Carnegie Mellon

Talk 3: Graph Mining Tools – Tensors, communities, parallelism

Christos Faloutsos
CMU

Overall Outline

- Introduction – Motivation
- Talk#1: Patterns in graphs; generators
- Talk#2: Tools (Ranking, proximity)
- ➡ Talk#3: Tools (Tensors, scalability)
- Conclusions

Outline

- Task 4: time-evolving graphs – tensors
- Task 5: community detection
- Task 6: virus propagation
- Task 7: scalability, parallelism and hadoop
- Conclusions

Thanks to

- Tamara Kolda (Sandia)

for the foils on tensor definitions, and on TOPHITS

Carnegie Mellon

Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

Examples of Matrices: Authors and terms

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation: Why tensors?

- Q: what is a tensor?

Motivation: Why tensors?

- A: N-D generalization of matrix:

KDD'09

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation: Why tensors?

- A: N-D generalization of matrix:

KDD'07

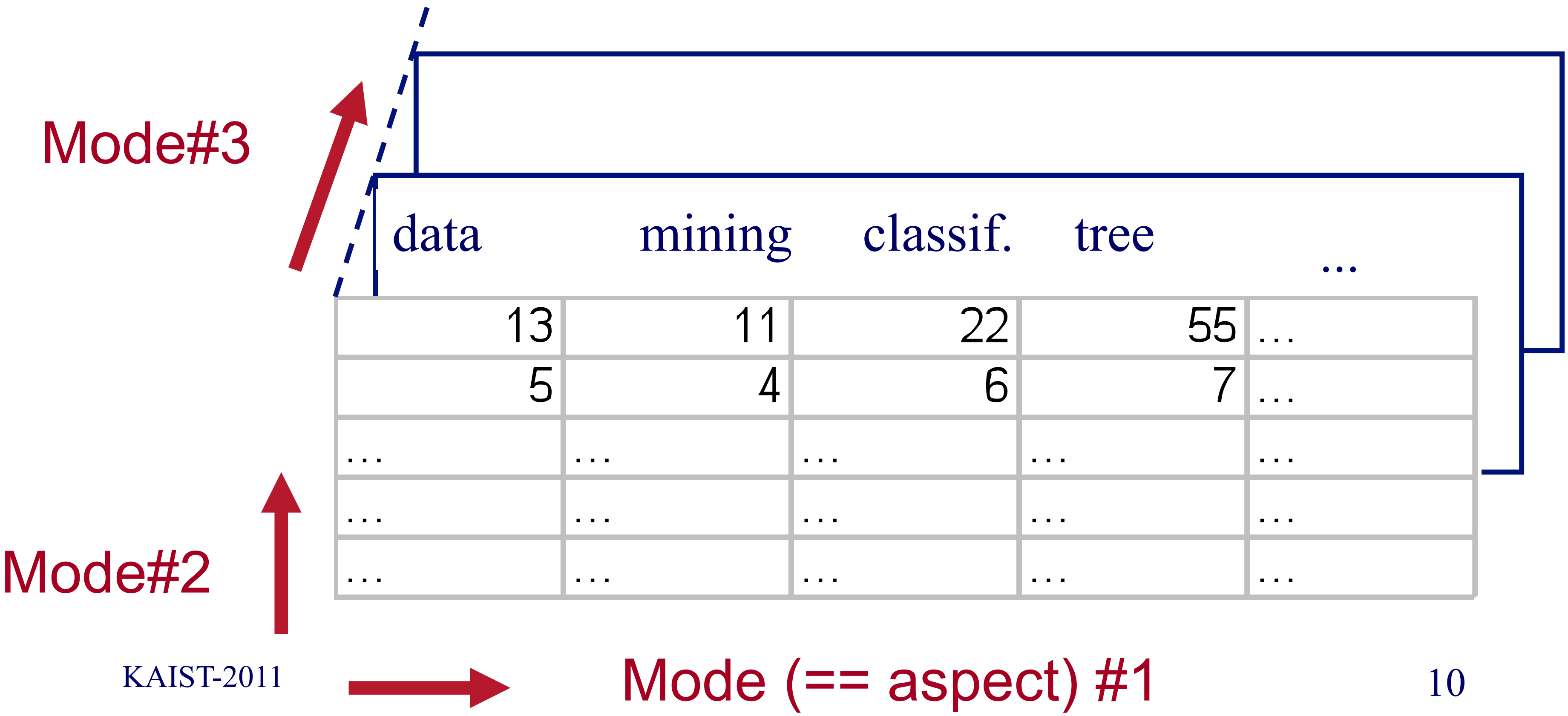
KDD'08

KDD'09

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Tensors are useful for 3 or more modes

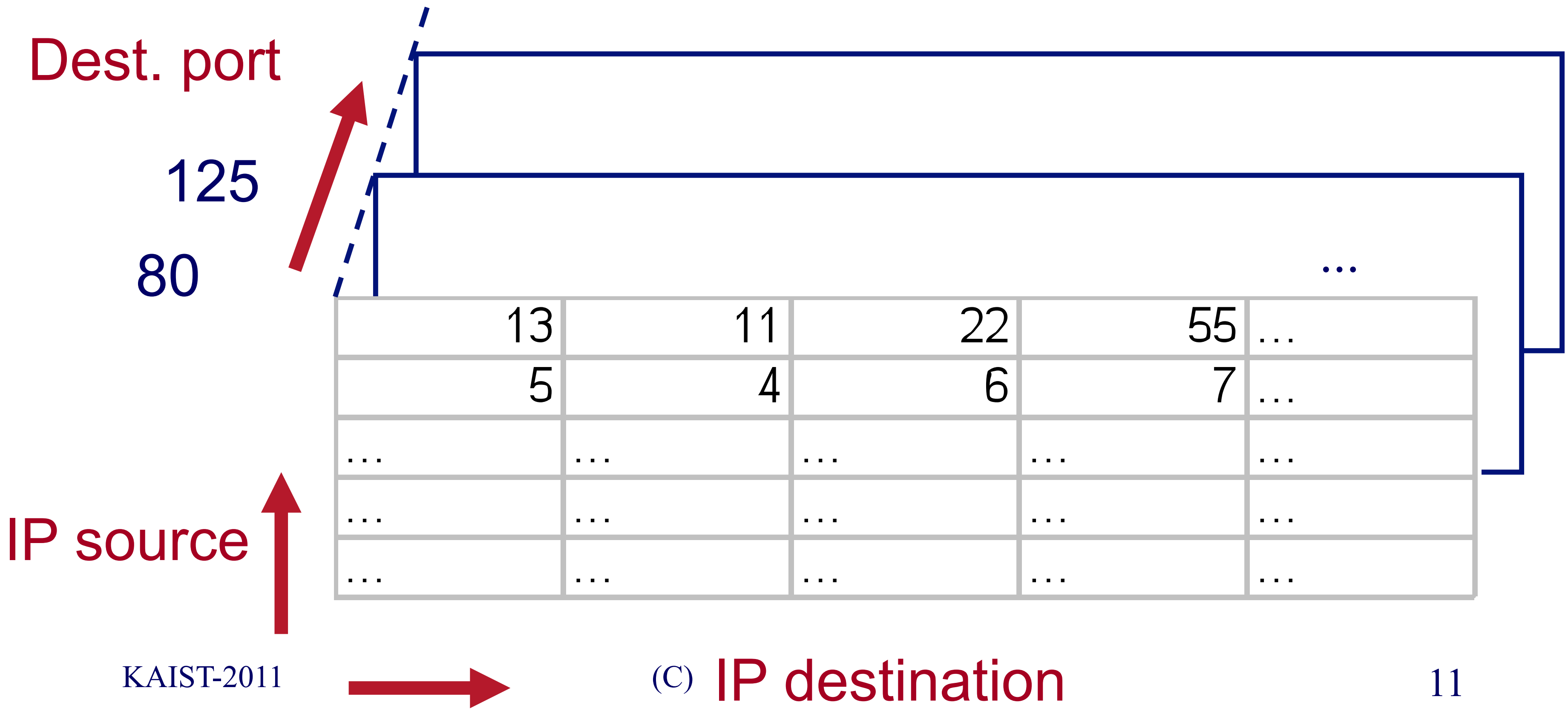
Terminology: 'mode' (or 'aspect'):



data	mining	classif.	tree	...
13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Notice

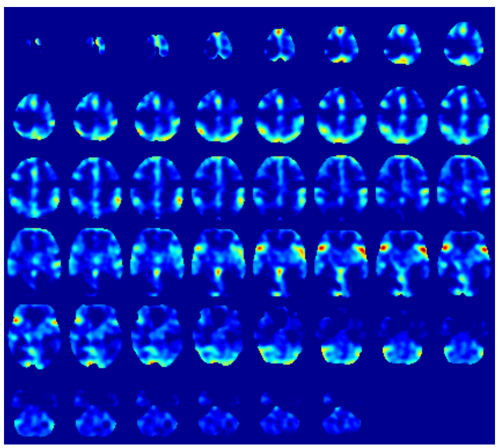
- 3rd mode does not need to be time
- we can have more than 3 modes



13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Notice

- 3rd mode does not need to be time
- we can have more than 3 modes
  - Eg, fFMRI: x,y,z, time, person-id, task-id



From DENLAB, Temple U.
(Prof. V. Megalooikonomou +)

http://denlab.temple.edu/bidms/cgi-bin/browse.cgi

Motivating Applications

- Why tensors are useful?
  - web mining (TOPHITS)
  - environmental sensors
  - Intrusion detection (src, dst, time, dest-port)
  - Social networks (src, dst, time, type-of-contact)
  - face recognition
  - etc …

Carnegie Mellon

Detailed outline

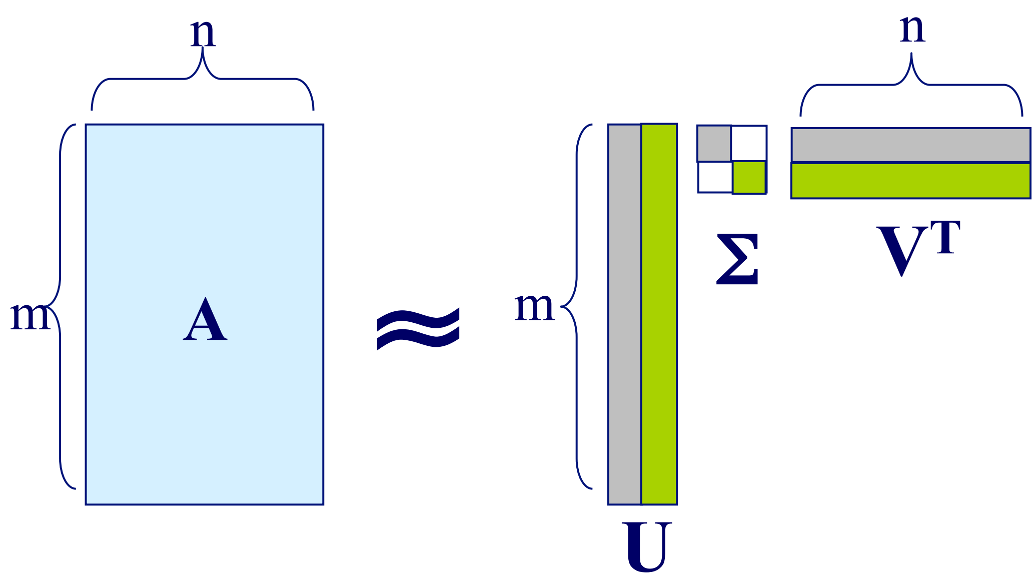
- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

Tensor basics

- Multi-mode extensions of SVD – recall that:

Reminder: SVD

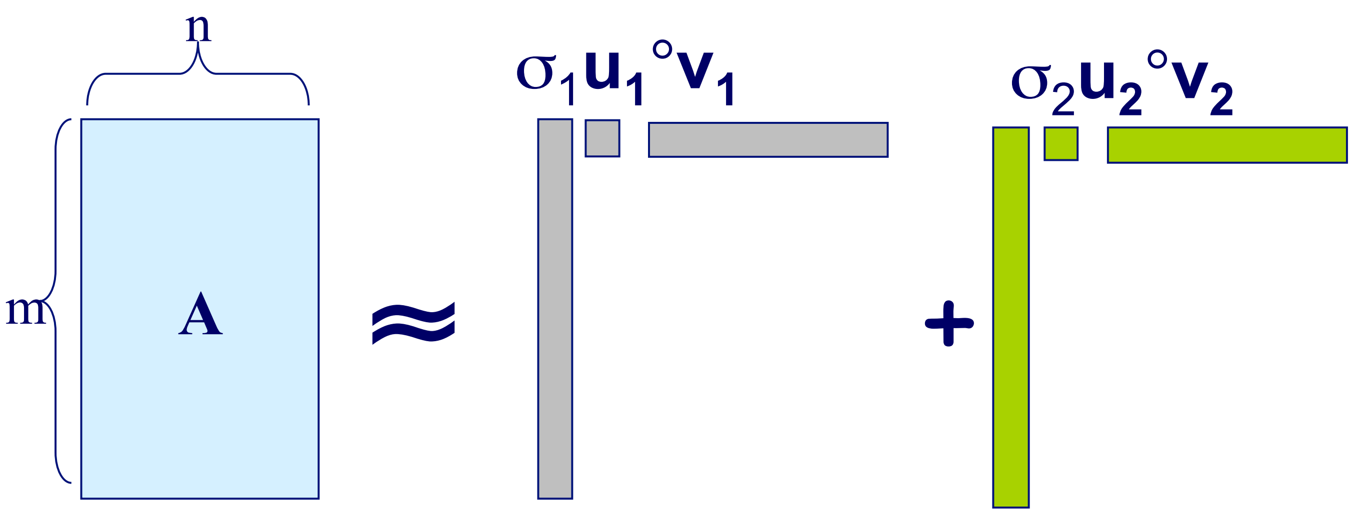
$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



– Best rank-k approximation in L2

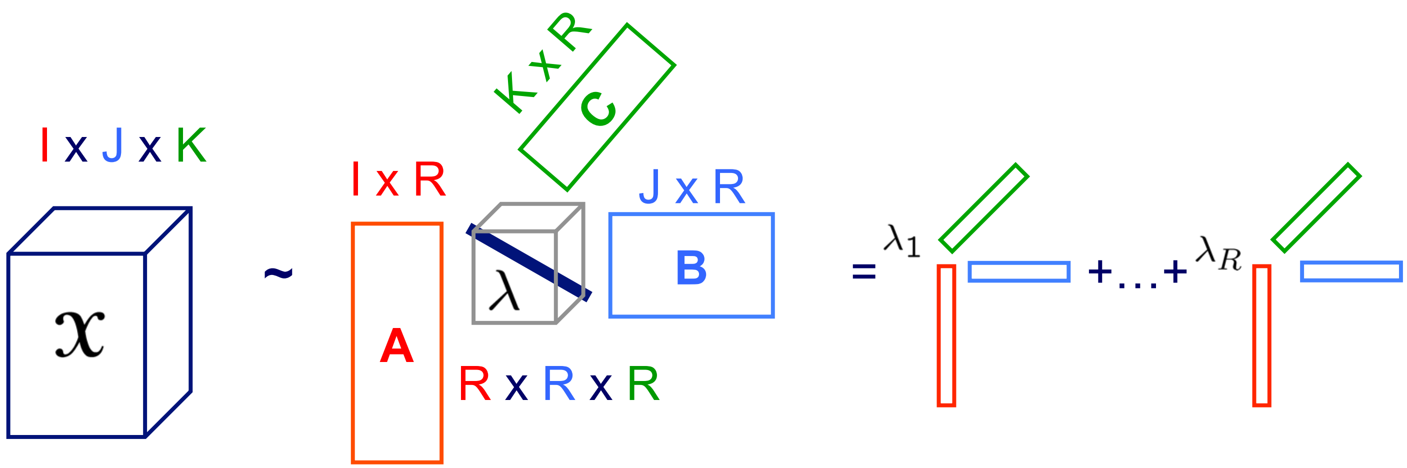
Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



– Best rank-k approximation in L2

Goal: extension to >=3 modes



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

Main points:

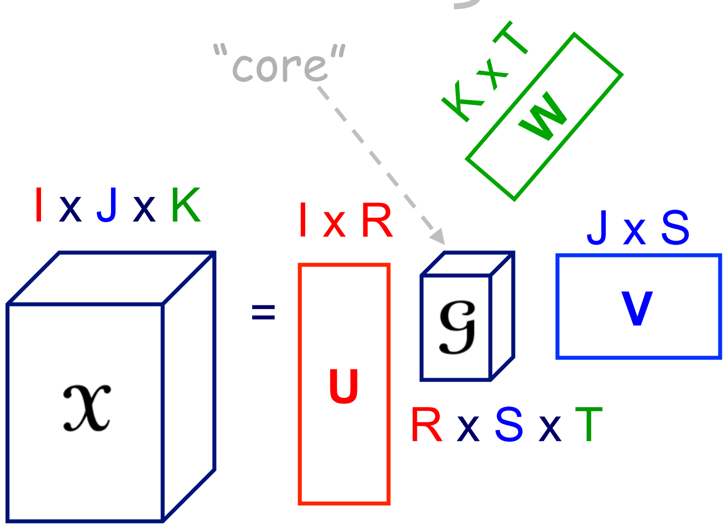
- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)

Specially Structured Tensors

- Tucker Tensor

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

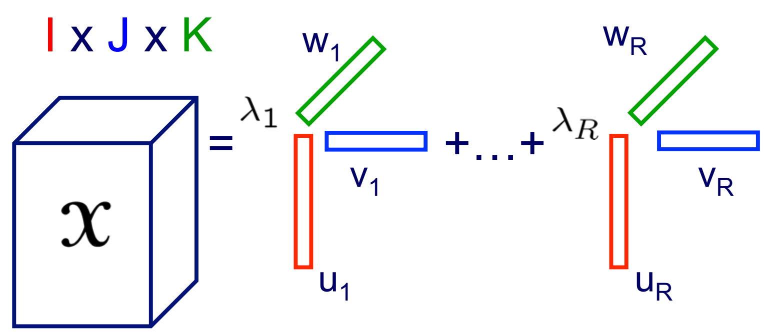
Our Notation



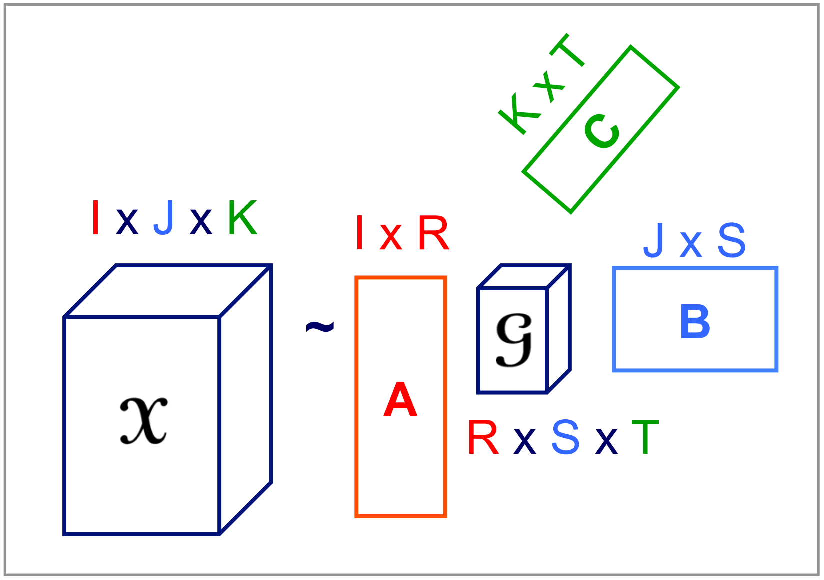
- Kruskal Tensor

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

Our Notation



Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- **G**: how groups relate to each other

Intuition behind core tensor

- 2-d case: co-clustering
- [Dhillon et al. Information-Theoretic Co-clustering, KDD'03]

Carnegie Mellon

eg, terms x documents

$$
m \; \overset{n}{\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}}
$$

$$
m \; \overset{k}{\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}} \; k \; \overset{l}{\begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}} \; l \; \overset{n}{\begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix}} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}
$$

Carnegie Mellon

med. doc

cs doc

med. terms

cs terms

common terms

term group x doc. group

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

term x term-group

doc x doc group

Tensor tools - summary

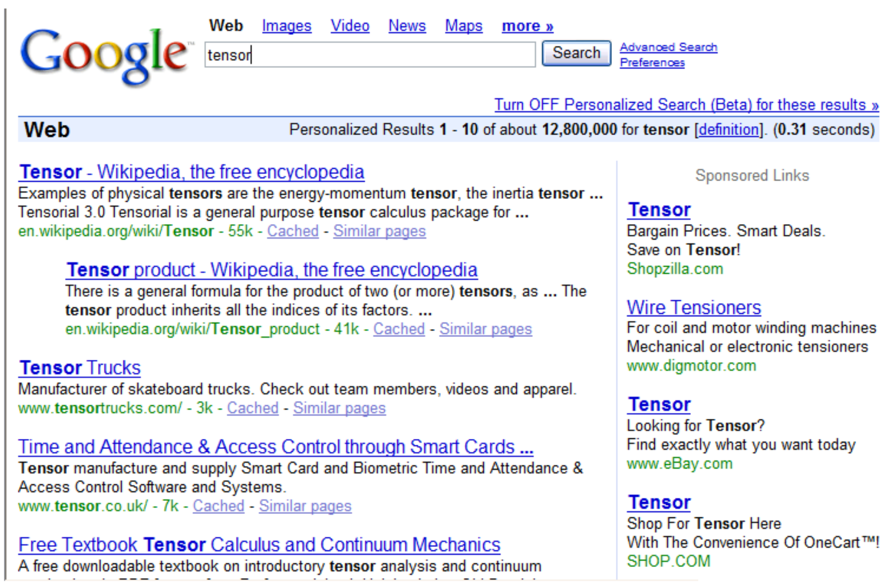
- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- (To solve: Alternating Least Squares)

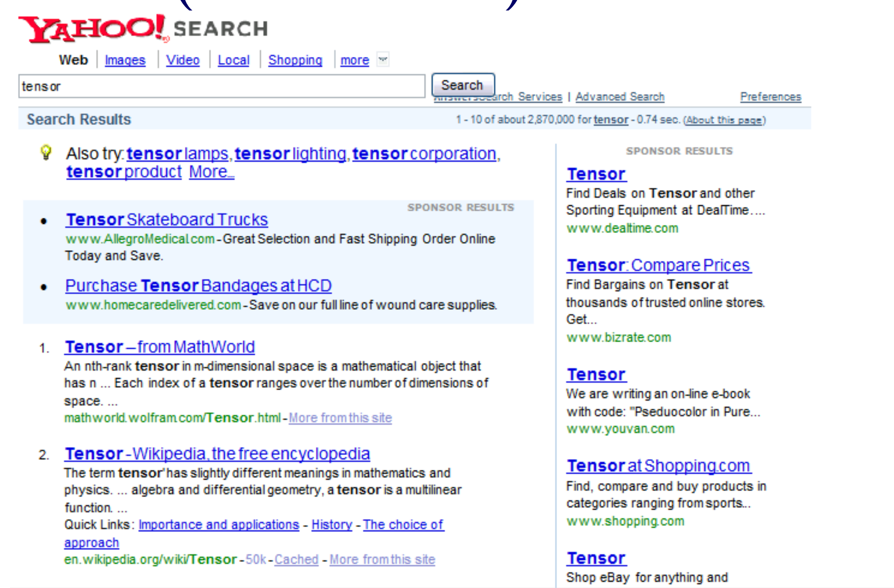
Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining

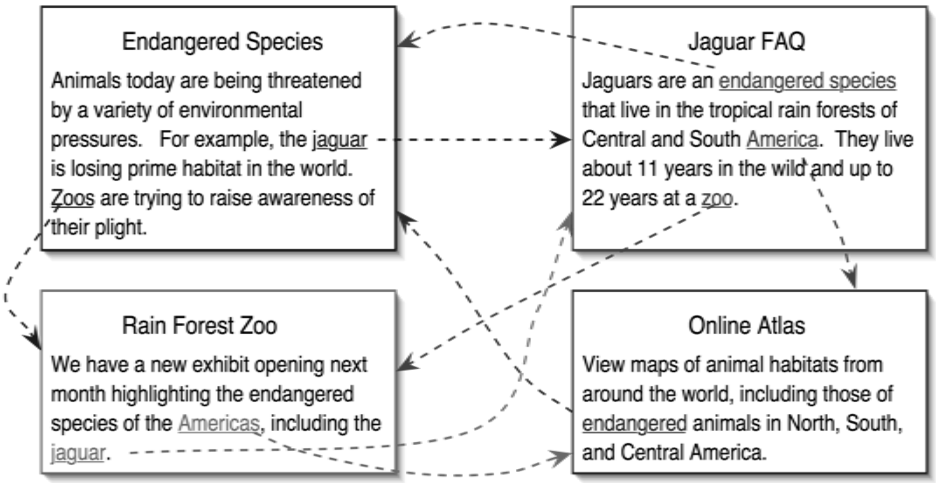
Web graph mining

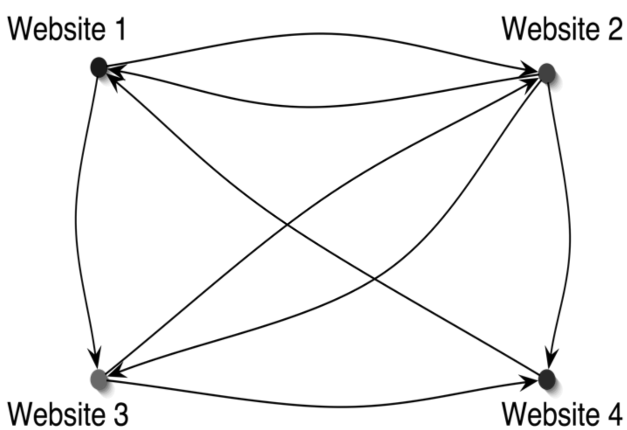
- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)





Kleinberg's Hubs and Authorities (the HITS method)

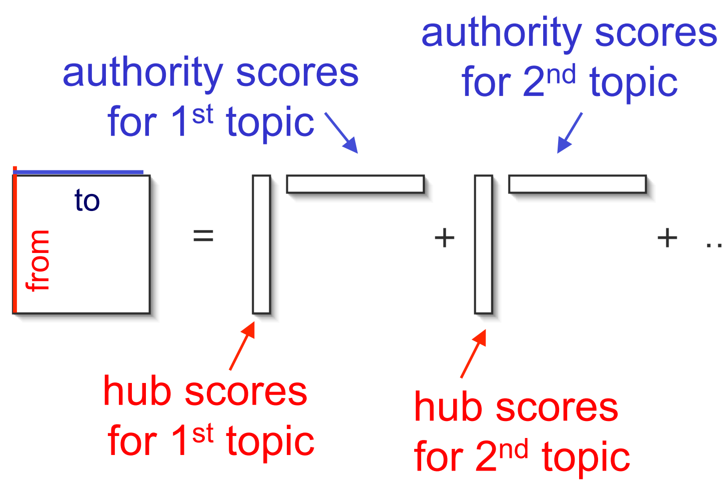




Sparse adjacency matrix and its SVD:

$$x_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{X} \approx \sum_r \sigma_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Kleinberg, JACM, 1999

HITS Authorities on Sample Data

1st Principal Factor	
.97	www.ibm.com
.24	www.alphawo
.08	www-128.ibm
.05	www.develop
.02	www.researc
.01	www.redbook
.01	news.com.co

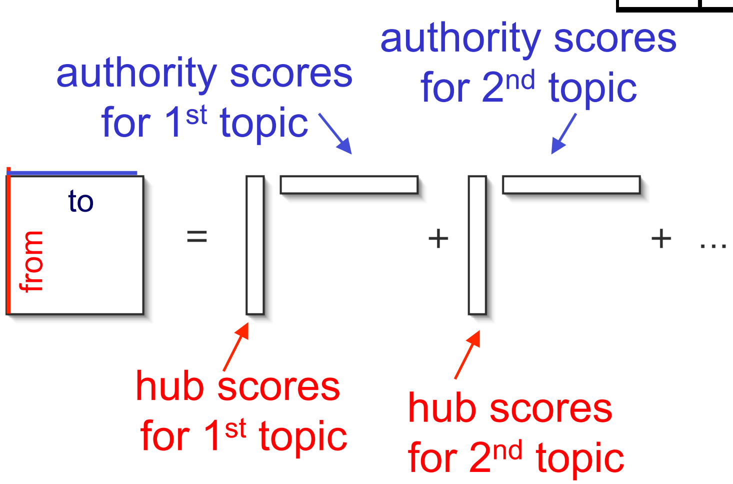
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh.edu
.06	www.lehighal
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

3rd Principal Factor	
.75	java.sun.com
.38	www.sun.com
.36	developers.sun.
.24	see.sun.com
.16	www.samag.co
.13	docs.sun.com
.12	blogs.sun.com
.08	sunsolve.sun.c
.08	www.sun-catalo
.08	news.com.com

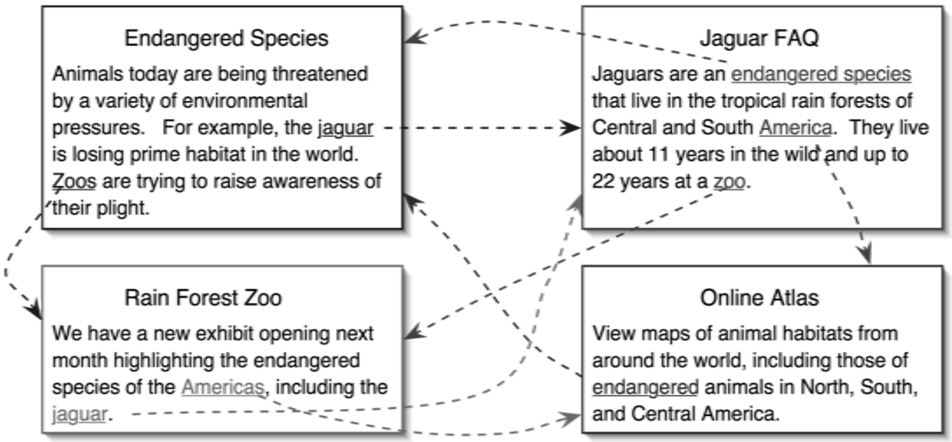
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

6th Principal Factor	
.97	mathpost.asu.edu
.18	math.la.asu.edu
.17	www.asu.edu
.04	www.act.org
.03	www.eas.asu.edu
.02	archives.math.utk.edu
.02	www.geom.uiuc.edu
.02	www.fulton.asu.edu
.02	www.amstat.org
.02	www.maa.org

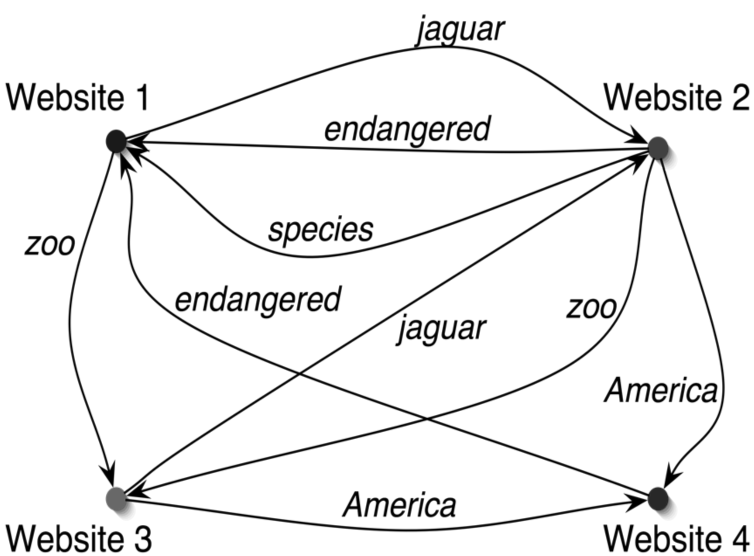
We started our crawl from http://www-neos.mcs.anl.gov/neos, and crawled 4700 pages, resulting in 560 cross-linked hosts.



