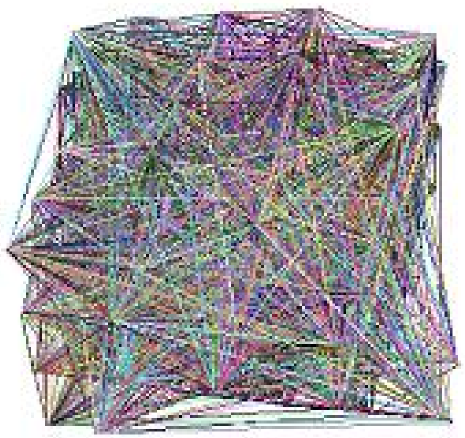
Graph Mining, self-similarity and power laws

Christos Faloutsos
Carnegie Mellon University

Overview

- Achievements
  - global patterns and ‘laws’ (static/dynamic)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

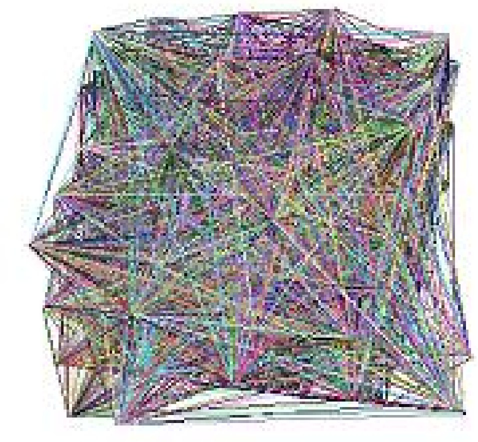
Motivating questions:



- How does the Internet look like?
- What constitutes a 'normal' social network?
- 'network value' of a customer? [Domingos+]
- which gene/species affects the others the most?

Problem #1 - topology

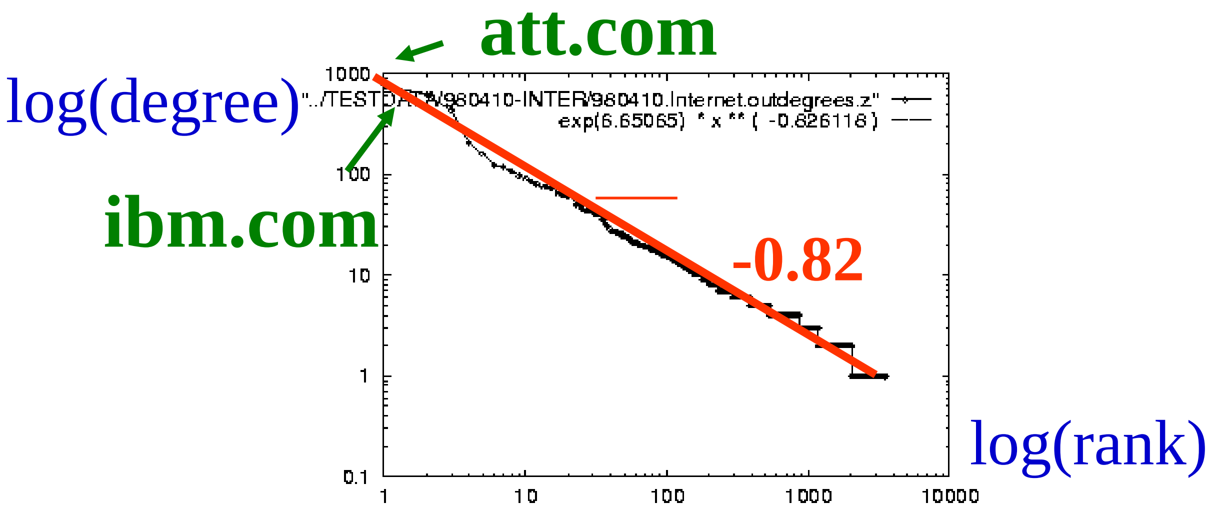
How does the Internet look like? Any rules?



Solution#1: Rank exponent *R*

- A1: Power law in the degree distribution [SIGCOMM99]

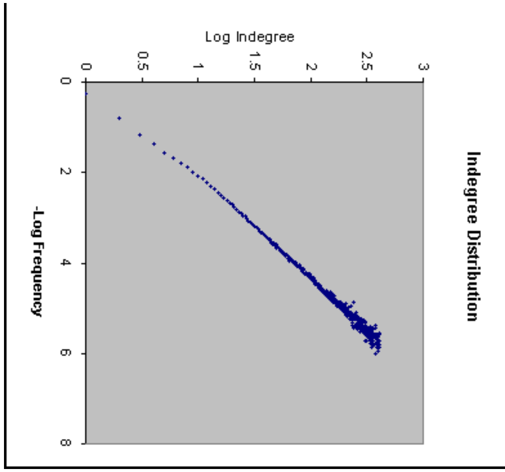
internet domains



Power laws

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log(freq)

