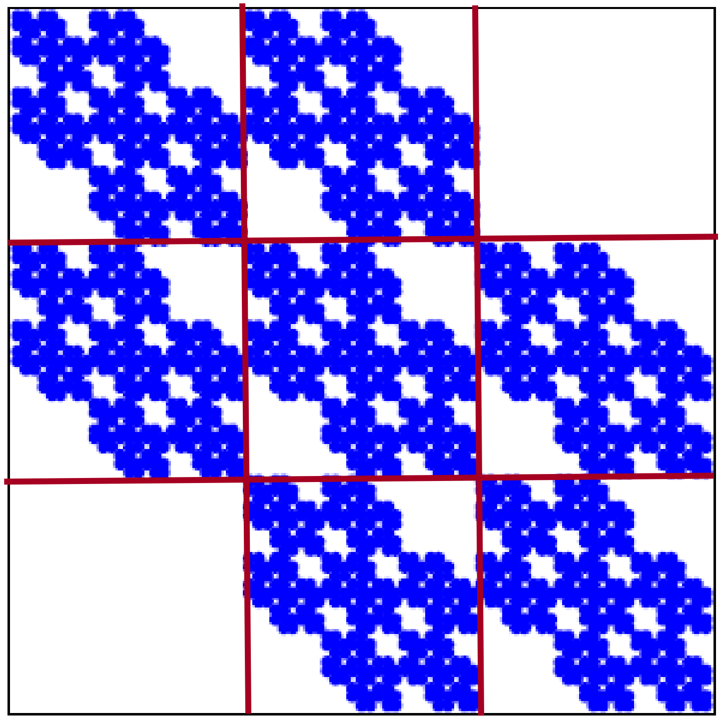
- Continuing multiplying with G_1 we obtain G_4 and so on …



G_4 adjacency matrix

Kronecker Product – a Graph

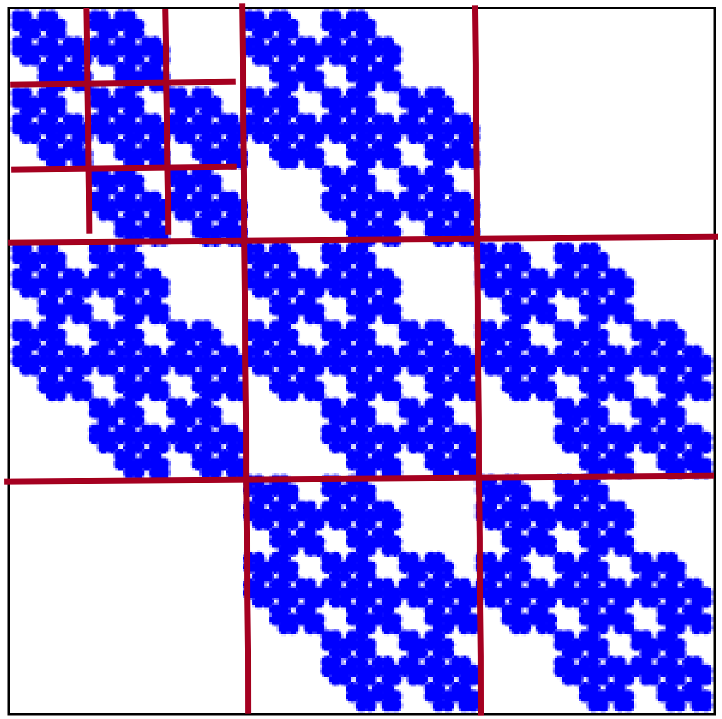
- Continuing multiplying with G_1 we obtain G_4 and so on …



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

skip

Stochastic Kronecker Graphs

- Create $N_1 \times N_1$ probability matrix P_1
- Compute the k^{th} Kronecker power P_k
- For each entry p_{uv} of P_k include an edge (u,v) with probability p_{uv}

0.4	0.2
0.1	0.3

P_1

Kronecker multiplication

0.16	0.08	0.08	0.04
0.04	0.12	0.02	0.06
0.04	0.02	0.12	0.06
0.01	0.03	0.03	0.09

P_k

Instance Matrix G_2

flip biased coins

Experiments

- How well can we match real graphs?
  - Arxiv: physics citations:
    - 30,000 papers, 350,000 citations
    - 10 years of data
  - U.S. Patent citation network
    - 4 million patents, 16 million citations
    - 37 years of data
  - Autonomous systems – graph of internet
    - Single snapshot from January 2002
    - 6,400 nodes, 26,000 edges
- We show both static and temporal patterns

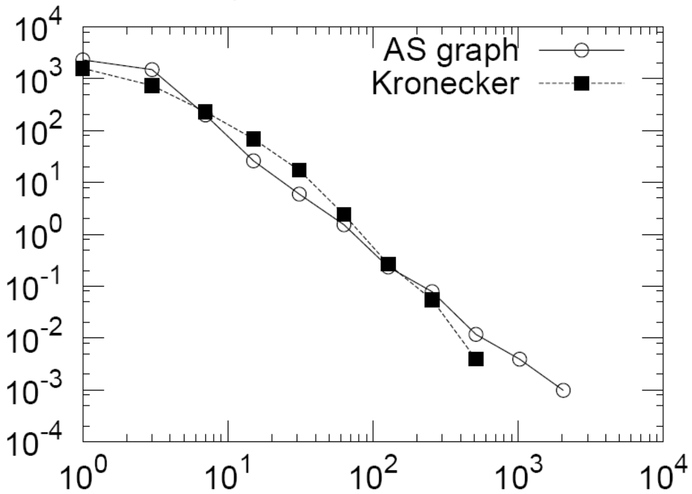
(Q: how to fit the parm's?)

A:

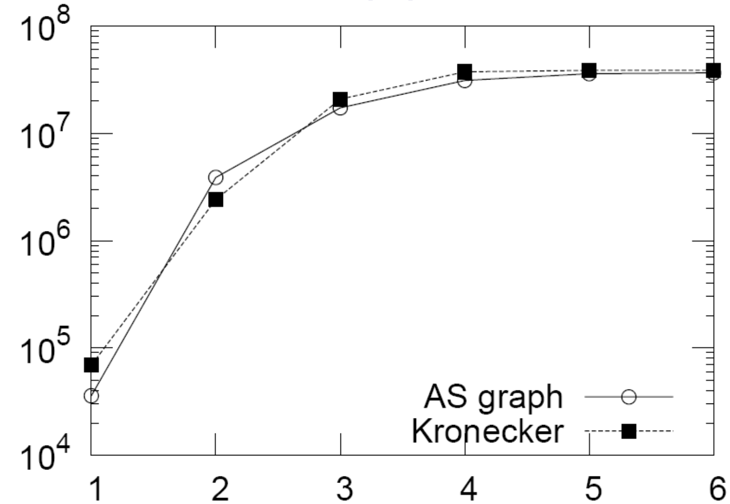
- Stochastic version of Kronecker graphs +
- Max likelihood +
- Metropolis sampling
- [Leskovec+, ICML'07]

Experiments on real AS graph

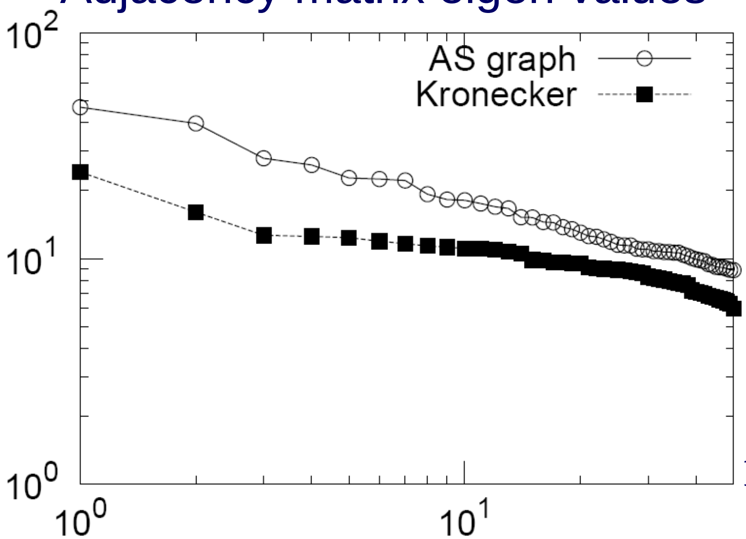
Degree distribution



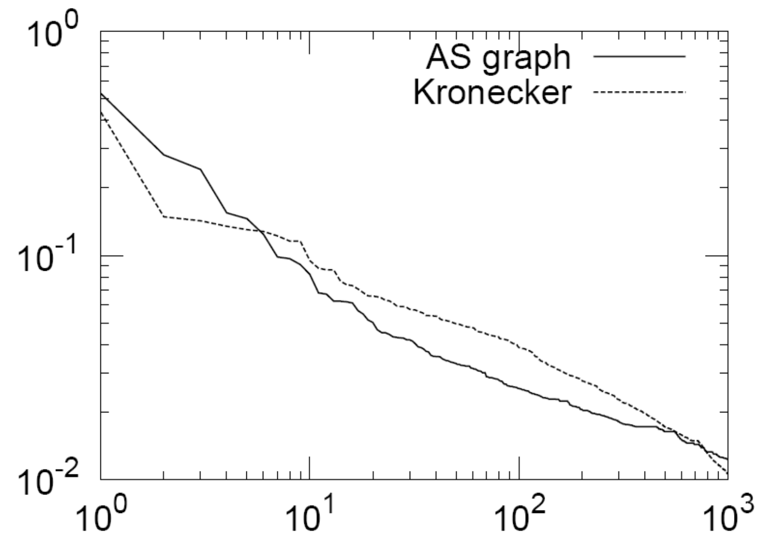
Hop plot



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)