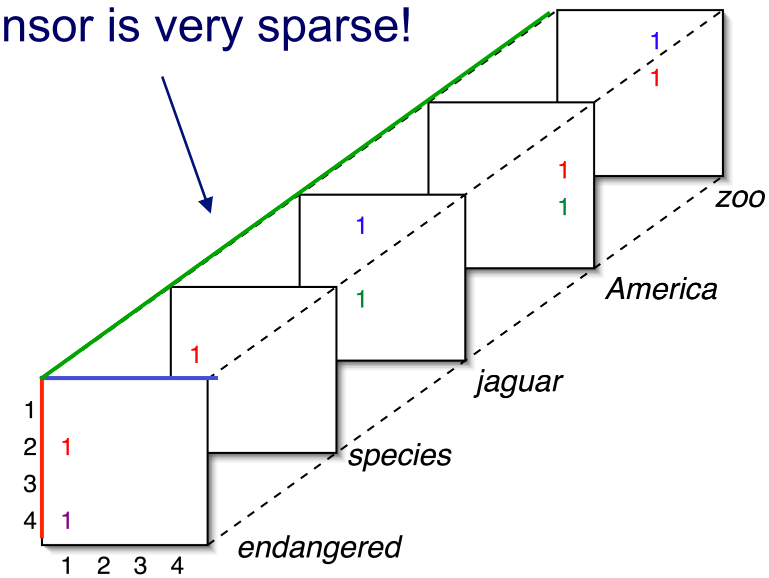
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$



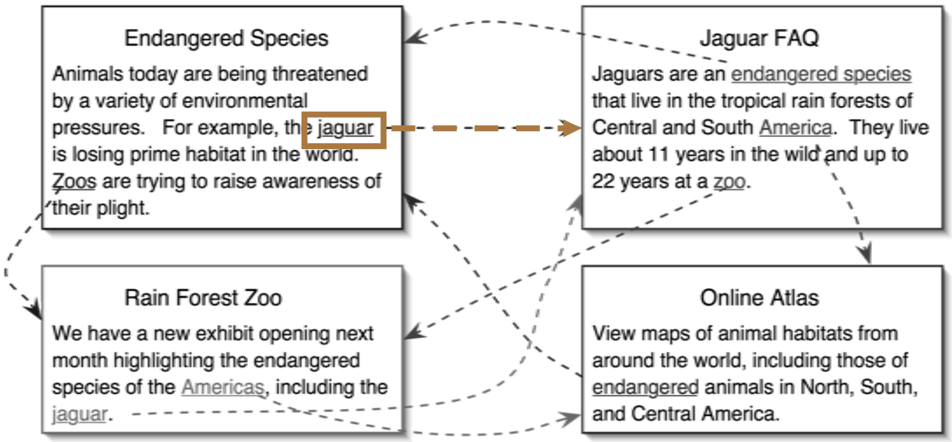
Observe that this tensor is very sparse!



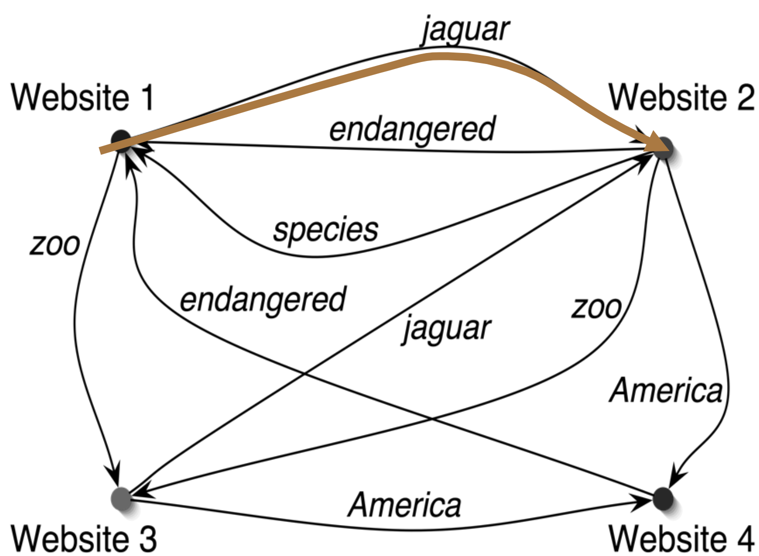
Kolda, Bader, Kenny, ICDM05

CarnegieMellon

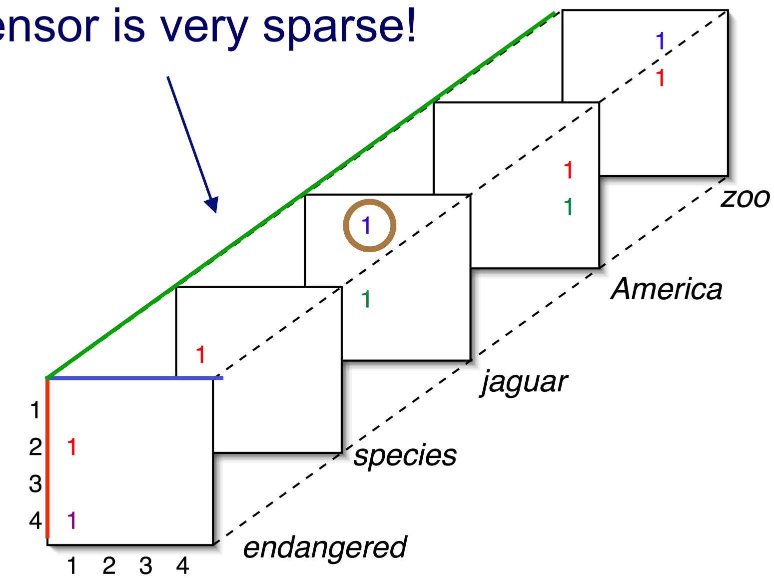
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \\ & \text{with term } k \\ 0 & \text{otherwise} \end{cases}$$

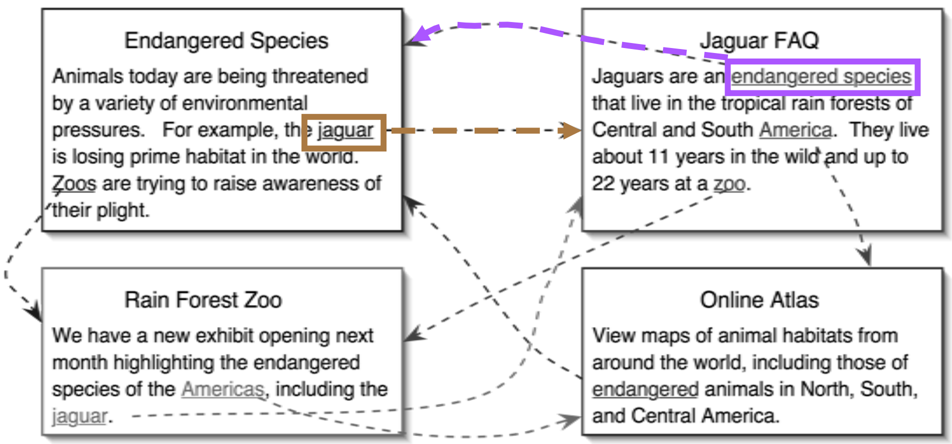


Observe that this tensor is very sparse!

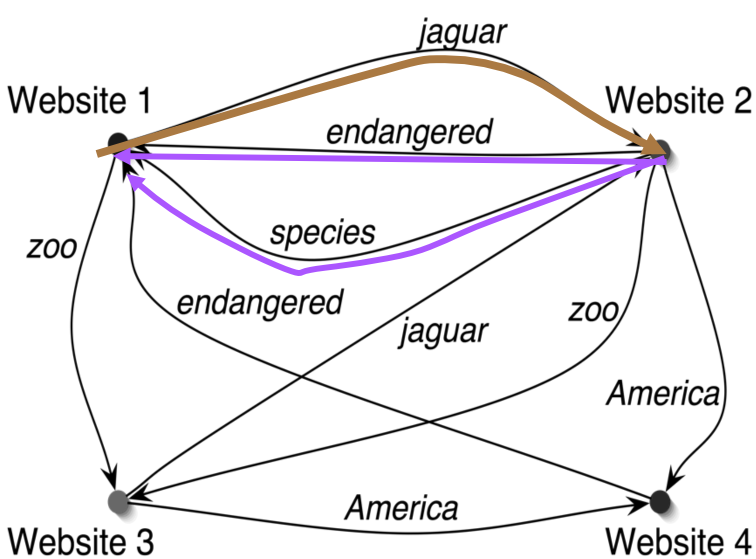


Kolda, Bader, Kenny, ICDM05

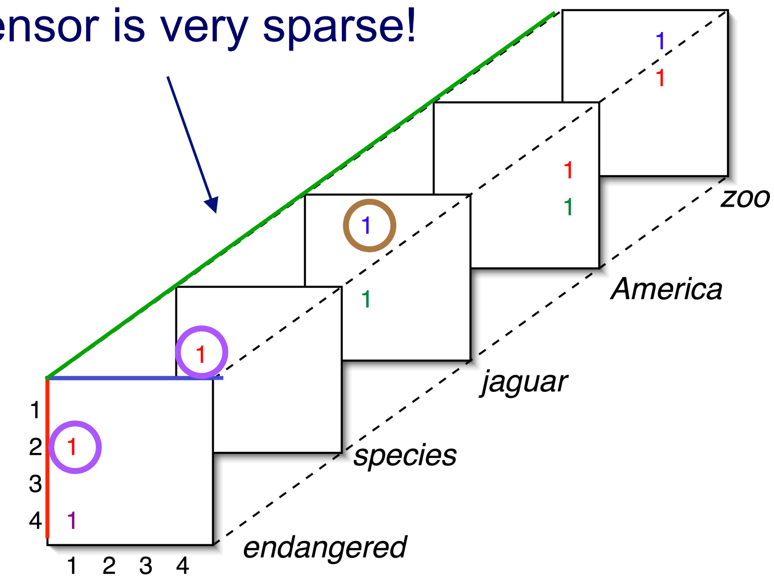
Three-Dimensional View of the Web



$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$



Observe that this tensor is very sparse!



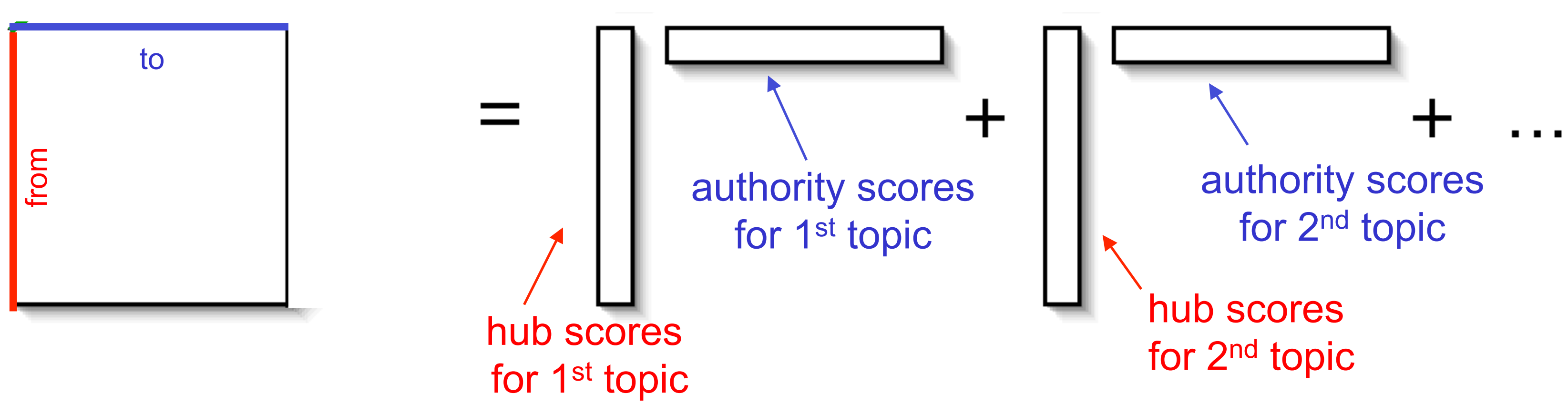
Kolda, Bader, Kenny, ICDM05

Carnegie Mellon

Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

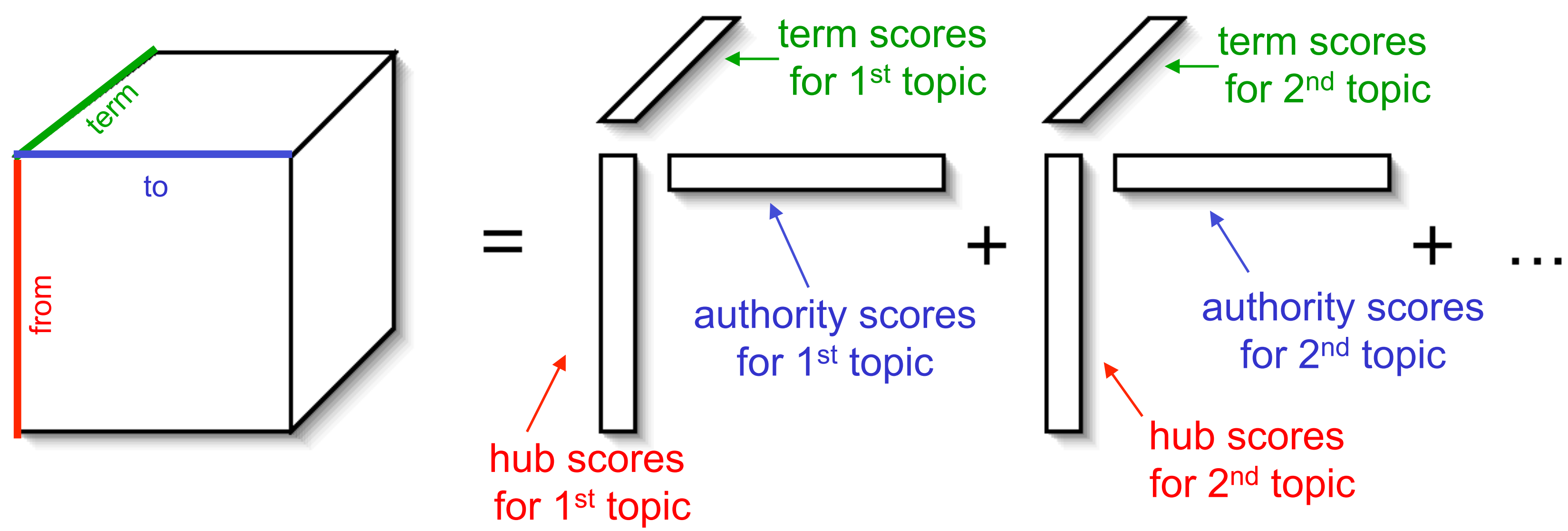
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r \circ \mathbf{t}_r$$



TOPHITS Terms & Authorities on Sample Data

1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN		
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-READABLE-TEXT		

2nd Principal Factor			
.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACULTY		
.16	SEARC		
.16	NEWS		
.16	LIBRAI		
.16	COMP		
.12	LEHIG		

3rd Principal Factor			
.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVI		
.12	WEBS		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESOU		
.11	TECHN		
.10	DOWN		

4th Principal Factor			
.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZEN		
.22	OTHER		
.19	CENTE		
.19	LANGU		
.15	U.S		
.15	PUBLI		
.14	CONS		
.13	FREE		

6th Principal Factor			
.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHITE		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRESI		
.11	OFFIC		

12th Principal Factor			
.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTWARE	.35	www.solver.com
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONTR		
.05	ILOG		

13th Principal Factor			
.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACROI		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor			
.50	WEATHER	.81	www.weather.gov
.24	OFFICE	.41	www.spc.noaa.gov
.23	CENTER	.30	lwf.ncdc.noaa.gov
.19	NO-RE		
.17	ORGA		
.15	NWS		
.15	SEVER		
.15	FIRE		
.15	POLIC		
.14	CLIMA		

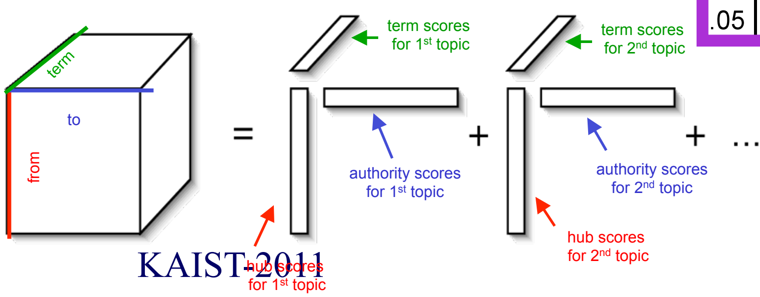
19th Principal Factor			
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

TOPHITS uses 3D analysis to find the dominant groupings of web pages and terms.

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k)+1} & \text{if } i \rightarrow j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

w_k = # unique links using term k

Tensor PARAFAC



Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Tensors provide elegant theory and algorithms
  - PARAFAC and Tucker: discover groups

References

- T. G. Kolda, B. W. Bader and J. P. Kenny. *Higher-Order Web Link Analysis Using Multilinear Algebra*. In: ICDM 2005, Pages 242-249, November 2005.
- Jimeng Sun, Spiros Papadimitriou, Philip Yu. *Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams*, Proc. of the Int. Conf. on Data Mining (ICDM), Hong Kong, China, Dec 2006

Resources

- See tutorial on tensors, KDD'07 (w/ Tamara Kolda and Jimeng Sun):

www.cs.cmu.edu/~christos/TALKS/KDD-07-tutorial

Carnegie Mellon

Tensor tools - resources

- Toolbox: from Tamara Kolda:

csmr.ca.sandia.gov/~tgkolda/TensorToolbox

- T. G. Kolda and B. W. Bader. ***Tensor Decompositions and Applications***. SIAM Review, Volume 51, Number 3, September 2009
csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/TensorReview-preprint.pdf

- T. Kolda and J. Sun: Scalable Tensor Decomposition for Multi-Aspect Data Mining (ICDM 2008)

Outline

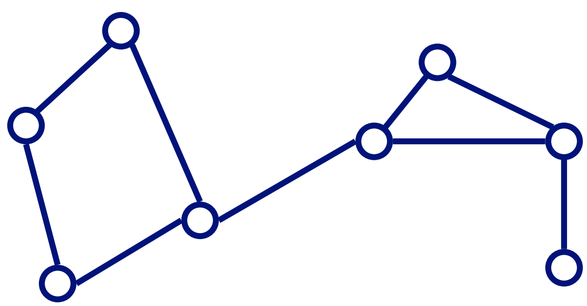
- Task 4: time-evolving graphs – tensors
- ➡ Task 5: community detection
- Task 6: virus propagation
- Task 7: scalability, parallelism and hadoop
- Conclusions

Detailed outline

- Motivation
- Hard clustering – *k* pieces
- Hard co-clustering – (*k*,*l*) pieces
- Hard clustering – optimal # pieces
- Observations

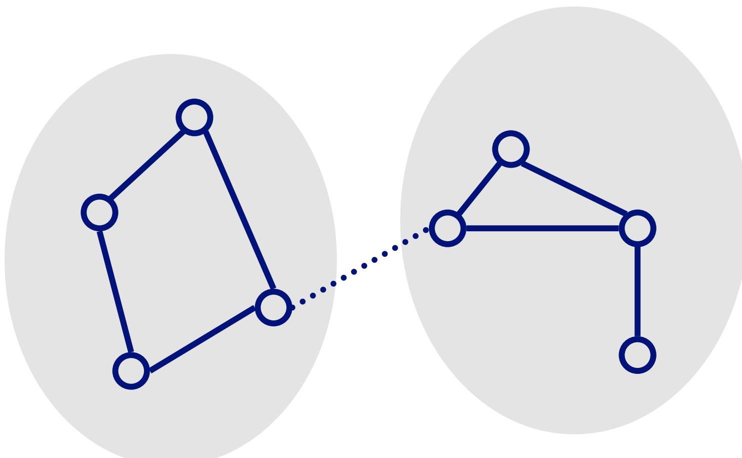
Problem

- Given a graph, and *k*
- Break it into *k* (disjoint) communities



Problem

- Given a graph, and *k*
- Break it into *k* (disjoint) communities

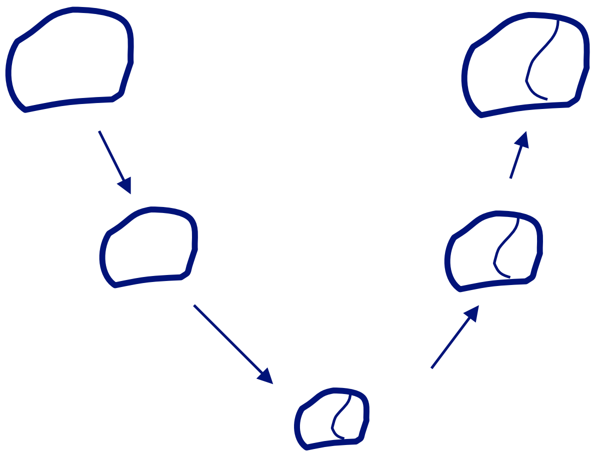


$k = 2$

Carnegie Mellon

Solution #1: METIS

- Arguably, the best algorithm
- Open source, at
  - http://www.cs.umn.edu/~metis
- and *many* related papers, at same url
- Main idea:
  - coarsen the graph;
  - partition;
  - un-coarsen



Carnegie Mellon

Solution #1: METIS

- G. Karypis and V. Kumar. *METIS 4.0: Unstructured graph partitioning and sparse matrix ordering system*. TR, Dept. of CS, Univ. of Minnesota, 1998.
- <and many extensions>

Solution #2

(problem: hard clustering, *k* pieces)

Spectral partitioning:

- Consider the 2nd smallest eigenvector of the (normalized) Laplacian

Solutions #3, …

Many more ideas:

- Clustering on the $\mathbf{A}^2$ (square of adjacency matrix) [Zhou, Woodruff, PODS'04]
- Minimum cut / maximum flow [Flake+, KDD'00]
- …

Detailed outline

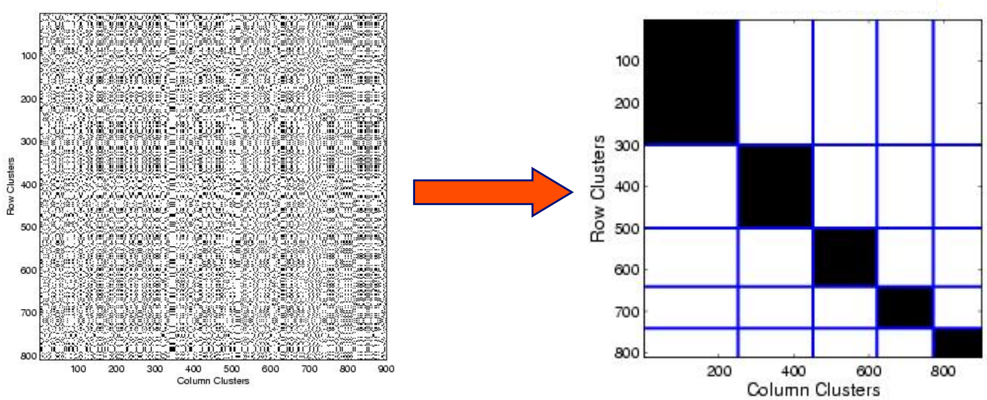
- Motivation
- Hard clustering – *k* pieces
- Hard co-clustering – (*k*,*l*) pieces
- Hard clustering – optimal # pieces
- Soft clustering – matrix decompositions
- Observations

Problem definition

- Given a bi-partite graph, and *k*, *l*
- Divide it into *k* row groups and *l* row groups
- (Also applicable to uni-partite graph)

Co-clustering

- Given data matrix and the number of row and column groups *k* and *l*
- Simultaneously
  - – Cluster rows into *k* disjoint groups
  - – Cluster columns into *l* disjoint groups



Carnegie Mellon

details

Co-clustering

- Let X and Y be discrete random variables
  - X and Y take values in $\{1, 2, \ldots, m\}$ and $\{1, 2, \ldots, n\}$
  - $p(X, Y)$ denotes the joint probability distribution—if not known, it is often estimated based on co-occurrence data
  - Application areas: text mining, market-basket analysis, analysis of browsing behavior, etc.
- Key Obstacles in Clustering Contingency Tables
  - High Dimensionality, Sparsity, Noise
  - Need for robust and scalable algorithms

Reference:

1. Dhillon et al. Information-Theoretic Co-clustering, KDD'03

eg, terms x documents

$$
\underset{m}{\overset{n}{\begin{bmatrix}
.05 & .05 & .05 & 0 & 0 & 0 \\
.05 & .05 & .05 & 0 & 0 & 0 \\
0 & 0 & 0 & .05 & .05 & .05 \\
0 & 0 & 0 & .05 & .05 & .05 \\
.04 & .04 & 0 & .04 & .04 & .04 \\
.04 & .04 & .04 & 0 & .04 & .04
\end{bmatrix}}}
$$

$$
\underset{m \times k}{\begin{bmatrix}
.5 & 0 & 0 \\
.5 & 0 & 0 \\
0 & .5 & 0 \\
0 & .5 & 0 \\
0 & 0 & .5 \\
0 & 0 & .5
\end{bmatrix}}
\underset{k \times l}{\begin{bmatrix}
.3 & 0 \\
0 & .3 \\
.2 & .2
\end{bmatrix}}
\underset{l \times n}{\begin{bmatrix}
.36 & .36 & .28 & 0 & 0 & 0 \\
0 & 0 & 0 & .28 & .36 & .36
\end{bmatrix}}
=
\begin{bmatrix}
.054 & .054 & .042 & 0 & 0 & 0 \\
.054 & .054 & .042 & 0 & 0 & 0 \\
0 & 0 & 0 & .042 & .054 & .054 \\
0 & 0 & 0 & .042 & .054 & .054 \\
.036 & .036 & 028 & .028 & .036 & .036 \\
.036 & .036 & .028 & .028 & .036 & .036
\end{bmatrix}
$$

Carnegie Mellon

med. doc

cs doc

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

med. terms

cs terms

common terms

term group x doc. group

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

doc x doc group

term x term-group

Co-clustering

Observations

- uses KL divergence, instead of L2
- the middle matrix is **not** diagonal
  - we saw that earlier in the Tucker tensor decomposition
- s/w at:

www.cs.utexas.edu/users/dml/Software/cocluster.html

Detailed outline

- Motivation
- Hard clustering – k pieces
- Hard co-clustering – (k,l) pieces
- Hard clustering – optimal # pieces
- Soft clustering – matrix decompositions
- Observations

Problem with Information Theoretic Co-clustering

- Number of row and column groups must be specified

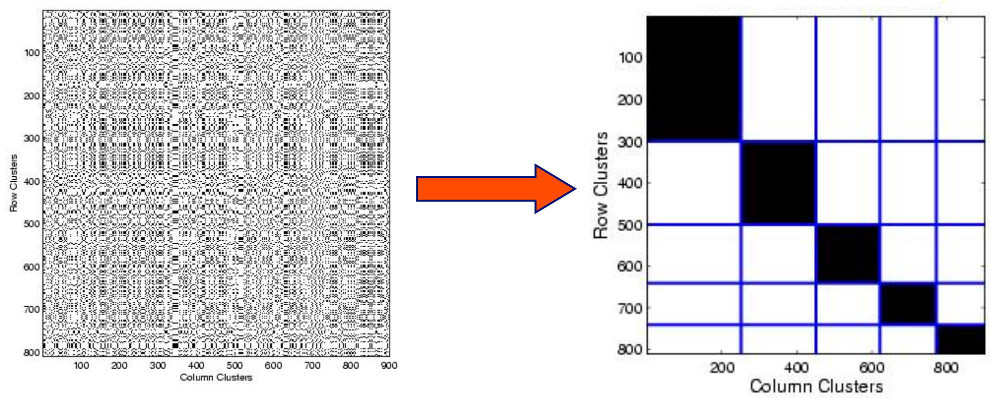
Desiderata:

✓ Simultaneously discover row and column groups

✗ Fully Automatic: No "magic numbers"

✓ Scalable to large graphs

Cross-association



Desiderata:

✓ Simultaneously discover row and column groups

✓ Fully Automatic: No "magic numbers"

✓ Scalable to large matrices

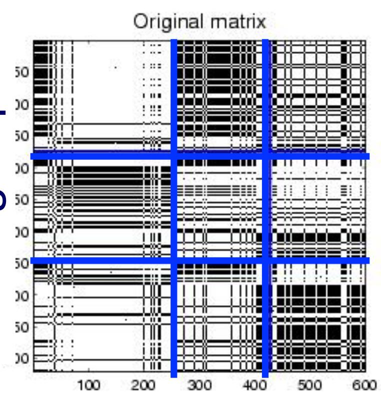
Reference:

1. Chakrabarti et al. Fully Automatic Cross-Associations, KDD'04

Carnegie Mellon

What makes a cross-association "good"?