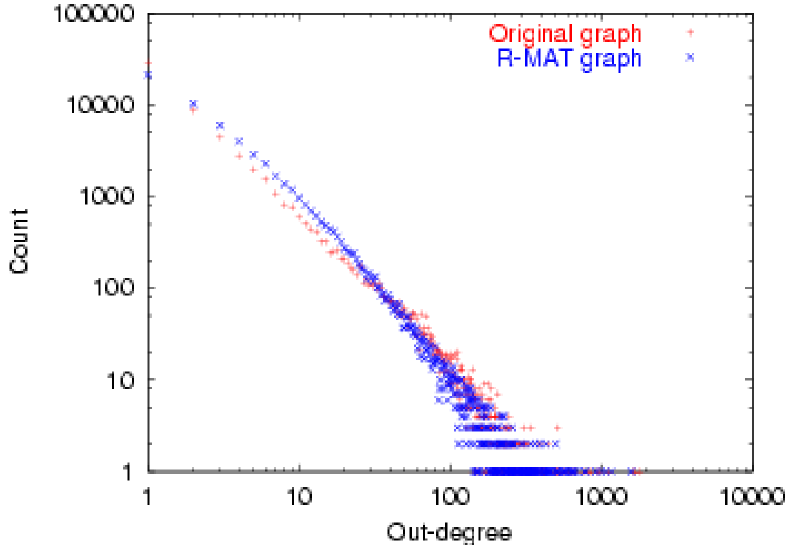


from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

log indegree

epinions.com

count

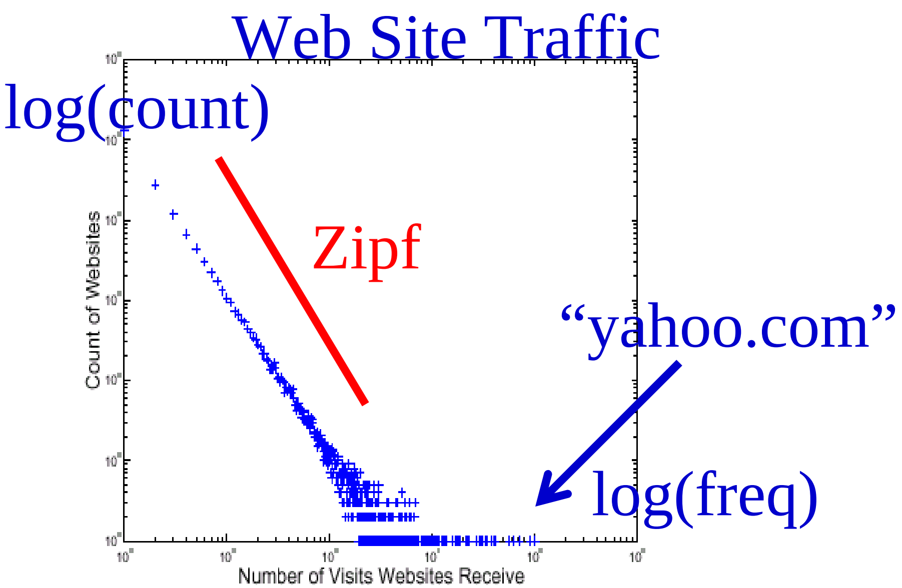


(out) degree

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

Even more power laws:

- web hit counts [w/ A. Montgomery]

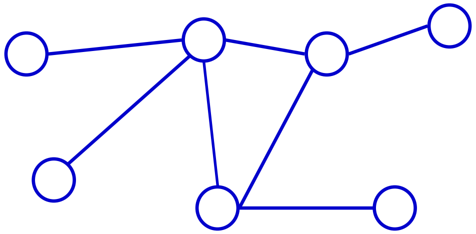


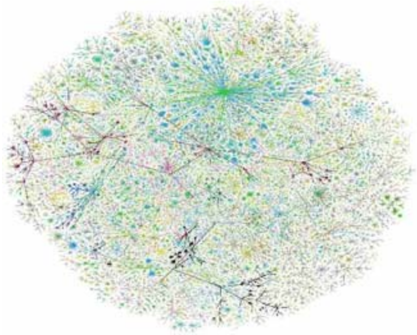
Overview

- Achievements
  - global patterns and 'laws' (static/**dynamic**)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

Problem#2: evolution

Given a graph:





- how will it look like, next year?

[from Lumeta: ISPs 6/1999]

Evolution of diameter?

- Prior analysis, on power-law-like graphs, hints that

  diameter ~ O(log(N)) or

  diameter ~ O(log(log(N)))
- i.e.., slowly increasing with network size
- Q: What is happening, in reality?

Evolution of diameter?

- Prior analysis, on power-law-like graphs, hints that

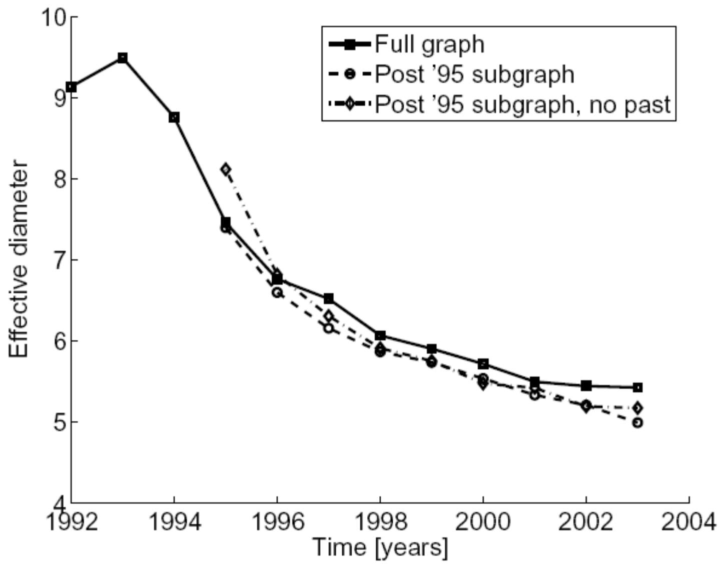
  diameter ~ O(log(N)) or

  diameter ~ O(log(log(N)))
- i.e.., slowly increasing with network size
- Q: What is happening, in reality?
- A: It **shrinks**(!!), towards a constant value

