Numerical Linear Algebra for Data and Link Analysis

Leonid Zhukov, Yahoo! Inc

David Gleich, Stanford

SIAM PP-2006, San Francisco, CA

Scientific computing

- **Engineering Computing**
  - continuum problems, PDE governed, control over discretization
  - 2D or 3D
  - Uniform distribution of node degrees

- **Information Retrieval**
  - Discrete data is given, no control over resolution
  - No associated physical space (coordinates)
  - Not planar, triangulation
  - Power-low degree distribution
  - “Small world” effect

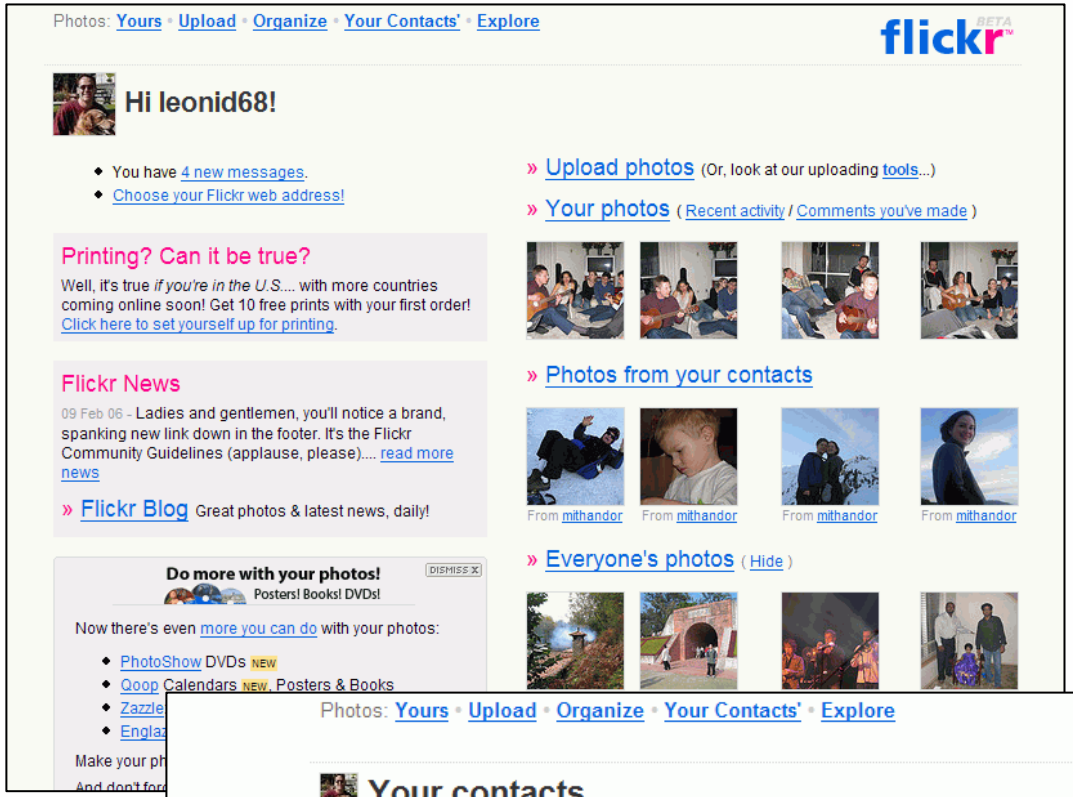
Large graphs

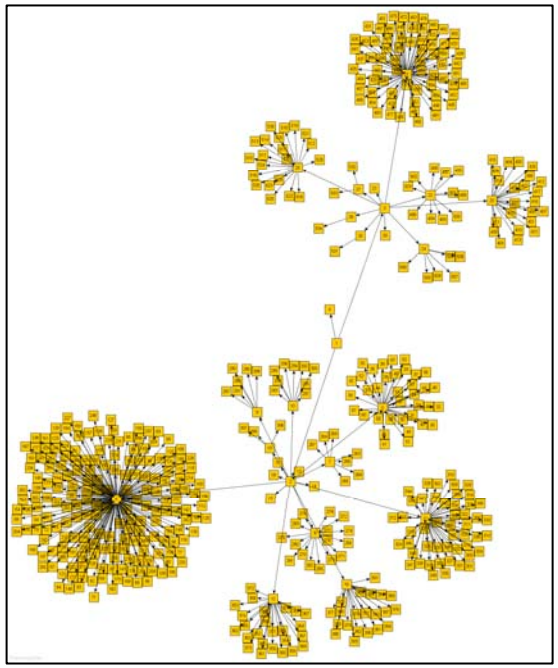
- **Explicit: graphs & networks**
  - Web graph
  - Internet
  - Yahoo! Photo Sharing (Flickr)
  - Yahoo! 360 (Social network)