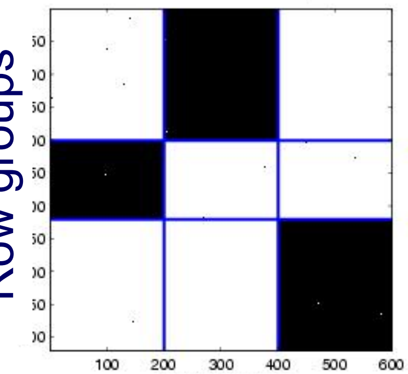
Row groups



Column groups

versus

Row groups

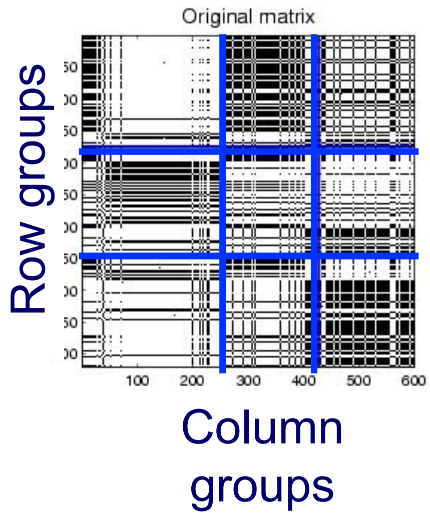


Column groups

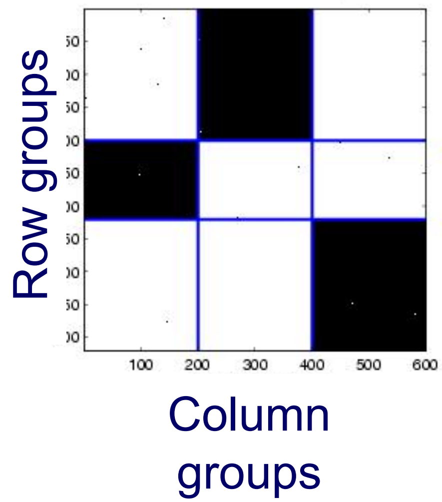
Why is this better?

Carnegie Mellon

What makes a cross-association "good"?



versus

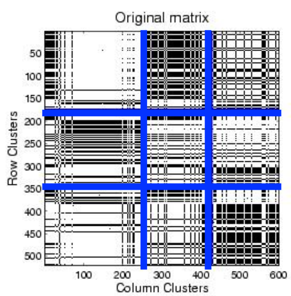


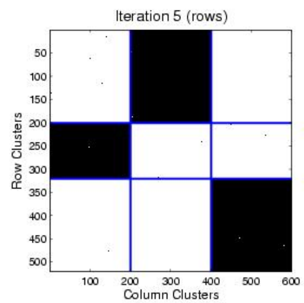
Why is this better?

simpler; easier to describe

easier to compress!

What makes a cross-association "good"?





Problem definition: given an encoding scheme

- decide on the # of col. and row groups *k* and *l*
- and reorder rows and columns,
- to achieve best compression

details

Main Idea

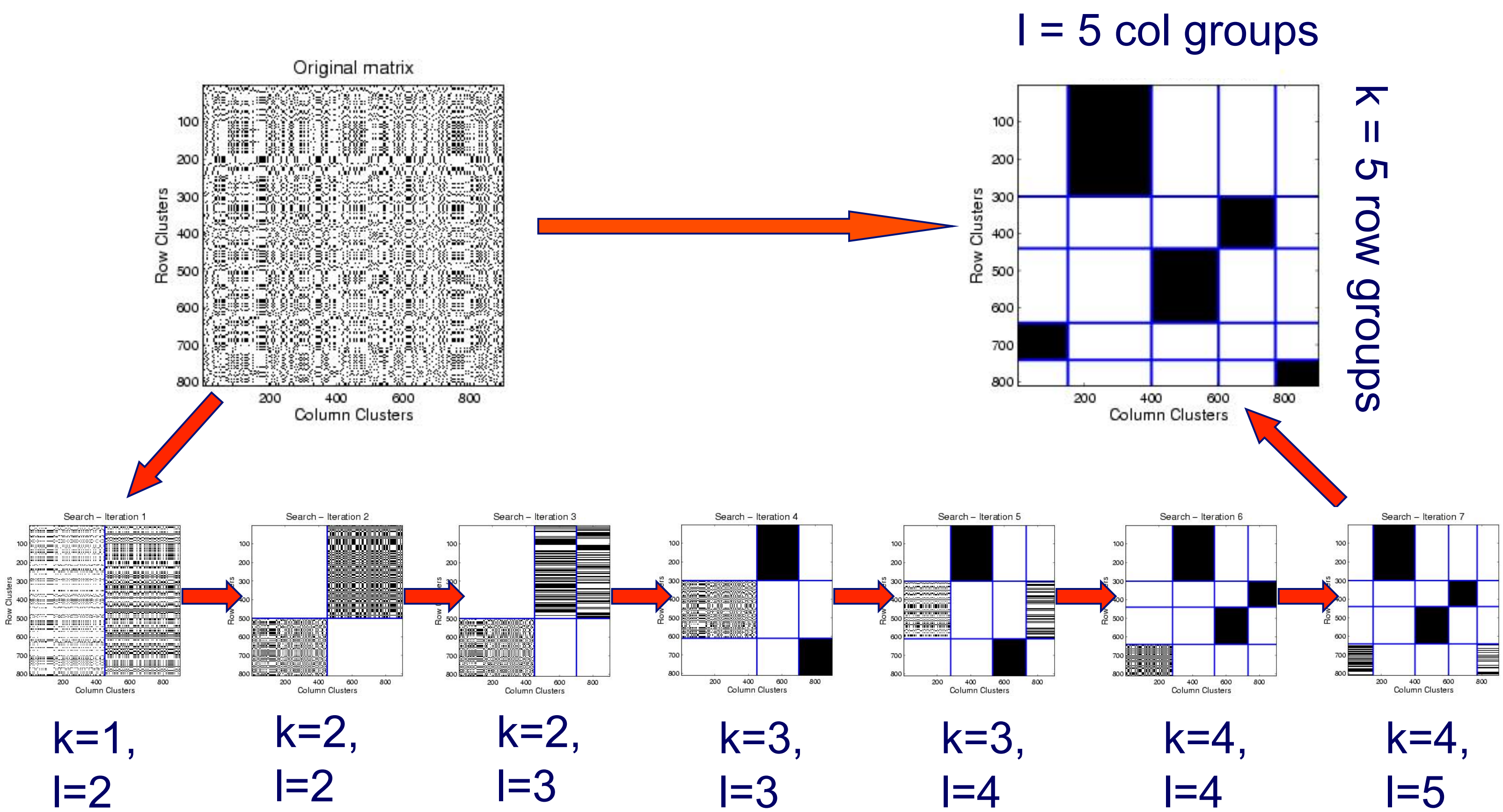
Good Compression → Better Clustering

Total Encoding Cost = $\underbrace{\sum_i size_i * H(x_i)}_{\text{Code Cost}}$ + $\underbrace{\text{Cost of describing cross-associations}}_{\text{Description Cost}}$

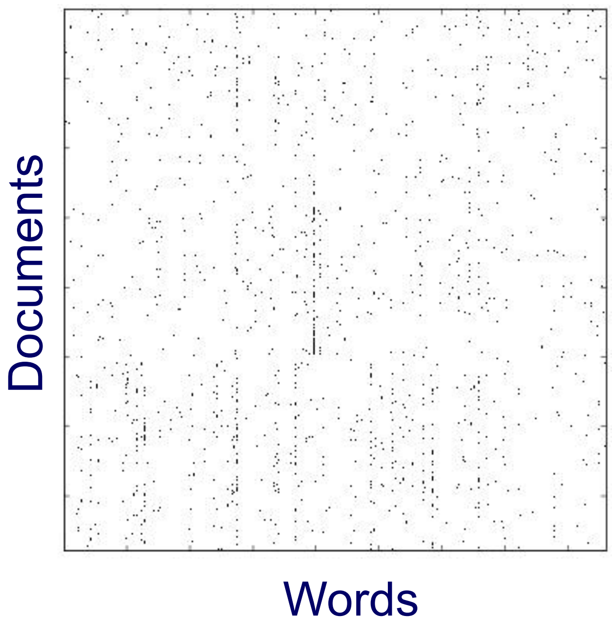
Minimize the total cost (# bits)

for lossless compression

Algorithm


l = 5 col groups
k = 5 row groups
Original matrix
Row Clusters
Column Clusters
Search – Iteration 1
Search – Iteration 2
Search – Iteration 3
Search – Iteration 4
Search – Iteration 5
Search – Iteration 6
Search – Iteration 7
k=1,
l=2
k=2,
l=2
k=2,
l=3
k=3,
l=3
k=3,
l=4
k=4,
l=4
k=4,
l=5

Experiments



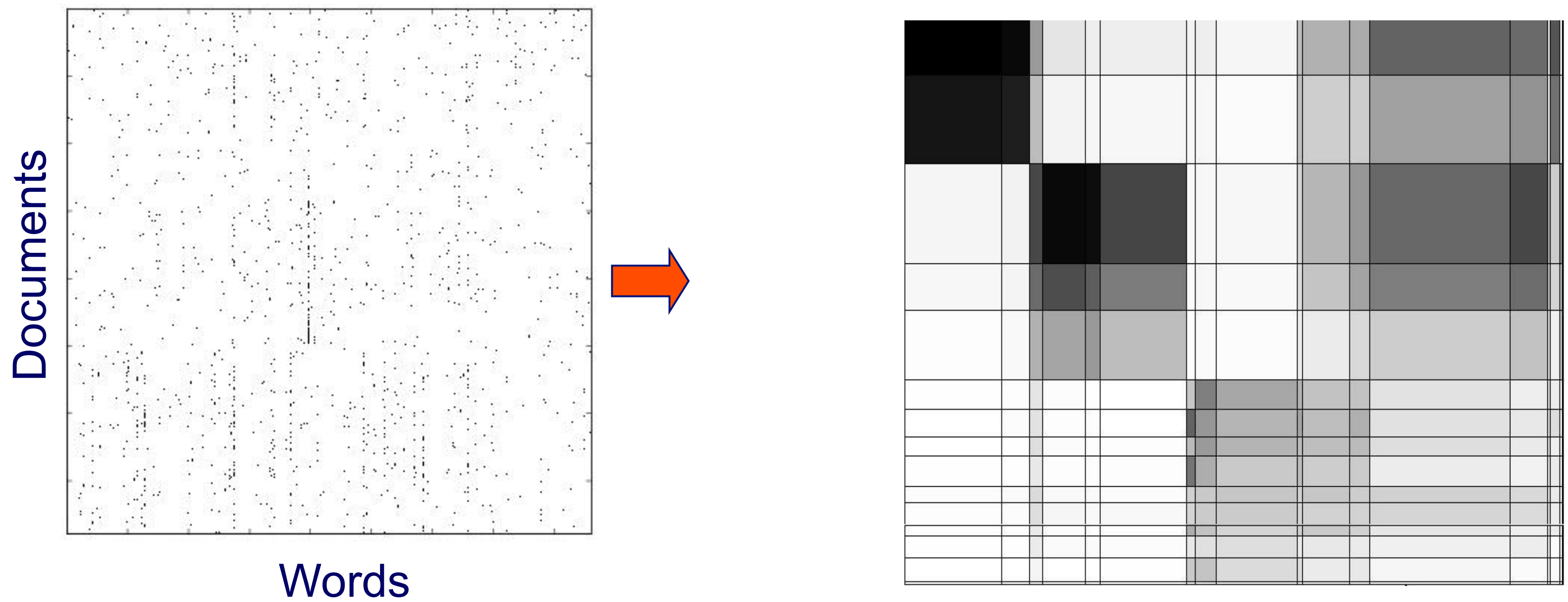
"CLASSIC"

- 3,893 documents
- 4,303 words
- 176,347 "dots"

Combination of 3 sources:

- MEDLINE (medical)
- CISI (info. retrieval)
- CRANFIELD (aerodynamics)

Experiments



“CLASSIC” graph of documents & words: k=15, l=19

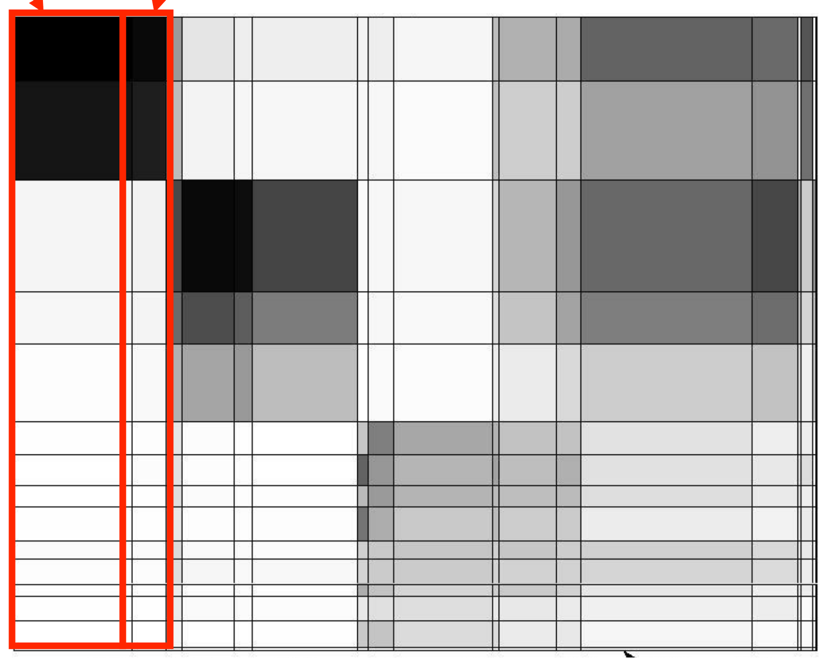
Carnegie Mellon

Experiments

insipidus, alveolar, aortic, death, prognosis, intravenous

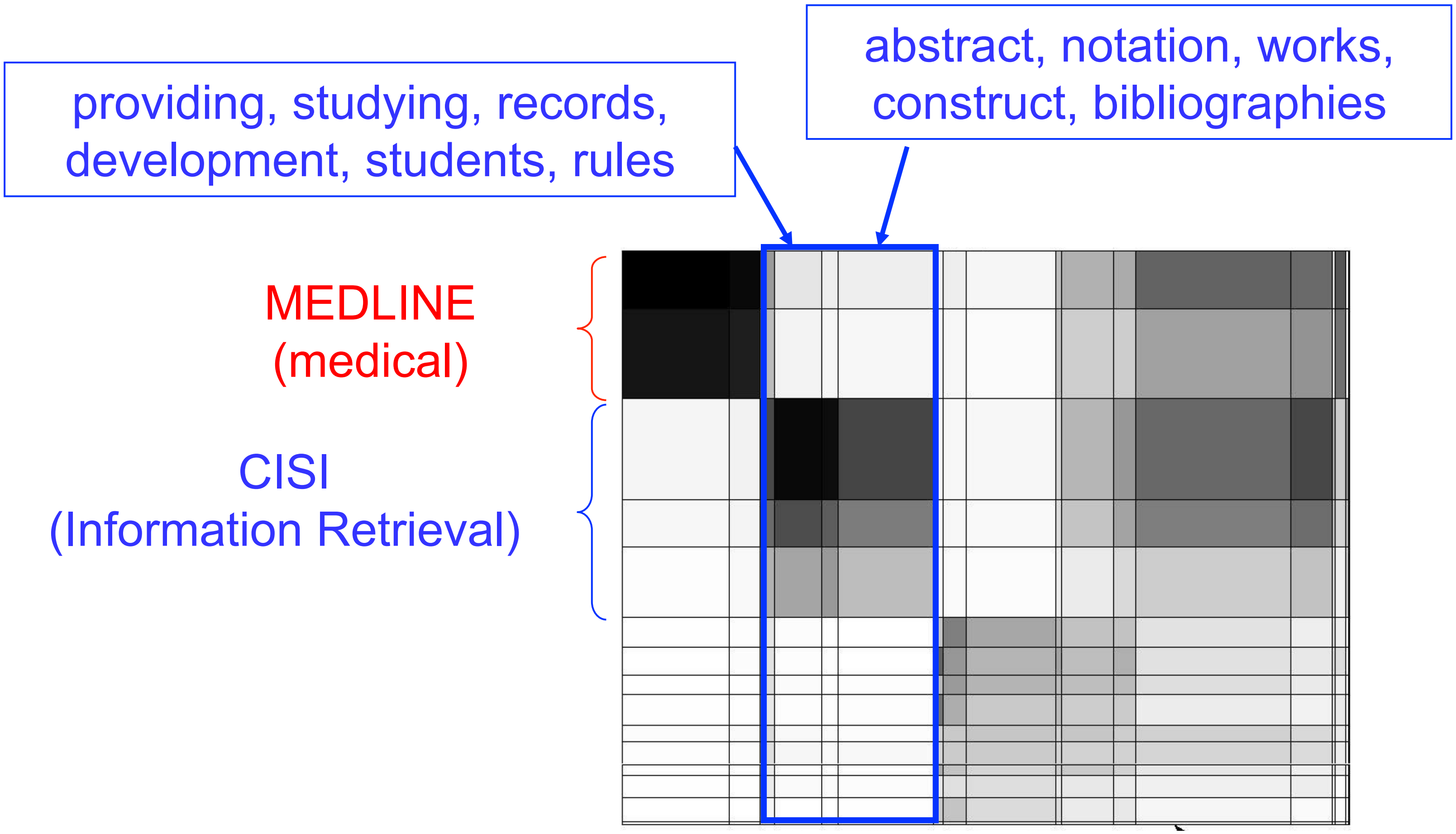
blood, disease, clinical, cell, tissue, patient

MEDLINE (medical)



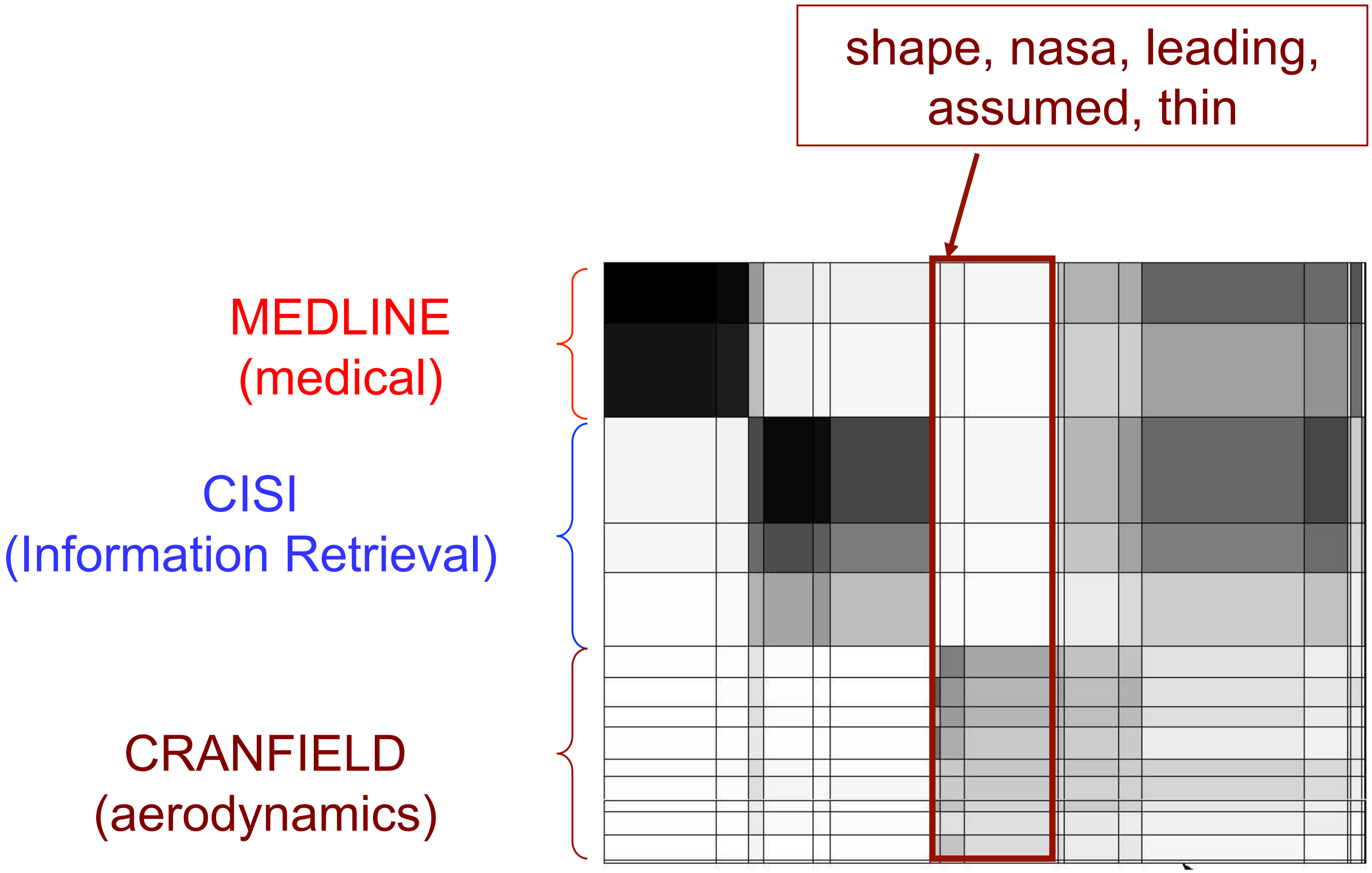
“CLASSIC” graph of documents & words: k=15, l=19

Experiments



"CLASSIC" graph of documents & words: k=15, l=19

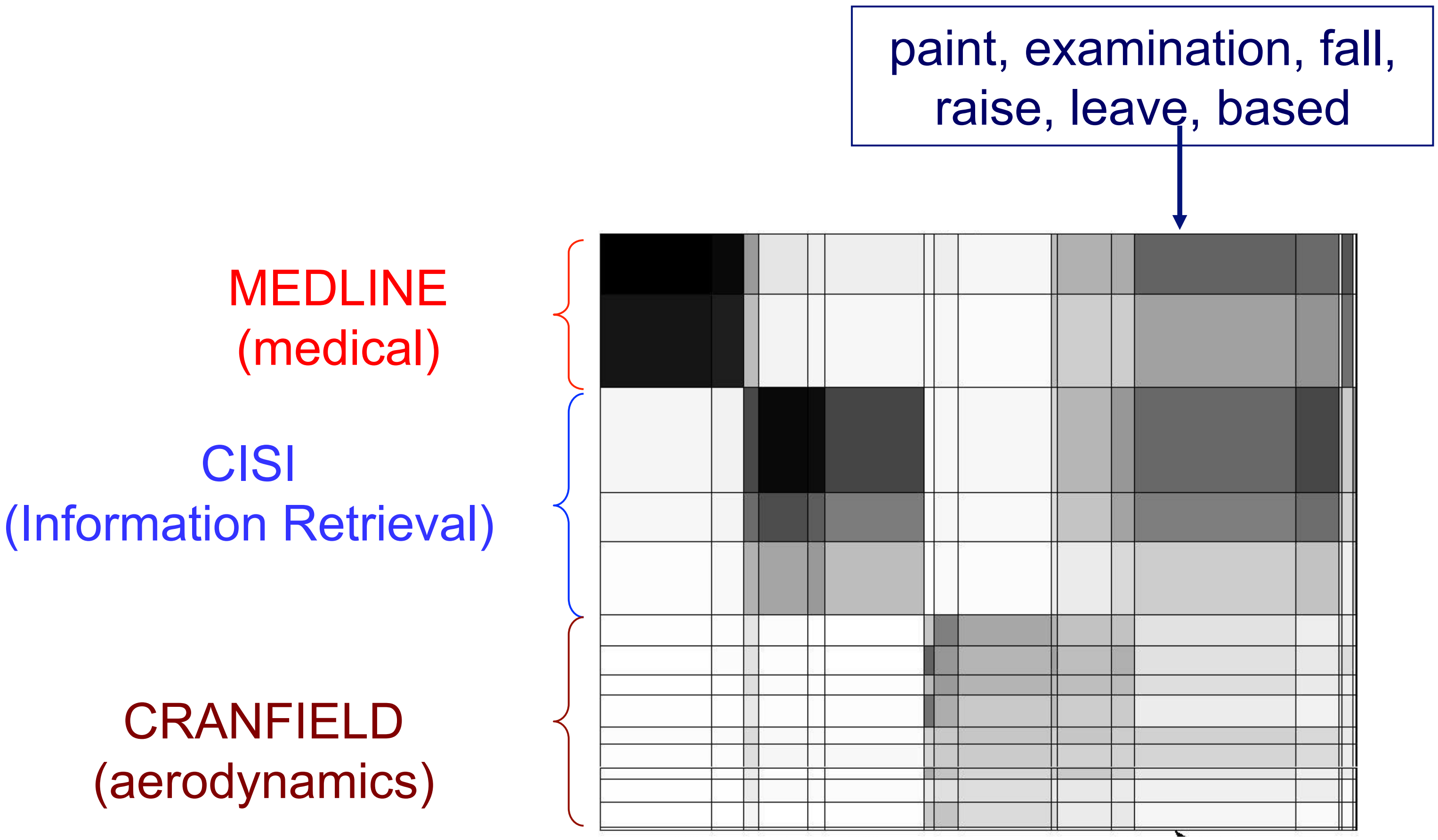
Experiments



"CLASSIC" graph of documents & words: k=15, l=19

Carnegie Mellon

Experiments



"CLASSIC" graph of documents & words: k=15, l=19

Algorithm

Code for cross-associations (matlab):

www.cs.cmu.edu/~deepay/mywww/software/CrossAssociations-01-27-2005.tgz

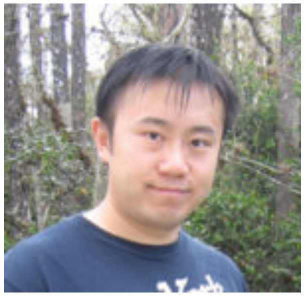
Variations and extensions:

- 'Autopart' [Chakrabarti, PKDD'04]
- www.cs.cmu.edu/~deepay

Carnegie Mellon

Algorithm

- Hadoop implementation [ICDM'08]



Spiros Papadimitriou, Jimeng Sun: DisCo: Distributed Co-clustering with Map-Reduce: A Case Study towards Petabyte-Scale End-to-End Mining. ICDM 2008: 512-521

Detailed outline

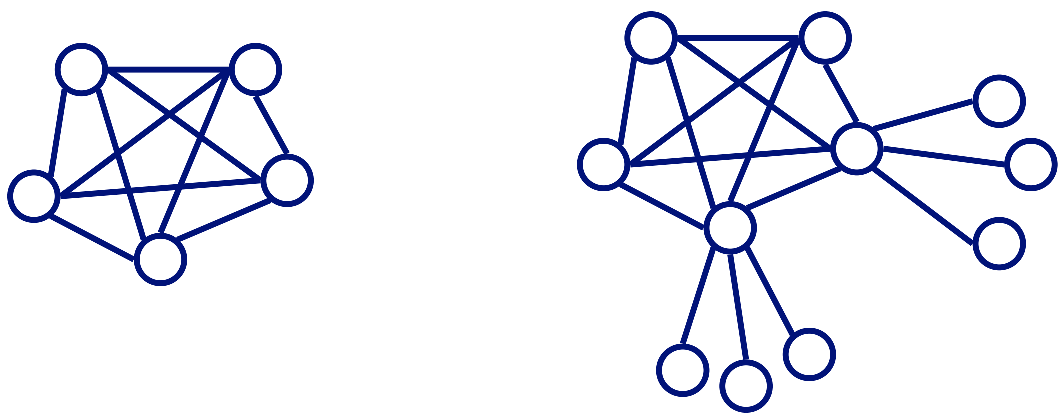
- Motivation
- Hard clustering – *k* pieces
- Hard co-clustering – (*k*,*l*) pieces
- Hard clustering – optimal # pieces
- Observations

Observation #1

- Skewed degree distributions – there are nodes with huge degree (>O(10^4), in facebook/linkedIn popularity contests!)