Shrinking diameter

ArXiv physics papers and their citations

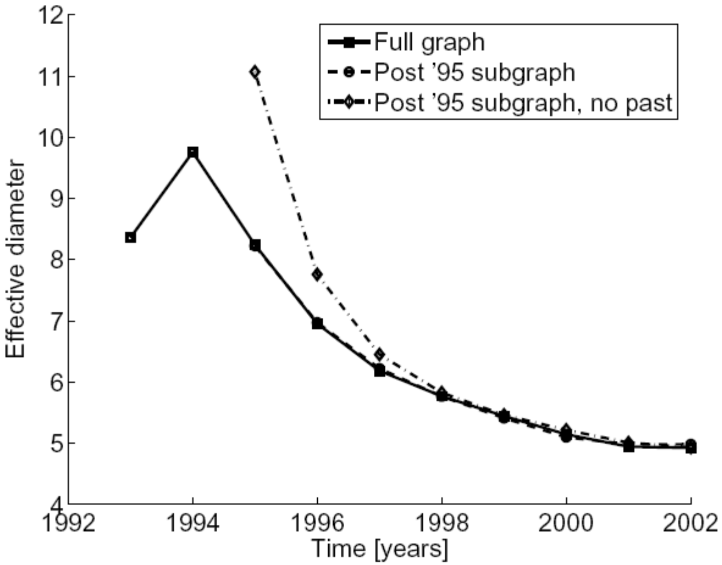
[Leskovec+05a]



(a) arXiv citation graph

Shrinking diameter

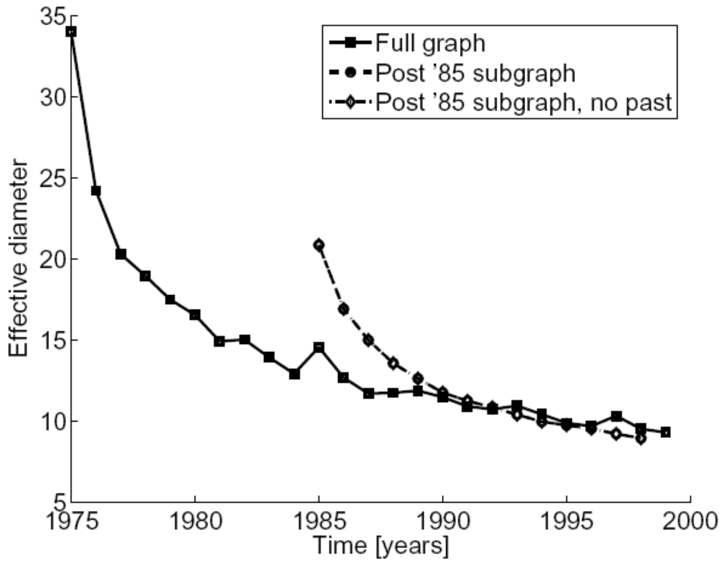
ArXiv: who wrote what



(b) Affiliation network

Shrinking diameter

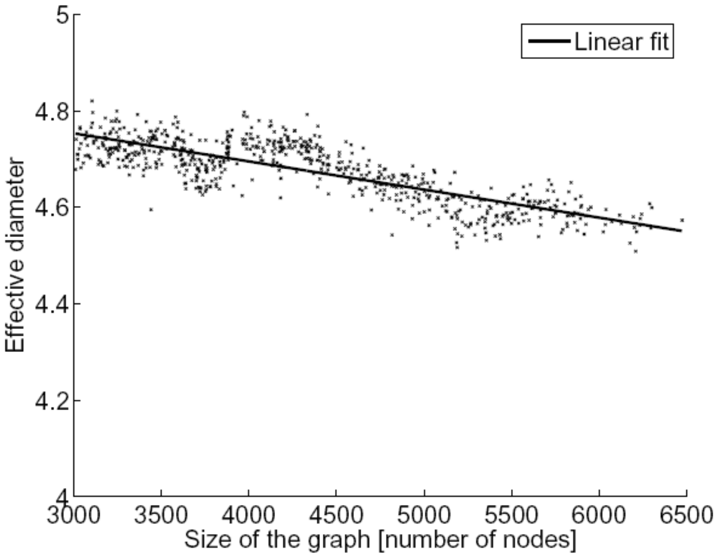
- U.S. patents citing each other



(c) Patents

Shrinking diameter

Autonomous systems



(d) AS

Temporal evolution of graphs

- N(t) nodes; 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 graphs

- N(t) nodes; E(t) edges at time t
- suppose that

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

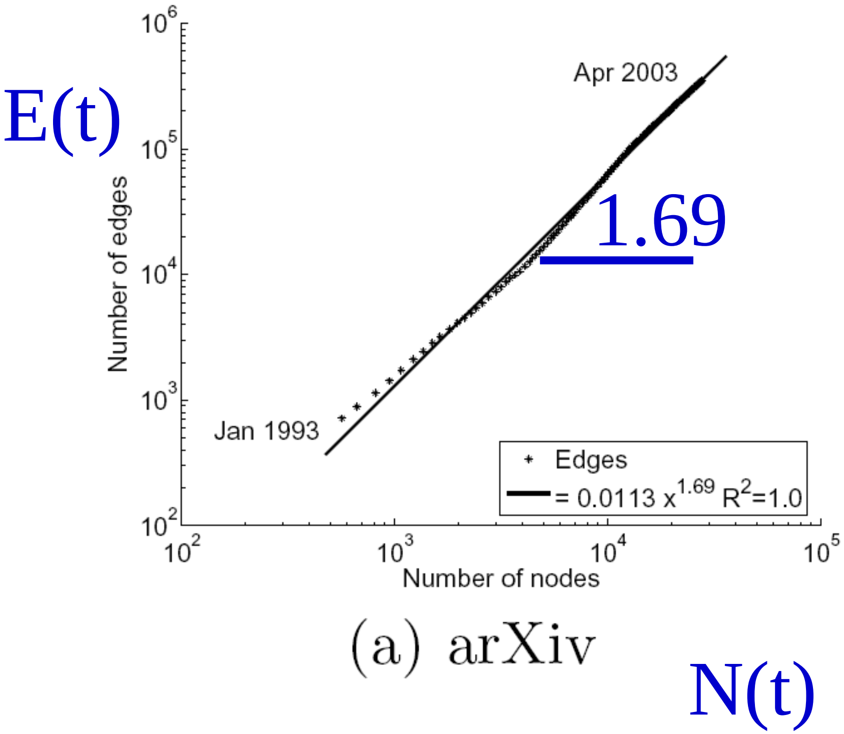
- Q: what is your guess for

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

- A: over-doubled!

