IIT Bombay Carnegie Mellon

Graph structures in data mining

Soumen Chakrabarti (*IIT-Bombay*)
Christos Faloutsos (*CMU*)

IIT Bombay Carnegie Mellon

Thanks

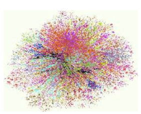
- Deepayan Chakrabarti (CMU)

- Michalis Faloutsos (UCR)

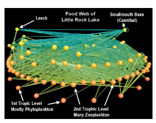
- George Siganos (UCR)

IIT Bombay Carnegie Mellon

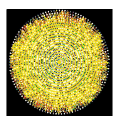
Introduction



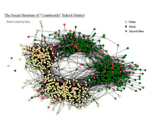
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



Protein Interactions
[genomebiology.com]



Friendship Network
[Moody '01]

Graphs are everywhere!

IIT Bombay Carnegie Mellon

Graph structures in KDD

- Physical networks
- Physical Internet
- Telephone lines
- Commodity distribution networks

IIT Bombay Carnegie Mellon

Networks derived from "behavior"

- Telephone call patterns
- Email, Blogs, Web, Databases, XML
- Language processing
- Web of trust, epinions

IIT Bombay Carnegie Mellon

Outline

Part 1: Topology, 'laws' and generators
Part 2: PageRank, HITS and eigenvalues
Part 3: Pairs, influence, communities

Motivating questions:

Part 1: Topology and generators

- What do real graphs look like?
- What properties of nodes, edges are important to model?
- What local and global properties are important to measure?
- How to model and generate realistic graphs?

Part 2: PageRank, HITS and eigenvalues

- How important is a node?
- Who is the best person/computer to immunize against a virus?
- Who is the best customer to advertise to?
- Who originated a raging rumor?

Part 3: Pairs, influence and communities

- How similar are two nodes?
- What does it mean to search for a node or a neighborhood?
- How do nodes influence their neighbors?
- Is "influence" a verb or a noun?

PART 1: Topology, laws and generators

Outline

Part 1: Topology, 'laws' and generators
- 'Laws' and patterns
- Generators
- Tools

Part 2: PageRank, HITS and eigenvalues

Part 3: Pairs, influence, communities

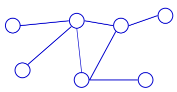
Motivating questions

- What do real graphs look like?
  - What properties of nodes, edges are important to model?
  - What local and global properties are important to measure?
- How to generate realistic graphs?

IIT Bombay Carnegie Mellon

Motivating questions

Given a graph:

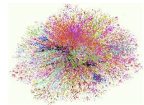


•Are there un-natural sub-graphs? (criminals' rings or terrorist cells)?

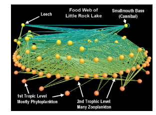
• How do P2P networks evolve?

IIT Bombay Carnegie Mellon

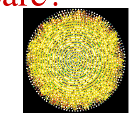
Why should we care?



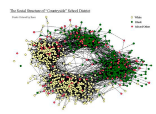
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



Protein Interactions
[genomebiology.com]



Friendship Network
[Moody '01]

IIT Bombay Carnegie Mellon

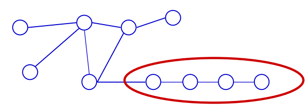
Why should we care?

- **A1: extrapolations**: how will the Internet/Web look like next year?
- **A2: algorithm design**: what is a realistic network topology,
  - to try a new routing protocol?
  - to study virus/rumor propagation, and immunization?

IIT Bombay Carnegie Mellon

Why should we care? (cont'd)

- **A3: Sampling**: How to get a 'good' sample of a network?
- **A4: Abnormalities**: is this sub-graph / sub-community / sub-network 'normal'? (what **is** normal?)



IIT Bombay Carnegie Mellon

Outline

Part 1: Topology, 'laws' and generators
- 'Laws' and patterns
- Generators
- Tools

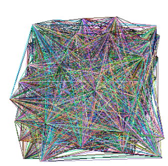
Part 2: PageRank, HITS and eigenvalues

Part 3: Pairs, influence, communities

IIT Bombay Carnegie Mellon

Topology

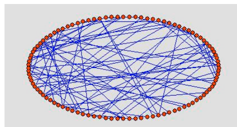
How does the Internet look like? Any rules?

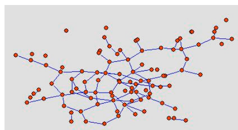


(Looks random – right?)

Are real graphs random?

- random (Erdos-Renyi) graph – 100 nodes, avg degree = 2
- before layout
- after layout
- No obvious patterns





(generated with: pajek http://vlado.fmf.uni-lj.si/pub/networks/pajek/)

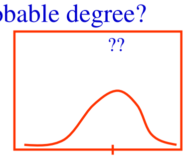
Laws and patterns

Real graphs are NOT random!!

- Diameter
- in- and out- degree distributions
- other (surprising) patterns

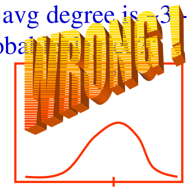
Laws – degree distributions

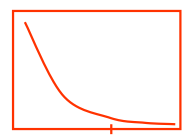
- Q: avg degree is ~2 - what is the most probable degree?



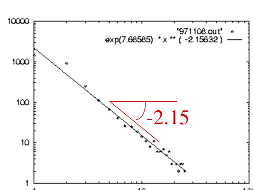
Laws – degree distributions

- Q: avg degree is ~3 - what is the most probable degree?