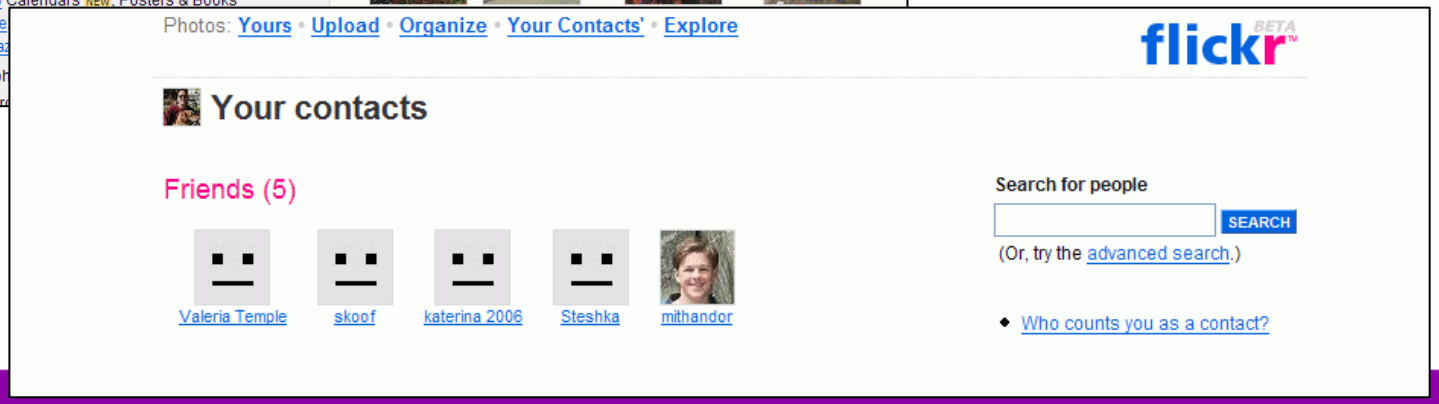
- **Implicit: transaction history, email, messenger**
  - Yahoo! Search marketing (Overture)
  - Yahoo! Mail
  - Yahoo Messenger

- **Constructed: affinity between data points**
  - Yahoo! Music (Launch)
  - Yahoo! Movies
  - Yahoo! ….