Densification Power Law

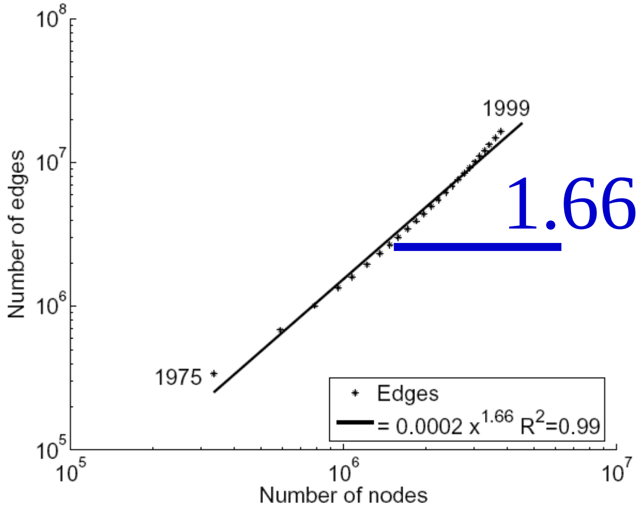
ArXiv: Physics papers and their citations



Densification Power Law

- U.S. Patents, citing each other

E(t)

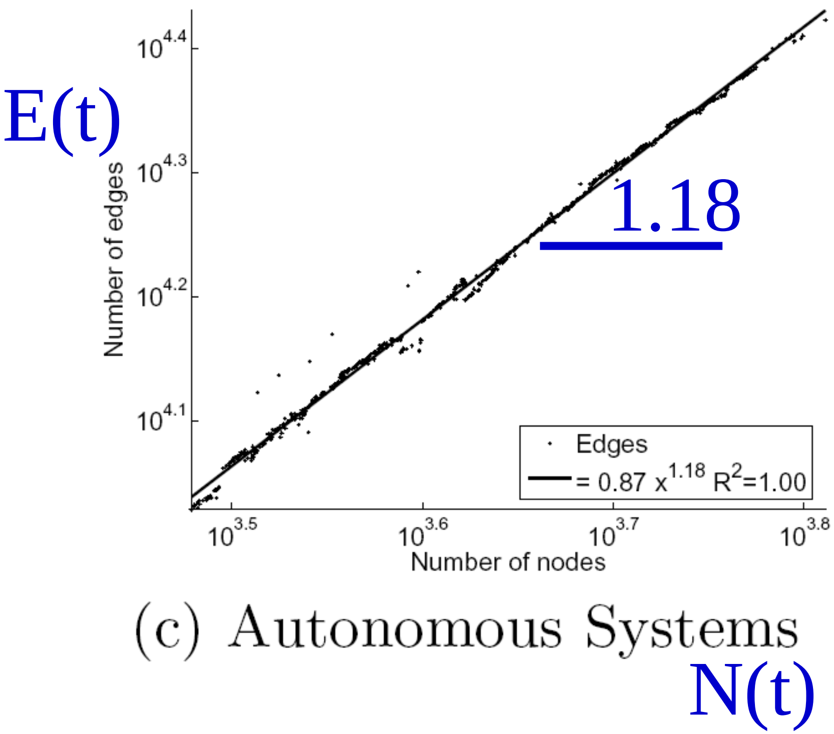


(b) Patents

N(t)

Densification Power Law

Autonomous Systems

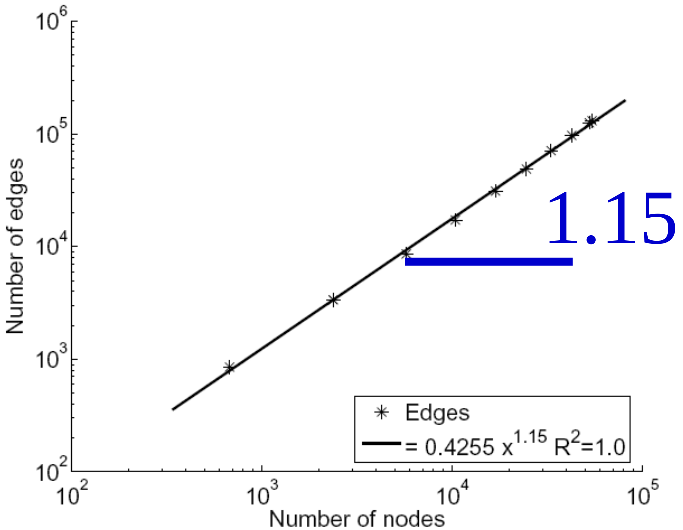


(c) Autonomous Systems

Densification Power Law

ArXiv: who wrote what

E(t)



(d) Affiliation network

N(t)

Summary of 'laws'

Static graphs

- power law degrees; power law eigenvalues
- communities (within communities)
- small diameters

Dynamic graphs

- shrinking diameter
- densification power law

Overview

- Achievements
  - global patterns and ‘laws’ (static/dynamic)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

Problem#3: Generators

- Q: what local behavior can generate such graphs?

Problem#3: Generators

Q: what local behavior can generate such graphs?

- A1: Preferential attachment [Barabasi+]
- A2: ‘copying’ model [Kleinberg+]
- A3: ‘forest-fire’ model [Leskovec+]
- A4: Kronecker [Leskovec+]
- A5: Economic reasons [Papadimitriou+]

Problem#3: Generators

Q: what local behavior can generate such graphs?