Observation #2

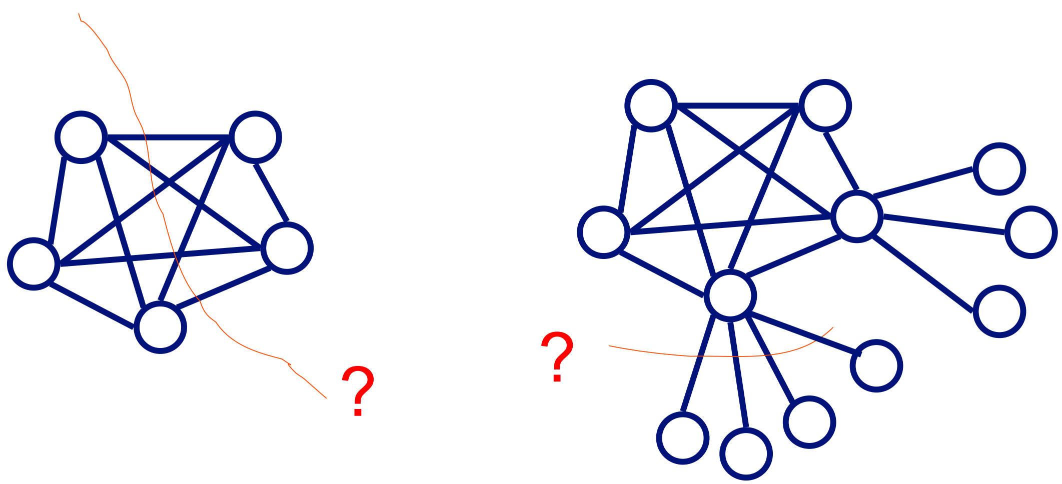
- Maybe there are no good cuts: ``jellyfish'' shape [Tauro+'01], [Siganos+,'06], strange behavior of cuts [Chakrabarti+'04], [Leskovec+,'08]



Carnegie Mellon

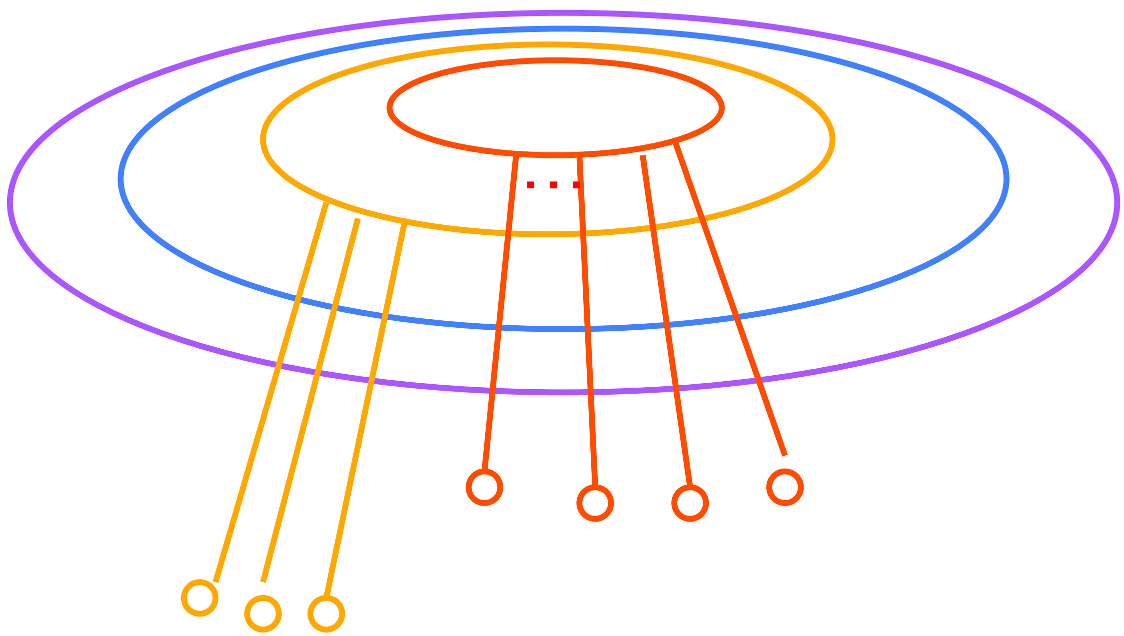
Observation #2

- Maybe there are no good cuts: ``jellyfish'' shape [Tauro+'01], [Siganos+,'06], strange behavior of cuts [Chakrabarti+,'04], [Leskovec+,'08]



Carnegie Mellon

Jellyfish model [Tauro+]



A Simple Conceptual Model for the Internet Topology, L. Tauro, C. Palmer, G. Siganos, M. Faloutsos, Global Internet, November 25-29, 2001

Jellyfish: A Conceptual Model for the AS Internet Topology G. Siganos, Sudhir L Tauro, M. Faloutsos, J. of Communications and Networks, Vol. 8, No. 3, pp 339-350, Sept. 2006.

Carnegie Mellon

Strange behavior of min cuts

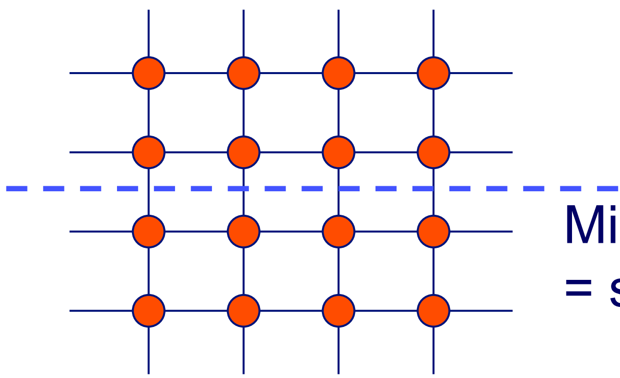
- 'negative dimensionality' (!)

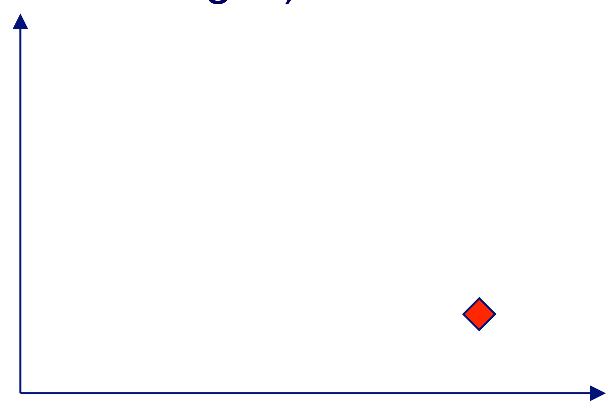
NetMine: New Mining Tools for Large Graphs, by D. Chakrabarti, Y. Zhan, D. Blandford, C. Faloutsos and G. Blelloch, in the SDM 2004 Workshop on Link Analysis, Counter-terrorism and Privacy

Statistical Properties of Community Structure in Large Social and Information Networks, J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. WWW 2008.

"Min-cut" plot

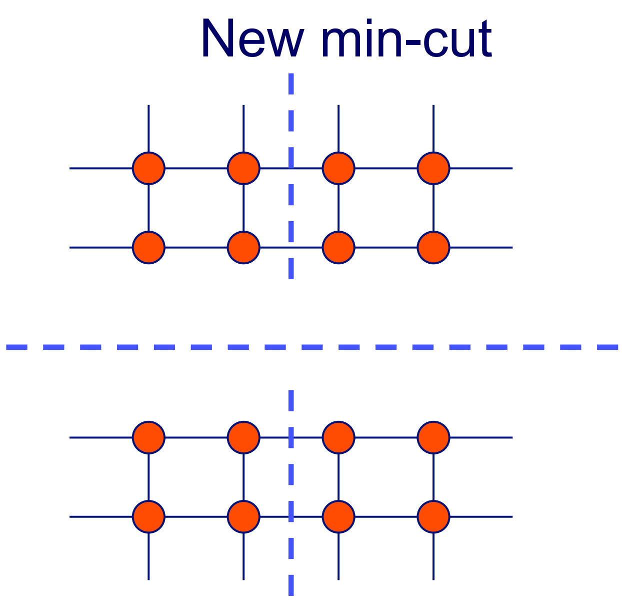
- Do min-cuts recursively.

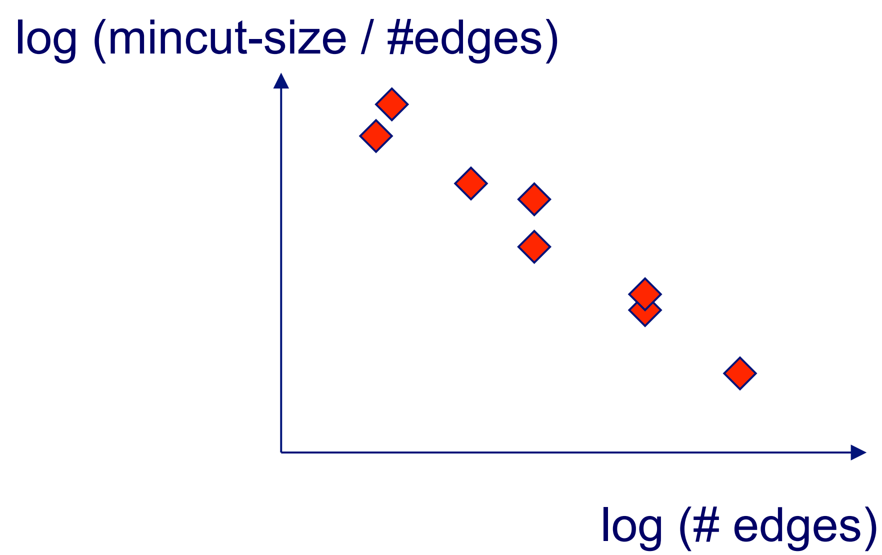




"Min-cut" plot

- Do min-cuts recursively.

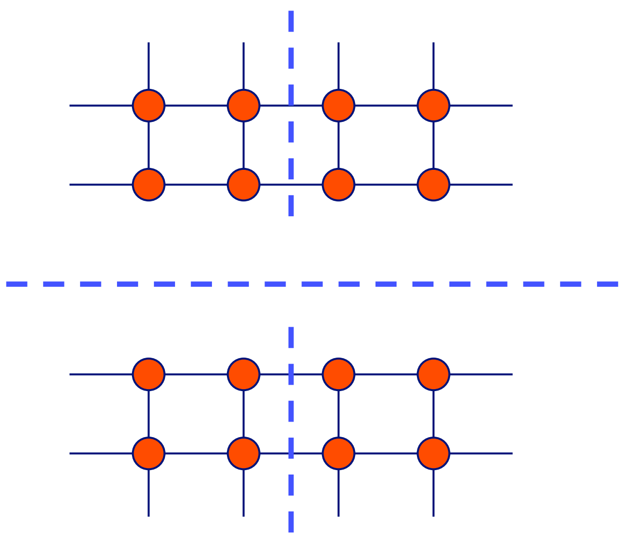




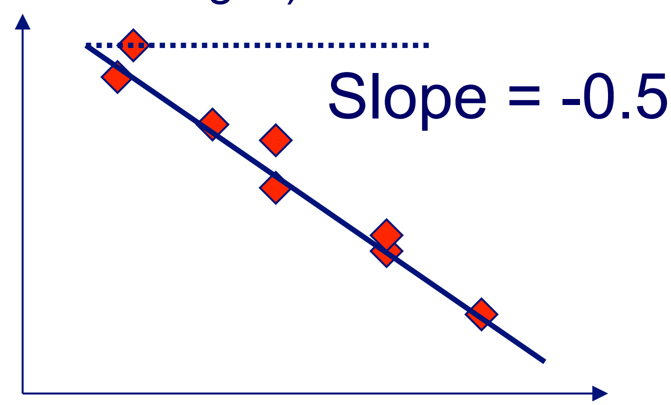
N nodes

"Min-cut" plot

- Do min-cuts recursively.



N nodes

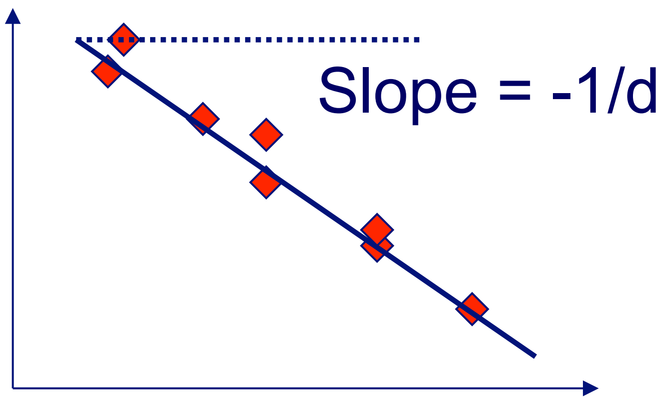


For a d-dimensional grid, the slope is -1/d

Carnegie Mellon

"Min-cut" plot

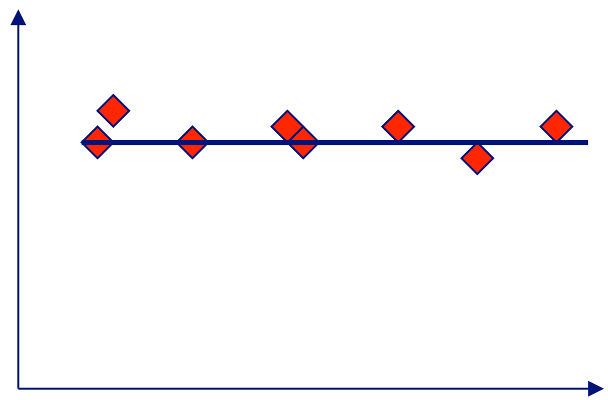
log (mincut-size / #edges)



log (# edges)

For a d-dimensional grid, the slope is -1/d

log (mincut-size / #edges)

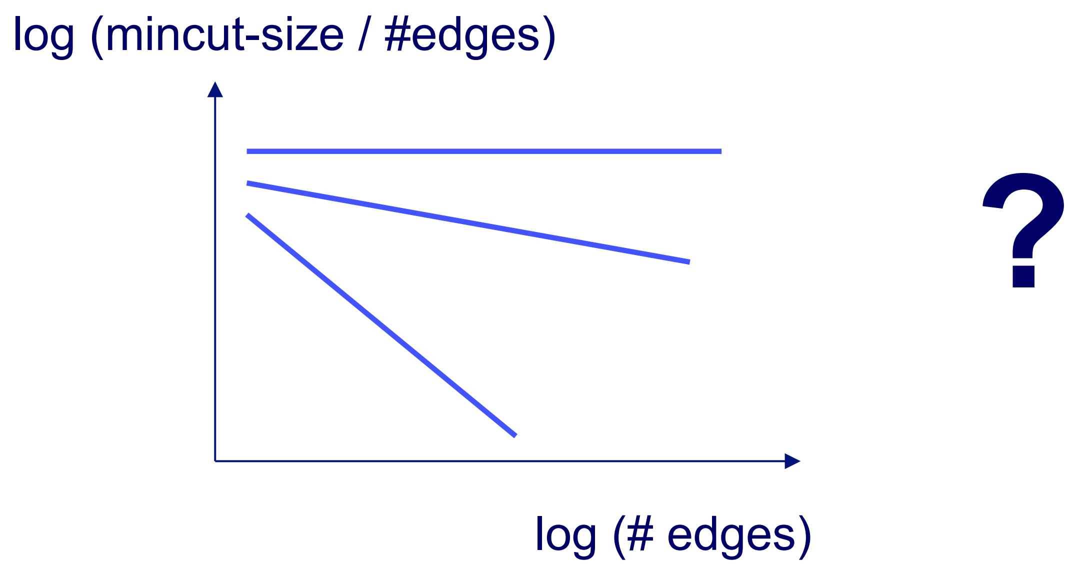


log (# edges)

For a random graph, the slope is 0

"Min-cut" plot

- What does it look like for a real-world graph?



Experiments

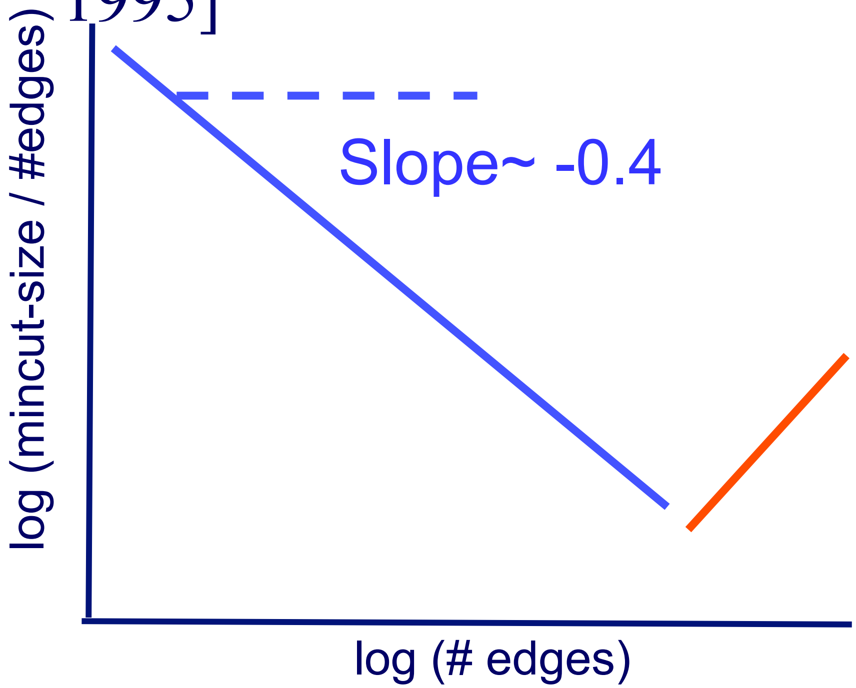
- Datasets:
  - Google Web Graph: 916,428 nodes and 5,105,039 edges
  - Lucent Router Graph: Undirected graph of network routers from www.isi.edu/scan/mercator/maps.html; 112,969 nodes and 181,639 edges
  - User → Website Clickstream Graph: 222,704 nodes and 952,580 edges

NetMine: New Mining Tools for Large Graphs, by D. Chakrabarti, Y. Zhan, D. Blandford, C. Faloutsos and G. Blelloch, in the SDM 2004 Workshop on Link Analysis, Counter-terrorism and Privacy

Carnegie Mellon

Experiments

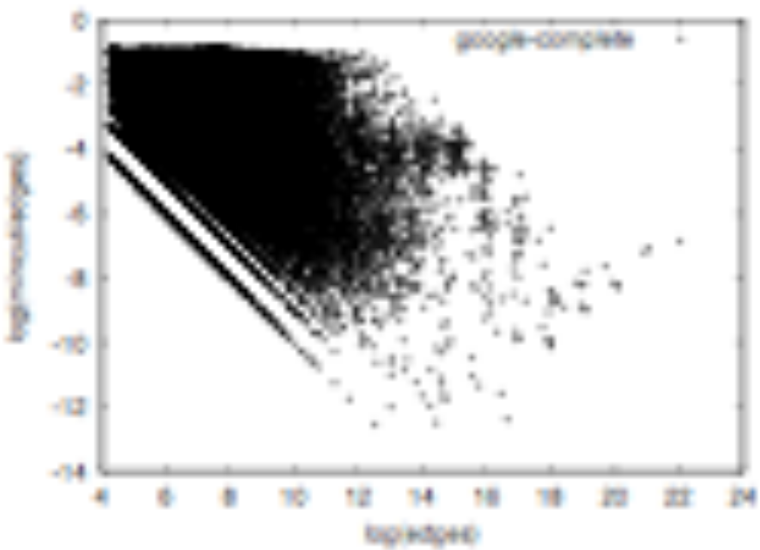
- Used the METIS algorithm [Karypis, Kumar, 1995]

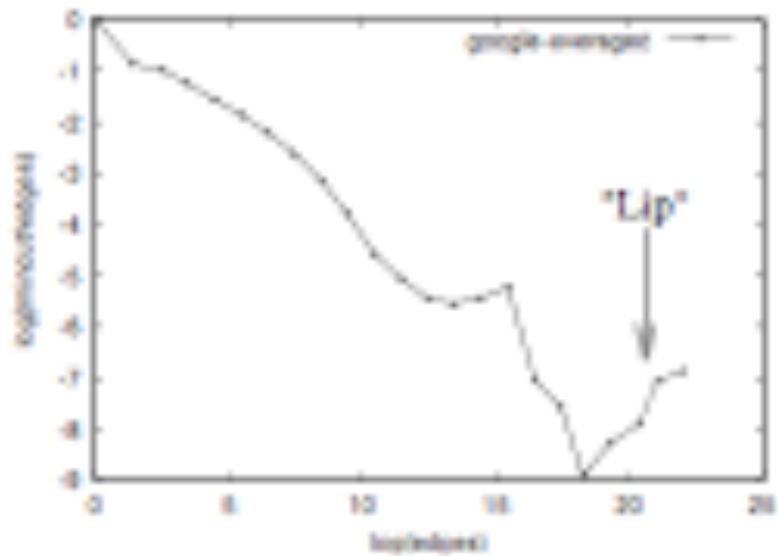


- Google Web graph
- Values along the y-axis are averaged
- We observe a “lip” for large edges
- Slope of -0.4, corresponds to a 2.5-dimensional grid!

Google graph

log (mincut-size / #edges)





Log(#edges)

Log(#edges)

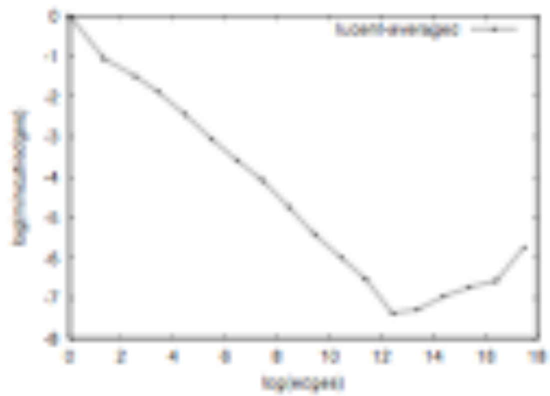
All min-cuts

averaged

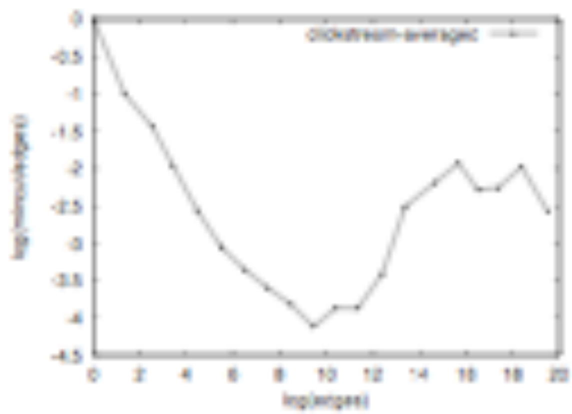
Experiments

- Same results for other graphs too…

Lucent Router graph



Clickstream graph



Carnegie Mellon

Conclusions – Practitioner's guide

- Hard clustering – *k* pieces — **METIS**
- Hard co-clustering – (*k*,*l*) pieces — **Co-clustering**
- Hard clustering – optimal # pieces — **Cross-associations**
- Observations — **'jellyfish': Maybe, there are no good cuts**

Outline

- Task 4: time-evolving graphs – tensors
- Task 5: community detection
- ➡ Task 6: virus propagation
- Task 7: scalability, parallelism and hadoop
- Conclusions

Detailed outline

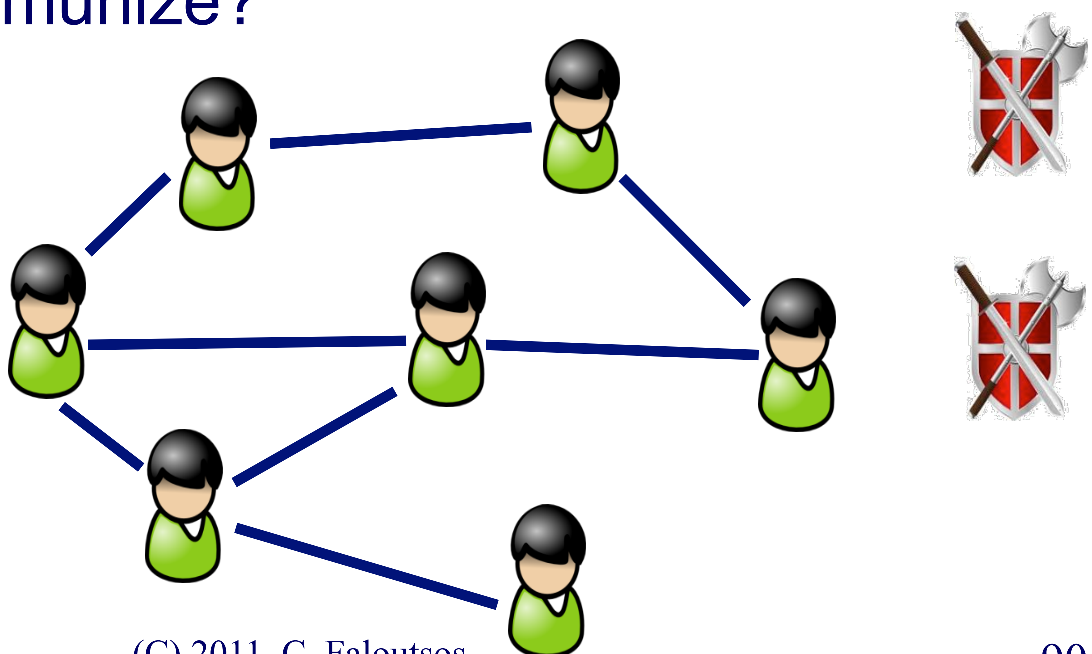
- Problem definition
- Analysis
- Experiments

Immunization and epidemic thresholds

- Q1: which nodes to immunize?
- Q2: will a virus vanish, or will it create an epidemic?

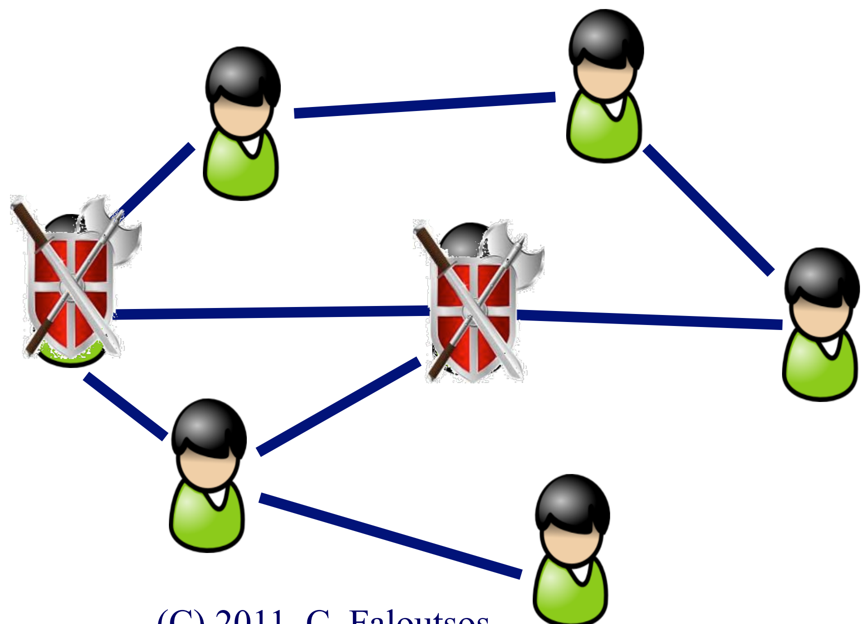
Q1: Immunization:

- Given
  - a network,
  - k vaccines, and
  - the virus details
- Which nodes to immunize?



Q1: Immunization:

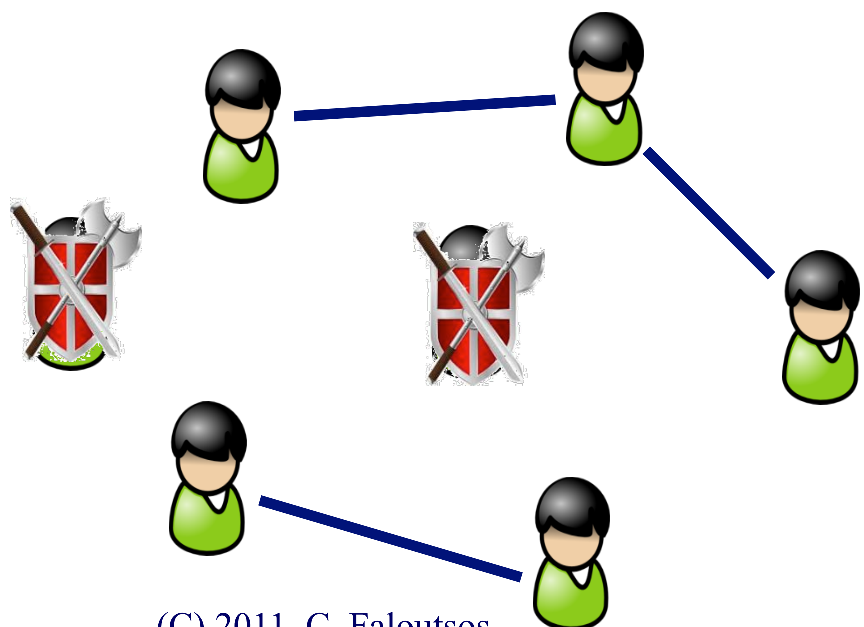
- Given
  - a network,
  - k vaccines, and
  - the virus details
- Which nodes to immunize?



Carnegie Mellon

Q1: Immunization:

- Given
  - a network,
  - k vaccines, and
  - the virus details
- Which nodes to immunize?

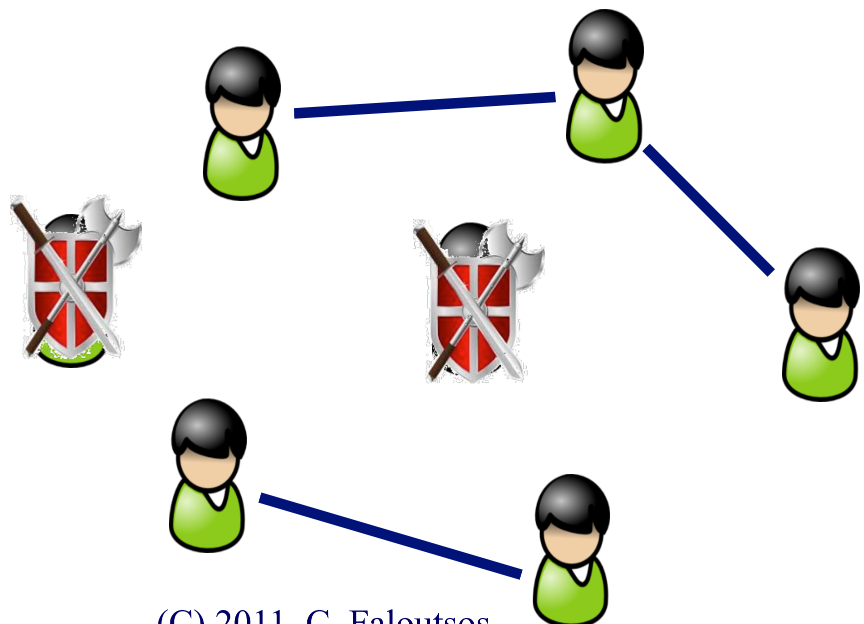


Carnegie Mellon

Q1: Immunization:

- Given
  - a network,
  - k vaccines, and
  - the virus details
- Which nodes to immunize?