I.Power-law: outdegree *O*



Exponent = slope

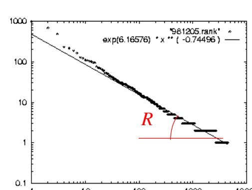
O = -2.15

Nov'97

The plot is linear in log-log scale [FFF'99]

$freq = degree^{(-2.15)}$

II.Power-law: rank *R*



Exponent = slope

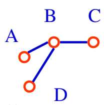
R = -0.74

Dec'98

- The plot is a line in log-log scale

III. Eigenvalues

- Let A be the adjacency matrix of graph
- The eigenvalue λ is:
  - $A\ \underline{v} = \lambda\ \underline{v}$, where $\underline{v}$ some vector
- Eigenvalues are strongly related to graph topology



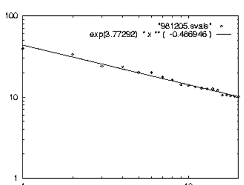
	A	B	C	D
A		1		
B	1		1	1
C		1		
D		1		

III. Eigenvalues

MUCH more on eigenvalues: in Part 2

III.Power-law: eigen *E*

Eigenvalue



Exponent = slope

E = -0.48

Dec'98

Rank of decreasing eigenvalue

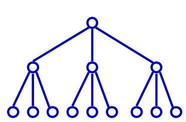
- Eigenvalues in decreasing order (first 20)
- [Mihail+, 02]: $R = 2 * E$

IV. The Node Neighborhood

- $N(h)$ = # of pairs of nodes within h hops

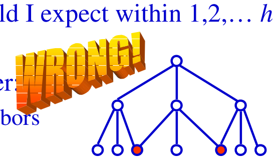
IV. The Node Neighborhood

- Q: average degree = 3 - how many neighbors should I expect within 1,2,… h hops?
- Potential answer:
  1 hop -> 3 neighbors
  2 hops -> 3 * 3
  …
  h hops -> 3^h



IV. The Node Neighborhood

- Q: average degree = 3 - how many neighbors should I expect within 1,2,… h hops?
- Potential answer:
  1 hop -> 3 neighbors
  2 hops -> 3 * 3
  …
  h hops -> 3^h



WE HAVE DUPLICATES!

IIT Bombay **Carnegie Mellon**

IV. The Node Neighborhood

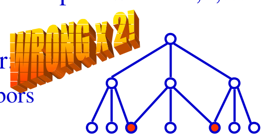
- Q: average degree = 3 - how many neighbors should I expect within 1,2,… *h* hops?
- Potential answer:

  1 hop -> 3 neighbors

  2 hops -> 3 * 3

  …

  h hops -> 3^h

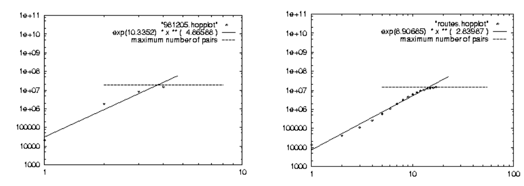


'avg' degree: meaningless!

KDD04

IIT Bombay **Carnegie Mellon**

IV. Power-law: hopplot *H*



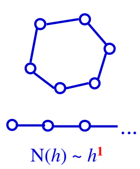
Pairs of nodes as a function of hops $N(h)= h^H$

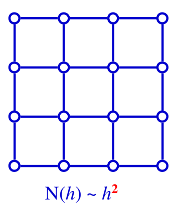
KDD04

IIT Bombay **Carnegie Mellon**

Observation

- Q: Intuition behind 'hop exponent'?
- A: 'intrinsic=fractal dimensionality' of the network





KDD04

IIT Bombay **Carnegie Mellon**

Hop plots

- More on fractal/intrinsic dimensionalities: very soon

KDD04

IIT Bombay **Carnegie Mellon**

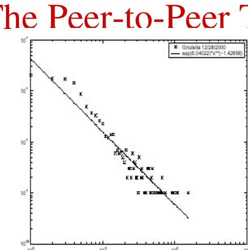
But:

- Q1: How about graphs from other domains?
- Q2: How about temporal evolution?

KDD04

IIT Bombay **Carnegie Mellon**

The Peer-to-Peer Topology



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

[Jovanovic+]

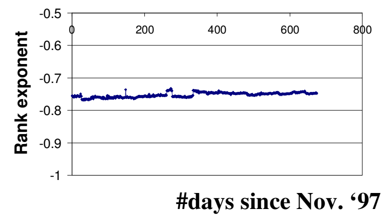
- Frequency versus degree
- Number of adjacent peers follows a power-law

KDD04

More Power laws

- Also hold for other web graphs [Barabasi+, ‘99], [Kumar+, ‘99] with additional ‘rules’ (bi-partite cores follow power laws)

Time Evolution: rank *R*



Domain level

The rank exponent has not changed! [Siganos+, ‘03]

Outline

Part 1: Topology, ‘laws’ and generators
- ‘Laws’ and patterns
  - Power laws for degree, eigenvalues, hop-plot
  - ???
- Generators
- Tools

Part 2: PageRank, HITS and eigenvalues

Part 3: Pairs, influence, communities

Any other ‘laws’?

Yes!

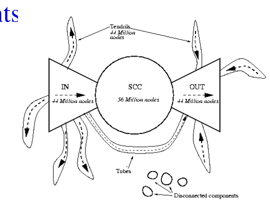
Any other ‘laws’?

Yes!

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

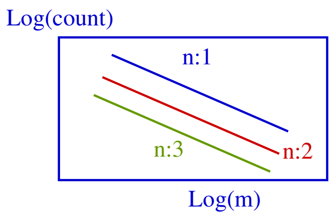
Any other ‘laws’?

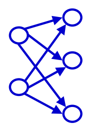
- Bow-tie, for the web [Kumar+ ‘99]
- IN, SCC, OUT, ‘tendrils’
- disconnected components



Any other 'laws'?

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

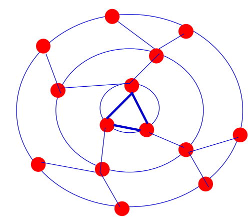




2:3 core
(m:n core)

Any other 'laws'?