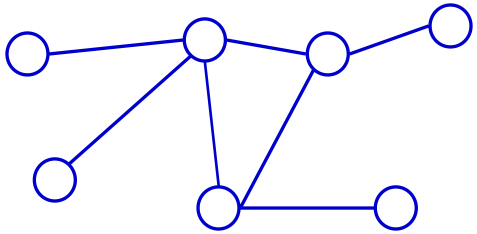
- A1: Preferential attachment — power law degree
- A2: 'copying' model — + communities
- A3: 'forest-fire' model — + DPL
- A4: Kronecker — + easily parallilizable
- A5: Economic reasons

Overview

- Achievements
  - global patterns and ‘laws’ (static/dynamic)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

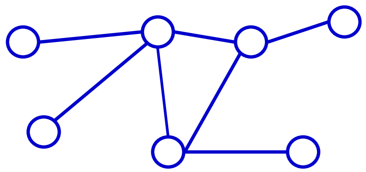
Problem#4: influence propagation

- how do influence/rumors/viruses propagate?
- what is the best customer to market to?



Problem#4: influence propagation

- how do influence/rumors/viruses propagate?
  - tipping point [Kleinberg+]
  - first **eigenvalue ->** epidemic threshold [Chakrabarti+]
- what is the best customer to market to?
  - network value of a customer [Domingos+]

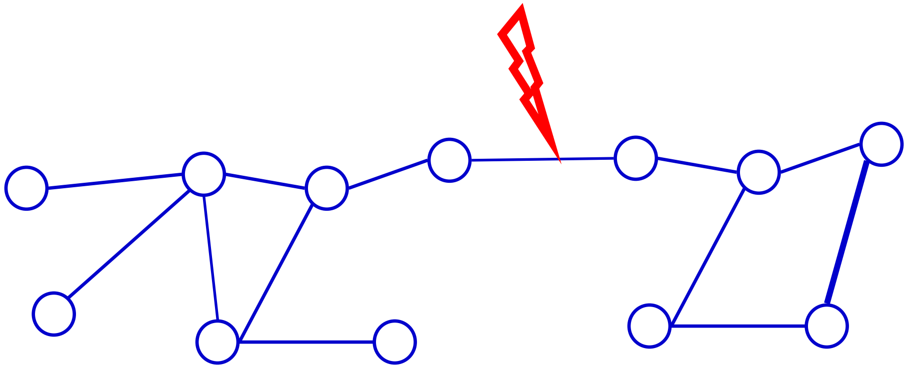


Overview

- Achievements
  - global patterns and ‘laws’ (static/dynamic)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

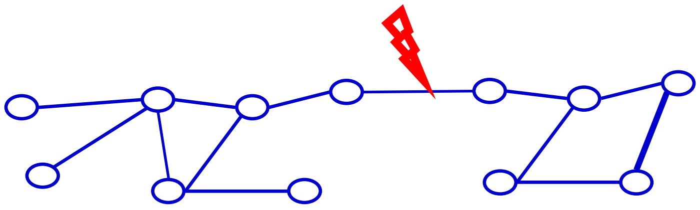
Problem #5: communities

- how to find ‘natural’ communities, quickly?
- how to find ‘strange’/suspicious/valuable edges?



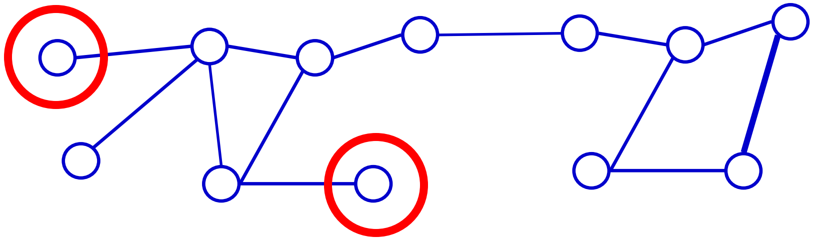
Problem #5: communities

- how to find ‘natural’ communities, quickly?
  – network flow [Flake+]
  – node/edge betweeness (~ ‘stress’)
  – cross-associations [Chakrabarti+]
  – 2^{nd} eigenvalue; METIS [Karypis+]
  – random walks [Newman]
  – etc etc etc



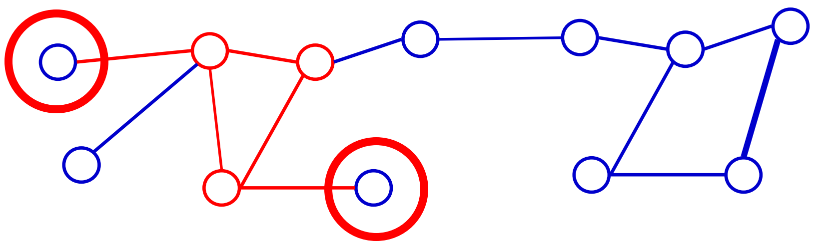
Problem #5: communities

- connection sub-graphs [Faloutsos+]



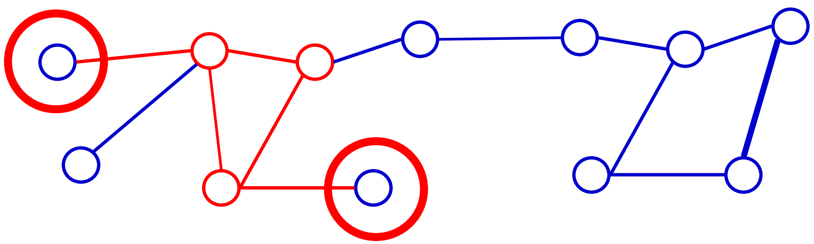
Problem #5: communities

- connection sub-graphs [Faloutsos+]