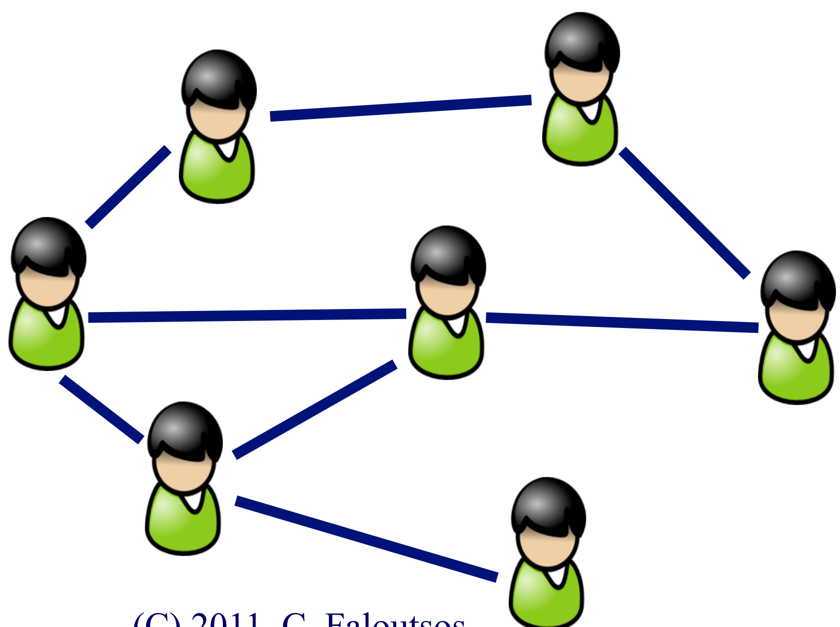
A: immunize the ones that maximally raise the `epidemic threshold' [Tong+, ICDM'10]



Q2: will a virus take over?

- Flu-like virus (no immunity, ‘SIS’)
- Mumps (life-time immunity, ‘SIR’)
- Pertussis (finite-length immunity, ‘SIRS’)

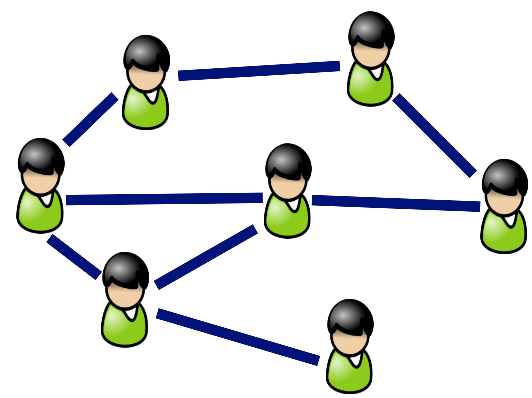
β: attack prob
δ: heal prob



Q2: will a virus take over?

- Flu-like virus (no immunity, 'SIS')
- Mumps (life-time immunity, 'SIR')
- Pertussis (finite-length immunity, 'SIRS')

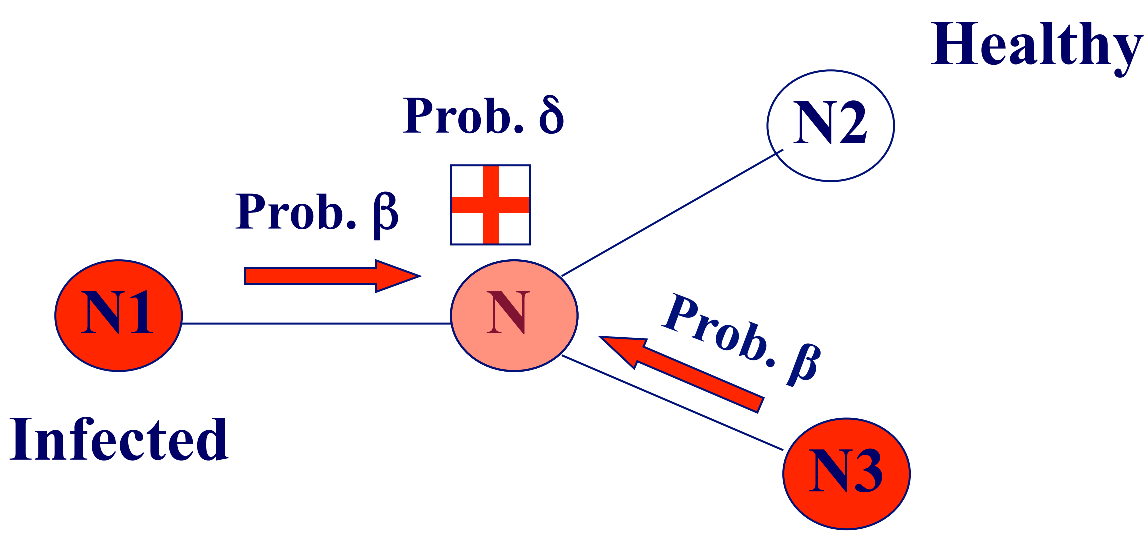
β: attack prob
δ: heal prob



A: depends on connectivity
(avg degree? Max degree?
variance? Something else?

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

$$\text{if strength } s = \beta / \delta < \tau$$

an epidemic can not happen

Thus,

- given a graph
- compute its epidemic threshold

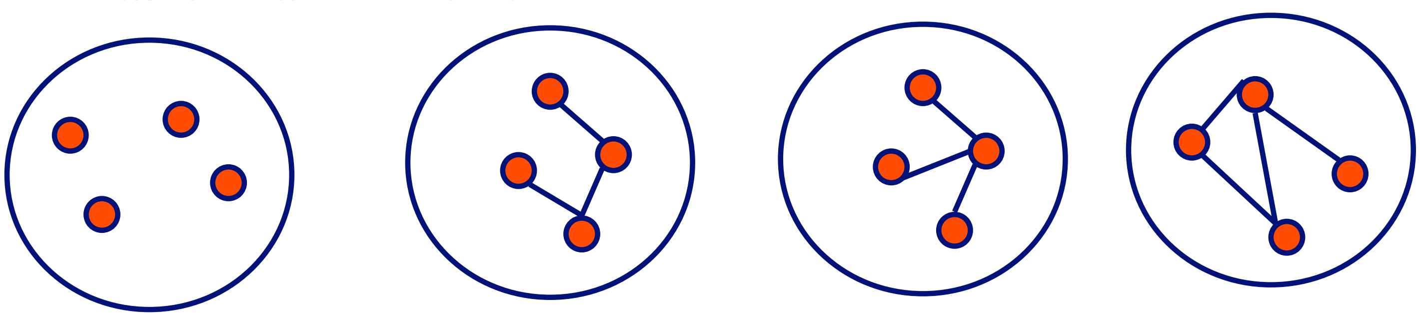
Detailed outline

- Problem definition
- ➡ Analysis
- Experiments

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

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

Carnegie Mellon

Epidemic threshold

- [Theorem] We have no epidemic, if

epidemic threshold

recovery prob.

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

attack prob.

largest eigenvalue of adj. matrix *A*

Proof: [Wang+03] (proof: for **SIS=flu only**)

Carnegie Mellon

Beginning of proof

details

Healthy @ t+1:
- (healthy or healed)
- and not attacked @ t

Let: p(i , t) = Prob node i is sick @ t+1

$$1 - p(i, t+1) = (1 - p(i, t) + p(i, t) * \delta) * \Pi_j (1 - \beta\, aji * p(j, t))$$

Below threshold, if the above *non-linear dynamical system* above is 'stable' (eigenvalue of Hessian < 1)

Epidemic threshold for various networks

Formula includes older results as special cases:

- Homogeneous networks [Kephart+White]
  - $\lambda_{1,A} = <k>; \tau = 1/<k>$ (<k> : avg degree)
- Star networks (d = degree of center)
  - $\lambda_{1,A} = sqrt(d); \tau = 1/ sqrt(d)$
- Infinite power-law networks
  - $\lambda_{1,A} = \infty; \tau = 0$; [Barabasi]

Epidemic threshold

- [Theorem 2] Below the epidemic threshold, the epidemic dies out exponentially

Recent generalization

- [Prakash+, arxiv '10]: similar threshold, for almost **all** virus propagation models (VPM)
  - SIS -> flu
  - SIR -> mumps
  - SIRS -> whooping cough (temporary immunity)
  - SIIR (-> HIV)
  - …

Carnegie Mellon

A2: will a virus take over?

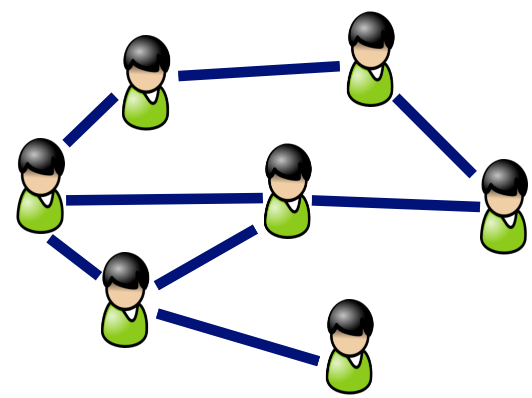
- For **all** typical virus propagation models (flu, mumps, pertussis, HIV, etc)
- The **only** connectivity measure that matters, is

  $1/\lambda_1$

  the first eigenvalue of the adj. matrix

  Proof for **all** VPM:

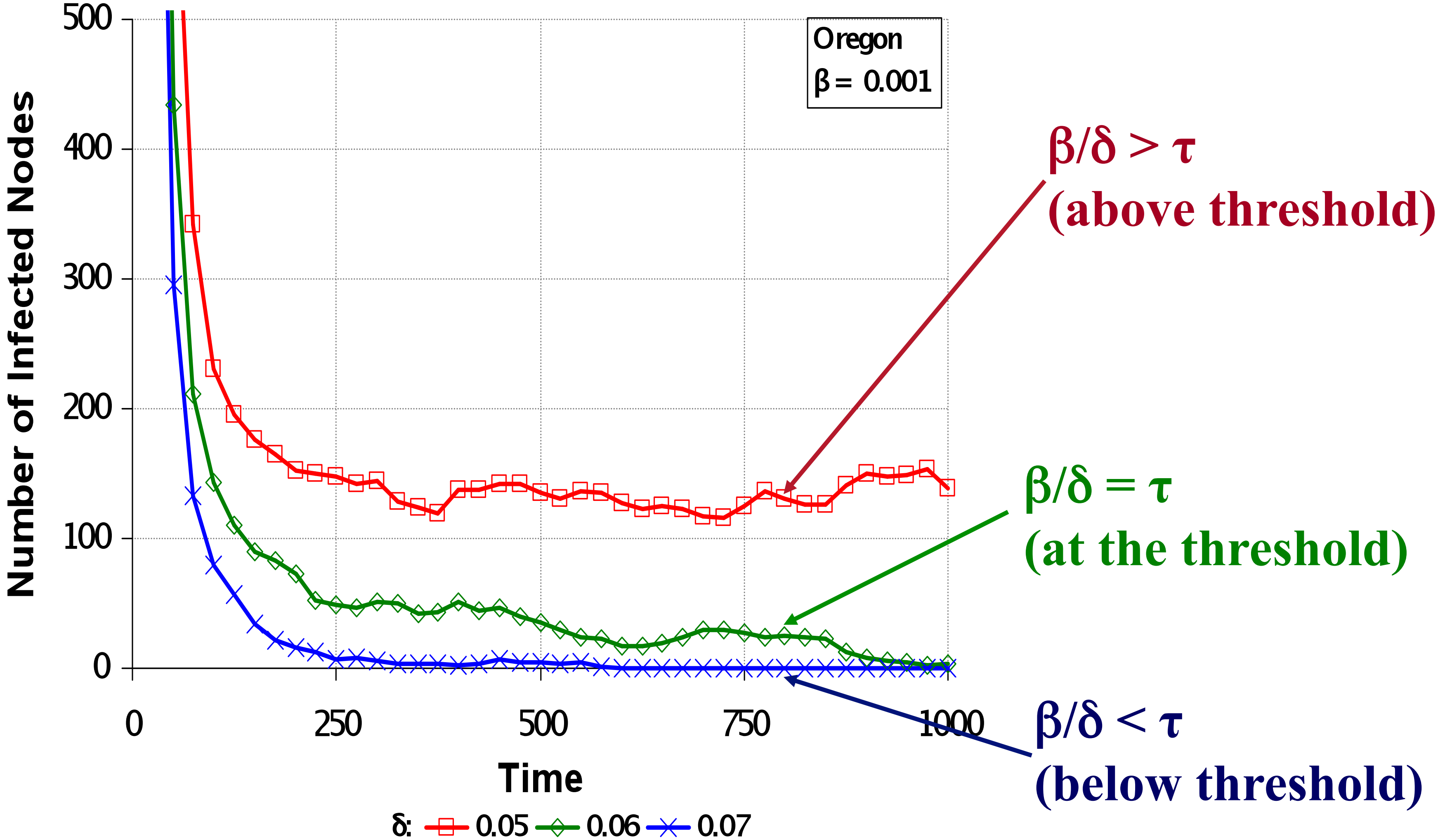
  [Prakash+, '10, arxiv]



Detailed outline

- Epidemic threshold
  - Problem definition
  - Analysis
  - Experiments

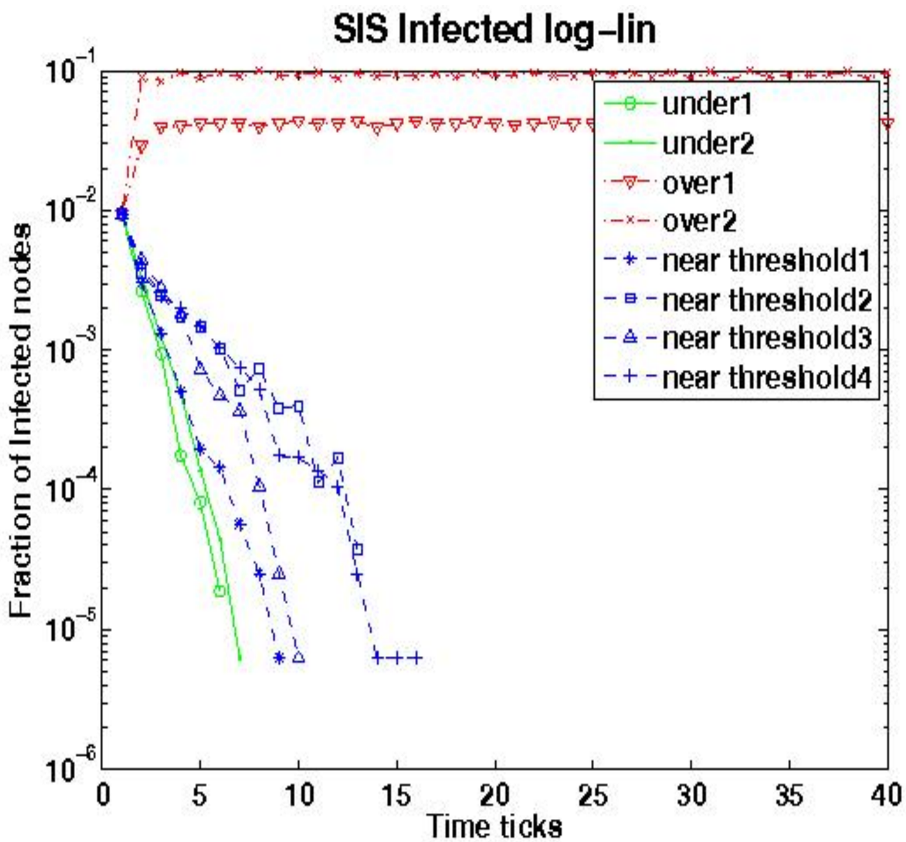
Experiments (Oregon)



SIS simulation - # infected nodes vs time

Log - Lin

#inf.
(log scale)



above

at

below

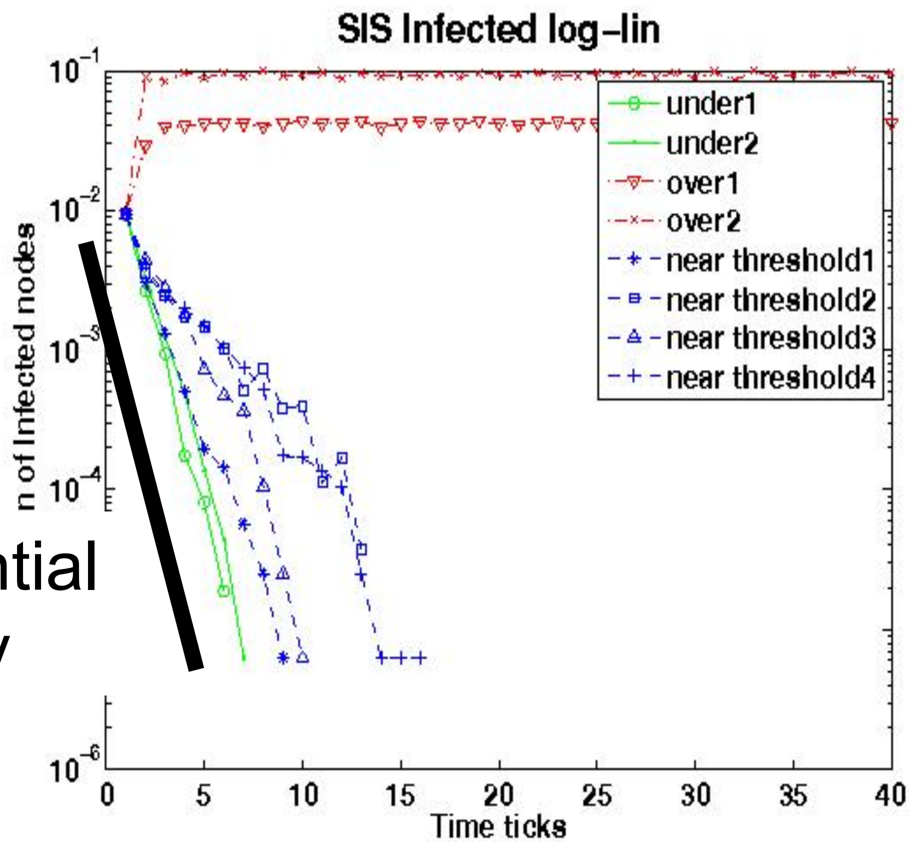
Time (linear scale)

SIS simulation - # infected nodes vs time

Log - Lin

#inf.
(log scale)

Exponential decay



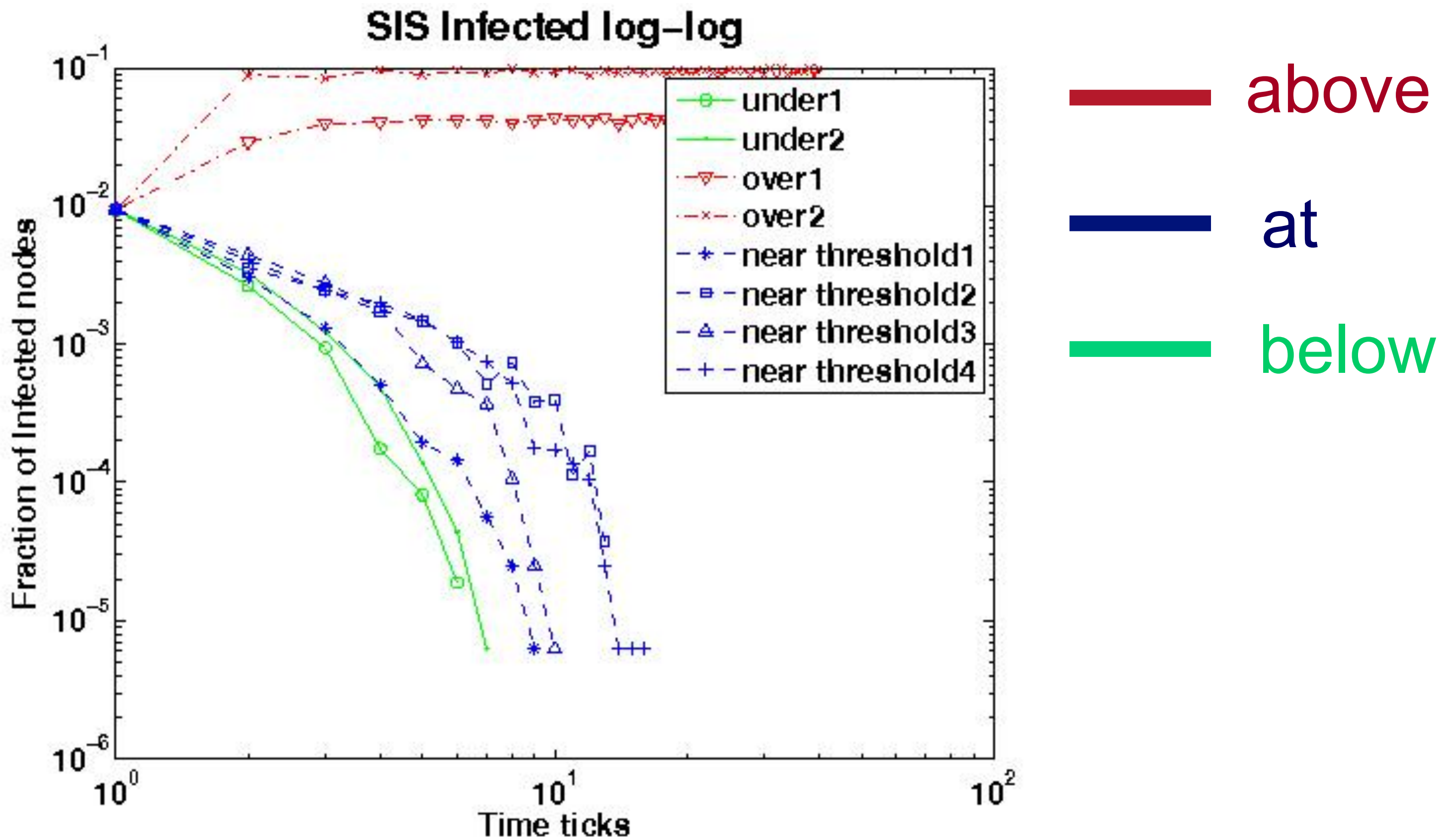
above
at
below

Time (linear scale)

SIS simulation - # infected nodes vs time

Log - Log

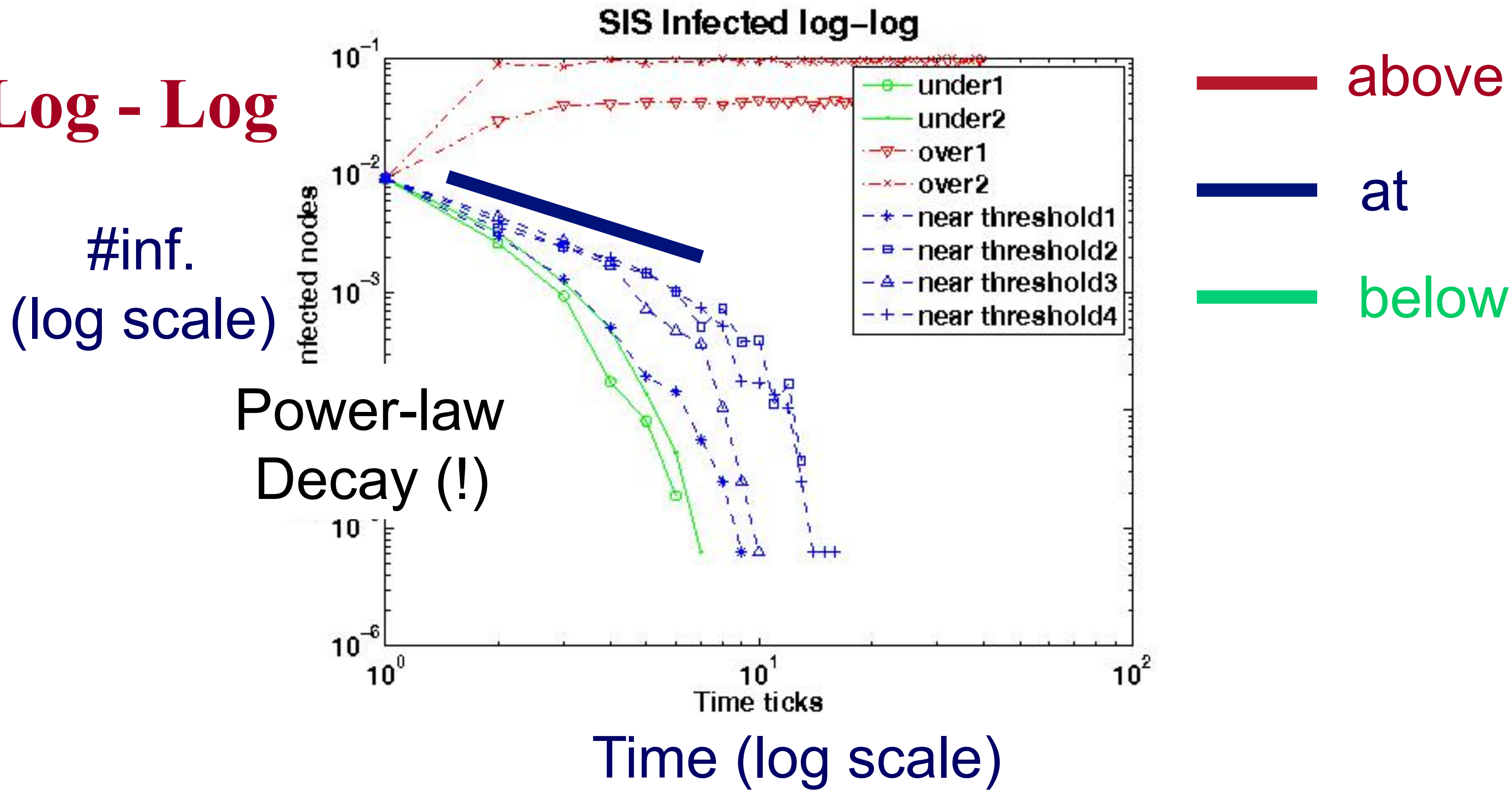
#inf.
(log scale)



Time (log scale)

SIS simulation - # infected nodes vs time

Log - Log

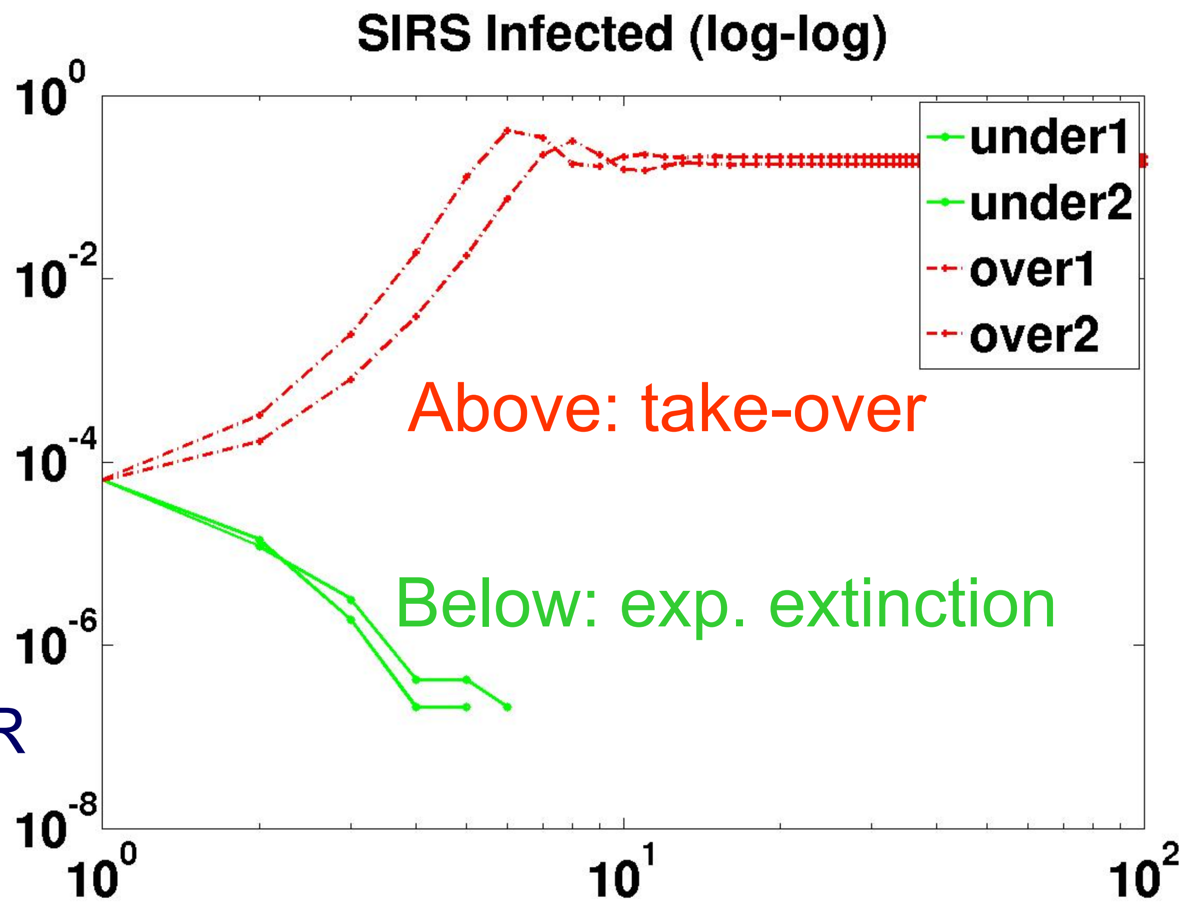


Carnegie Mellon

How about other VPMs?

A2: will a virus take over? (SIRS case)

Fraction of infected



Time ticks

Graph:
Portland, OR
31M links
1.5M nodes

Conclusions

$\lambda_{1,A}$: Eigenvalue of adjacency matrix determines the survival of (almost) **any** virus

- measure of connectivity (~ # paths)
- Can answer 'what-if' scenarios
  - May guide immunization policies
- Can help us avoid expensive simulations

References

- D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, *Epidemic Thresholds in Real Networks*, in ACM TISSEC, 10(4), 2008
- Ganesh, A., Massoulie, L., and Towsley, D., 2005. The effect of network topology on the spread of epidemics. In *INFOCOM*.

References (cont'd)

- Hethcote, H. W. 2000. The mathematics of infectious diseases. *SIAM Review 42*, 599–653.
- Hethcote, H. W. AND Yorke, J. A. 1984. *Gonorrhea Transmission Dynamics and Control*. Vol. 56. Springer. Lecture Notes in Biomathematics.