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



Summary of 'laws'

- Power laws for degree distributions
- ………….... for eigenvalues, bi-partite cores
- Small diameter ('6 degrees')
- 'Bow-tie' for web; 'jelly-fish' for internet

Outline

Part 1: Topology, 'laws' and generators
- 'Laws' and patterns
- Generators
- Tools

Part 2: PageRank, HITS and eigenvalues

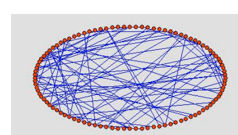
Part 3: Pairs, influence, communities

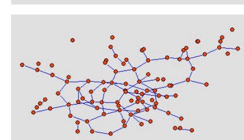
Generators

- How to generate random, realistic graphs?
  - Erdos-Renyi model: beautiful, but unrealistic
  - degree-based generators
  - process-based generators

Erdos-Renyi

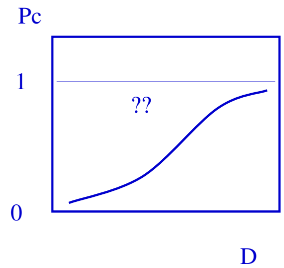
- random graph – 100 nodes, avg degree = 2
- Fascinating properties (phase transition)
- But: unrealistic (Poisson degree distribution != power law)





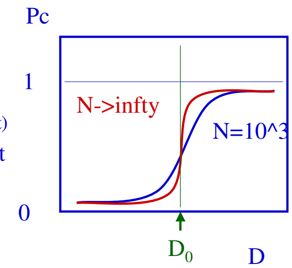
E-R model & Phase transition

- vary avg degree D
- watch Pc =
  Prob(there is a giant connected component)
- How do you expect it to be?



E-R model & Phase transition

- vary avg degree D
- watch Pc =
  Prob(there is a giant connected component)
- How do you expect it to be?



Degree-based

- Figure out the degree distribution (eg., ‘Zipf’)
- Assign degrees to nodes
- Put edges, so that they match the original degree distribution

Process-based

- Barabasi; Barabasi-Albert: Preferential attachment -> power-law tails!
  - ‘rich get richer’
- [Kumar+]: preferential attachment + mimick
  - Create ‘communities’

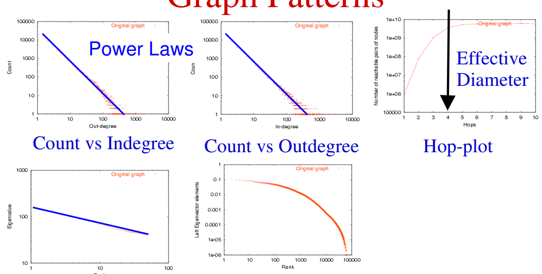
Process-based (cont’d)

- [Fabrikant+, ‘02]: H.O.T.: connect to closest, high connectivity neighbor
- [Pennock+, ‘02]: Winner does NOT take all

R-MAT

- Recursive MATrix generator [Chakrabarti+,’04]
- Goals:
  - Power-law in- and out-degrees
  - Power law eigenvalues
  - Small diameter
  - Few parameters

Graph Patterns

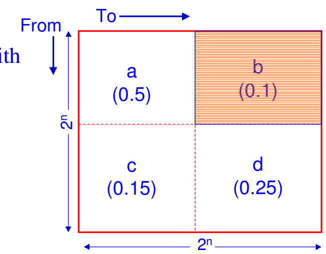


Count vs Indegree | Count vs Outdegree | Hop-plot

Eigenvalue vs Rank | "Network values" vs Rank

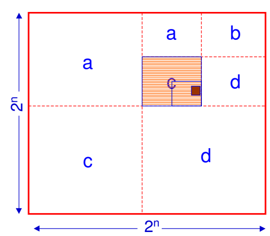
R-MAT

- Subdivide the adjacency matrix
- choose a quadrant with probability (a,b,c,d)



R-MAT

- Recurse till we reach a 1*1 cell



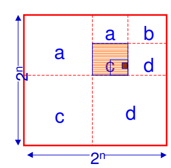
R-MAT

by construction:

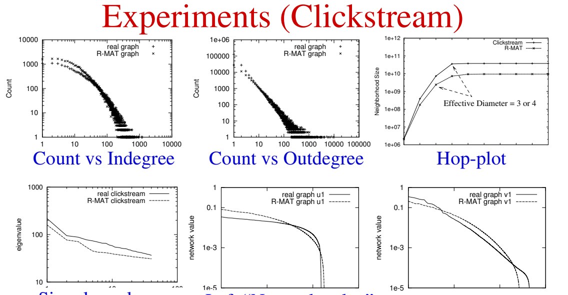
- rich-get-richer for in-degree
- .... for out-degree
- communities within communities

and

- small diameter



Experiments (Clickstream)



Count vs Indegree | Count vs Outdegree | Hop-plot

Singular value vs Rank | Left "Network value" | Right "Network value"

R-MAT matches it well

Conclusions

'Laws' and patterns:

- Power laws for degrees, eigenvalues, 'communities'/cores
- Small diameter
- Bow-tie; jelly-fish

Conclusions, cont'd

Generators
- Preferential attachment (Barabasi)
- Variations
- Recursion – RMAT

Conclusions, cont'd

Tools
- Power laws – rank/frequency plots
- Self-similarity / recursion / fractals
- 'correlation integral' = hop-plot

Resources

Generators:
- RMAT (deepay@cs.cmu.edu)
- BRITE http://www.cs.bu.edu/brite/
- INET: http://topology.eecs.umich.edu/inet

Other resources

Visualization - graph algo's:
- Graphviz: http://www.graphviz.org/
- pajek: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Kevin Bacon web site:
http://www.cs.virginia.edu/oracle/

Outline

Part 1: Topology, 'laws' and generators
- 'Laws' and patterns
- Generators
- ➡ Tools

Part 2: PageRank, HITS and eigenvalues

Part 3: Pairs, influence, communities

Outline

Part 1: Topology, 'laws' and generators
- 'Laws' and patterns
- Generators
- ➡ Tools: power laws and fractals
  - Why so many power laws?
  - Self-similarity, power laws, fractal dimension

IIT Bombay Carnegie Mellon

Power laws

- Q1: Why so many?
- A1:
- Q2: Are they only in graph-related settings?
- A2:

KDD04 © 2004, S. Chakrabarti and C. Faloutsos 67

IIT Bombay Carnegie Mellon

Power laws

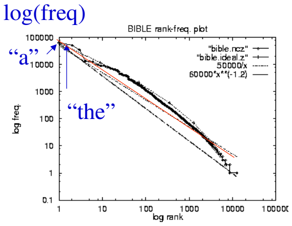
- Q1: Why so many?
- A1: self-similarity; 'rich-get-richer'
- Q2: Are they only in graph-related settings?
- A2: NO!