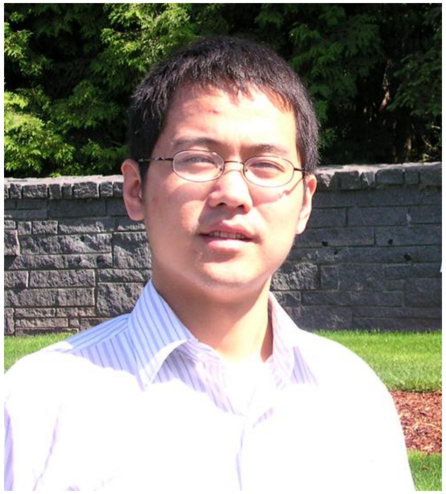
- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- ✓ Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the ‘master-mind’?
- Problem#5: Track communities over time

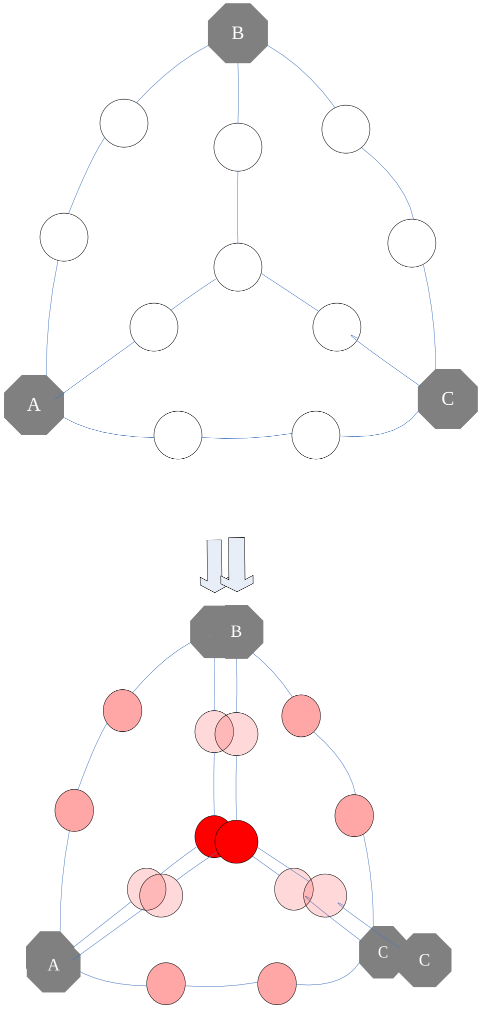
Problem#4: MasterMind – 'CePS'

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



Center-Piece Subgraph(Ceps)

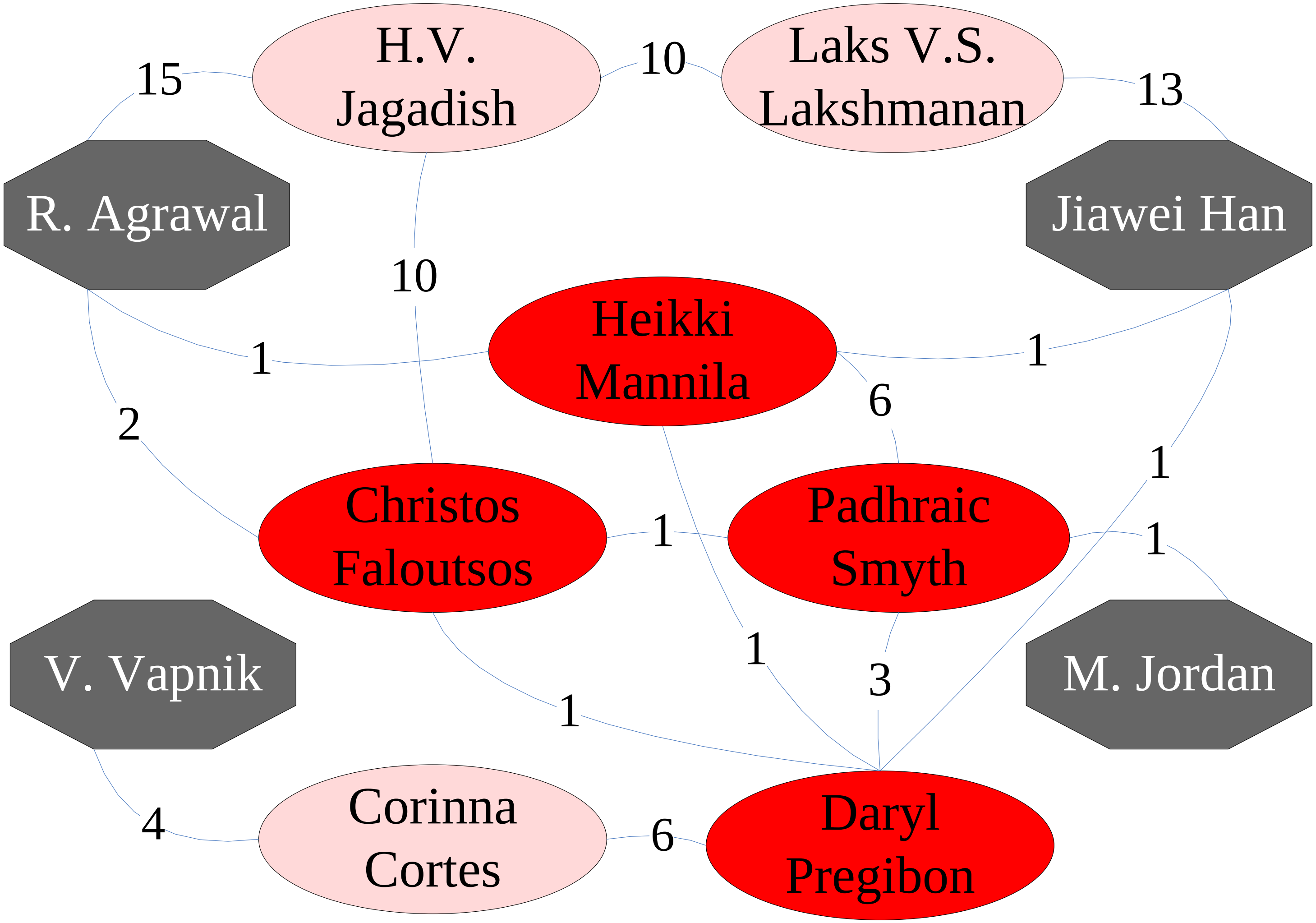
- **Given** Q query nodes
- **Find** Center-piece ($\leq b$)