References (cont'd)

- Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*, in SRDS 2003 (pages 25-34), Florence, Italy

Outline

- Task 4: time-evolving graphs – tensors
- Task 5: community detection
- Task 6: virus propagation
- ➡ Task 7: scalability, parallelism and hadoop
- Conclusions

Scalability

- How about if graph/tensor does not fit in core?

- How about handling huge graphs?

Scalability

- How about if graph/tensor does not fit in core?
- [‘MET’: Kolda, Sun, ICMD’08, best paper award]
- How about handling huge graphs?

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?

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

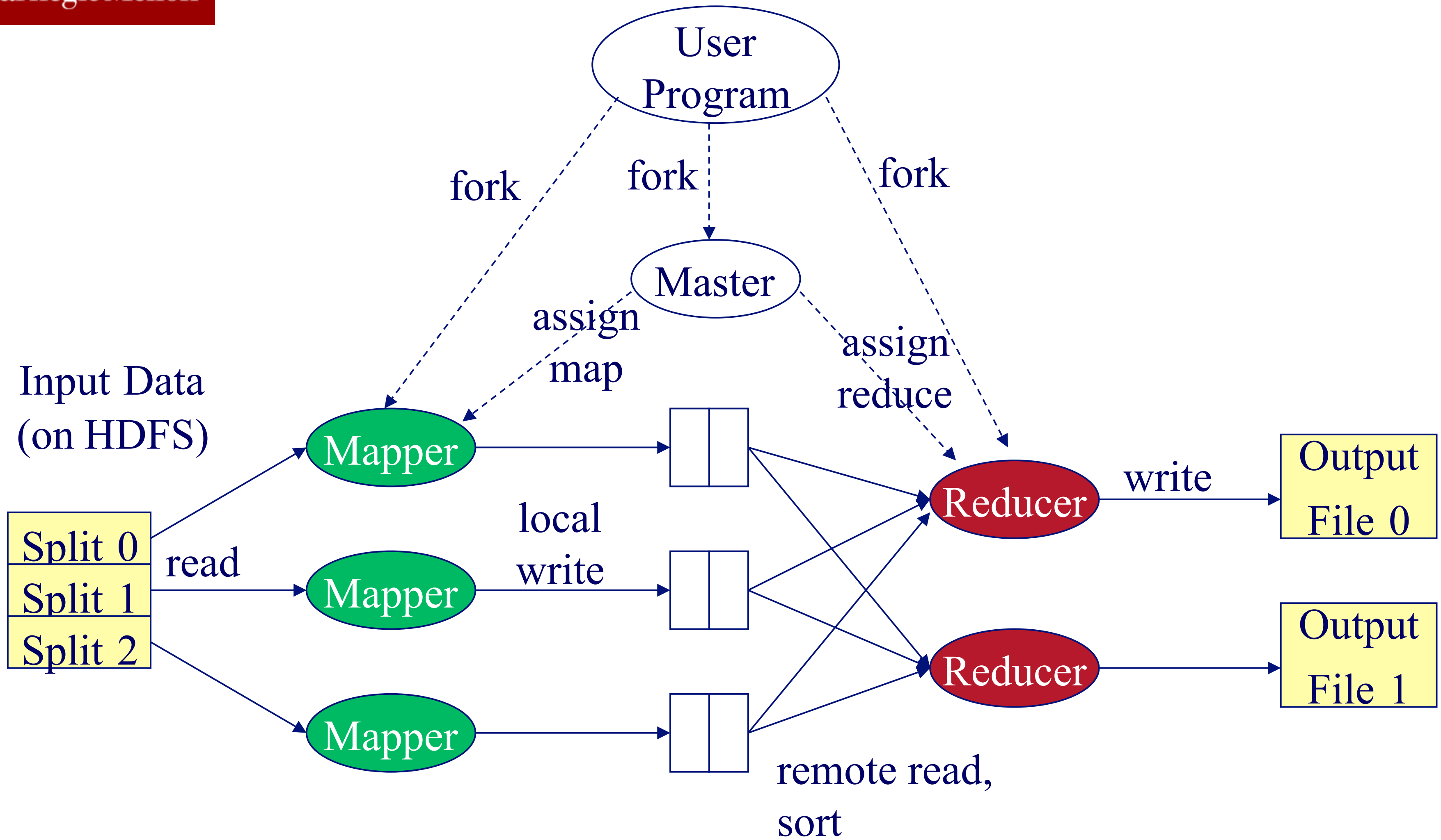
- master-slave architecture; n-way replication (default n=3)
- 'group by' of SQL (in parallel, fault-tolerant way)
- e.g, find histogram of word frequency
  - compute local histograms
  - then merge into global histogram

```
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
  - compute local histograms
  - then merge into global histogram

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



By default: 3-way replication;
Late/dead machines: ignored, **transparently** (!)

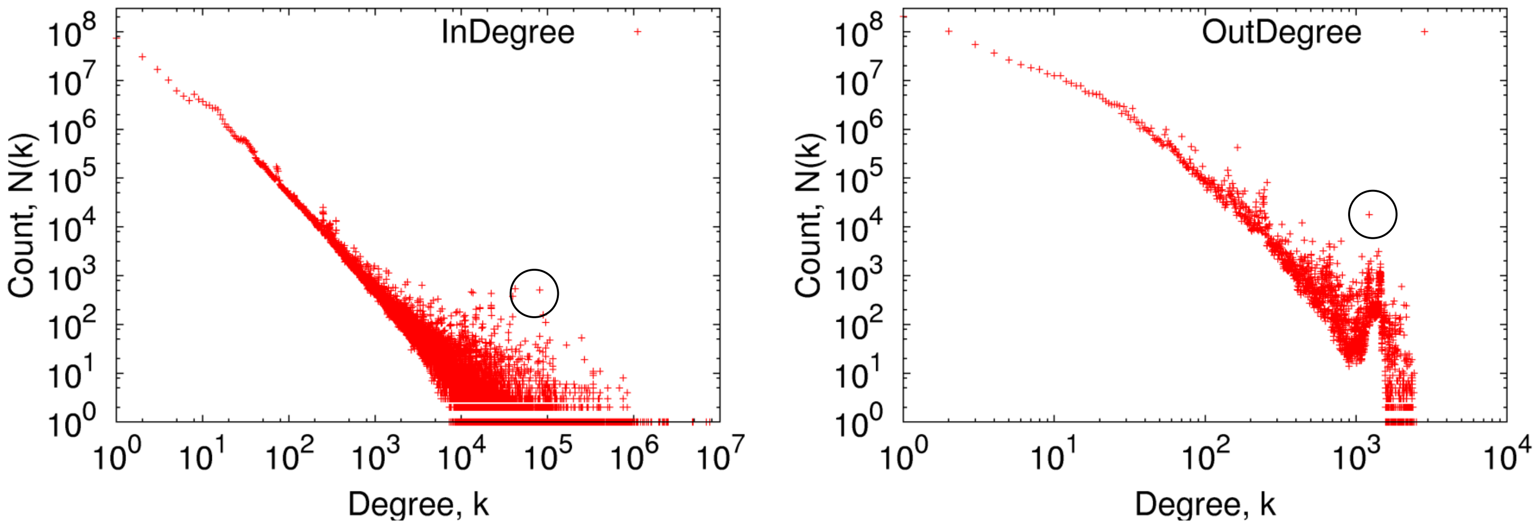
D.I.S.C.

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

Analysis of a large graph

~200Gb (Yahoo crawl) - Degree Distribution:

- in 12 minutes with 50 machines
- Many (link spams ?) at out-degree 1200



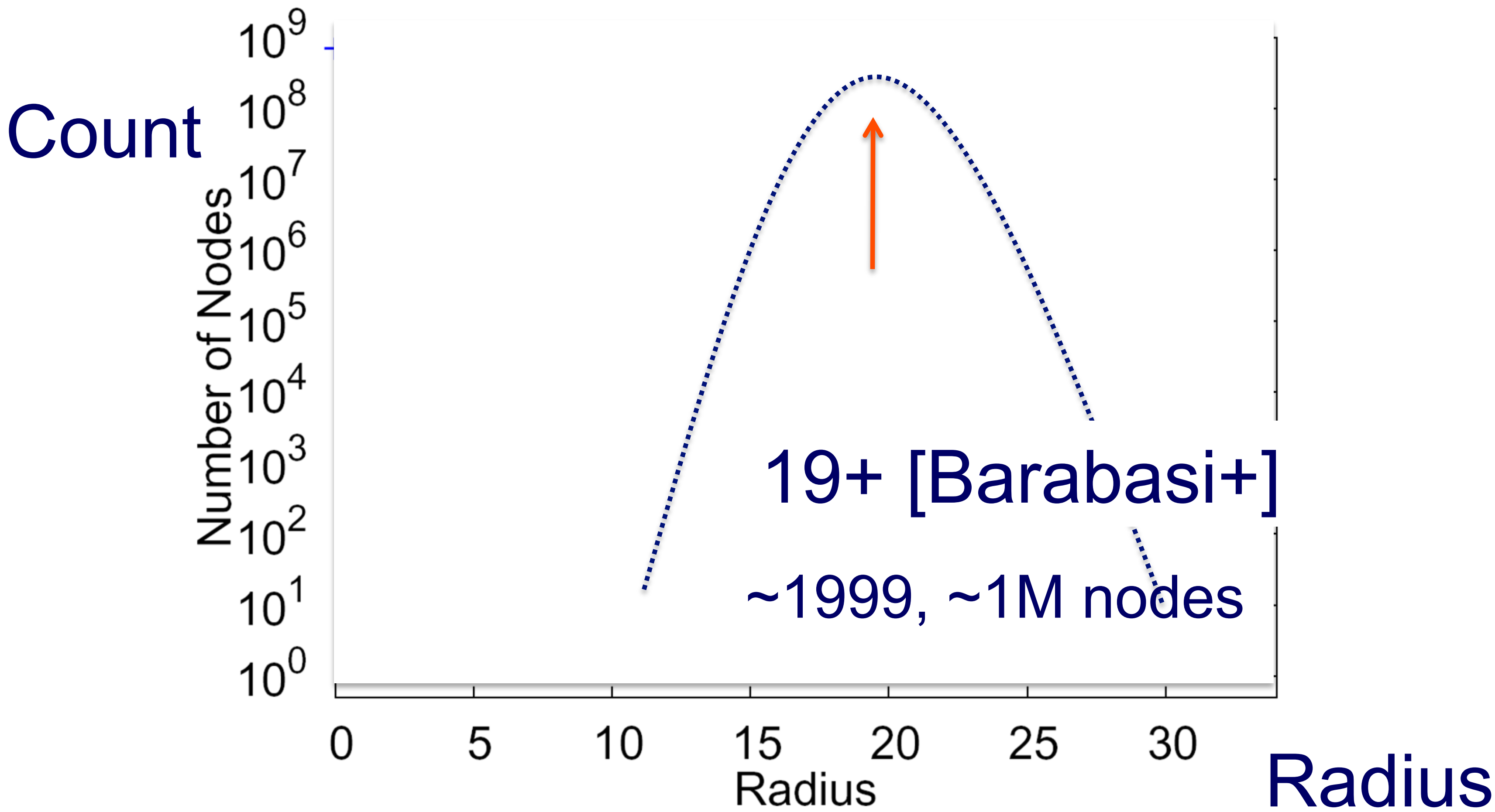
Outline – Algorithms & results

	Centralized	Hadoop/ PEGASUS
Degree Distr.	old	old
Pagerank	old	old
➡ Diameter/ANF	old	**DONE**
Conn. Comp	old	**DONE**
Triangles	**DONE**	
Visualization	**STARTED**	

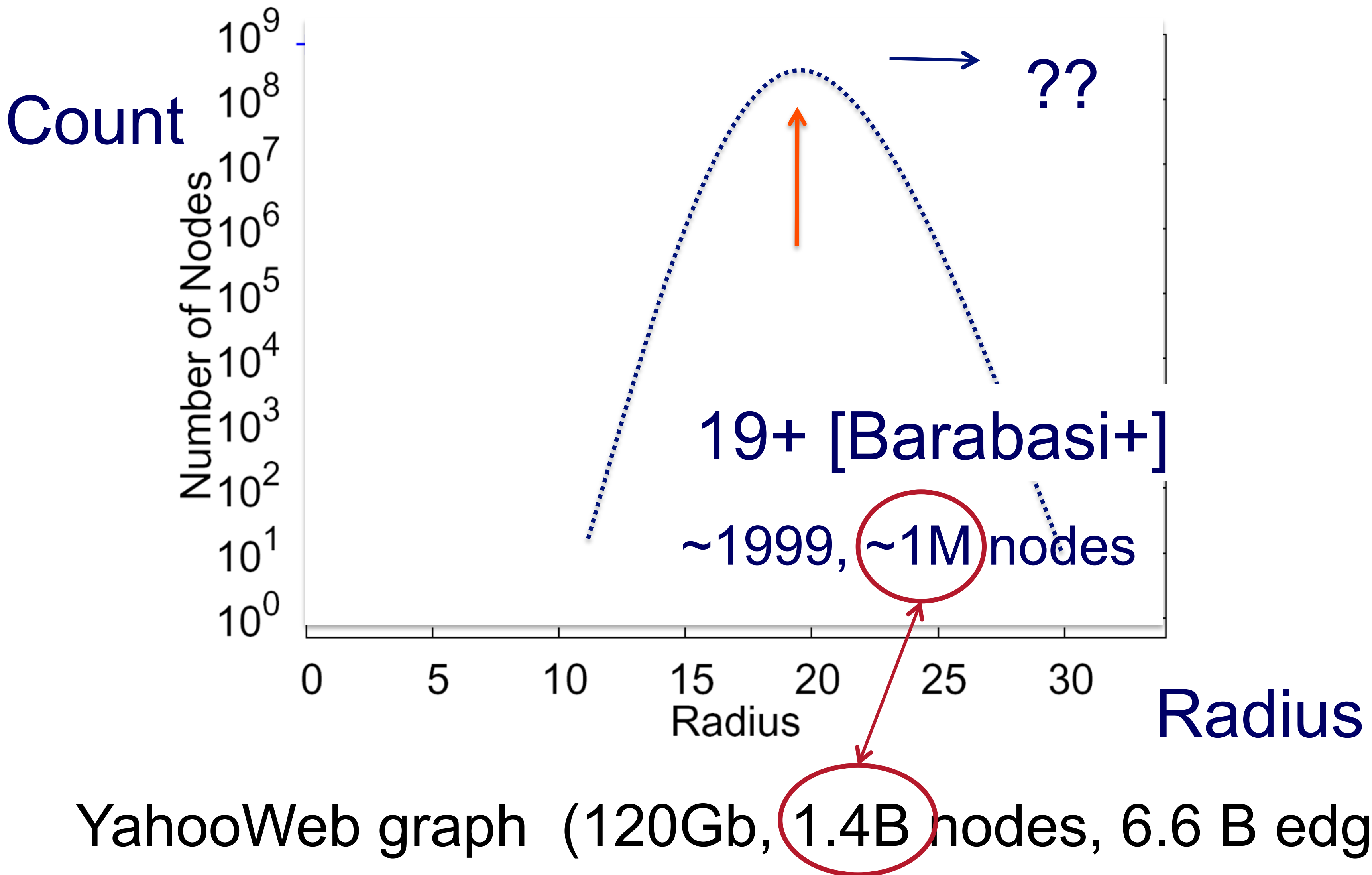
Carnegie Mellon

HADI for diameter estimation

- *Radius Plots for Mining Tera-byte Scale Graphs* **U Kang**, Charalampos Tsourakakis, Ana Paula Appel, Christos Faloutsos, Jure Leskovec, SDM'10
- Naively: diameter needs **O(N**2)** space and up to O(N**3) time – **prohibitive** (N~1B)
- Our HADI: linear on E (~10B)
  - Near-linear scalability wrt # machines
  - Several optimizations -> 5x faster

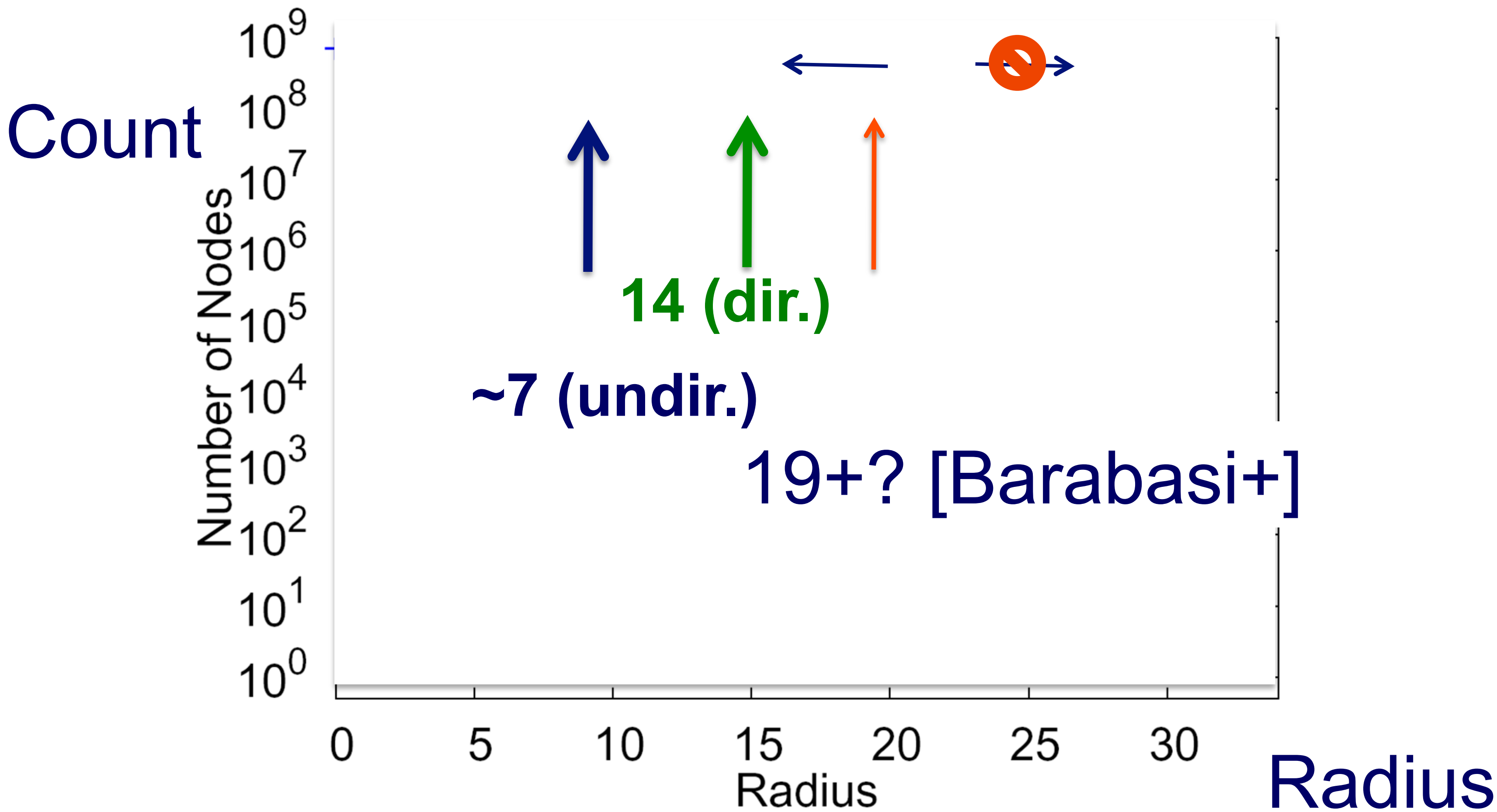


Count
Number of Nodes
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
5
10
15
20
25
30
Radius
19+ [Barabasi+]
~1999, ~1M nodes
Radius



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

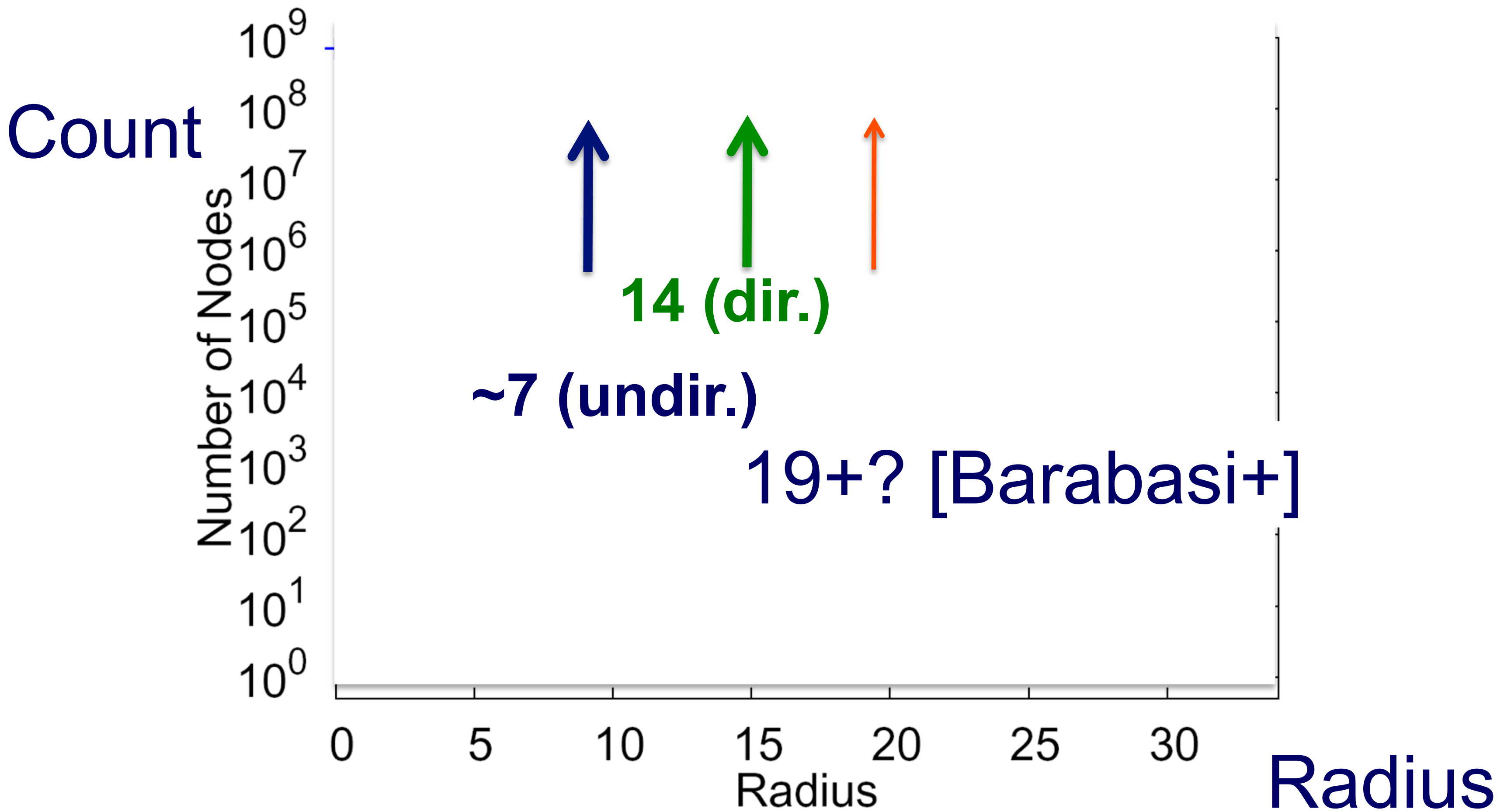
- Largest publicly available graph ever studied.



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

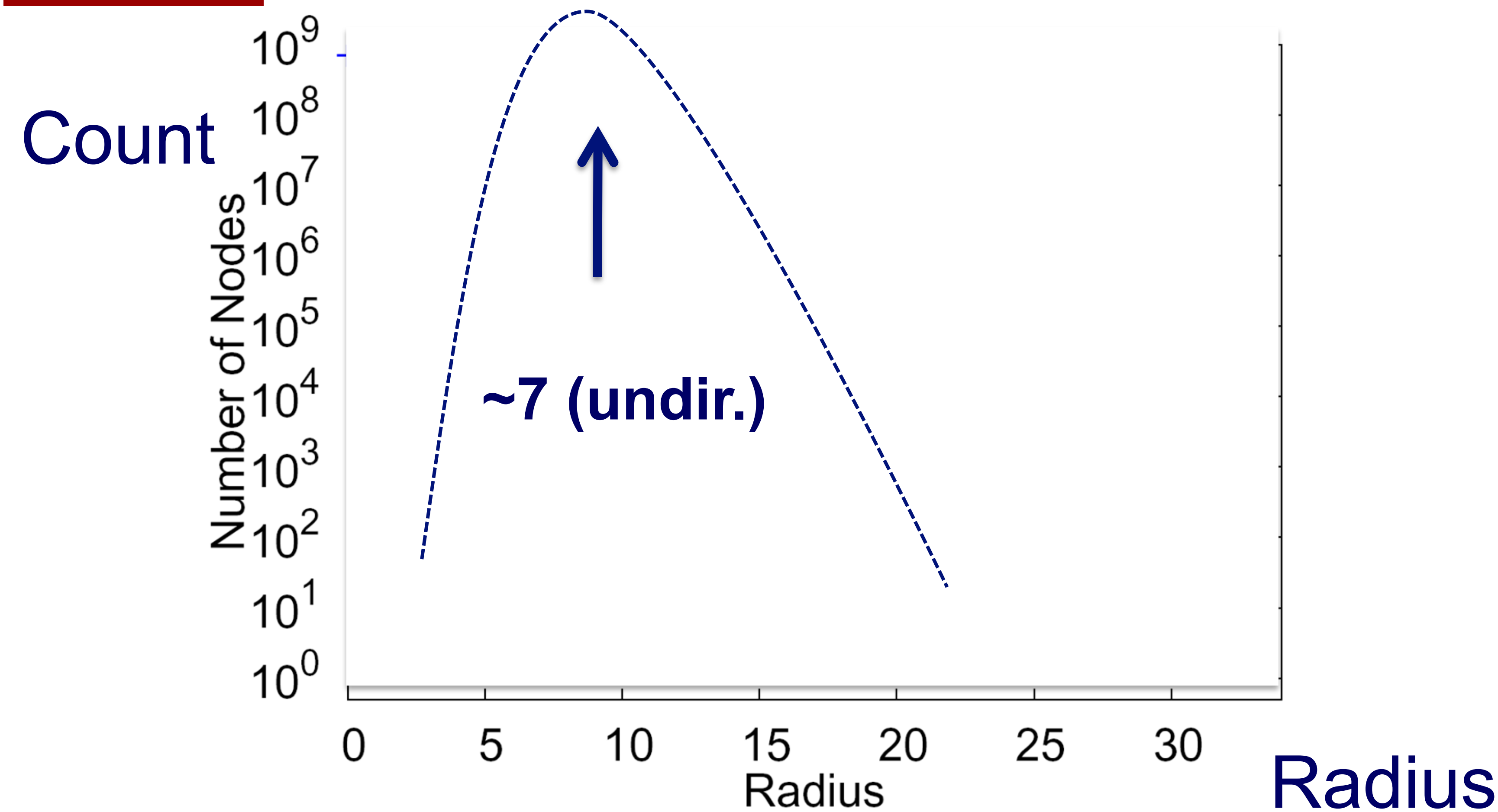
- Largest publicly available graph ever studied.

Carnegie Mellon



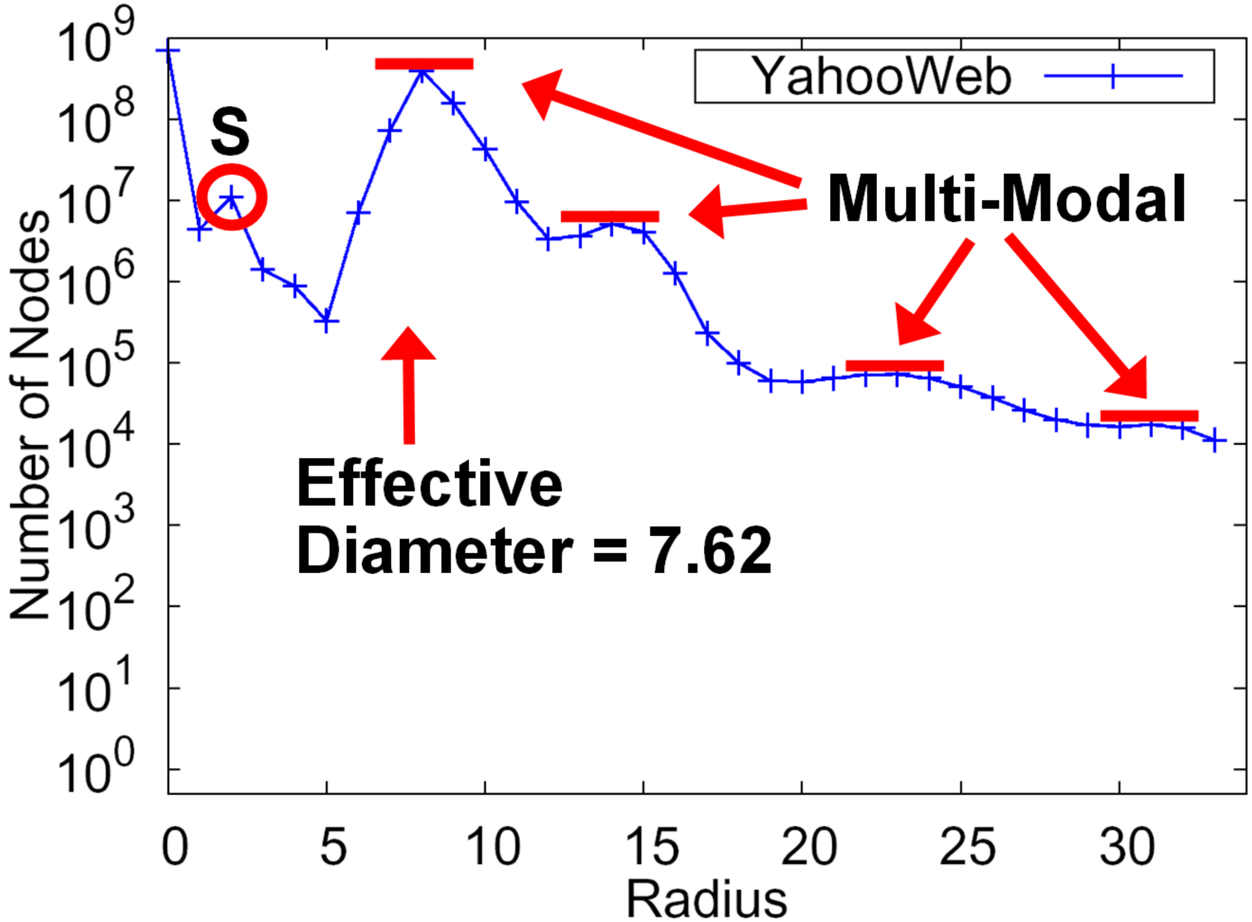
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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



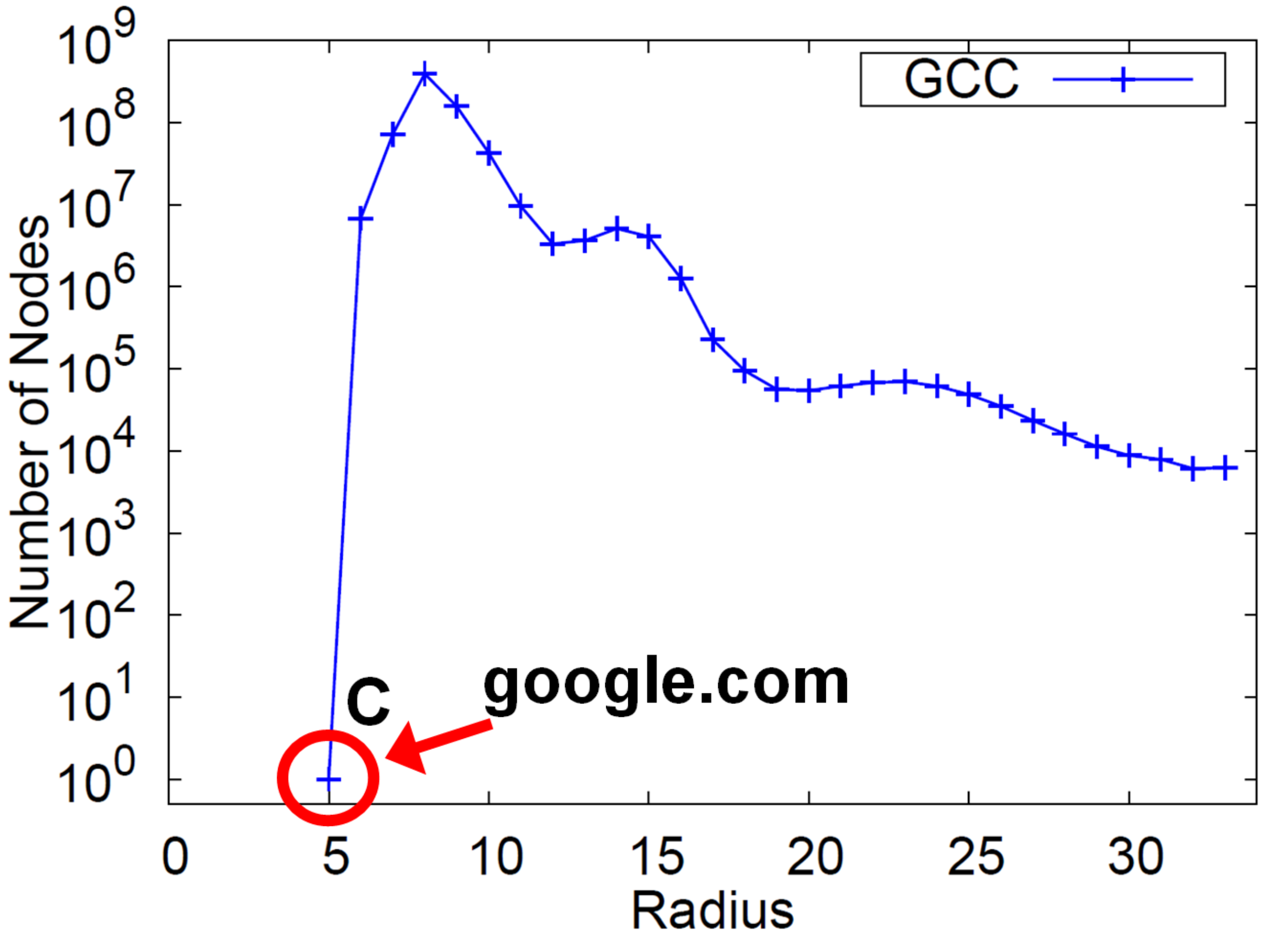