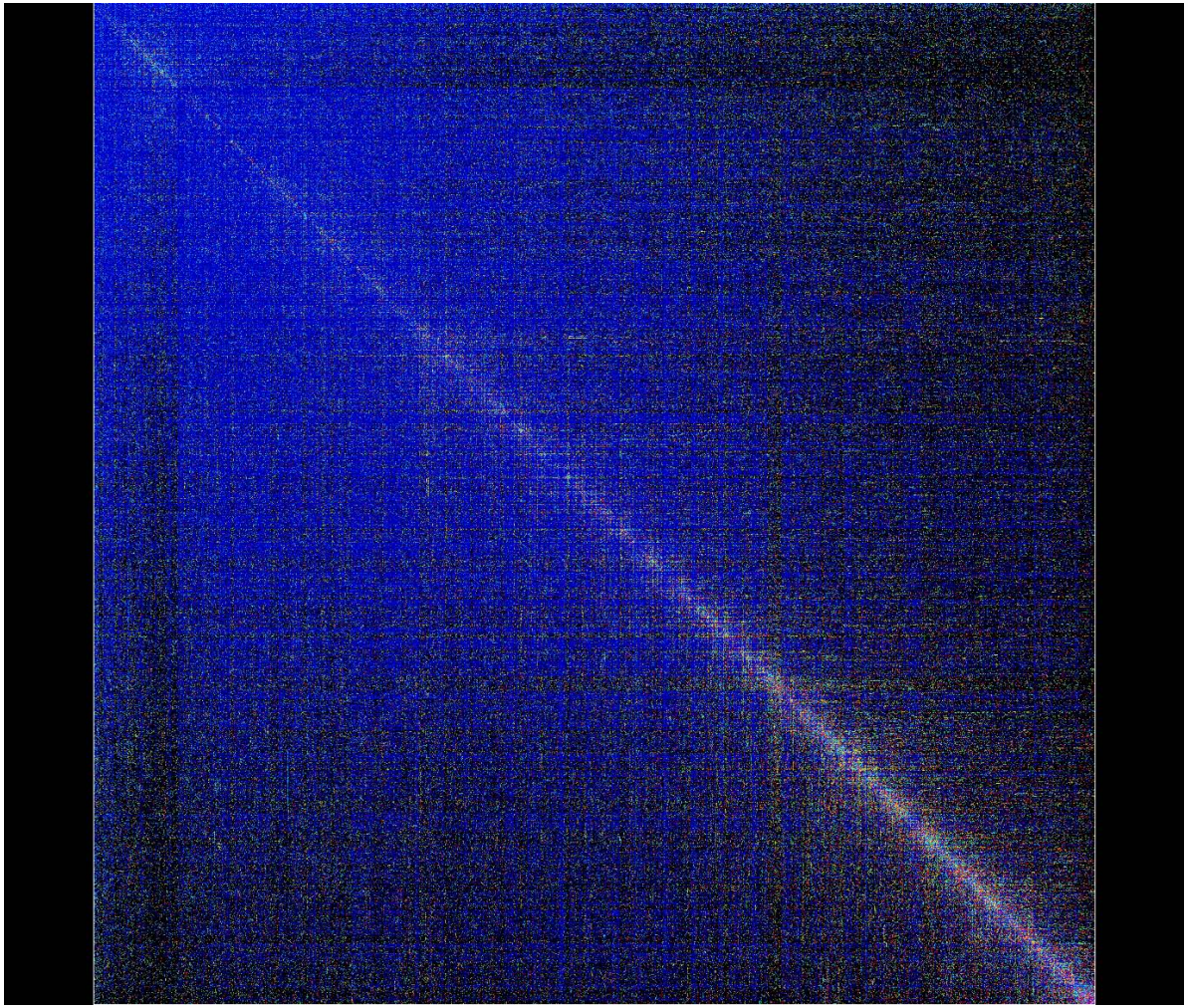
Flickr – social network





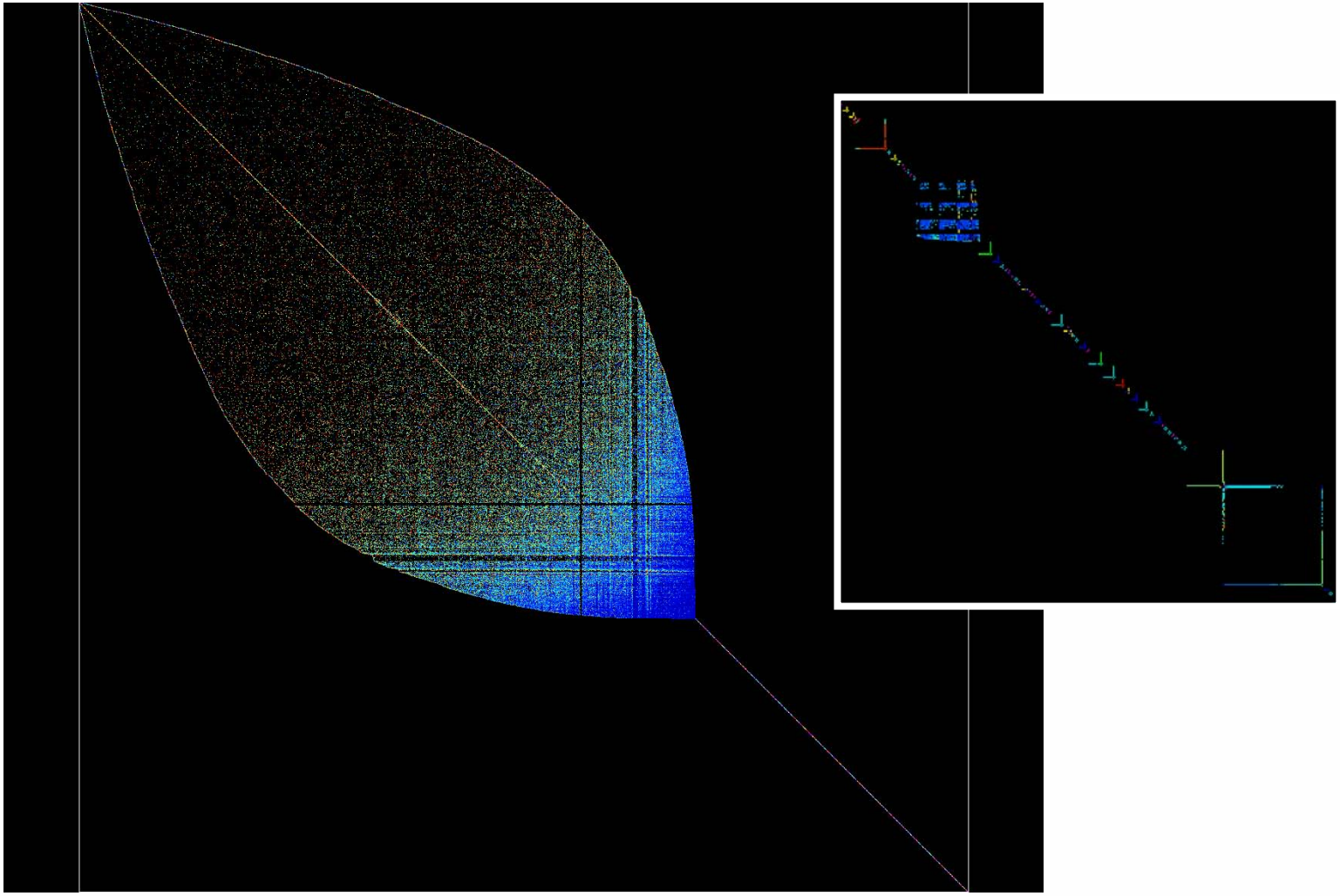


Flickr: adjacency matrix



580,000 users, 3,500,000 links

Flickr: reverse Cuthill-McKee



580,000 users, 3,500,000 links

Flickr – a graph

Flickr: some stats

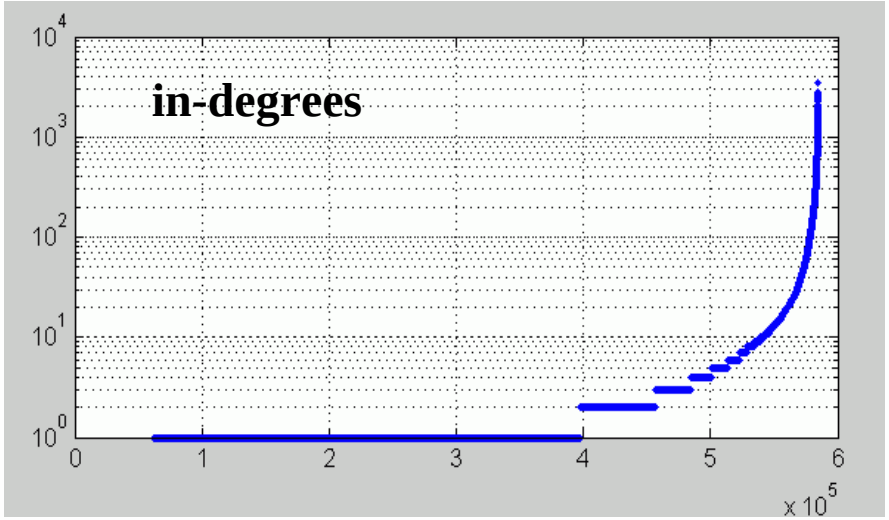
- # nodes = 584,207
- # edges (nnz) = 3,555,115

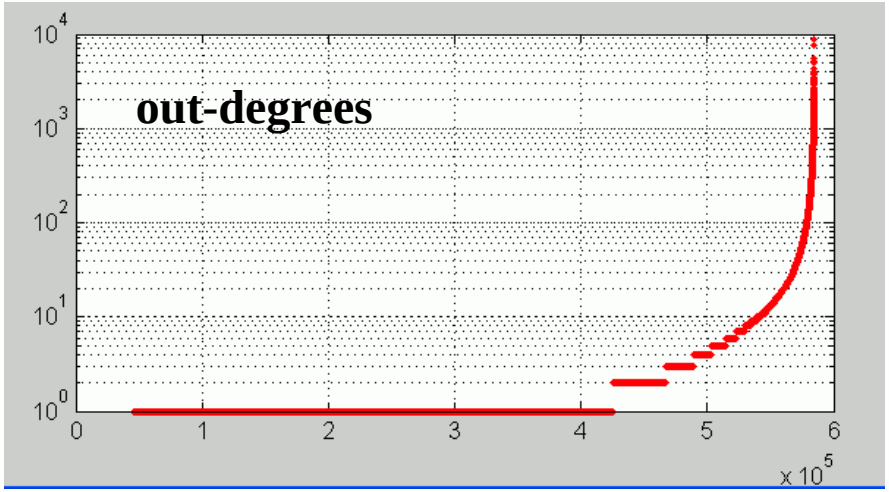
- max in-degree = 3531
- max out-degree = 8976
- < in-degree > = < out-degree > = 6
- diameter = 18

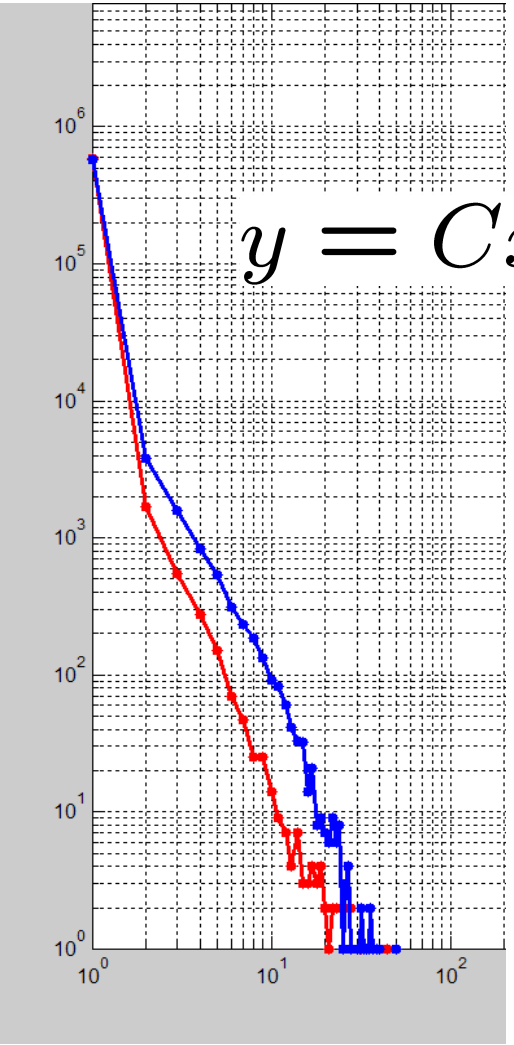
- # strongly connected components = 152,324
- largest strongly connected components = 274,649 374 186 155 11

- # connected component = 43,189
- largest connected components = 404,893 378 112 108 103
- highest core number = 249 (size 668)

Flickr: scale free graph







Data and links analysis

- **Graph partitioning:**
  - “balanced”, for data distribution
  - “natural boundaries”, for clustering / communities

- **Numerical computing on graphs:**
  - Node ranking methods (PageRank)
  - Dimensionality reduction (SVD, HITS)
  - Low dimensional embeddings
  - Graph interpolation / regularization

What do we compute?