Problem #5: communities

- connection sub-graphs [Faloutsos+]
- BANKS system [Chakrabarti+]
- ObjectRank [Papakonstantinou+]

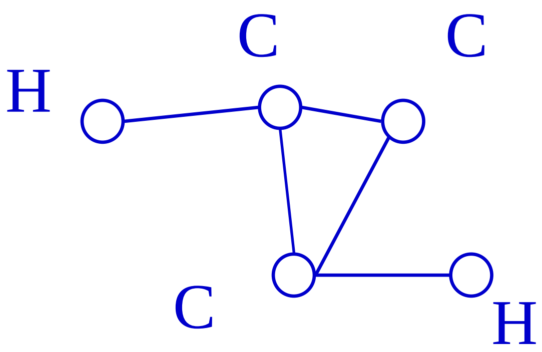


Overview

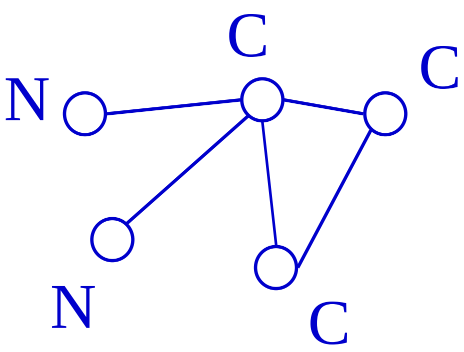
- Achievements
  - global patterns and ‘laws’ (static/dynamic)
  - generators
  - influence propagation
  - communities; graph partitioning
  - local patterns: frequent subgraphs
- Challenges

Problem #6: local patterns

- Which sub-graphs are common/frequent?



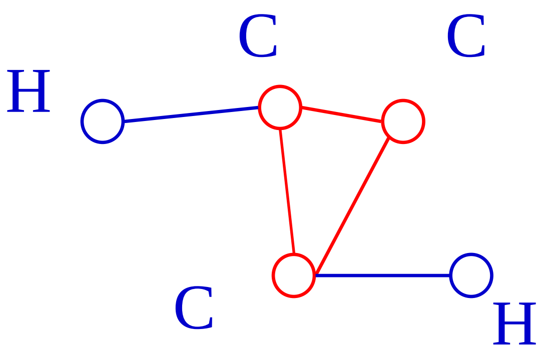
molecule #1



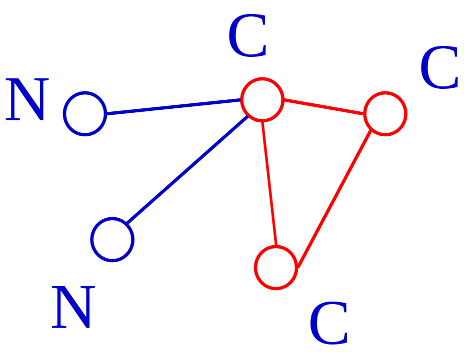
molecule #M

Problem #6: local patterns

- Which sub-graphs are common/frequent?



molecule #1



molecule #M

Problem #6: local patterns

- Clever extensions of ‘Association Rules’ (= frequent itemsets / market basket analysis)
  - [Jiawei Han+]
  - [Jian Pei+]
  - [G. Karypis]
  - [M. Zaki+]
  - ...

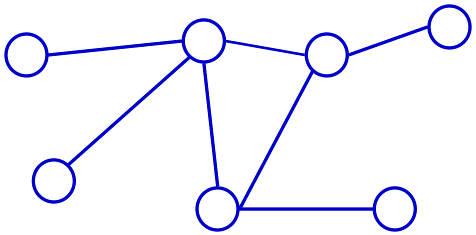
Overview

- Achievements
- Challenges
  - time evolving graphs
  - multi-graphs

Time evolving graphs

- Q: what will happen next? (eg., on a traffic matrix?)

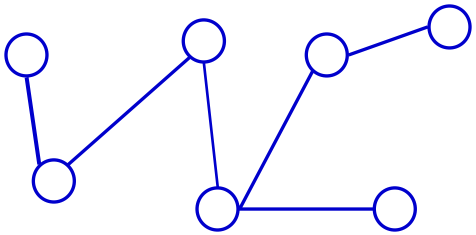
1/1/2005



Time evolving graphs

- Q: what will happen next? (eg., on a traffic matrix?)

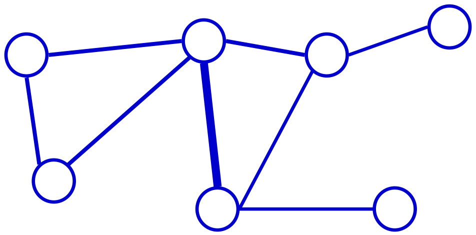
1/2/2005



Time evolving graphs

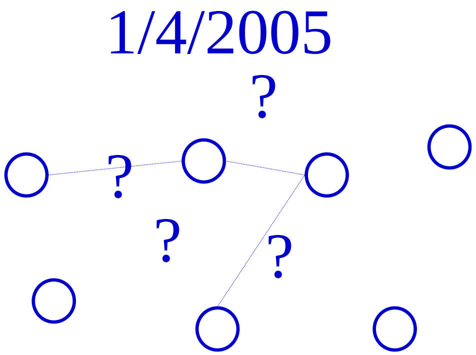
- Q: what will happen next? (eg., on a traffic matrix?)

1/3/2005



Time evolving graphs

- Q: what will happen next? (eg., on a traffic matrix?)