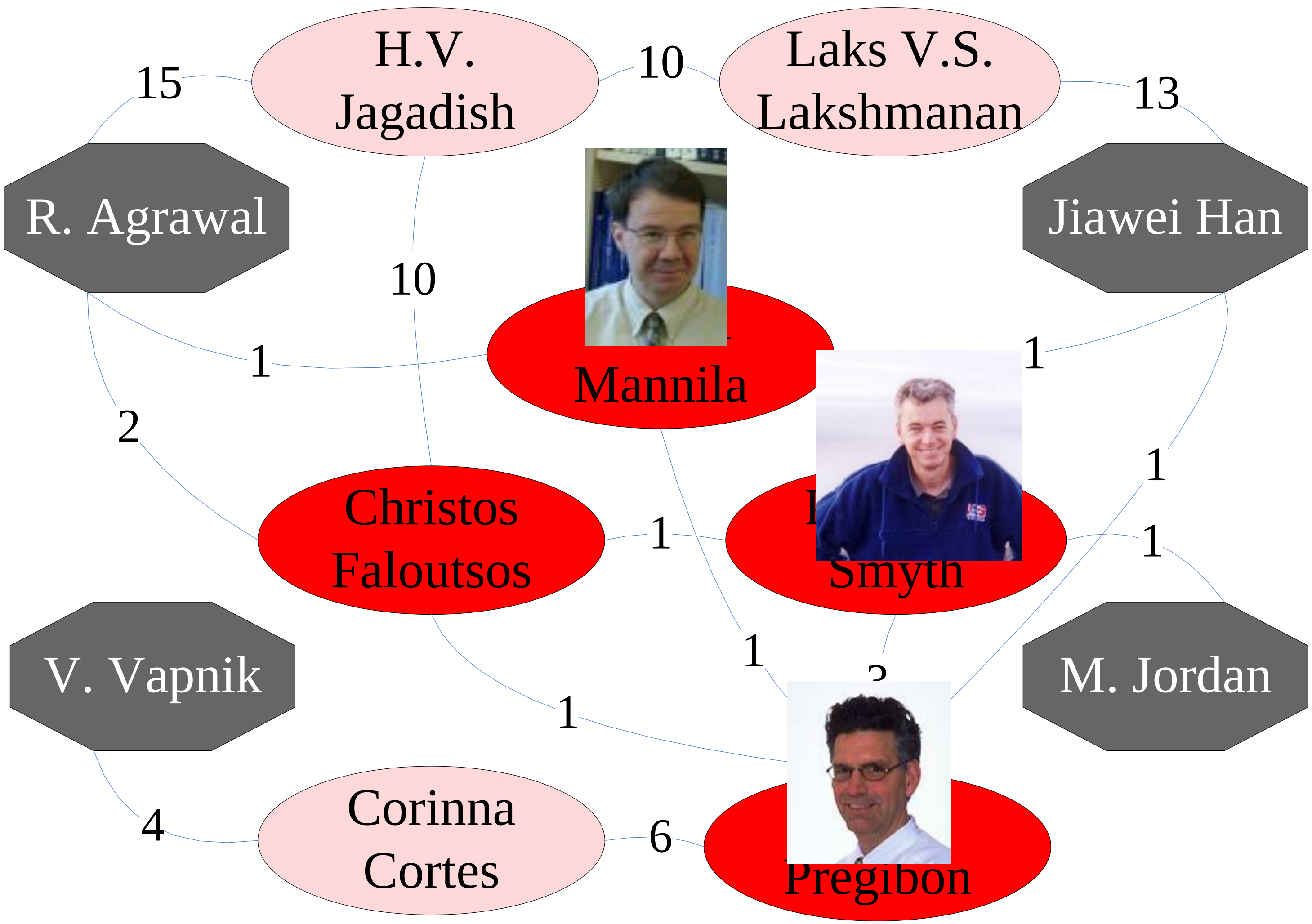
- App.
  - Social Networks
  - Law Inforcement, …
- Idea:
  - Proximity -> random walk with restarts

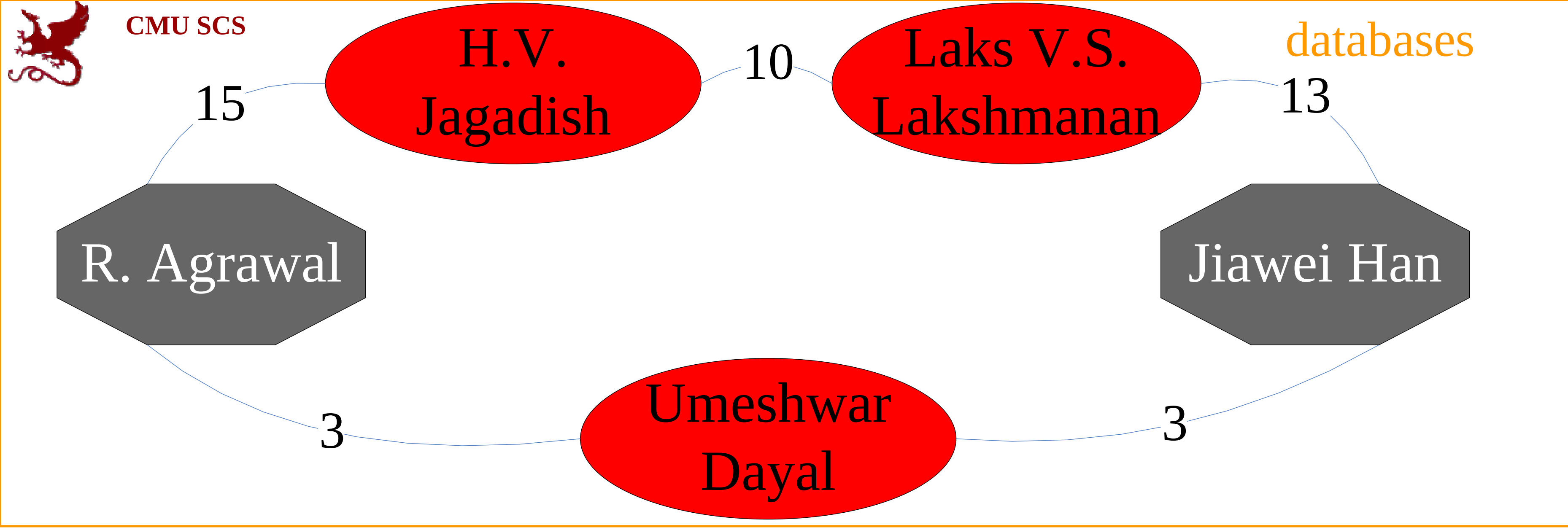
Case Study: AND query

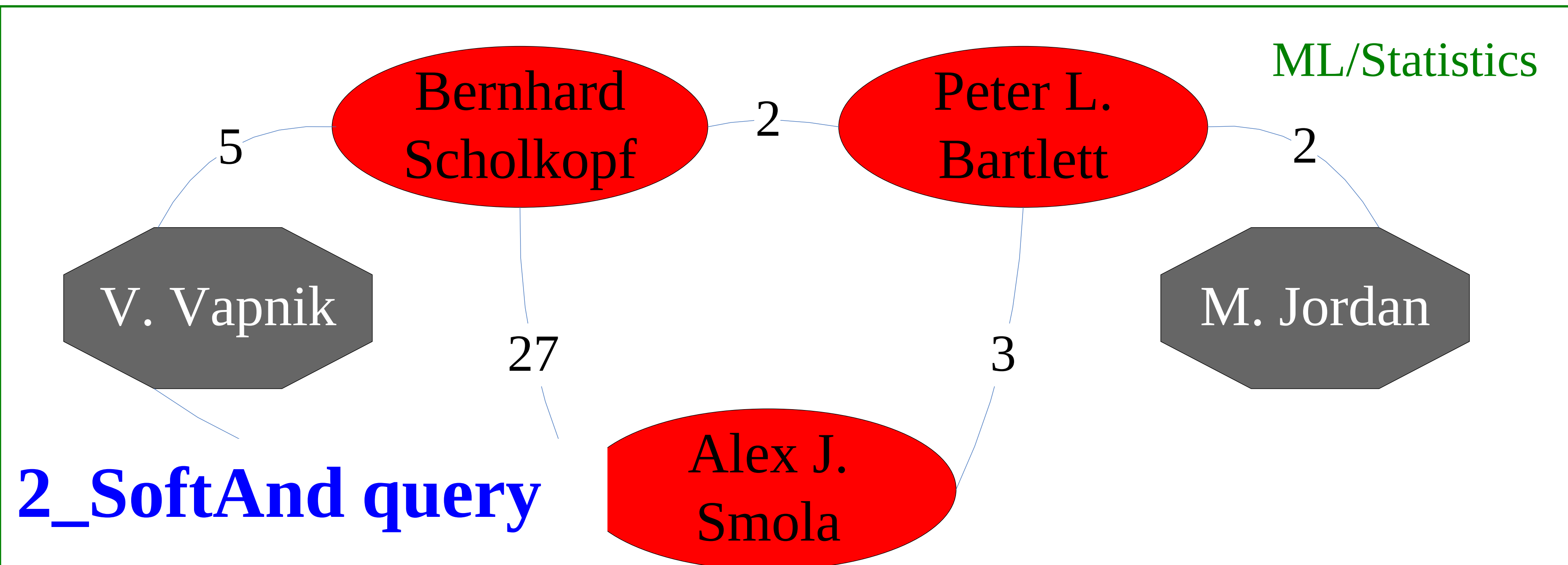
Case Study: AND query



Case Study: AND query

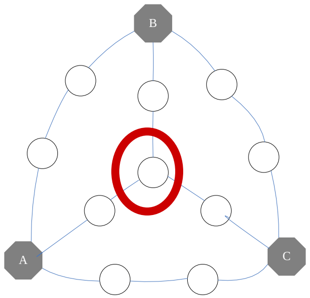






2_SoftAnd query

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, b.p. award)

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; **Tensors**
- Other projects (Virus propagation, e-bay 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