- Probability matrix:

$$P = D_{out}^{-1} A$$

- Graph Laplacian:

$$L = D - A, \qquad L_n = D^{-1/2} L D^{1/2}$$

$$L_d = \Pi - \frac{P^T \Pi + \Pi P}{2}$$

- SVD/LSI:

$$A = U S V^T$$

Computing eigenvalues / eigenvectors, linear systems

Yahoo! Search Marketing (Overture)

Web | Images | Video | Directory | Local | News | Shopping

YAHOO! SEARCH epson Search

My Web BETA Subscriptions (New) Shortcuts Advanced Search Preferences

Search Results Results 1 - 10 of about 28,800,000 for epson - 0.02 sec. (About this page)

Also try: epson printers, epson driver, epson p-2000, epson scanners More...

SPONSOR RESULTS

- **Epson** Go with an industry favorite in business printers-Hewlett-Packard. Reliable, high-quality laser, color and all-in-one printers. Compare HP printers head-to-head with **Epson**.
  www.hp.com
- **Epson Compatible Ink Cartridges - $6.95** High-quality **Epson** compatible inkjet cartridges from Inkjetcartridge.com. Toll-free sales and support. Yahoo five-star award-winning service.
  www.inkjetcartridge.com
- **75% off Epson Ink - Free Shipping** Up to 75% off **Epson** ink and toner, plus a one-year quality guarantee and free shipping. Save an extra 5% with coupon code over15. Free promotional items with every purchase.
  www.123inkjets.com

Y! **Epson** - Ink Jet Printers - Scanners - Projectors
Yahoo! Shortcut - About

1. **Epson**
   creates and sells digital imaging products including printers, digital cameras, scanners, and projectors. Also offers ink jet cartridges, paper, and a variety of other supplies and accessories.
   Category: B2B Imaging Equipment
   www.**epson**.com - 8k - Cached - More from this site - Save - Block

SPONSOR RESULTS

Epson Batteries and Chargers - eBatts
eBatts sells **Epson** batteries and chargers. Batteries and chargers for...
www.ebatts.com

Save on **Epson** Inks
Buy 2 **Epson** printer ink cartridges, get 1 cartridge free. Free 2 day...
www.clickinks.com

Epson Inkjet Cartridges - Save
Save up to 80% on **Epson** inkjet cartridges. We offer a 100% money-back...
www.printpal.com

Epson Multimedia Projectors
Shop with us to save money on multimedia projection equipment.
www.projectorsforsale.com