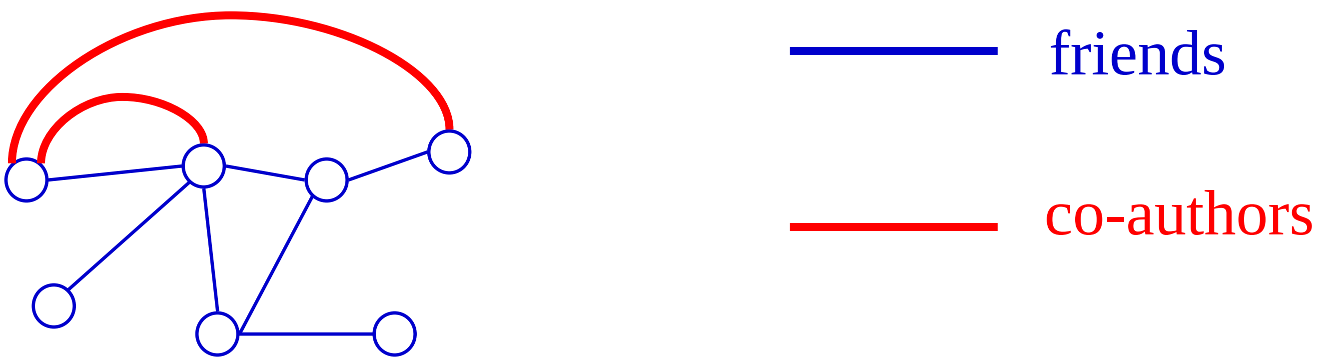


Overview

- Achievements
- Challenges
  - time evolving graphs
  - multi-graphs

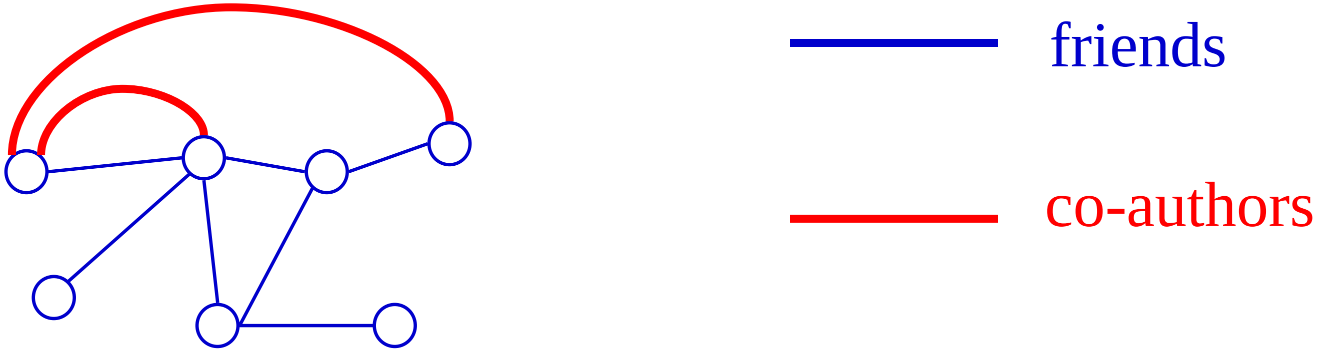
Multi-graphs

- Patterns/outliers?