KDD04 © 2004, S. Chakrabarti and C. Faloutsos 68

IIT Bombay Carnegie Mellon

A famous power law: Zipf's law

log(freq)



log(rank)

- Bible - **rank** vs **frequency** (log-log)

KDD04 © 2004, S. Chakrabarti and C. Faloutsos 69

IIT Bombay Carnegie Mellon

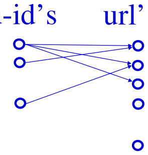
Power laws, cont'ed

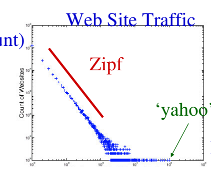
- length of file transfers [Bestavros+]
- web hit counts [Huberman]
- Click-stream data [Montgomery+01]

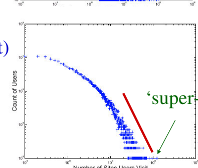
KDD04 © 2004, S. Chakrabarti and C. Faloutsos 70

IIT Bombay Carnegie Mellon

Click-stream data





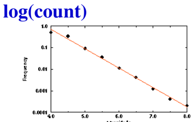


KDD04 © 2004, S. Chakrabarti and C. Faloutsos 71

IIT Bombay Carnegie Mellon

More power laws

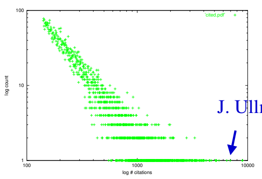
- duration of UNIX jobs; of UNIX file sizes
- Energy of earthquakes (Gutenberg-Richter law) [simscience.org]



KDD04 © 2004, S. Chakrabarti and C. Faloutsos 72

Lotka's law

(Lotka's law of publication count); and citation counts: (*citeseer.nj.nec.com* 6/2001)

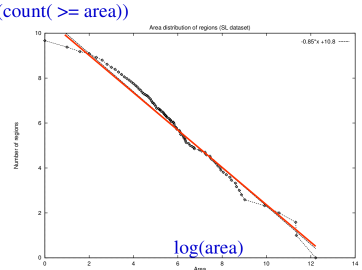


Korcak's law

Scandinavian lakes

Any pattern?

Korcak's law



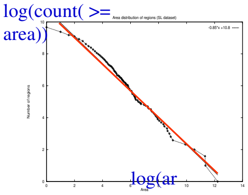
CCDF=NCDF:

Scandinavian lakes area vs complementary cumulative count (log-log axes)

Korcak's law

Similar laws for

- islands
- connected components, at phase transition [Schroeder, '91]

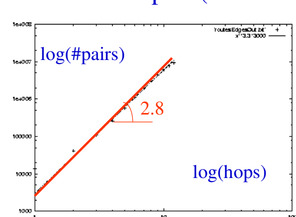


Power laws

- Q1: Why so many?
- A1: self-similarity; 'rich-get-richer'
- Q2: Are they only in graph-related settings?
- A2: NO!

Recall: Hop Plot

- Internet routers: how many neighbors within *h* hops? (= **correlation integral**!)

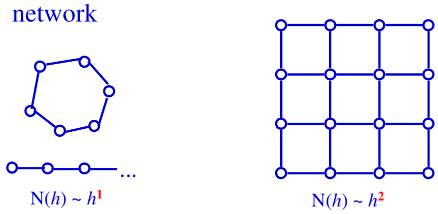


Reachability function: number of neighbors within r hops, vs r (log-log).

Mbone routers, 1995

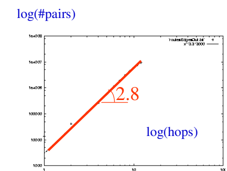
Observation

- Q: Intuition behind 'hop exponent'?
- A: 'intrinsic=fractal dimensionality' of the network



Non-integer dimensionality??

- Q3: How is it possible?
- A3:
- Q4: What does it mean?
- A4:



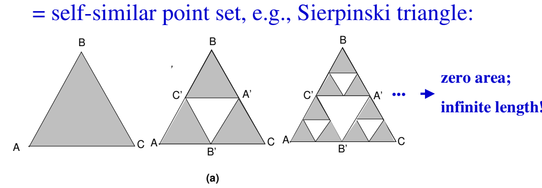
Non-integer dimensionality??

- Q3: How is it possible?
- A3: Through recursion!
- Q4: What does it mean?
- A4: There are groups (quasi-cliques / communities) in every scale

For example: a famous set of points:

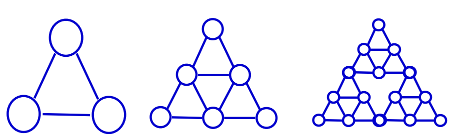
A famous fractal

= self-similar point set, e.g., Sierpinski triangle:



A famous fractal

equivalent graph:

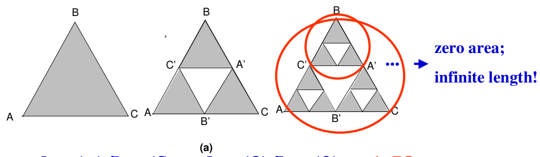


Definitions (cont'd)

- Paradox: Infinite perimeter ; Zero area!
- 'dimensionality': between 1 and 2
- actually: Log(3)/Log(2) = 1.58...

Dfn of fd:

ONLY for a perfectly self-similar point set:



=log(n)/log(f) = log(3)/log(2) = 1.58

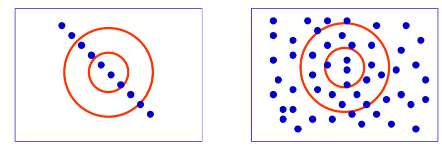
Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: nn (<= r) ~ r^1

('power law': y=x^a)

- Q: fd of a plane?
- A: nn (<= r) ~ r^2

fd== slope of (log(nn) vs log(r))



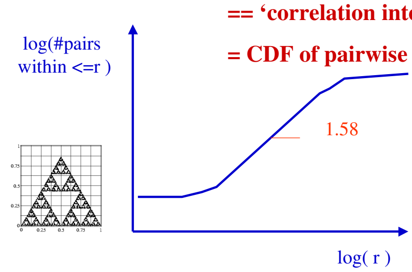
Intrinsic ('fractal') dimension

Algorithm, to estimate it?