Epson Compatible Ink

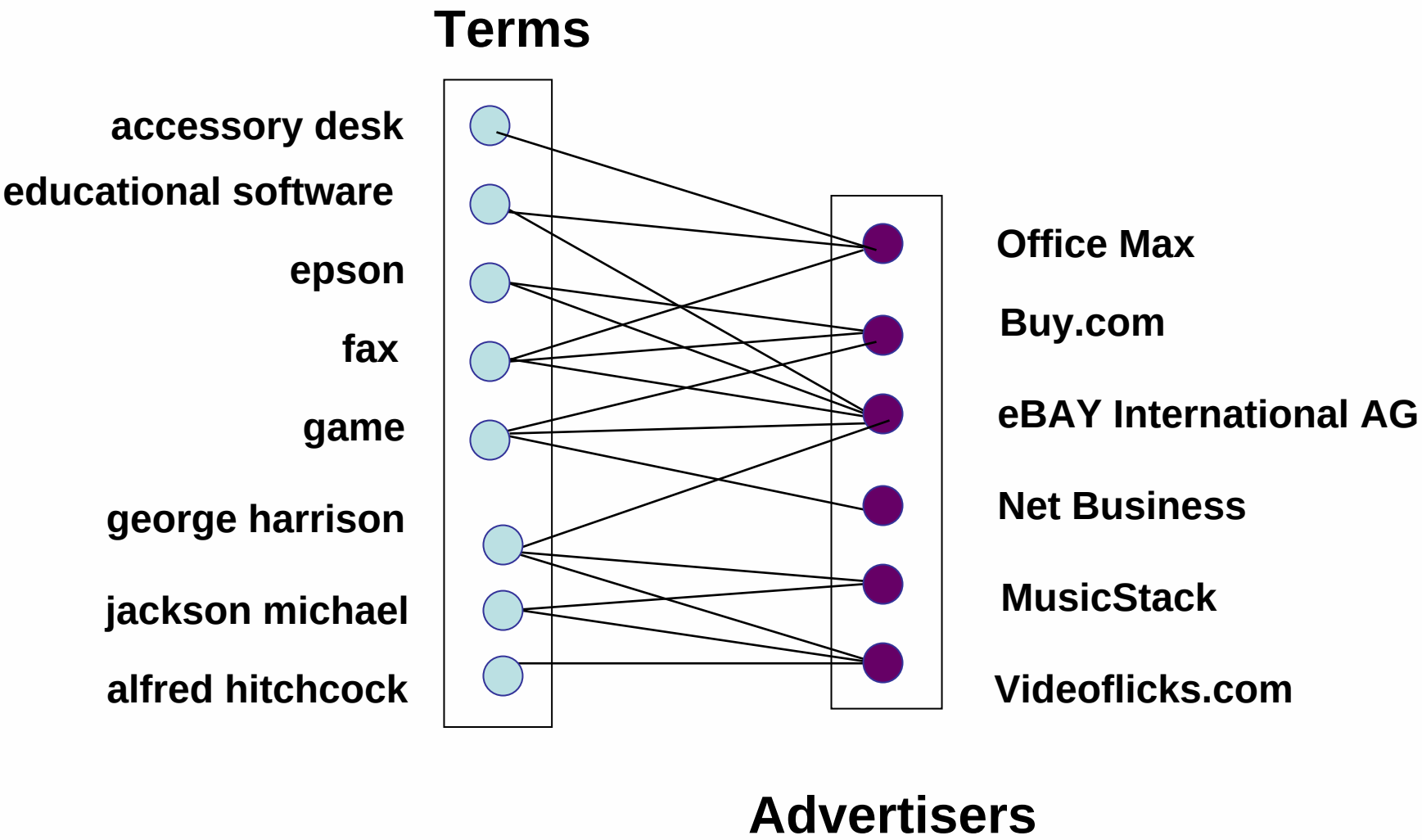
View Bids Tool

1. **Epson**
   Go with an industry favorite in business printers-Hewlett-Packard. Reliable, high-quality laser, color and all-in-one printers. Compare HP printers head-to-head with Epson.
   www.hp.com
   (Advertiser's Max Bid: $1.80)
2. **Epson Compatible Ink Cartridges - $6.95**
   High-quality Epson compatible inkjet cartridges from Inkjetcartridge.com. Toll-free sales and support. Yahoo five-star award-winning service.
   www.inkjetcartridge.com
   (Advertiser's Max Bid: $1.79)
3. **75% off Epson Ink - Free Shipping**
   Up to 75% off Epson ink and toner, plus a one-year quality guarantee and free shipping. Save an extra 5% with coupon code over15. Free promotional items with every purchase.
   www.123inkjets.com
   (Advertiser's Max Bid: $1.79)

Overture data

search	*$ bid*	*advertiser*
accessory desk	.83	Office Max
alfred hitchcock	.01	Videoflicks.com
educational software	.13	Buy.com
educational software	.4	eBAY International AG
educational software	.17	OfficeMax
epson	.28	Buy.com
epson	.4	eBAY International AG
fax	.13	Buy.com
fax	.4	eBAY International AG
fax	.38	OfficeMax
game	.02	Net Business
game	.25	Buy.com
george harrison	.15	eBAY International AG
george harrison	.05	MusicStack
george harrison	.01	Videoflicks.com
jackson michael	.05	MusicStack
jackson michael	.01	Videoflicks.com