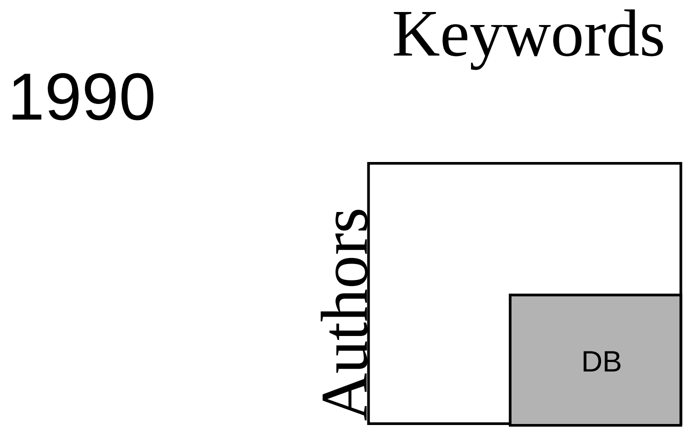
- ✓ Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

Tensors for time evolving graphs

- [Jimeng Sun+ KDD’06]
- [“ , SDM’07]
- [CF, Kolda, Sun, SDM’07 tutorial]

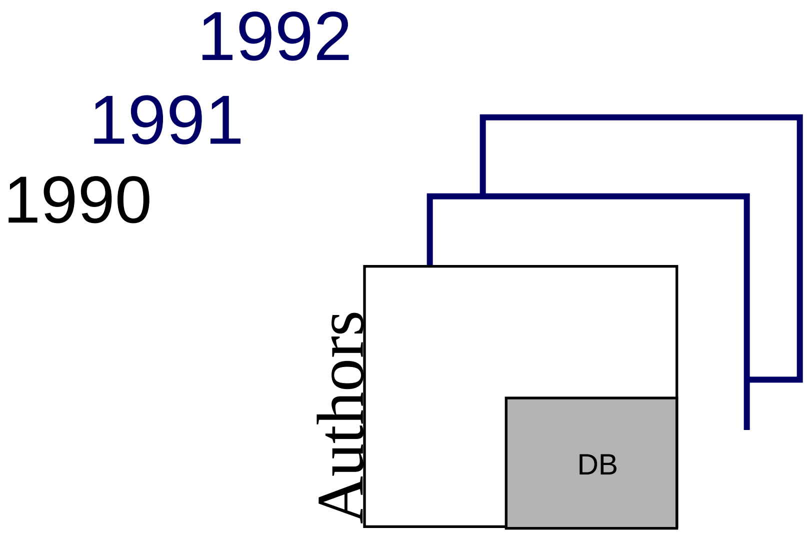
Social network analysis

- Static: find community structures



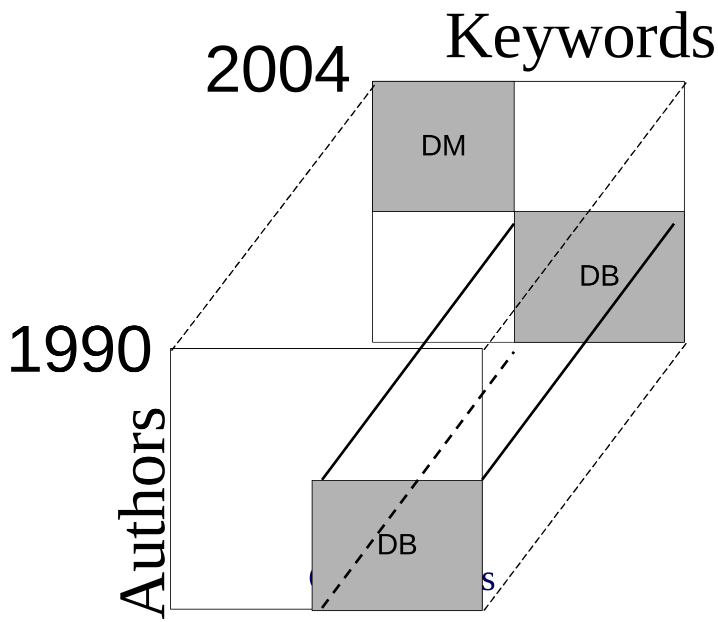
Social network analysis

- Static: find community structures

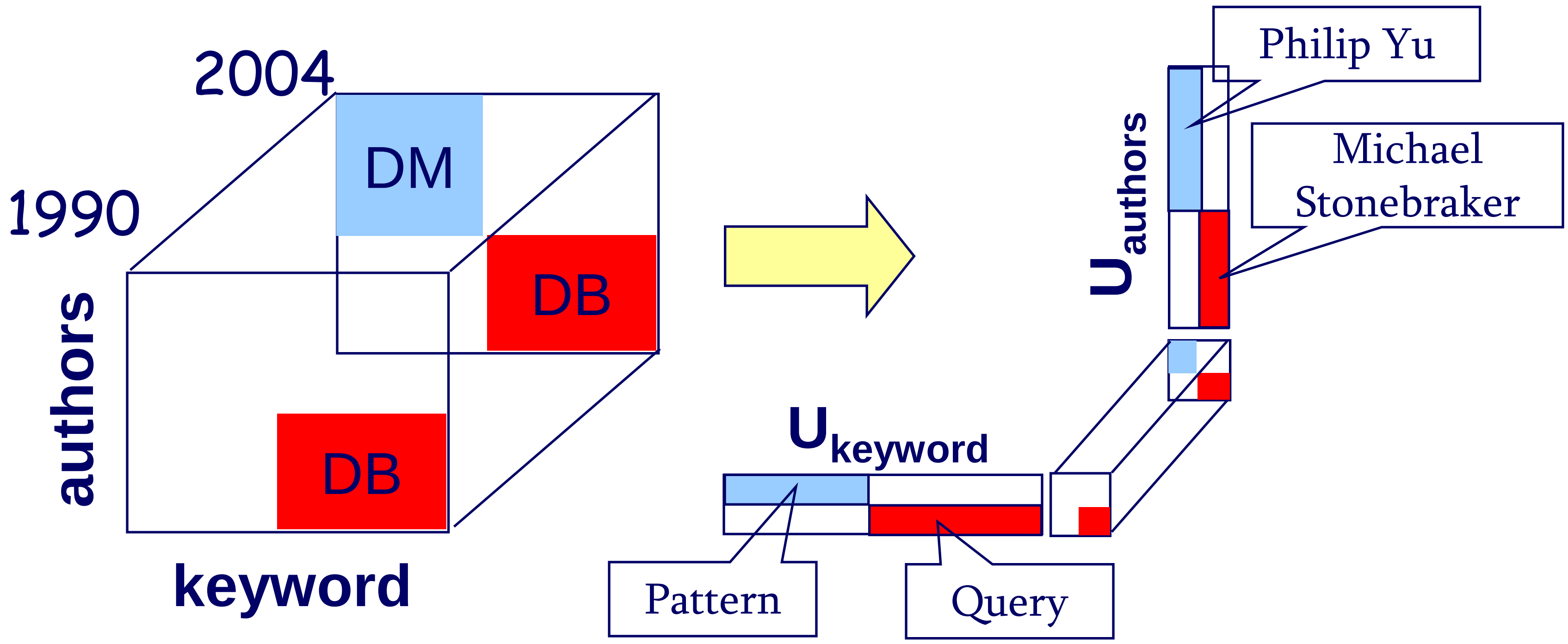


Social network analysis

- Static: find community structures
- Dynamic: monitor community structure evolution; spot abnormal individuals; abnormal time-stamps



Application 1: Multiway latent semantic indexing (LSI)



- Projection matrices specify the clusters
- Core tensors give cluster activation level

Bibliographic data (DBLP)

- Papers from VLDB and KDD conferences
- Construct 2nd order tensors with yearly windows with <author, keywords>
- Each tensor: 4584×3741
- 11 timestamps (years)

Multiway LSI

Authors	Keywords	Year
michael carey, michael stonebraker, h. jagadish, hector garcia-molina	queri,parallel,optimization,concurr, objectorient	1995
surajit chaudhuri,mitch[illegible] cherniack,michael stonebraker,ugur etintemel	[illegible]tribut,systems,view,storage,servic,process,cache	2004
jiawei han,jian pei,philip s. yu, jianyong wang,charu c. aggarw[illegible]	streams,pattern,support, cluster, [illegible] gener,queri	2004

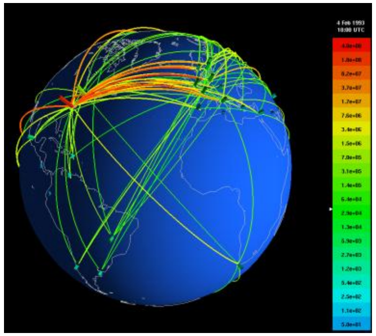
DB

DM

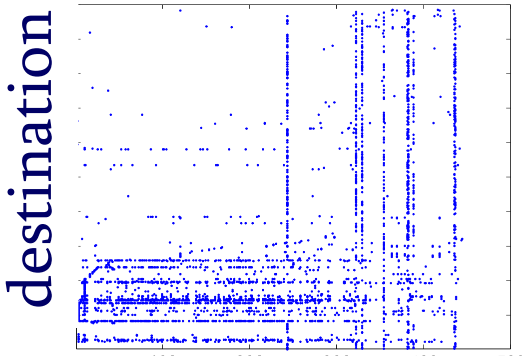
- Two groups are correctly identified: Databases and Data mining
- People and concepts are drifting over time