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

Q: Shape?

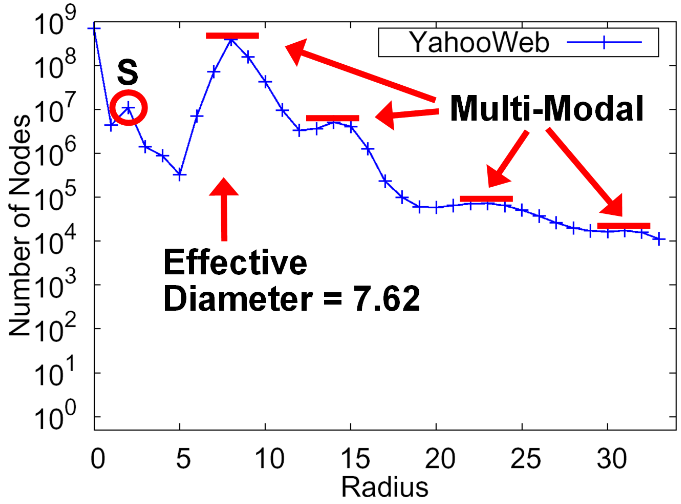


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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



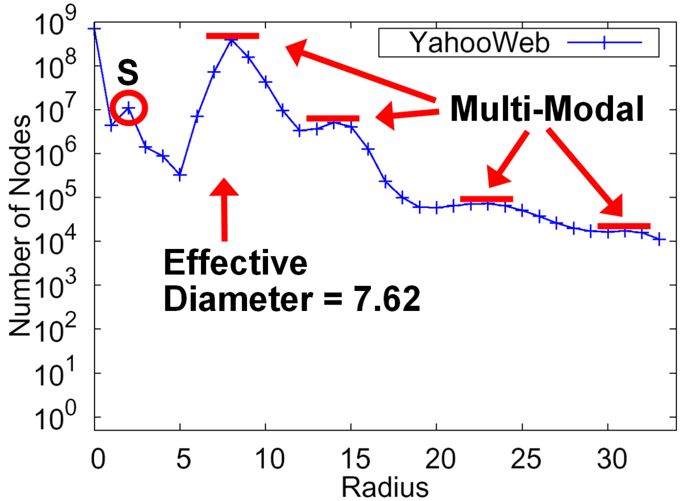
Radius Plot of **GCC** of YahooWeb.



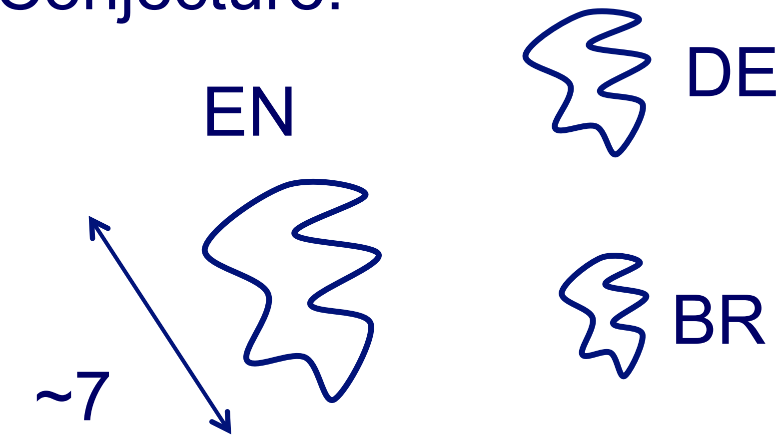
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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

Carnegie Mellon

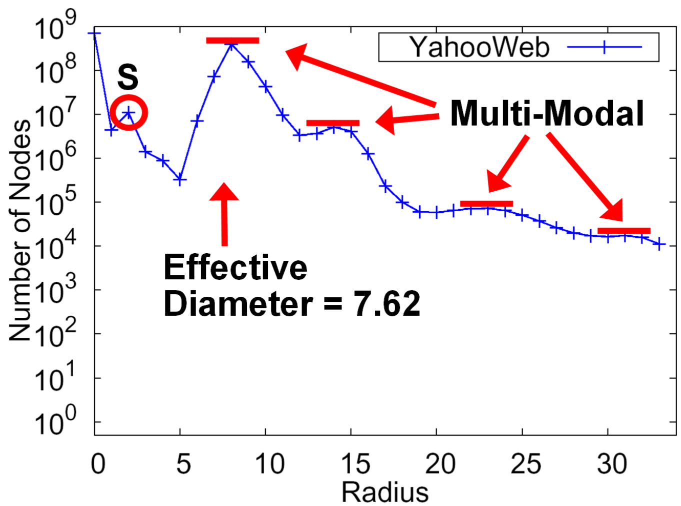


Conjecture:

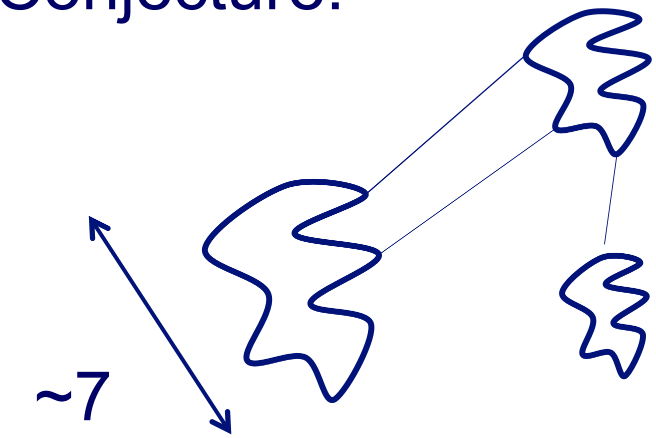


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

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



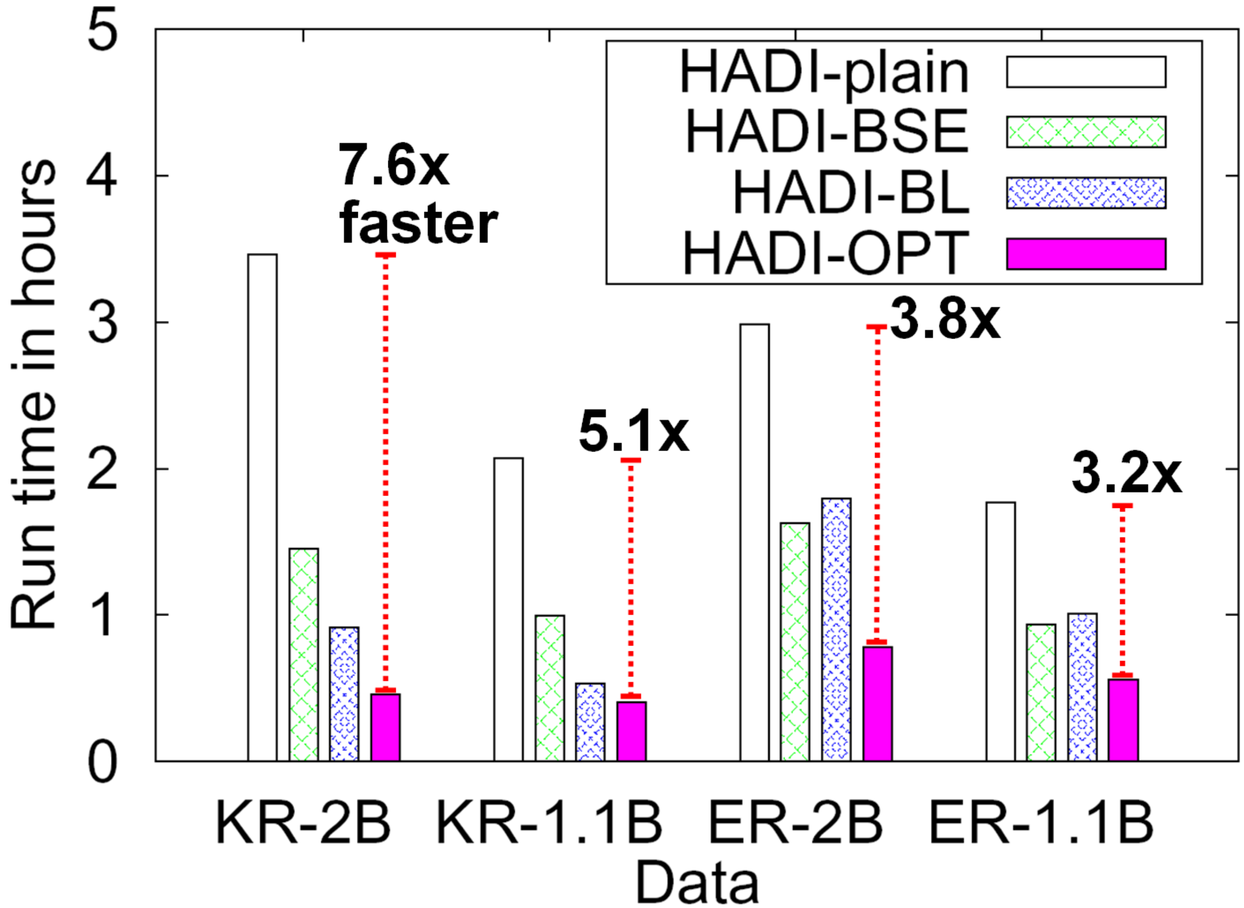
Conjecture:



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

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

details



Running time - Kronecker and Erdos-Renyi Graphs with billions edges.

Outline – Algorithms & results

	Centralized	Hadoop/ PEGASUS
Degree Distr.	old	old
Pagerank	old	old
Diameter/ANF	old	**DONE**
➡ Conn. Comp	old	**DONE**
Triangles	**DONE**	
Visualization	**STARTED**	

Carnegie Mellon

Generalized Iterated Matrix Vector Multiplication (GIMV)

PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations.
U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos.
(ICDM) 2009, Miami, Florida, USA.
Best Application Paper (runner-up).

Generalized Iterated Matrix Vector Multiplication (GIMV)

details

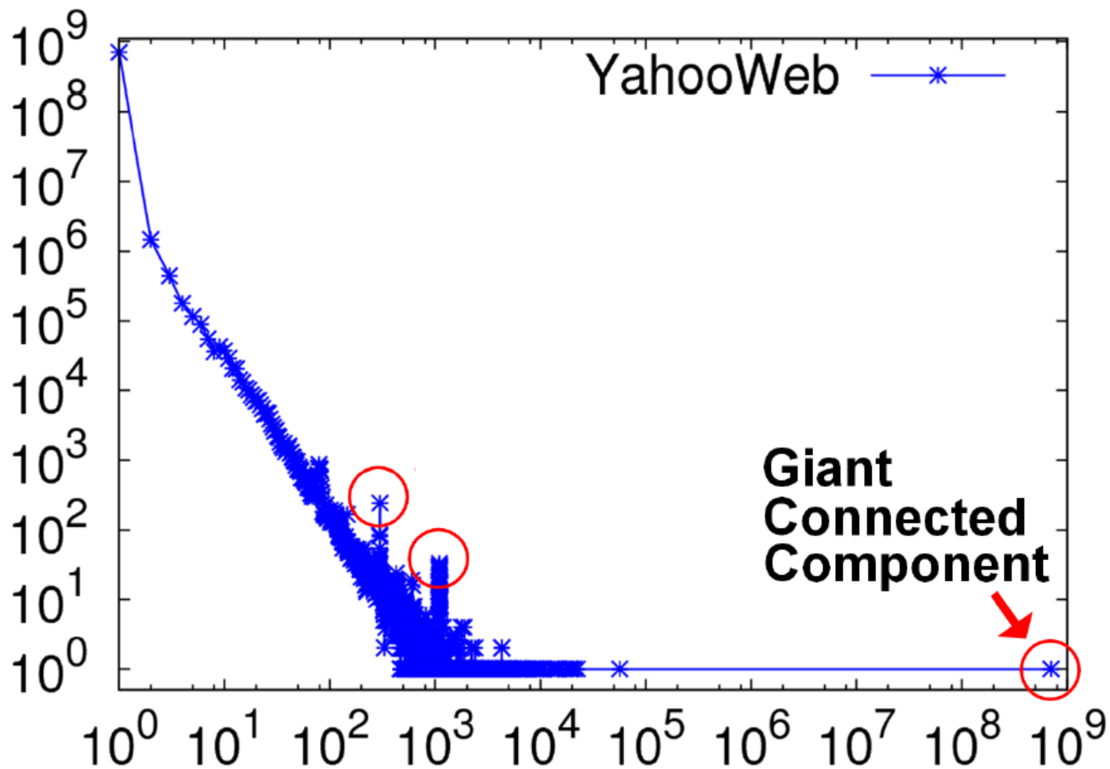
- PageRank
- proximity (RWR)
- Diameter
- Connected components
- (eigenvectors,
- Belief Prop.
- …)

Matrix – vector Multiplication (iterated)

Example: GIM-V At Work

- Connected Components – 4 observations:

Count

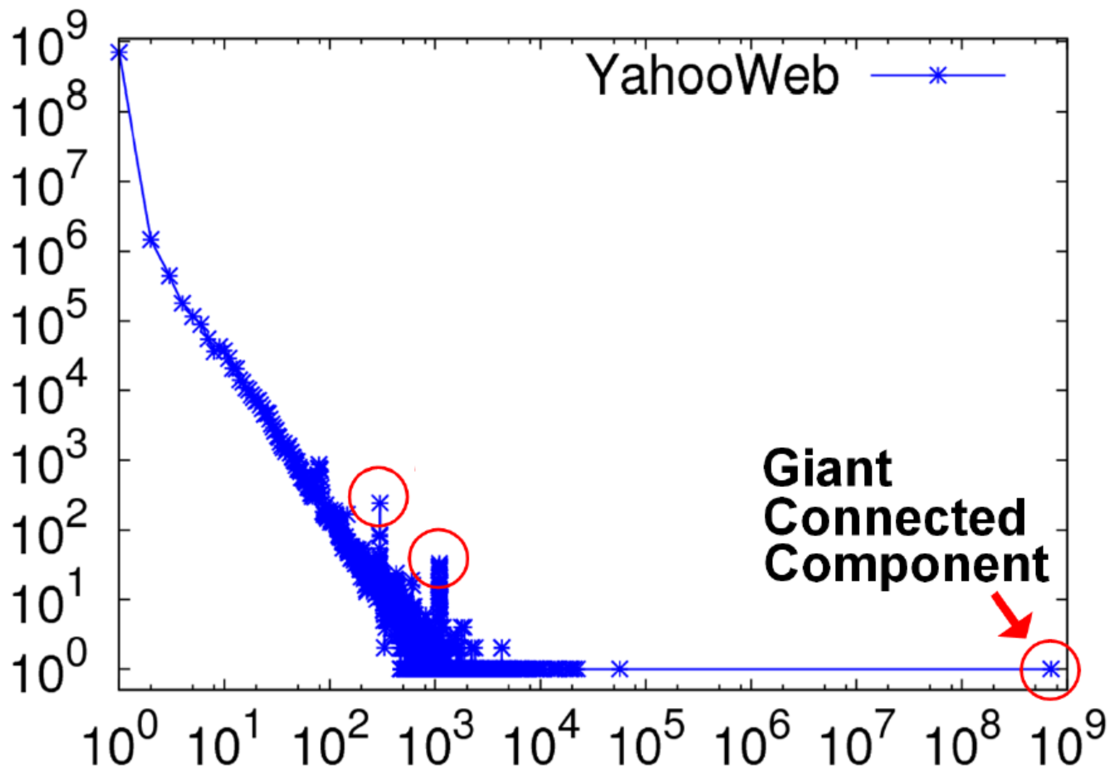


Size

Example: GIM-V At Work

- Connected Components

Count

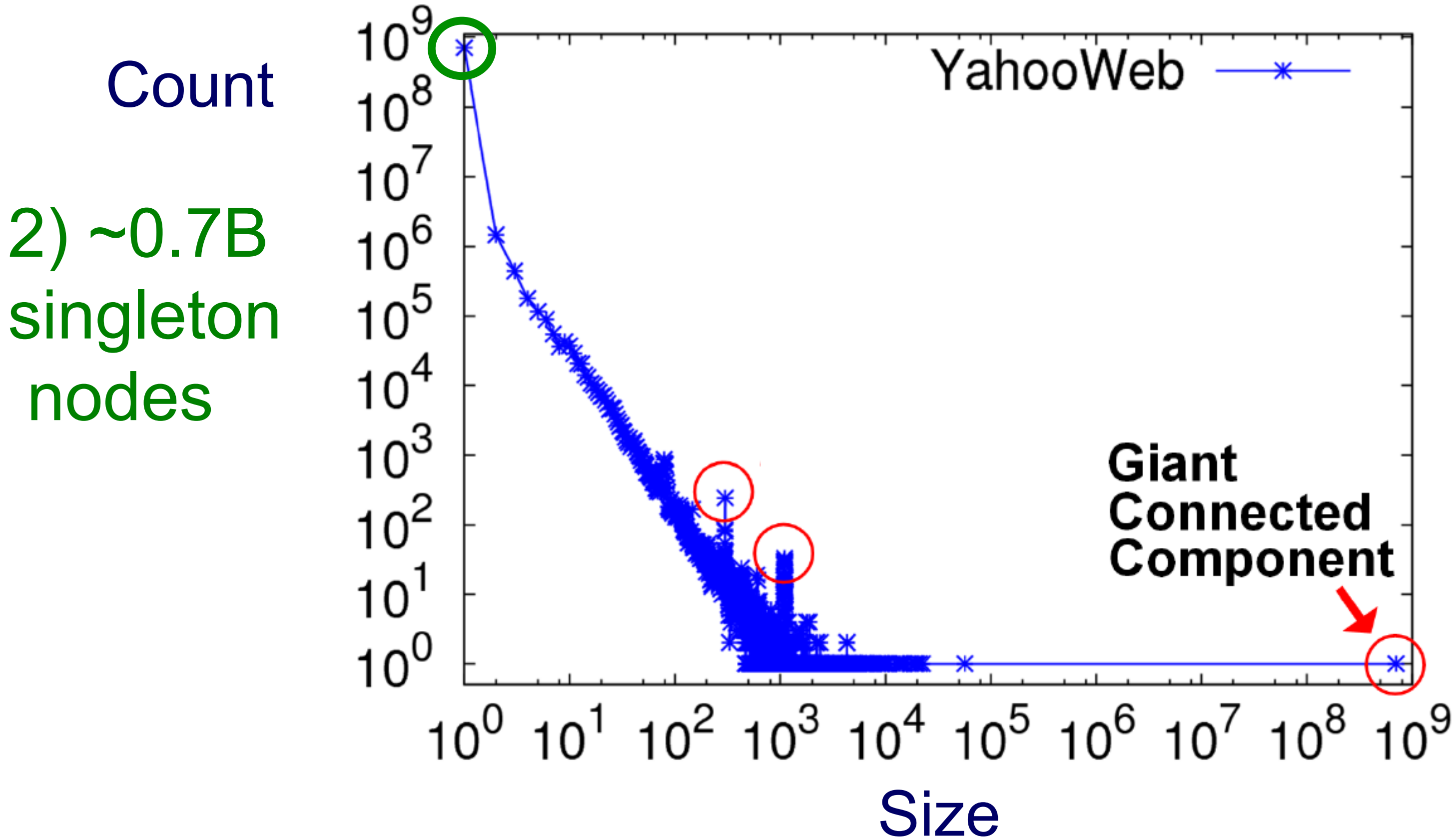


Size

1) 10K x larger than next

Example: GIM-V At Work

- Connected Components

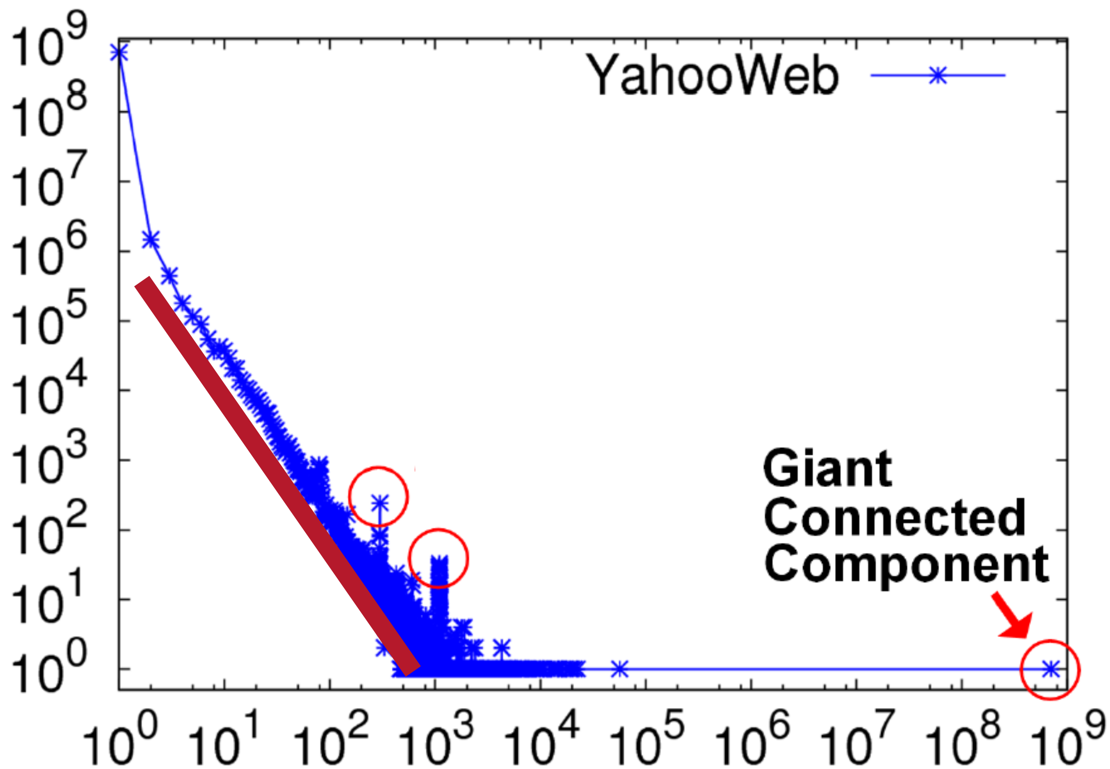


Example: GIM-V At Work

- Connected Components

Count

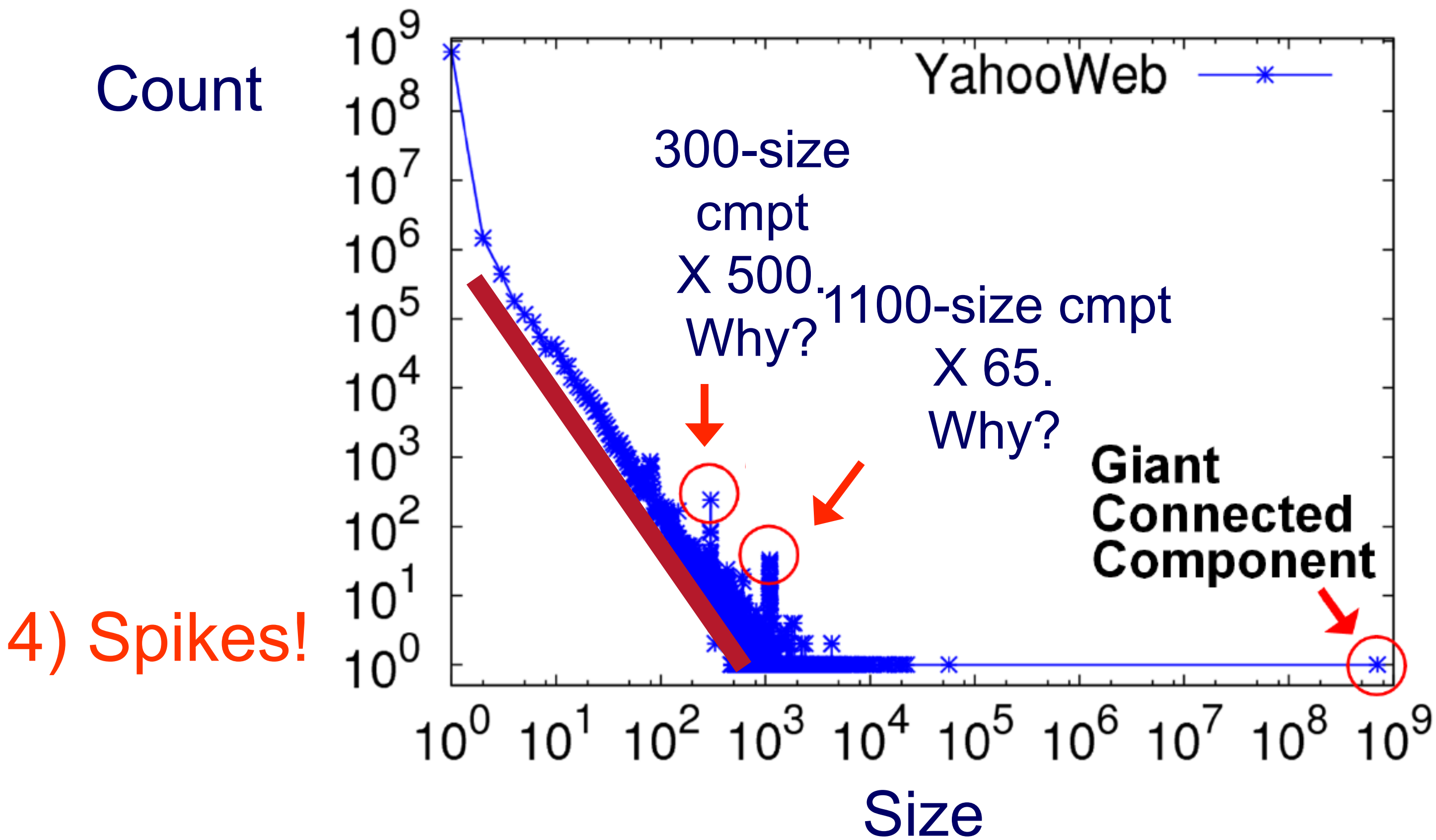
3) SLOPE!



Size

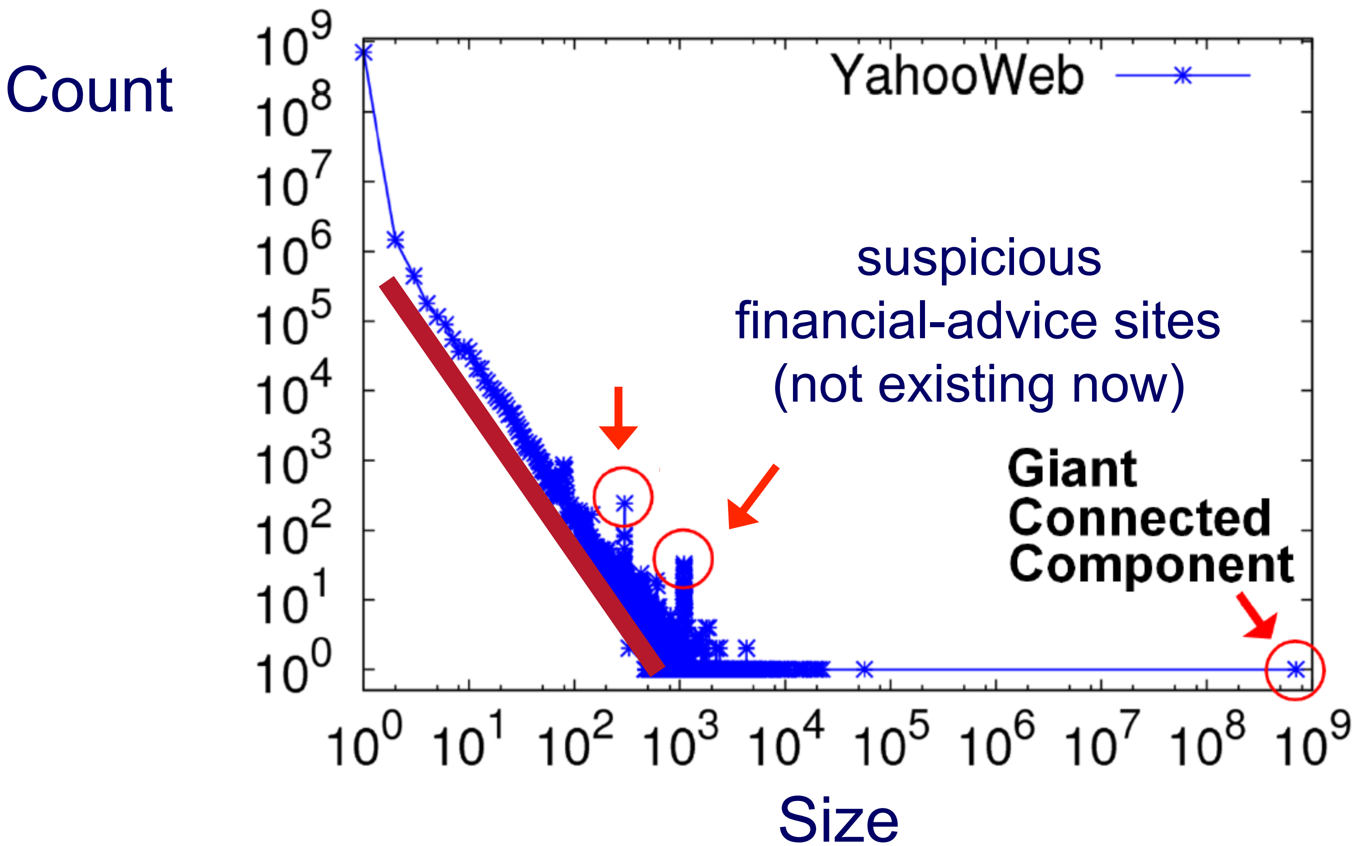
Example: GIM-V At Work

- Connected Components



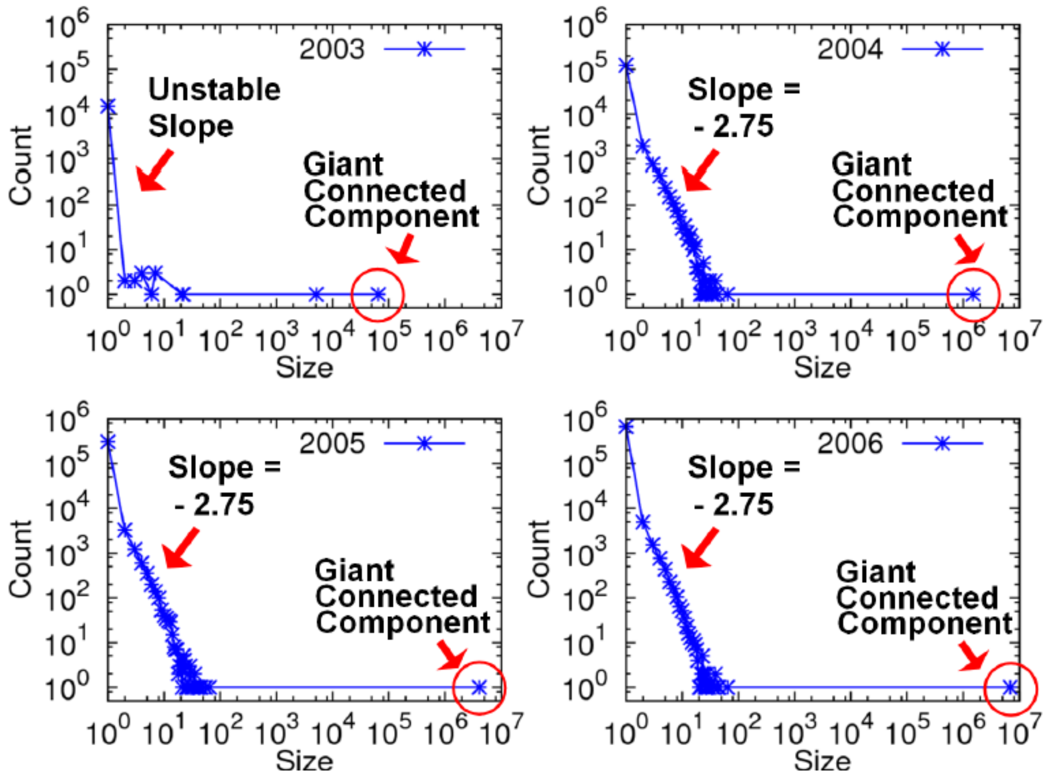
Example: GIM-V At Work

- Connected Components



GIM-V At Work

- Connected Components over Time
- **LinkedIn: 7.5M nodes and 58M edges**



Stable tail slope after the gelling point

Conclusions

- Hadoop: promising architecture for Tera/ Peta scale graph mining

Resources:

- http://hadoop.apache.org/core/
- http://hadoop.apache.org/pig/

  Higher-level language for data processing

References

- Jeffrey Dean and Sanjay Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters,* OSDI'04
- Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, Andrew Tomkins: *Pig latin: a not-so-foreign language for data processing*. SIGMOD 2008: 1099-1110