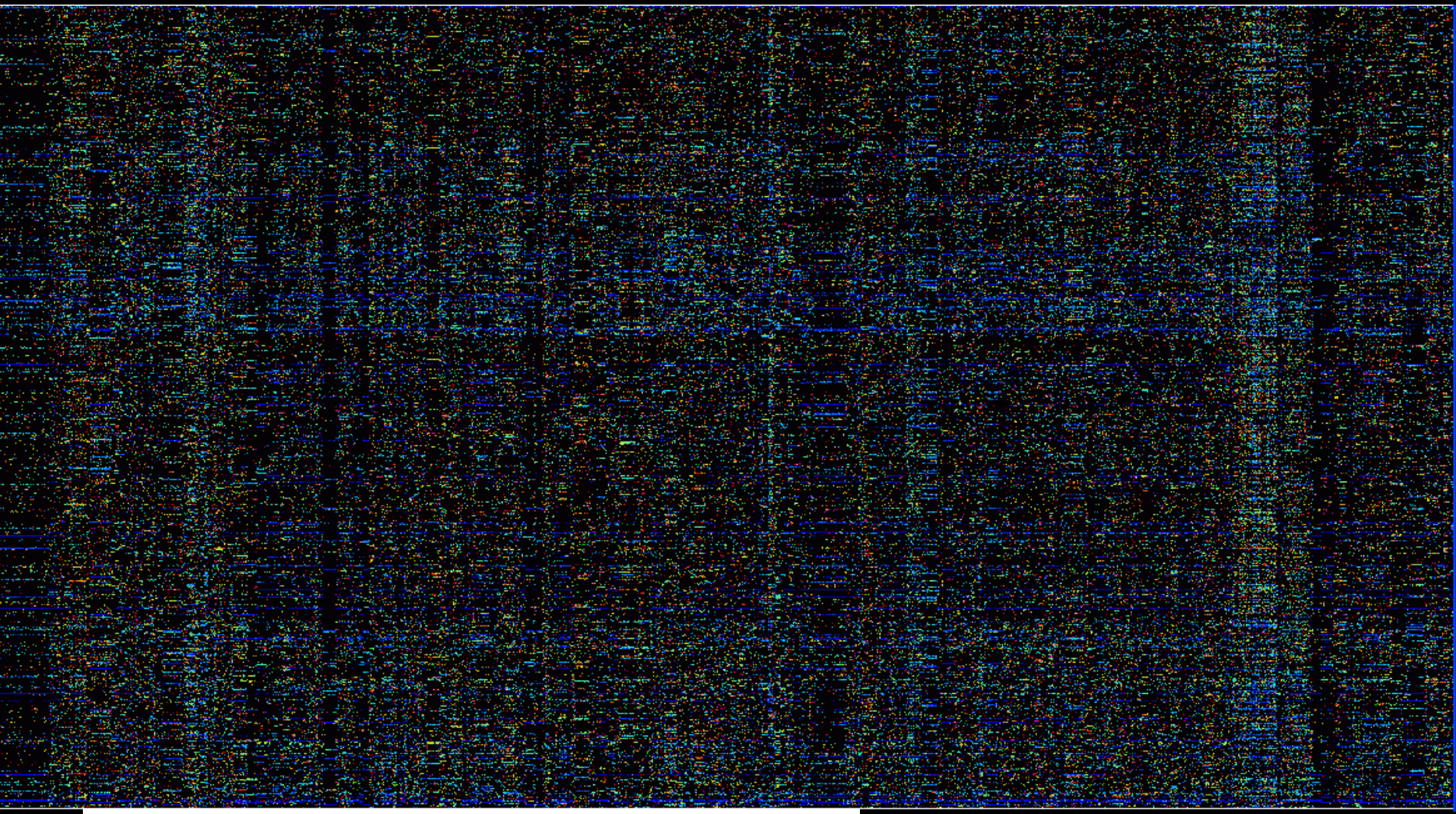
Overture bid graph



Overture bid matrix

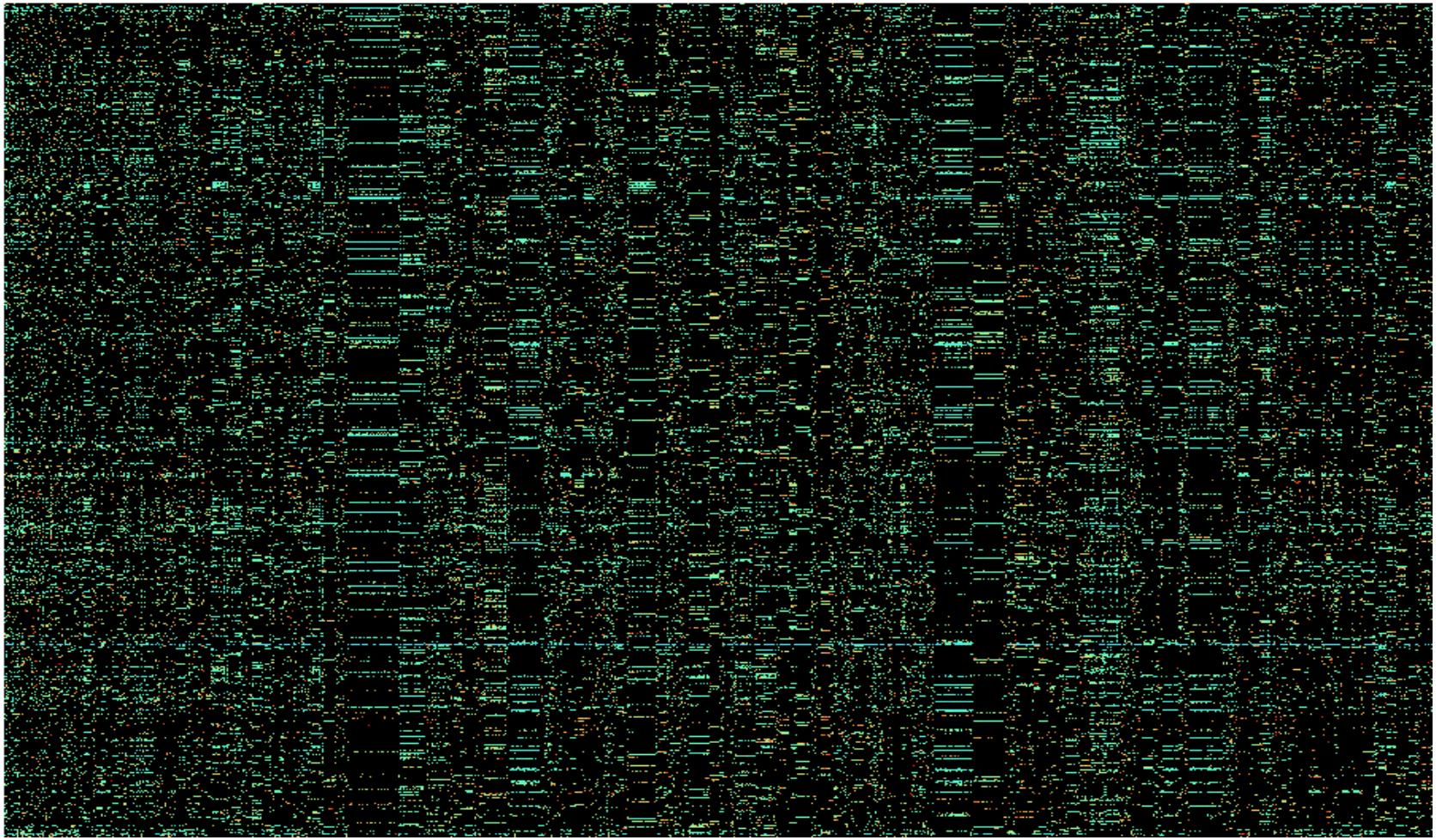
	1	2	3	4	5	6
accessory desk		■				
game	■		■			
fax		■	■	■		
educational software		■	■	■		
epson			■	■		
george harrison				■	■	■
jackson michael					■	■
alfred hitchcock						■