Network forensics

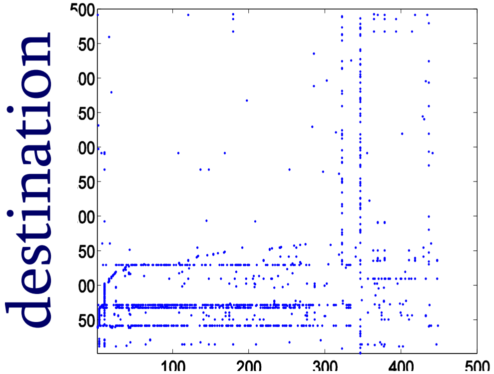
- Directional network flows
- A large ISP with 100 POPs, each POP 10Gbps link capacity [Hotnets2004]
  - 450 GB/hour with compression
- Task: Identify abnormal traffic pattern and find out the cause



abnormal traffic



normal traffic



(with Prof. Hui Zhang and Dr. Yinglian Xie)

Conclusions

Tensor-based methods (WTA/DTA/STA):

- spot patterns and anomalies on time evolving graphs, and
- on streams (monitoring)

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: Track communities over time

Outline

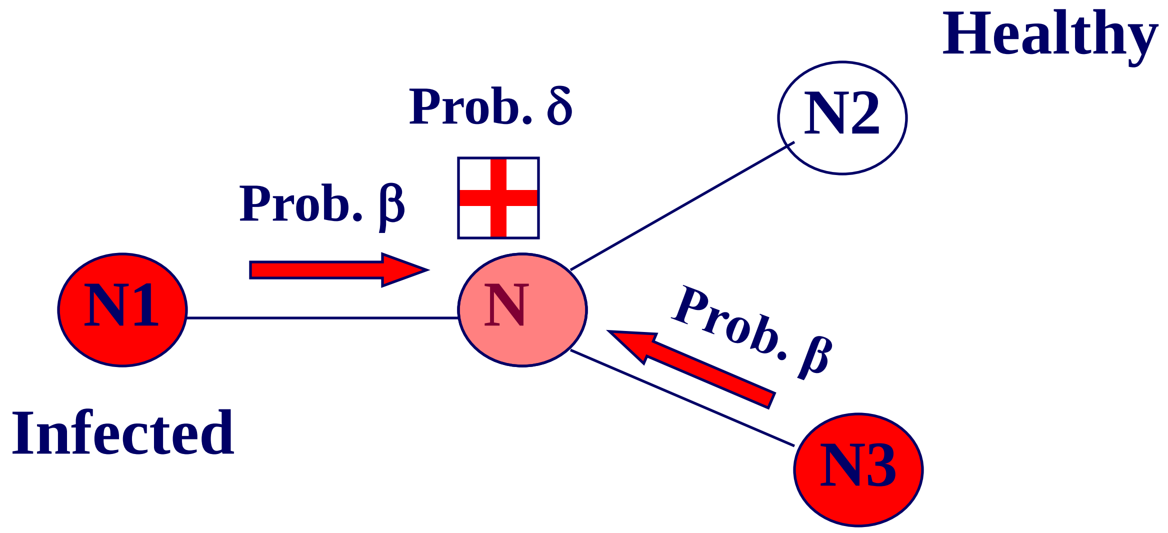
- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (**Virus propagation**, e-bay fraud detection, blogs)
- Conclusions

Virus propagation

- How do viruses/rumors propagate?
- Blog influence?
- Will a flu-like virus linger, or will it become extinct soon?

The model: SIS

- 'Flu' like: Susceptible-Infected-Susceptible
- Virus 'strength' s= β/δ



Epidemic threshold τ

of a graph: the value of τ, such that

if strength $s = \beta / \delta < \tau$

an epidemic can not happen

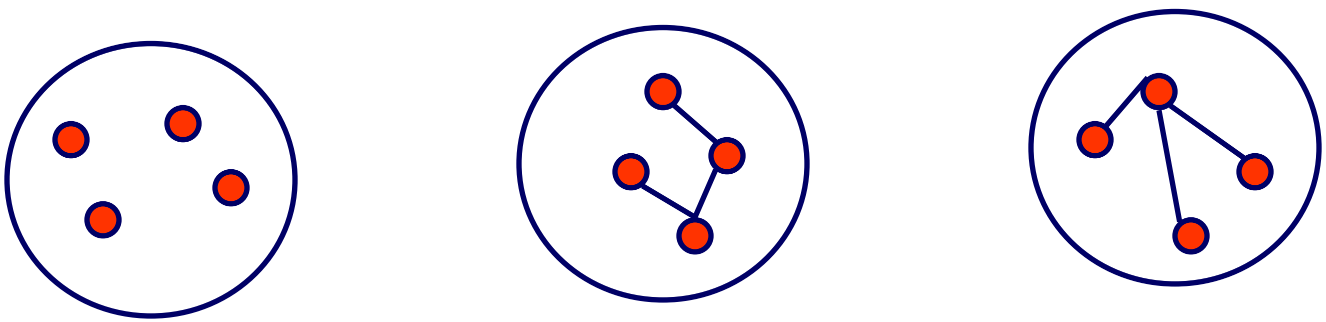
Thus,

- given a graph
- compute its epidemic threshold

Epidemic threshold τ

What should τ depend on?

- avg. degree? and/or highest degree?
- and/or variance of degree?
- and/or third moment of degree?
- and/or diameter?



Epidemic threshold

- [Theorem] We have no epidemic, if

$$\boxed{\beta/\delta < \tau = 1/\,\lambda_{1,A}}$$

Epidemic threshold

- [Theorem] We have no epidemic, if

recovery prob.

epidemic threshold

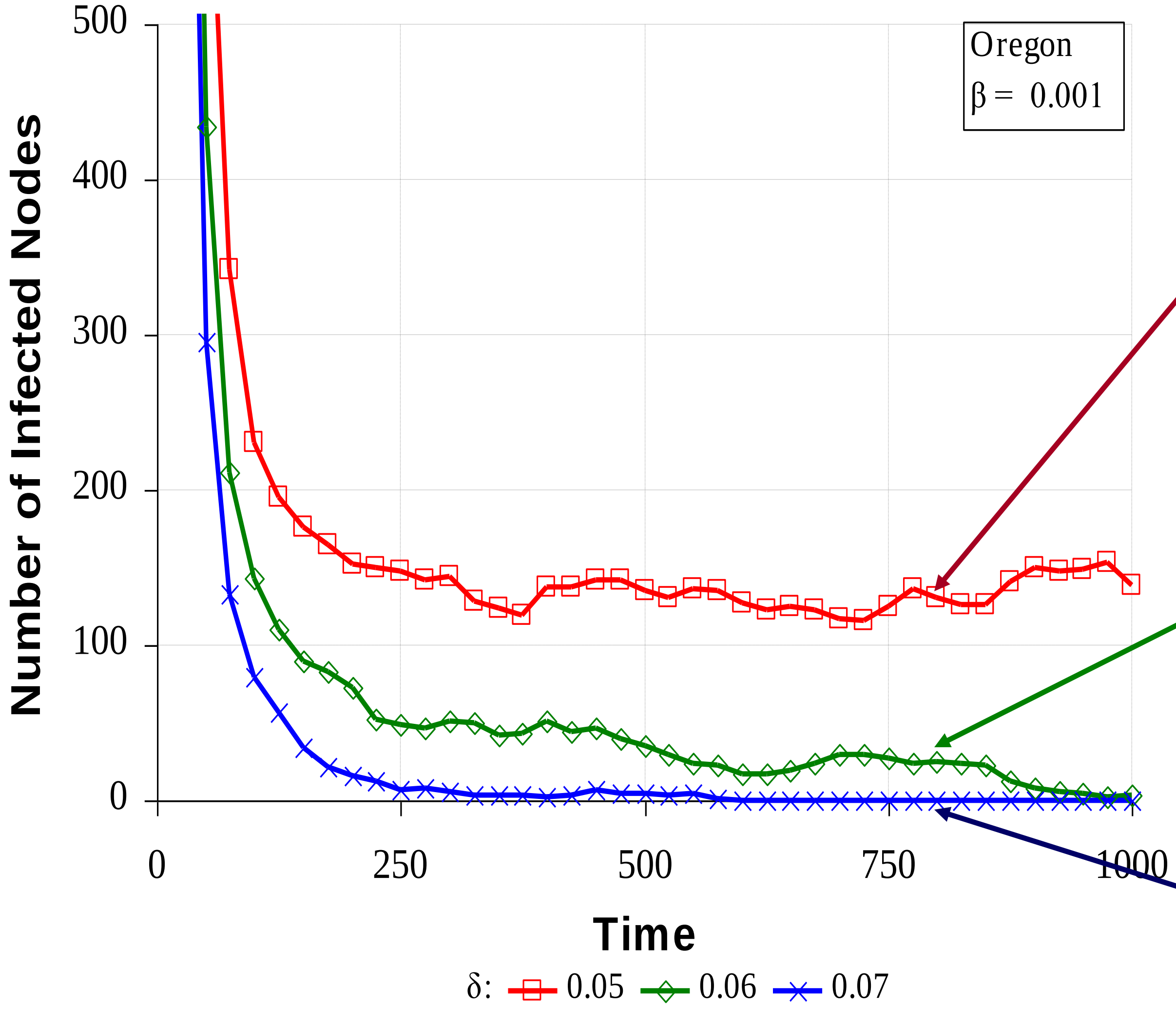
$$\boxed{\beta/\delta < \tau = 1/\lambda_{1,\boldsymbol{A}}}$$

attack prob.