Sierpinsky triangle

== 'correlation integral'

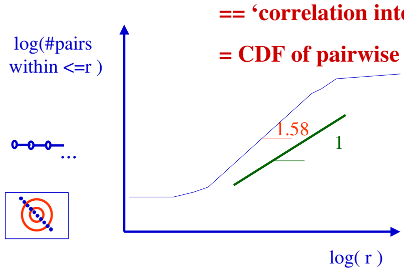
= CDF of pairwise distances



Line

== 'correlation integral'

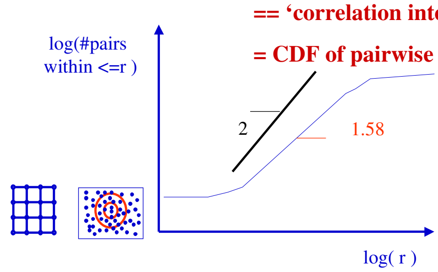
= CDF of pairwise distances



2-d (Plane)

== 'correlation integral'

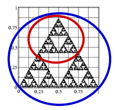
= CDF of pairwise distances



Fractals and power laws

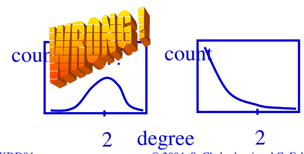
They are related concepts:

- fractals <=>
- self-similarity <=>
- scale-free <=>
- power laws ($y = x^a$)



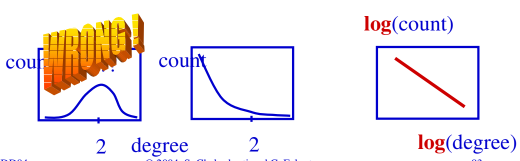
Conclusions

- Real settings/graphs: skewed distributions
  - ‘mean’ is meaningless
  - slope of power law, instead



Conclusions

- Real settings/graphs: skewed distributions
  - ‘mean’ is meaningless
  - slope of power law, instead

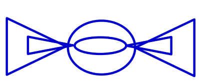


Conclusions: Tools:

- rank-frequency plot (a’la Zipf)
- NCDF, PDF in log-log
- Correlation integral (= neighborhood function)

Conclusions (cont’d)

- Recursion/self-similarity
  - May reveal non-obvious patterns (e.g., bow-ties within bow-ties within bow-ties) [Dill+, ‘01]



“To iterate is human, to recurse is divine”

References

- [Aiello+, '00] William Aiello, Fan R. K. Chung, Linyuan Lu: *A random graph model for massive graphs*. STOC 2000: 171-180
- [Albert+] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi: *Diameter of the World Wide Web*, Nature 401 130-131 (1999)
- [Barabasi, '03] Albert-Laszlo Barabasi *Linked: How Everything Is Connected to Everything Else and What It Means* (Plume, 2003)

References, cont'd

- [Barabasi+, '99] Albert-Laszlo Barabasi and Reka Albert. *Emergence of scaling in random networks*. Science, 286:509--512, 1999
- [Broder+, '00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. *Graph structure in the web*, WWW, 2000

References, cont'd

- [Chakrabarti+, '04] *RMAT: A recursive graph generator*, D. Chakrabarti, Y. Zhan, C. Faloutsos, SIAM-DM 2004
- [Dill+, '01] Stephen Dill, Ravi Kumar, Kevin S. McCurley, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins: *Self-similarity in the Web*. VLDB 2001: 69-78

References, cont'd

- [Fabrikant+, '02] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. *Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet*. ICALP, Malaga, Spain, July 2002
- [FFF, 99] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "*On power-law relationships of the Internet topology*," in SIGCOMM, 1999.

References, cont'd

- [Jovanovic+, '01] M. Jovanovic, F.S. Annexstein, and K.A. Berman. *Modeling Peer-to-Peer Network Topologies through "Small-World" Models and Power Laws*. In TELFOR, Belgrade, Yugoslavia, November, 2001
- [Kumar+ '99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins: *Extracting Large-Scale Knowledge Bases from the Web*. VLDB 1999: 639-650

References, cont'd

- [Leland+, '94] W. E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic*, IEEE Transactions on Networking, 2, 1, pp 1-15, Feb. 1994.
- [Mihail+, '02] Milena Mihail, Christos H. Papadimitriou: *On the Eigenvalue Power Law*. RANDOM 2002: 254-262

References, cont'd

- [Milgram '67] Stanley Milgram: *The Small World Problem*, Psychology Today 1(1), 60-67 (1967)
- [Montgomery+, '01] Alan L. Montgomery, Christos Faloutsos: *Identifying Web Browsing Trends and Patterns*. IEEE Computer 34(7): 94-95 (2001)

IIT Bombay **Carnegie Mellon**

References, cont'd

- [Palmer+, '01] Chris Palmer, Georgos Siganos, Michalis Faloutsos, Christos Faloutsos and Phil Gibbons *The connectivity and fault-tolerance of the Internet topology* (NRDM 2001), Santa Barbara, CA, May 25, 2001
- [Pennock+, '02] David M. Pennock, Gary William Flake, Steve Lawrence, Eric J. Glover, C. Lee Giles: *Winners don't take all: Characterizing the competition for links on the web* Proc. Natl. Acad. Sci. USA 99(8): 5207-5211 (2002)

IIT Bombay **Carnegie Mellon**

References, cont'd

- [Schroeder, '91] Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W H Freeman & Co., 1991