Overall Conclusions

- Real graphs exhibit surprising **patterns** (power laws, shrinking diameter, super-linearity on edge weights, triangles etc)
- **SVD**: a powerful tool (HITS, PageRank)
- Several other tools: **tensors**, METIS, …
  - But: good communities might **not** exist…
- Immunization: **first** eigenvalue
- Scalability: **hadoop**/parallelism

Our goal:

Open source system for mining huge graphs:

PEGASUS project (PEta GrAph mining System)

- www.cs.cmu.edu/~pegasus
- code and papers

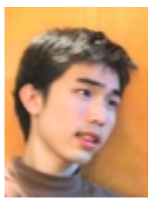
Carnegie Mellon

Project info

www.cs.cmu.edu/~pegasus

Akoglu, Leman

Chau, Polo



Kang, U

Koutra, Danae

McGlohon, Mary

Prakash, Aditya

Tong, Hanghang

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

Extra material

- E-bay fraud detection
- Outlier detection

Detailed outline

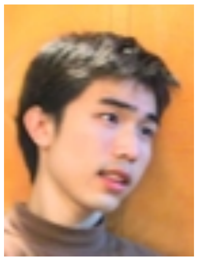
extra

- Fraud detection in e-bay
- Anomaly detection

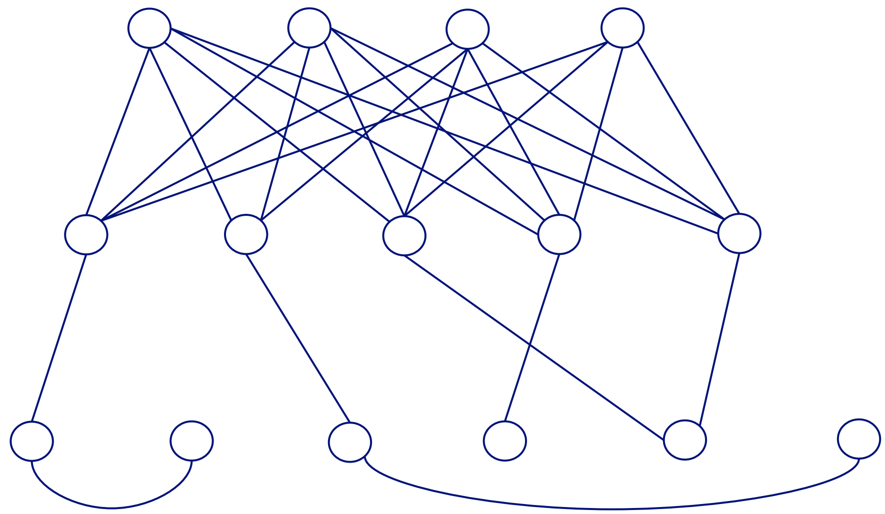
Carnegie Mellon

E-bay Fraud detection

extra



w/ Polo Chau &
Shashank Pandit, CMU

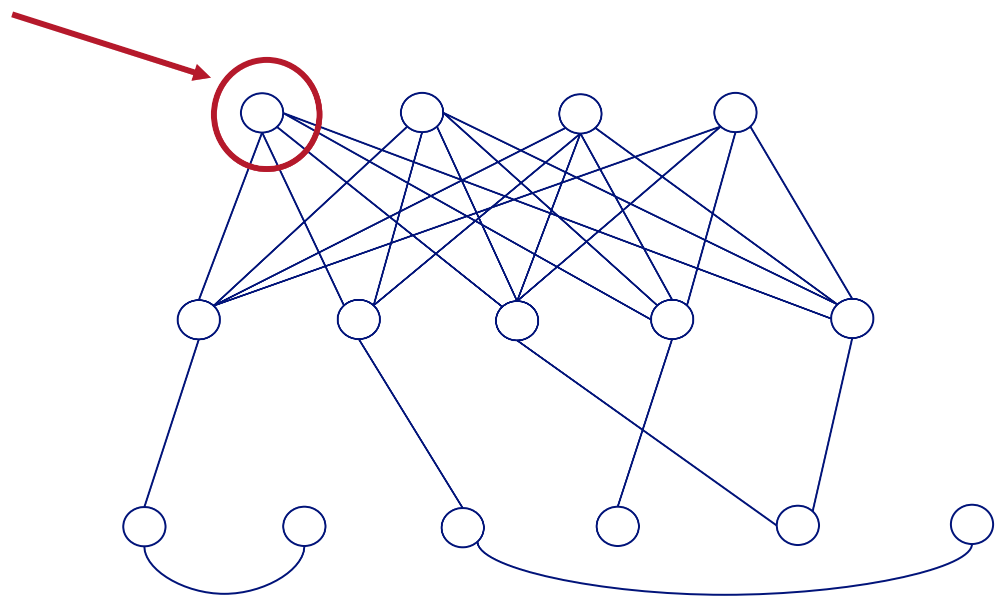


NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks, S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos *(WWW'07),* pp. 201-210

E-bay Fraud detection

extra

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

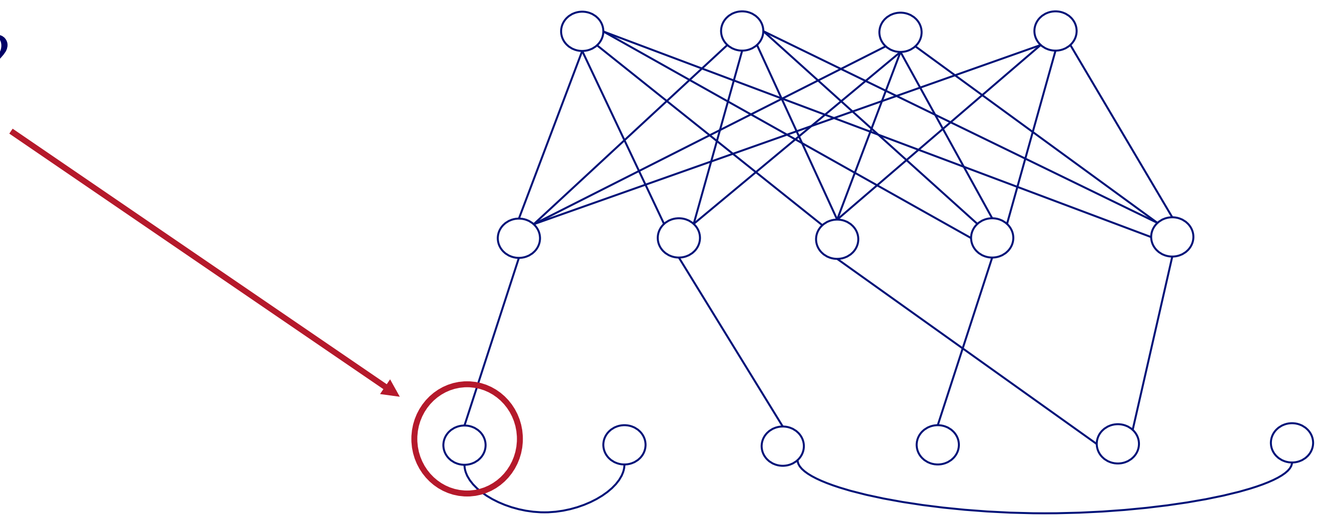


Carnegie Mellon

E-bay Fraud detection

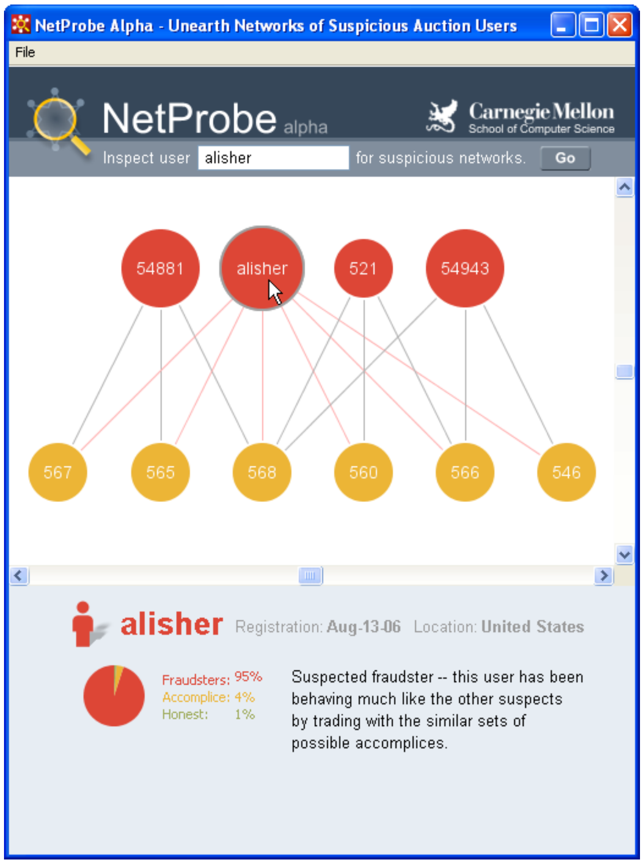
extra

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

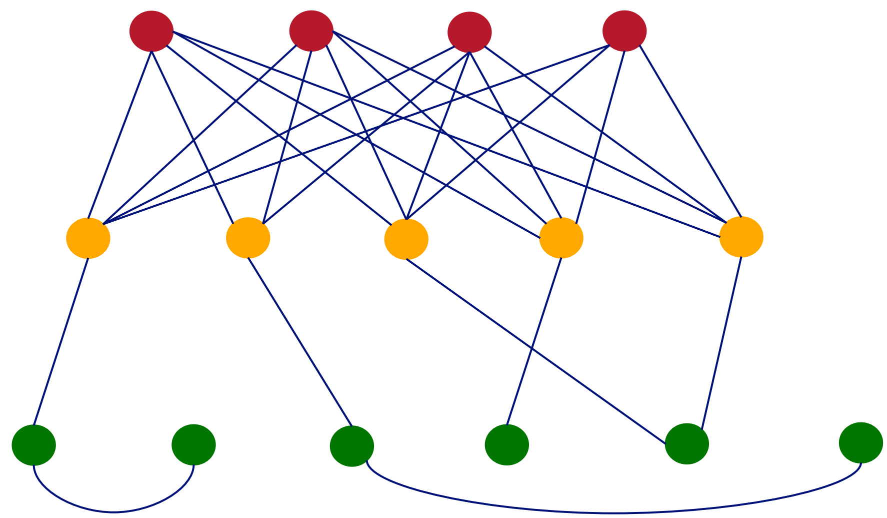


E-bay Fraud detection - NetProbe

extra



Belief Propagation gives:



Popular press

USA TODAY

The Washington Post

Los Angeles Times

And less desirable attention:

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

Extra material

- E-bay fraud detection
- Outlier detection

Carnegie Mellon

OddBall: Spotting Anomalies in Weighted Graphs

Leman Akoglu, Mary McGlohon, Christos Faloutsos

Carnegie Mellon University

School of Computer Science

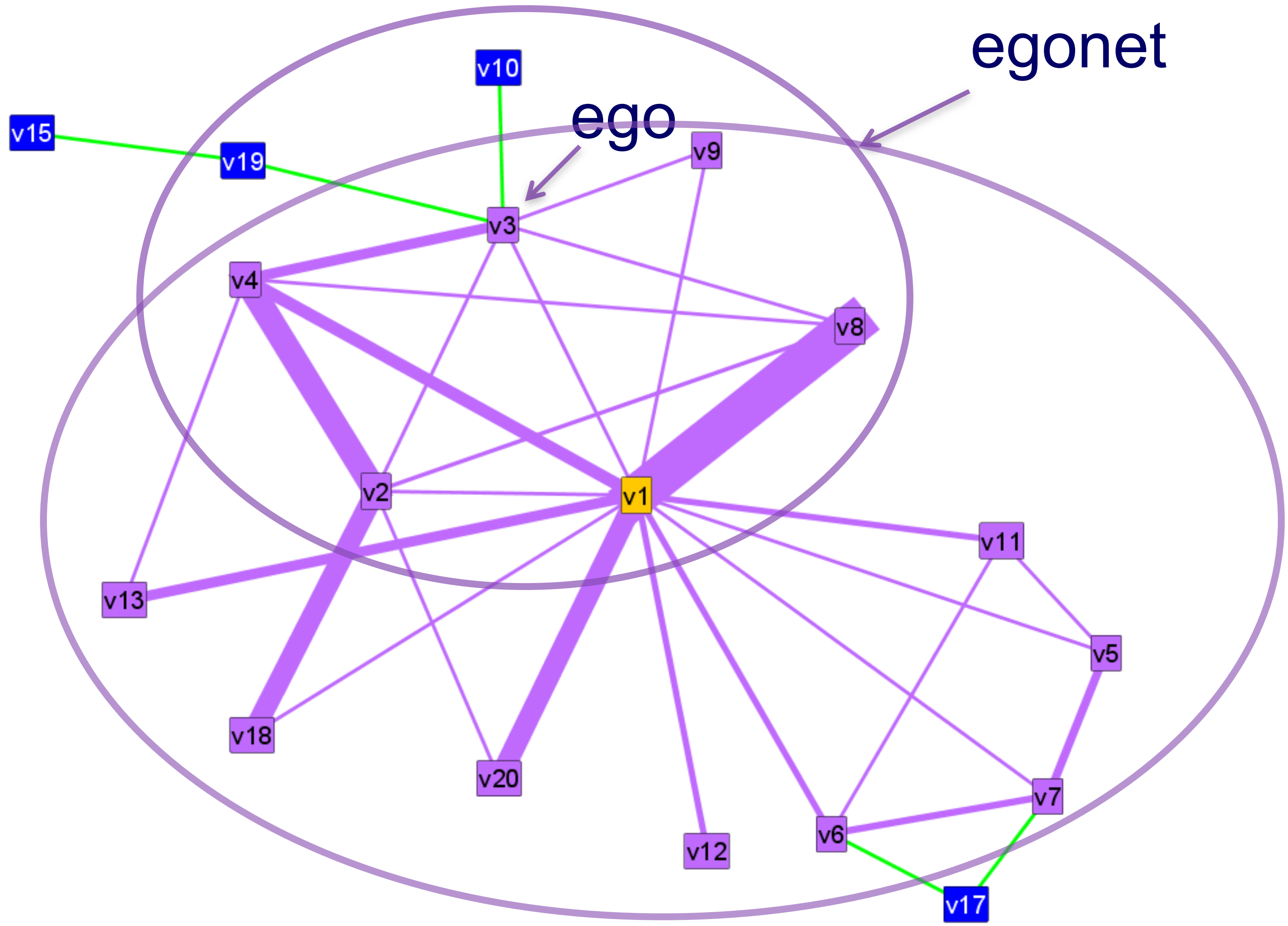
PAKDD 2010, Hyderabad, India

Main idea

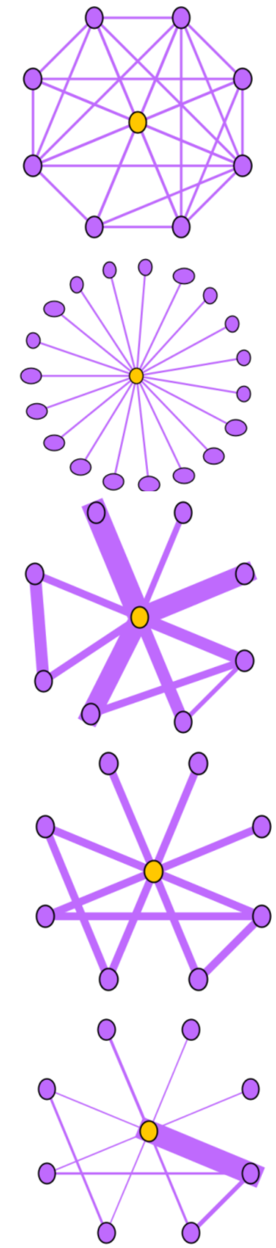
For each node,

- extract 'ego-net' (=1-step-away neighbors)
- Extract features (#edges, total weight, etc etc)
- Compare with the rest of the population

What is an egonet?

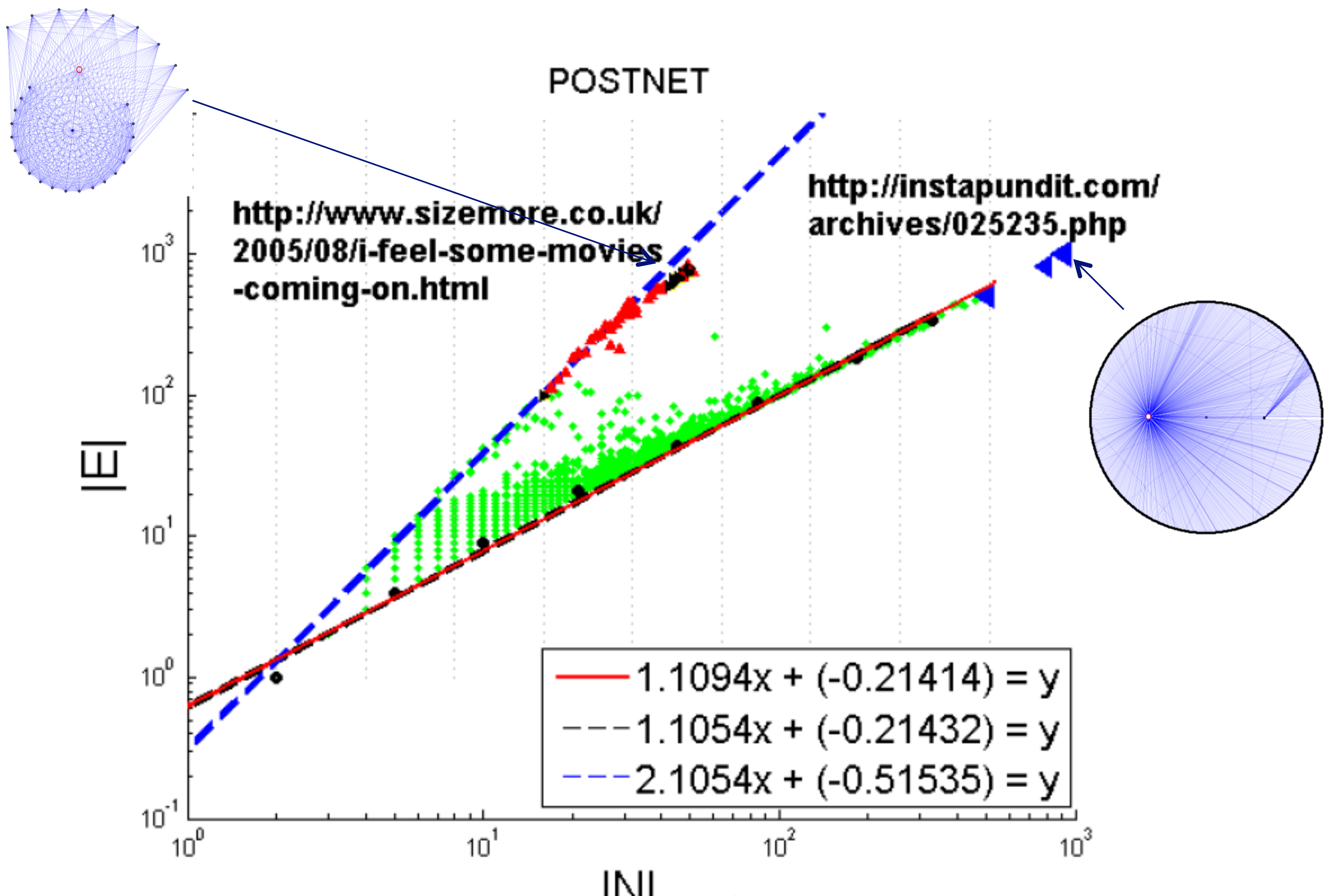


Selected Features



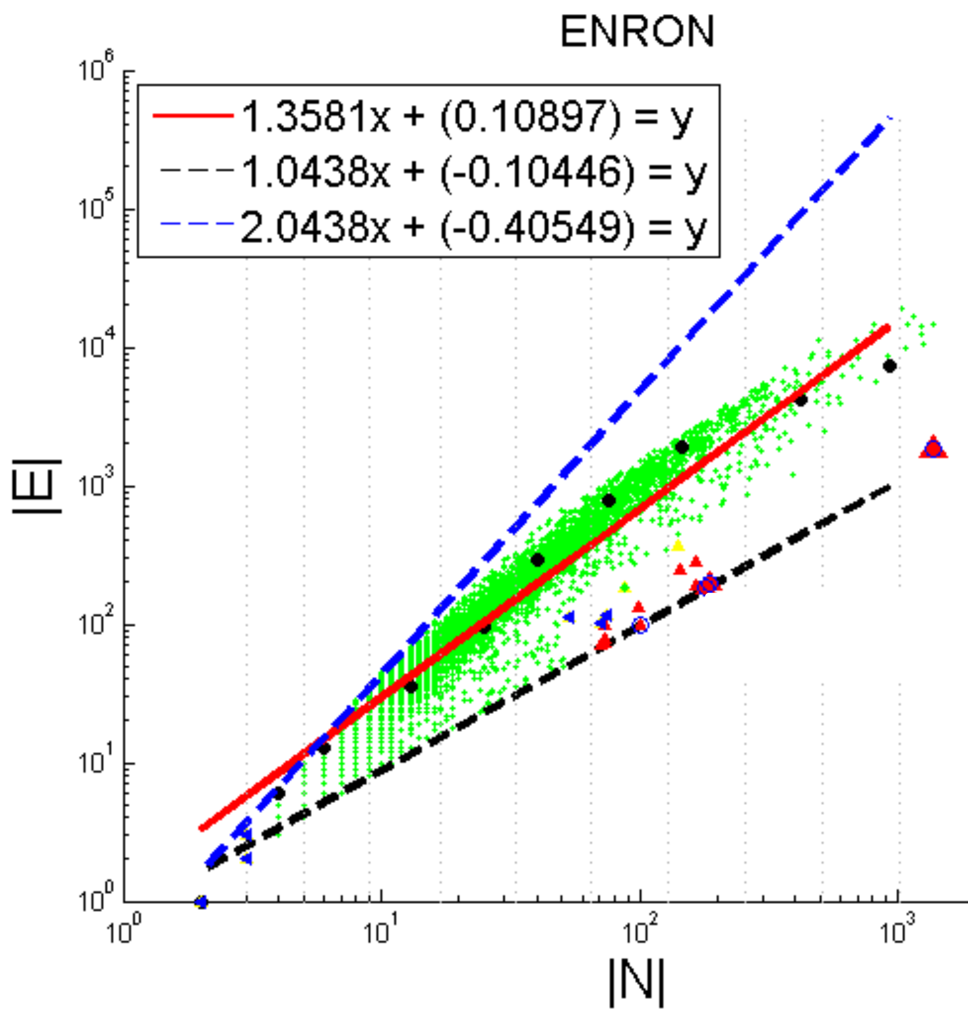
- N_i: number of neighbors (degree) of ego i
- E_i: number of edges in egonet i
- W_i: total weight of egonet i
- $\lambda_{w,i}$: principal eigenvalue of the weighted adjacency matrix of egonet I

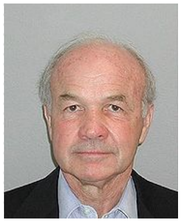
Near-Clique/Star



Carnegie Mellon

Near-Clique/Star





Carnegie Mellon

END