largest eigenvalue of adj. matrix ***A***

Proof: [Wang+03]

Experiments (Oregon)



$\beta/\delta > \tau$ (above threshold)

$\beta/\delta = \tau$ (at the threshold)

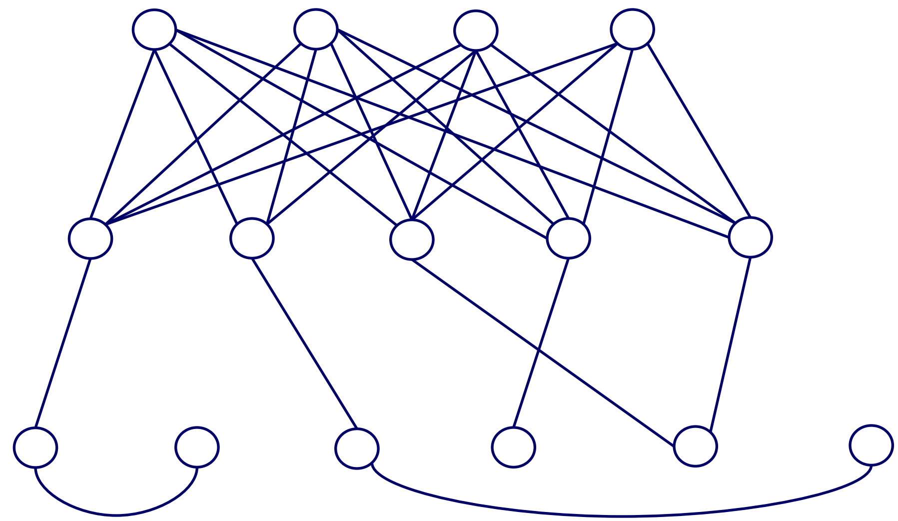
$\beta/\delta < \tau$ (below threshold)

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, **e-bay fraud detection,** blogs)
- Conclusions

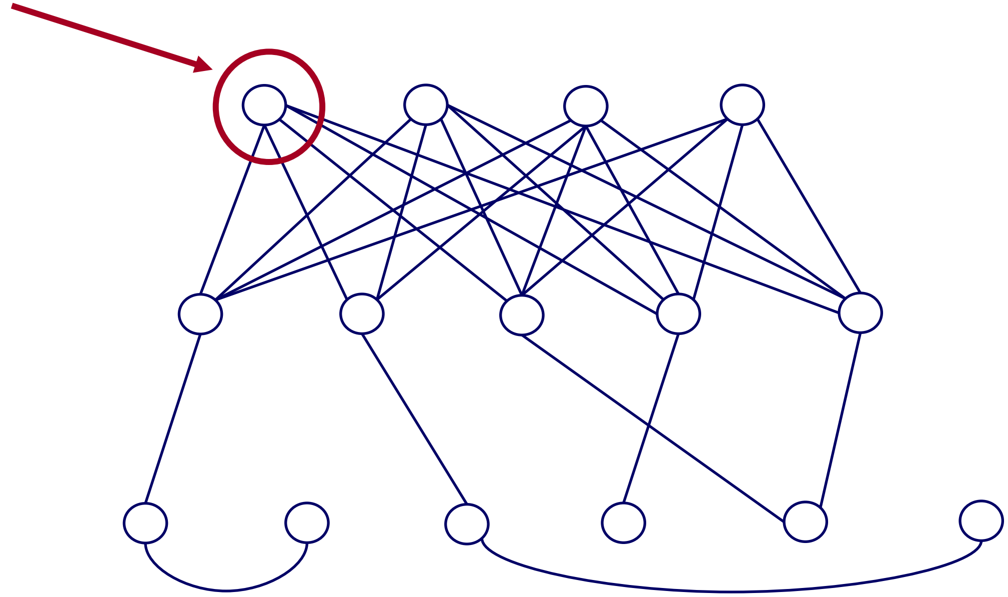
E-bay Fraud detection

w/ Polo Chau &
Shashank Pandit, CMU



E-bay Fraud detection

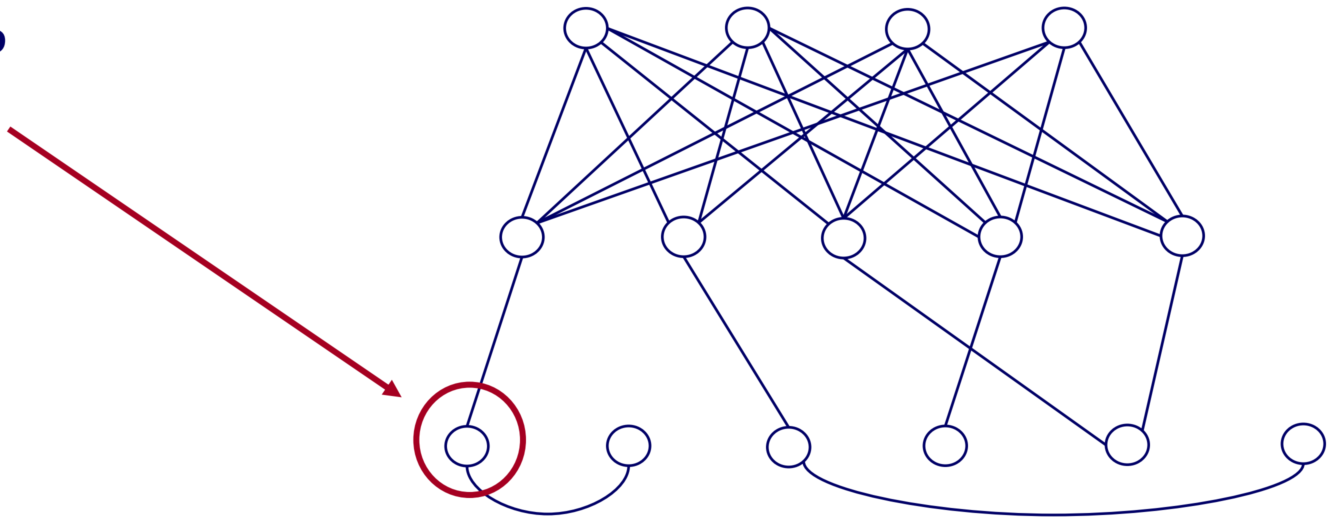
- lines: positive feedbacks
- would you buy from him/her?