IIT Bombay **Carnegie Mellon**

References, cont'd

- [Siganos+, '03] G. Siganos, M. Faloutsos, P. Faloutsos, C. Faloutsos *Power-Laws and the AS-level Internet Topology*, Transactions on Networking, August 2003.
- [Watts+ Strogatz, '98] D. J. Watts and S. H. Strogatz *Collective dynamics of 'small-world' networks*, Nature, 393:440-442 (1998)
- [Watts, '03] Duncan J. Watts *Six Degrees: The Science of a Connected Age* W.W. Norton & Company; (February 2003)

IIT Bombay **Carnegie Mellon**

PART 2: PageRank, HITS, and eigenvalues

IIT Bombay **Carnegie Mellon**

Outline

Part 1: Topology, 'laws' and generators

➡ Part 2: PageRank, HITS and eigenvalues

Part 3: Pairs, influence, communities

IIT Bombay **Carnegie Mellon**

Part 2: PageRank, HITS and eigenvalues

- How important is a node?
- Who is the best person/computer to immunize against a virus?
- Who is the best customer to advertise to?
- Who originated a raging rumor?

Outline

Part 1: Topology, 'laws' and generators

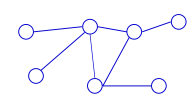
➡ Part 2: PageRank, HITS and eigenvalues

- Eigenvalues and PageRank
- SVD and HITS
- Virus propagation

Part 3: Pairs, influence, communities

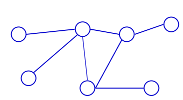
Motivating problem

Given a graph, find its most interesting/central node



Motivating problem

Given a graph, find its most interesting/central node

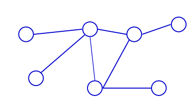


A node is important, if it is connected with important nodes (recursive, but OK!)

Motivating problem – pageRank solution

Given a graph, find its most interesting/central node

Proposed solution: Random walk; spot most 'popular' node (-> steady state prob. (ssp))



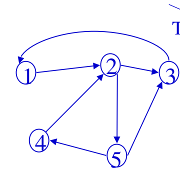
A node has high ssp, if it is connected with high ssp nodes (recursive, but OK!)

Notational conventions

- bold capitals -> matrix (eg. $\mathbf{A}$, $\mathbf{U}$, Λ, $\mathbf{V}$)
- bold lower-case -> <u>column</u> vector (eg., $\mathbf{x}$, $\mathbf{v}_1$, $\mathbf{u}_3$)
- regular lower-case -> scalars (eg., λ_1 , λ_r)

(Simplified) PageRank algorithm

- Let **A** be the transition matrix (= adjacency matrix); let $\mathbf{A}^T$ become column-normalized - then



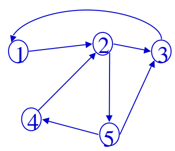
From / To

$$\mathbf{A}^T: \begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix} = \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}$$

IIT Bombay Carnegie Mellon

(Simplified) PageRank algorithm

- $A^T p = p$



$$\overset{A^T}{\begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix}} \overset{p}{\begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}} \overset{=}{=} \overset{p}{\begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}}$$

IIT Bombay Carnegie Mellon

(Simplified) PageRank algorithm

- $A^T p = 1 * p$
- thus, **p** is the eigenvector that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)

IIT Bombay Carnegie Mellon

(Simplified) PageRank algorithm

- In short: imagine a particle randomly moving along the edges
- compute its steady-state probabilities (ssp)

Full version of algo: with occasional random jumps – see later

IIT Bombay Carnegie Mellon

Formal definition

If **A** is a (n x n) square matrix
(λ , **x**) is an **eigenvalue/eigenvector** pair of **A** if

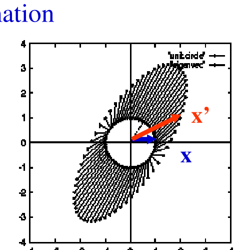
$$\boxed{A\,x = \lambda\,x}$$

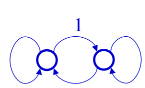
IIT Bombay Carnegie Mellon

Intuition

- **A** as vector transformation

$$\overset{x'}{\begin{bmatrix} 2 \\ 1 \end{bmatrix}} = \overset{A}{\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}} \overset{x}{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}$$



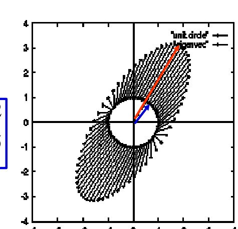


IIT Bombay Carnegie Mellon

Intuition

- By defn., eigenvectors remain parallel to themselves ('**fixed points**')

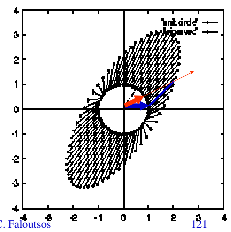
$$\overset{\lambda_1}{3.62} * \overset{v_1}{\begin{bmatrix} 0.52 \\ 0.85 \end{bmatrix}} = \overset{A}{\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}} \overset{v_1}{\begin{bmatrix} 0.52 \\ 0.85 \end{bmatrix}}$$



IIT Bombay Carnegie Mellon

Convergence

- Usually, fast:

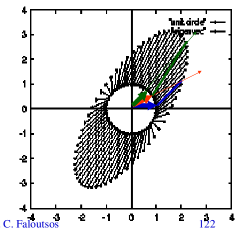


KDD04 121

IIT Bombay Carnegie Mellon

Convergence

- Usually, fast:



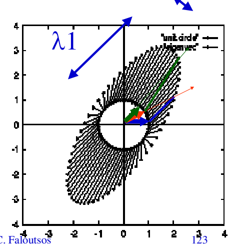
KDD04 122

IIT Bombay Carnegie Mellon

Convergence

- Usually, fast:
- depends on ratio
  $\lambda 1 : \lambda 2$

$\lambda 2$



KDD04 123

IIT Bombay Carnegie Mellon

Our wish list:

- How important is a node?
- Who is the best person/computer to immunize against a virus?
- Who is the best customer to advertise to?
- Who originated a raging rumor?