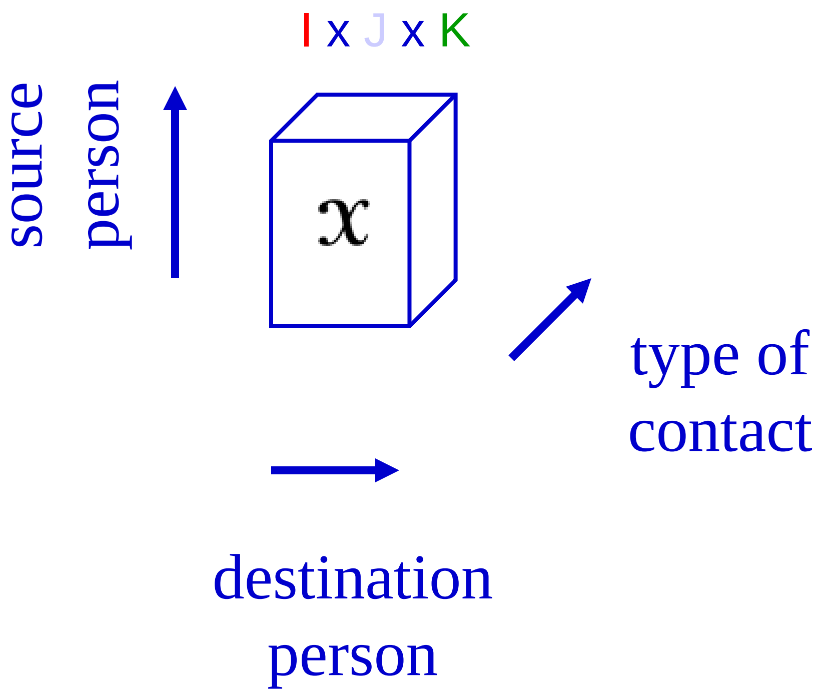


Multi-graphs

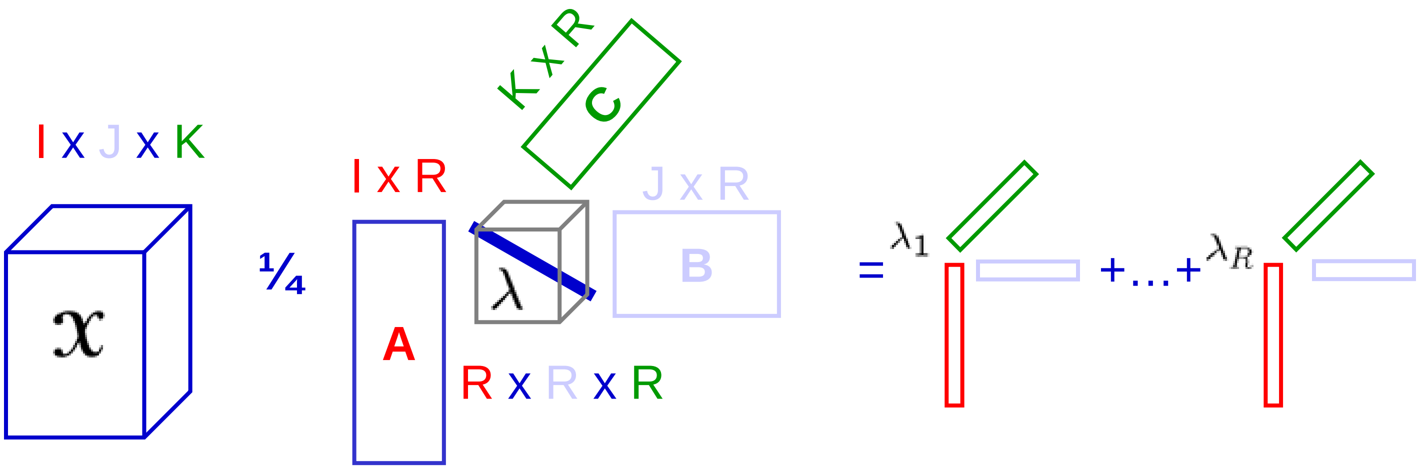
- Relational Learning
- link prediction, feature extraction [Jensen+] etc
- Dis-ambiguation, de-duplication [Getoor+] etc



Promising solution: tensors



Tensors: ~SVD, for >=3 modes



$$\mathcal{X} \approx [\![\lambda\,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r\, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Foil: from Tamara Kolda (Sandia)

Specially Structured Tensors

- Tucker Tensor

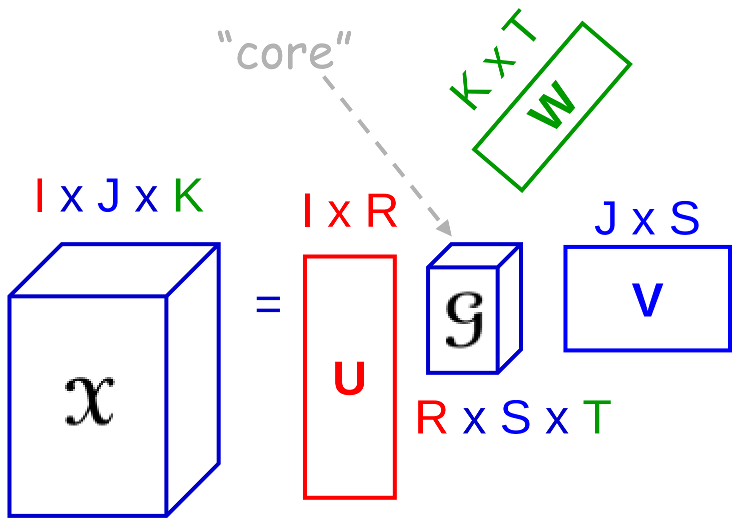
$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$$
$$= \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t$$
$$\equiv [\![\mathcal{G} \,;\, \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

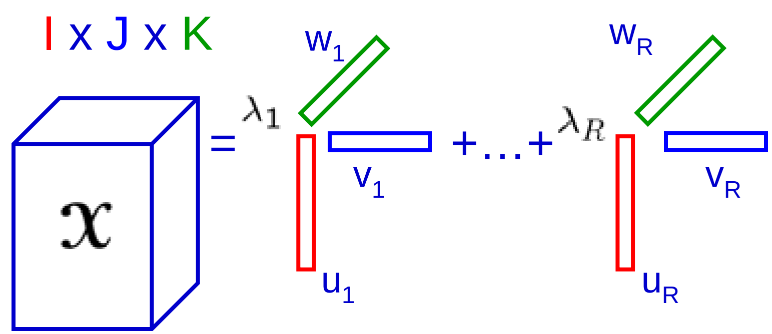
Our Notation

- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$$
$$\equiv [\![\lambda \,;\, \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

Our Notation





Foil: from Tamara Kolda (Sandia)

Conclusions - achievements

- Surprising patterns in graphs
  – power laws; communities
  – small/shrinking diameters
- simple, local behavior can lead to such patterns (eg., preferential attachment, etc)
- fast algorithms for communities/partitioning
- fast algorithms for ‘frequent subgraphs’

Conclusions - next steps

- multi-graphs
- time-evolving graphs
- scalability
- (graph sampling)
- (large graph visualization)

Conclusions - philosophically:

- deep connections with self-similarity, cellular automata (~agents), and ‘fractals’

Resources

- Manfred Schroeder "*Chaos, Fractals and Power Laws*", 1991
- A-L. Barabasi, "*Linked*", 2002
- D. Watts, "*Six Degrees*", 2004

References

- R. Albert, H. Jeong, and A.-L. Barabási, *Diameter of the World Wide Web*. Nature 401, 130-131 (1999)
- A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. *Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet*. Int. Colloquium on Automata, Languages, and Programming (ICALP), Malaga, Spain, July 2002.

References

- [sigcomm99] Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, *What does the Internet look like? Empirical Laws of the Internet Topology*, SIGCOMM 1999
- [Leskovec 05] Jure Leskovec, Jon M. Kleinberg, Christos Faloutsos: *Graphs over time: densification laws, shrinking diameters and possible explanations*. KDD 2005: 177-187

References

- [brite] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. *BRITE: An Approach to Universal Topology Generation*. MASCOTS '01
- Xifeng Yan, Xianghong Jasmine Zhou, Jiawei Han: *Mining closed relational graphs with connectivity constraints*. KDD 2005: 324-333

Thank you!

Contact info:

christos <at> cs.cmu.edu

www. cs.cmu.edu /~christos

(w/ papers, datasets, code, etc)