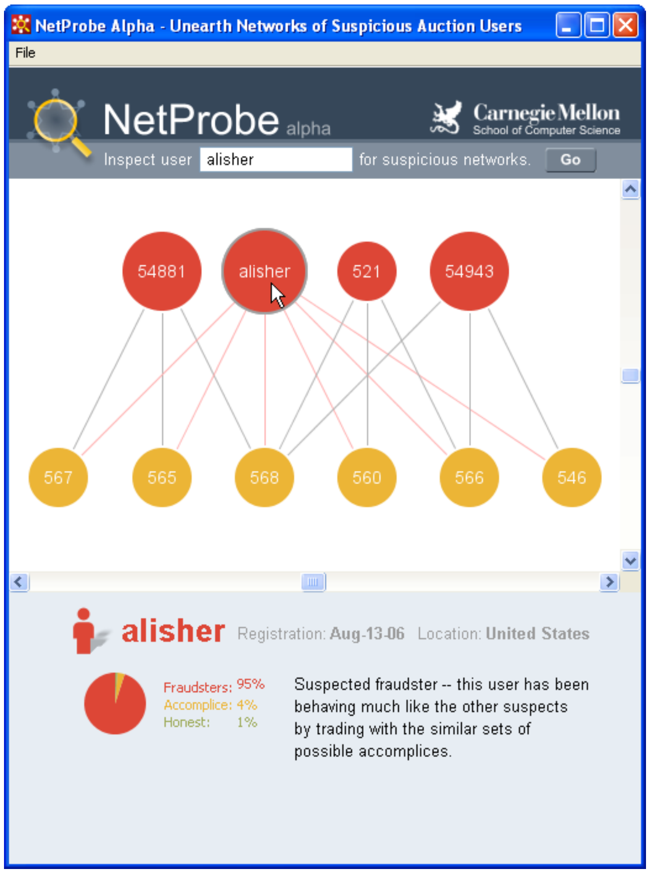
E-bay Fraud detection

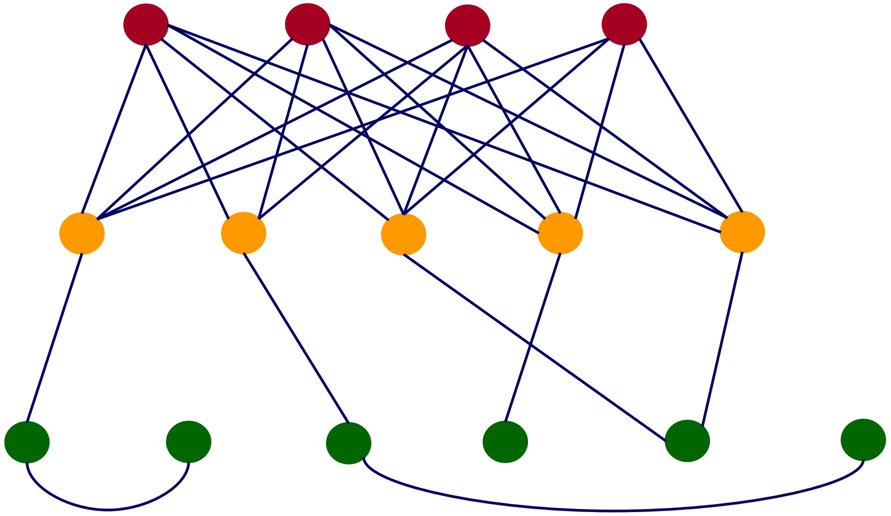
- lines: positive feedbacks
- would you buy from him/her?

- or him/her?



E-bay Fraud detection - NetProbe





Outline

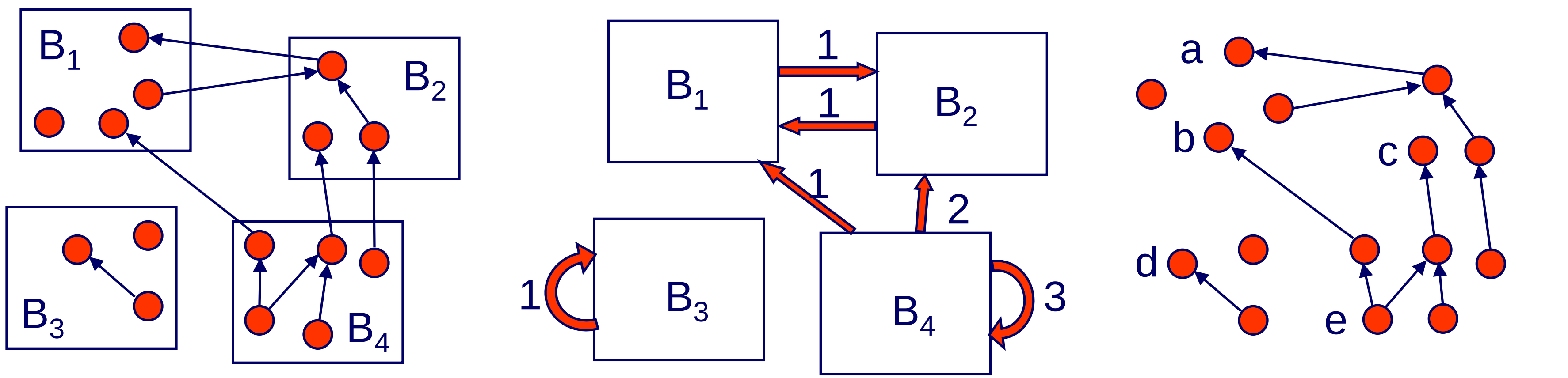
- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection**, <u>blogs</u>**)
- Conclusions

Blog analysis

- with Mary McGlohon (CMU)
- Jure Leskovec (CMU)
- Natalie Glance (now at Google)
- Mat Hurst (now at MSR)

[SDM'07]

Cascades on the Blogosphere



Blogosphere
blogs + posts

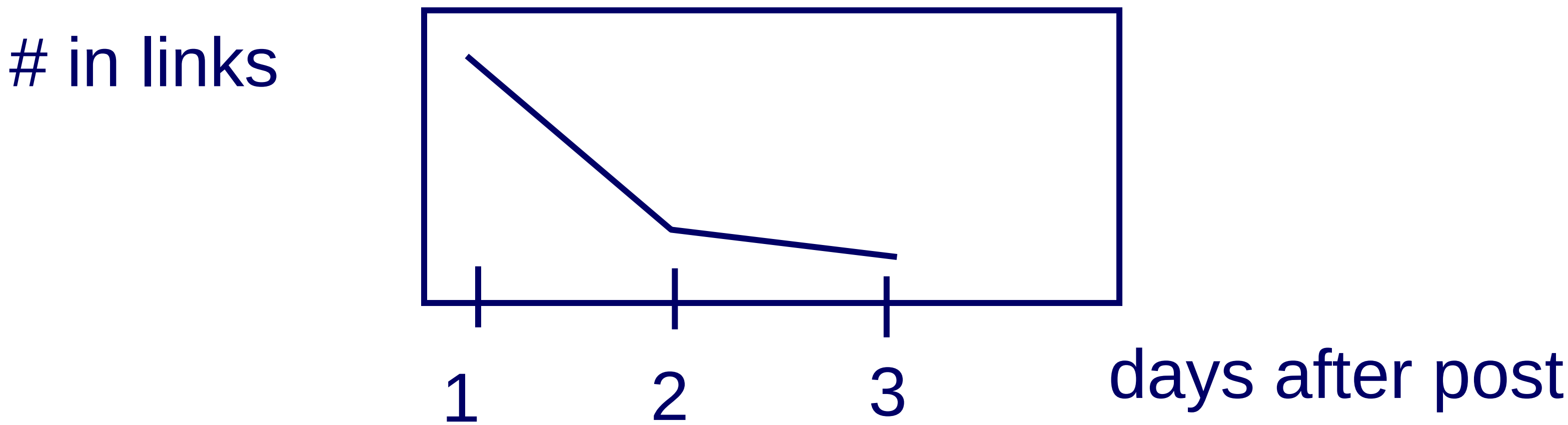
Blog network
links among blogs

Post network
links among posts

Q1: popularity-decay of a post?

Q2: degree distributions?

Q1: popularity over time

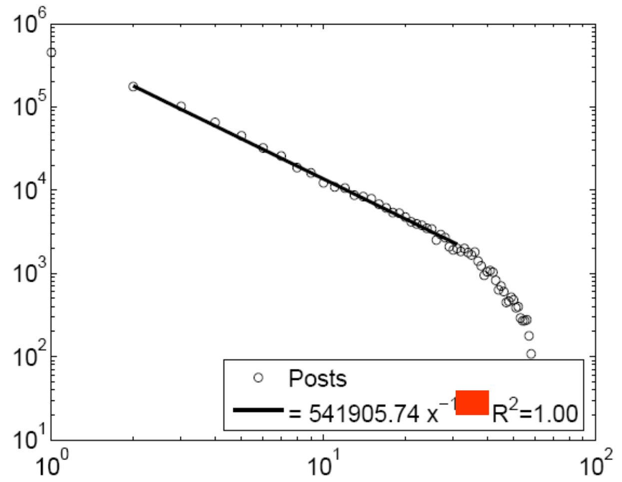


Post popularity drops-off – exponentially?

Days after post

Q1: popularity over time

in links
(**log**)



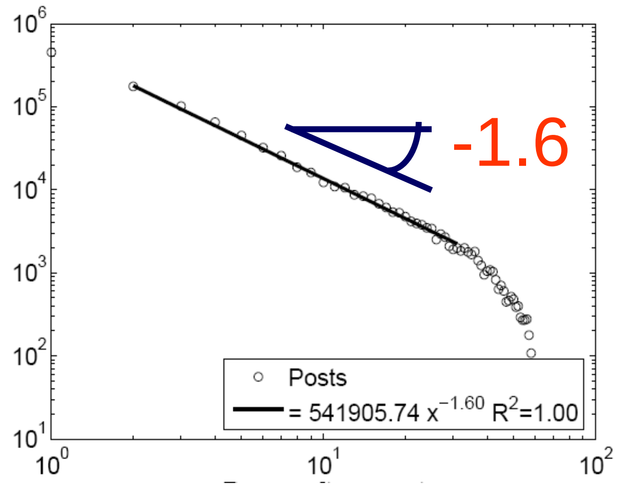
days after post
(**log**)

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

Days after post

Q1: popularity over time

in links (log)



days after post (log)

Post popularity drops-off – exponentially?