ssp values answer these questions

KDD04 124

IIT Bombay Carnegie Mellon

Outline

Part 1: Topology, 'laws' and generators

Part 2: PageRank, HITS and eigenvalues

- Eigenvalues and PageRank
- SVD and HITS
- Virus propagation

Part 3: Pairs, influence, communities

KDD04 125

IIT Bombay Carnegie Mellon

SVD vs eigenvalues

- very similar, but not identical
- Motivating example: HITS/Kleinberg algo:

KDD04 126

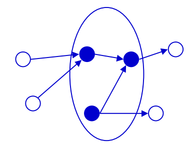
Kleinberg’s algorithm

- Problem dfn: given the web and a query
- find the most ‘authoritative’ web pages for this query

Step 0: find all pages containing the query terms
Step 1: expand by one move forward and backward

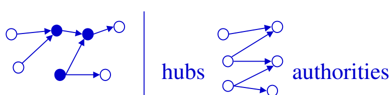
Kleinberg’s algorithm

- Step 1: expand by one move forward and backward



Kleinberg’s algorithm

- give high score (= ‘authorities’) to nodes that many important nodes point to
- give high importance score (‘hubs’) to nodes that point to good ‘authorities’)



Kleinberg’s algorithm

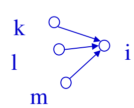
Observations

- recursive definition!
- each node (say, ‘i’-th node) has both an authoritativeness score a_i and a hubness score h_i

Kleinberg’s algorithm

Let **A** be the adjacency matrix:
the (i,j) entry is 1 if the edge from i to j exists
Let **h** and **a** be [n x 1] vectors with the ‘hubness’ and ‘authoritativiness’ scores.
Then:

Kleinberg’s algorithm



Then:

$$a_i = h_k + h_l + h_m$$

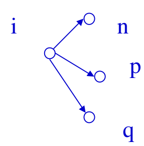
that is

a_i = Sum (h_j) over all j that (j,i) edge exists

or

$\mathbf{a} = \mathbf{A}^T \mathbf{h}$

Kleinberg's algorithm



symmetrically, for the 'hubness':

$$h_i = a_n + a_p + a_q$$

that is

h_i = Sum (q_j) over all j that (i,j) edge exists

or

h = A a

Kleinberg's algorithm

In conclusion, we want vectors **h** and **a** such that:

$$\mathbf{h} = \mathbf{A}\,\mathbf{a}$$

$$\mathbf{a} = \mathbf{A}^T\,\mathbf{h}$$

That is:

$$\mathbf{a} = \mathbf{A}^T\mathbf{A}\,\mathbf{a}$$

Kleinberg's algorithm

a is a right- singular vector of the adjacency matrix **A** (by dfn!)

== eigenvector of $\mathbf{A}^T\mathbf{A}$

Starting from random **a'** and iterating, we'll eventually converge

(Q: to which of all the eigenvectors? why?)

Kleinberg's algorithm

(Q: to which of all the eigenvectors? why?)

A: to the one of the strongest eigenvalue

Kleinberg's algorithm - results

Eg., for the query 'java':

0.328 www.gamelan.com

0.251 java.sun.com

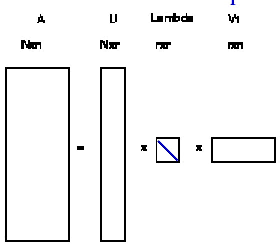
0.190 www.digitalfocus.com ("the java developer")

SVD: formal definitions

- Let **A** be a matrix (eg., adjacency matrix of a graph)

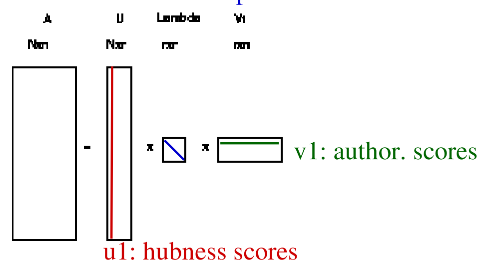
SVD - Definition

- $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$ - example:



SVD - Definition

- $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$ - example:



Skip

SVD - Definition

$\boxed{\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]}\,\Lambda_{[r \times r]}\,(\mathbf{V}_{[m \times r]})^T}$

- **A**: n x m matrix (eg., n documents, m terms)
- **U**: n x r matrix (n documents, r concepts)
- **Λ**: r x r diagonal matrix (strength of each 'concept') (r : rank of the matrix)
- **V**: m x r matrix (m terms, r concepts)

Skip

SVD - Properties

THEOREM [Press+92]: always possible to decompose matrix **A** into $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$, where

- **U, Λ, V**: unique (*)
- **U**, **V**: column orthonormal (ie., columns are unit vectors, orthogonal to each other)
  - $\mathbf{U}^T \mathbf{U} = \mathbf{I}$; $\mathbf{V}^T \mathbf{V} = \mathbf{I}$ (**I:** identity matrix)
- **Λ:** singular values are positive, and sorted in decreasing order

SVD – other uses:

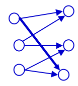
- LSI (Latent Semantic Indexing) [Deerwester+]
- PCA (Principal Component Analysis) [Jolliffe]
- Karhunen-Loeve transform [Fukunaga], [Duda+Hart]
- Low-rank approximation, dim. Reduction
- Over- and under-constraint linear systems

SVD – other uses (cont'd):

- Graph partitioning (on 'Laplacian')
- + MANY MORE …

SVD - Interpretation

- $\mathbf{A = U\, \Lambda\, V^T}$ - example:



$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{sys} & \text{db.} & \text{os} & \text{lung} & \text{liver} \end{array} \\
\begin{array}{c} \text{CS.} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
=
\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\times
\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
\times
\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

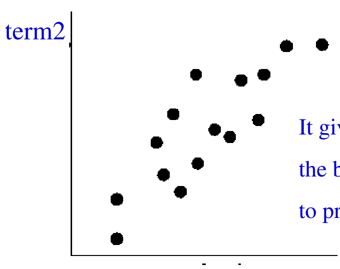
SVD - Interpretation

- $\mathbf{A = U\, \Lambda\, V^T}$ - example:

U: doc-to-concept similarity matrix

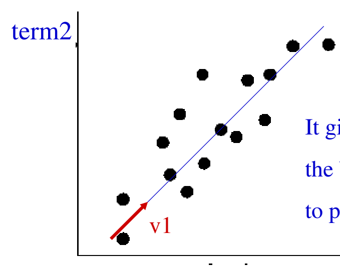
$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{sys} & \text{db.} & \text{os} & \text{lung} & \text{liver} \end{array} \\
\begin{array}{c} \text{CS.} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
=
\begin{array}{c}
\begin{array}{cc} \text{CS-concept} & \text{MD-concept} \end{array} \\
\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\end{array}
\times
\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
\times
\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

SVD - interpretation:



It gives
the best hyperplane
to project on

SVD - interpretation:



It gives
the best hyperplane
to project on

Outline

Part 1: Topology, 'laws' and generators

Part 2: PageRank, HITS and eigenvalues
- Eigenvalues and PageRank
- SVD and HITS
- Virus propagation

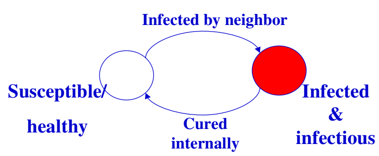
Part 3: Pairs, influence, communities

Problem definition

- Q1: How does a virus spread across an arbitrary network?
- Q2: will it create an epidemic?

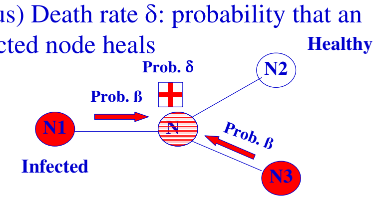
Framework

- Susceptible-Infected-Susceptible (SIS) model
  - Cured nodes immediately become susceptible



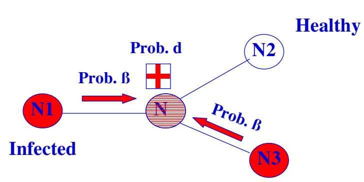
The model

- (virus) Birth rate β: probability than an infected neighbor attacks
- (virus) Death rate δ: probability that an infected node heals



The model

- Virus 'strength' s= β/δ



Epidemic threshold τ

of a graph, defined as the value of τ, such that

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

an epidemic can not happen

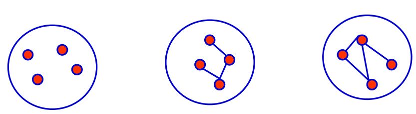
Thus,

- 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?



Epidemic threshold

- [Theorem] We have no epidemic, if

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

IIT Bombay Carnegie Mellon

Epidemic threshold

- [Theorem] We have no epidemic, if

recovery prob. epidemic threshold

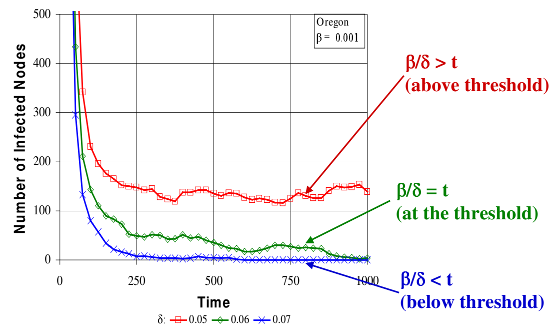
$$\boxed{\beta/\delta < \tau = 1/\lambda_{1,A}}$$

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

Proof: [Wang+03]

IIT Bombay Carnegie Mellon

Experiments (Oregon)



IIT Bombay Carnegie Mellon

Our wish list:

- ✓ How important is a node?
- Who is the best person/computer to immunize against a virus?
- ✓ Who is the best customer to advertise to?
- Who originated a raging rumor?

ssp values answer these questions

IIT Bombay Carnegie Mellon

Our wish list:

- ✓ How important is a node?
- ✓ Who is the best person/computer to immunize against a virus? Highest diff in λ1
- ✓ Who is the best customer to advertise to?
- ✓ Who originated a raging rumor? Probably, highest ssp

Virus prop. helps answer the rest

IIT Bombay Carnegie Mellon

Conclusions

eigenvalues/eigenvectors: vital for

- PageRank,
- virus propagation,
- (graph partitioning)

IIT Bombay Carnegie Mellon

Conclusions, cont'd

SVD

- closely related: HITS/Kleinberg
- (and also LSI, KLT, PCA, Least squares, ...)

Both are **extremely useful**, **well understood** tools for graphs / matrices.

IIT Bombay **Carnegie Mellon**

Resources: Software and urls

- SVD packages: in **many** systems (matlab, mathematica, LINPACK, LAPACK)
- stand-alone, free code: SVDPACK from Michael Berry

  http://www.cs.utk.edu/~berry/projects.html

IIT Bombay **Carnegie Mellon**

Books

- Faloutsos, C. (1996). *Searching Multimedia Databases by Content*, Kluwer Academic Inc.
- Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer Verlag.

IIT Bombay **Carnegie Mellon**

Books

- [Press+92] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery: *Numerical Recipes in C*, Cambridge University Press, 1992, 2nd Edition. (Great description, intuition and code for SVD)

IIT Bombay **Carnegie Mellon**

References

- Berry, Michael: http://www.cs.utk.edu/~lsi/
- Brin, S. and L. Page (1998). *Anatomy of a Large-Scale Hypertextual Web Search Engine*. 7th Intl World Wide Web Conf.

IIT Bombay **Carnegie Mellon**

References (cont'd)

- [Foltz+92] Foltz, P. W. and S. T. Dumais (Dec. 1992). "*Personalized Information Delivery: An Analysis of Information Filtering Methods*." Comm. of ACM (CACM) 35(12): 51-60.

IIT Bombay **Carnegie Mellon**

References (cont'd)

- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press.
- Kleinberg, J. (1998). *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.

References (cont'd)

- [Wang+03] Yang Wang, Deepayan Chakrabarti, Chenxi Wang and Christos Faloutsos: *Epidemic Spreading in Real Networks: an Eigenvalue Viewpoint,* SRDS 2003, Florence, Italy.

BREAK!