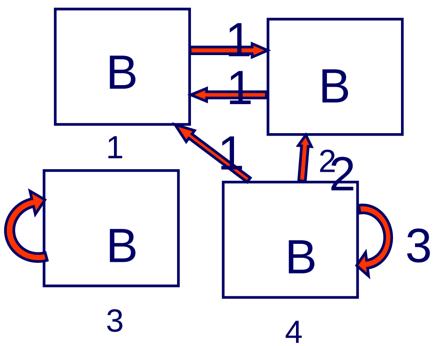
POWER LAW!

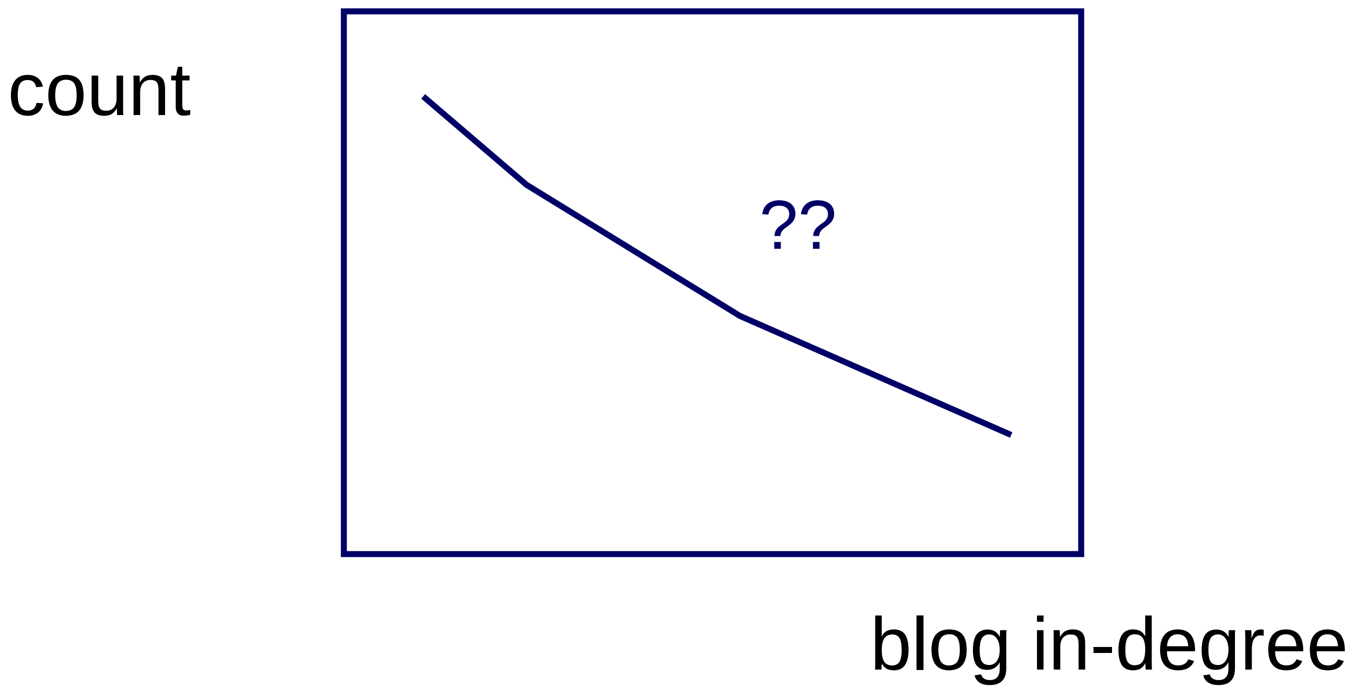
Exponent? -1.6 (close to -1.5: Barabasi's stack model)

Days after post

Q2: degree distribution

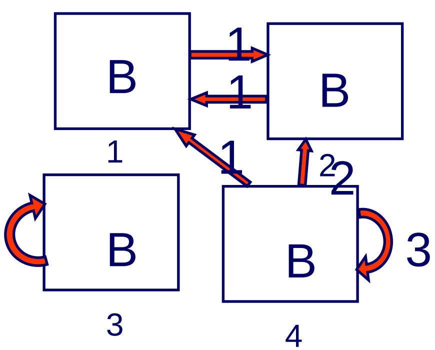
44,356 nodes, 122,153 edges. Half of blogs belong to largest connected component.



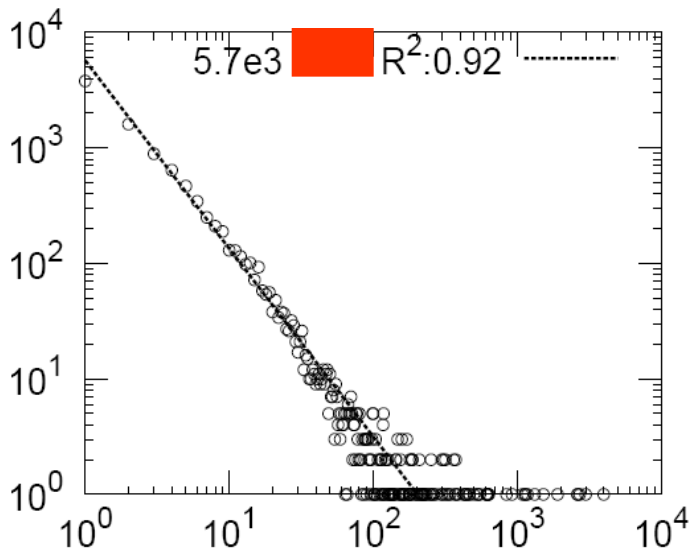


Q2: degree distribution

44,356 nodes, 122,153 edges. Half of blogs belong to largest connected component.



count

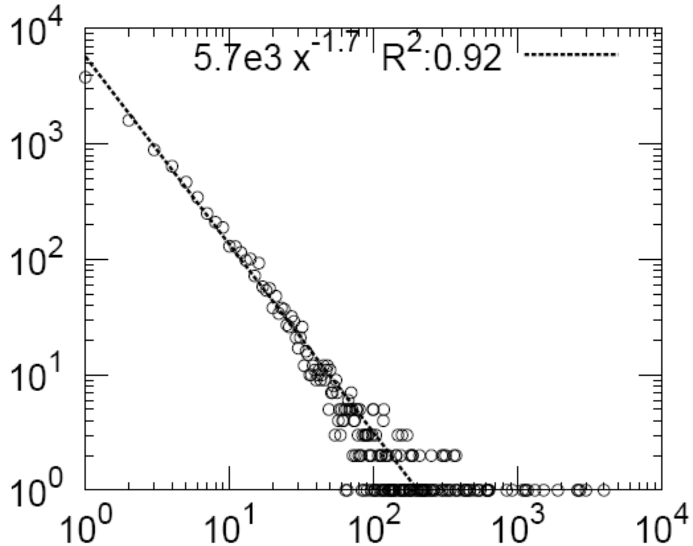


blog in-degree

Q2: degree distribution

44,356 nodes, 122,153 edges. Half of blogs belong to largest connected component.

count



blog in-degree

in-degree slope: -1.7
out-degree: -3
'rich get richer'

Next steps:

- edges with categorical attributes and/or time-stamps
- nodes with attributes

- **scalability** (hadoop – PetaByte scale)
  - first eigenvalue; diameter [done]
  - rest eigenvalues; community detection [to be done]
  - modularity, anomalies etc etc
- visualization (-> summarization)

E.g.: self-* system @ CMU

- >200 nodes
- target: 1 PetaByte

D.I.S.C.

- ‘Data Intensive Scientific Computing’ [R. Bryant, CMU]
  - ‘big data’
  - http://www.cs.cmu.edu/~bryant/pubdir/cmu-cs-07-128.pdf

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/

2’ intro to hadoop

- master-slave architecture; n-way replication (default n=3)
- ‘group by’ of SQL (in parallel, fault-tolerant way)
- e.g, find histogram of word frequency
  - slaves compute local histograms
  - master merges into global histogram

```
select course-id, count(*)
from ENROLLMENT
group by course-id
```

2' intro to hadoop

- master-slave architecture; n-way replication (default n=3)
- 'group by' of SQL (in parallel, fault-tolerant way)
- e.g, find histogram of word frequency
  - slaves compute local histograms
  - master merges into global histogram

select course-id, count(*)
from ENROLLMENT
group by course-id

reduce

map

OVERALL CONCLUSIONS

- Graphs: Self-similarity and power laws work, when textbook methods fail!
- New patterns (shrinking diameter!)
- New generator: Kronecker
- SVD / tensors / RWR: valuable tools
- hadoop/mapReduce for scalability

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)