1. Net Business
2. Office Max
3. Buy.com
4. eBAY
5. MusicStack
6. Videoflicks.com



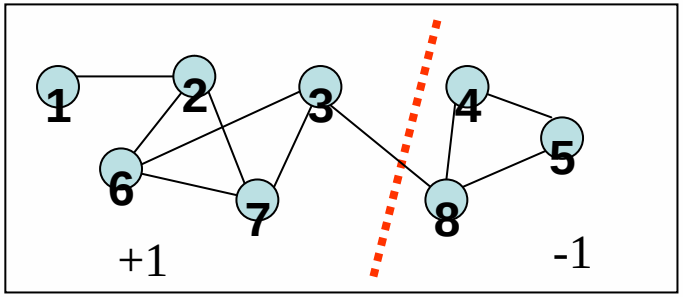
170,077 terms, 25,971 advertisers, 1,307,316 bids

Overture bid matrix: 18-core



2000 advertisers , 3000 search terms, 92,347 bids

Spectral graph partitioning



- assign each node indicator $p_i = \pm 1$
- partition p = {-1,-1,-1,-1, +1, +1, +1}

$$cut(V_1, V_2) = \frac{p^T L p}{4}$$

Relaxation procedure: $p_i = \{-1, 1\}^N$ **=>** $x_i \in [-1, 1], x_i \in R^1$

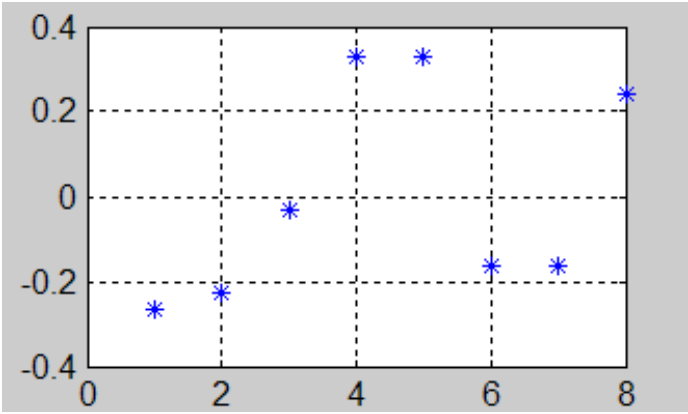
Quadratic optimization: $E(x) = \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{4}, \quad \sum_i x_i^2 = N, \quad (x^T e) = 0$

Min cut solution : $\mathbf{L}\mathbf{x} = \lambda \mathbf{x}$

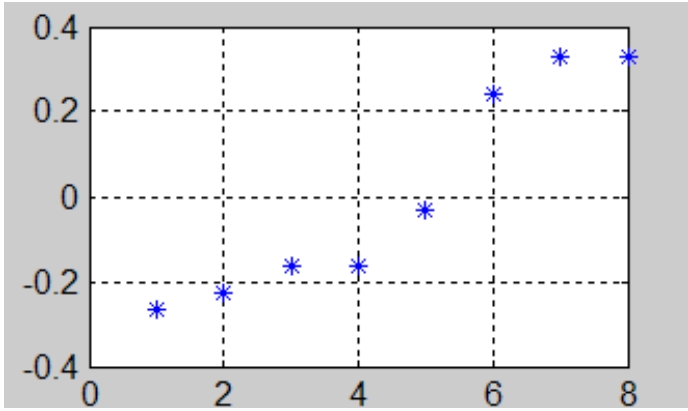
Rounding off: $p_i = +1, \quad if \ x_i > 0; \quad p_1 = -1, \quad if \ x_i < 0$

Spectral: ordering

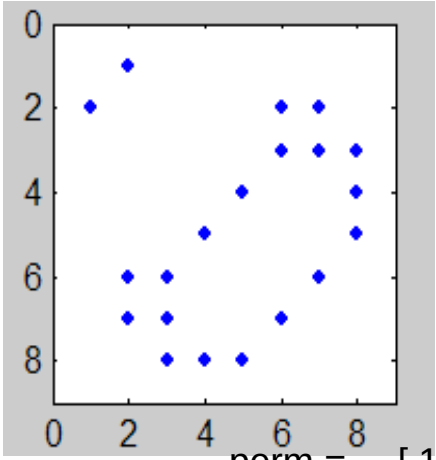
Eigenvector



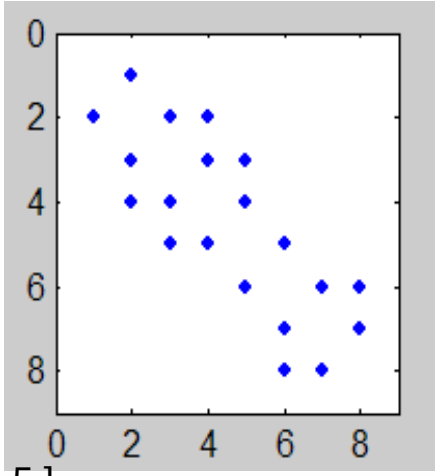
Eigenvector – sorted



Adjacency matrix



Adjacency matrix – re-ordered

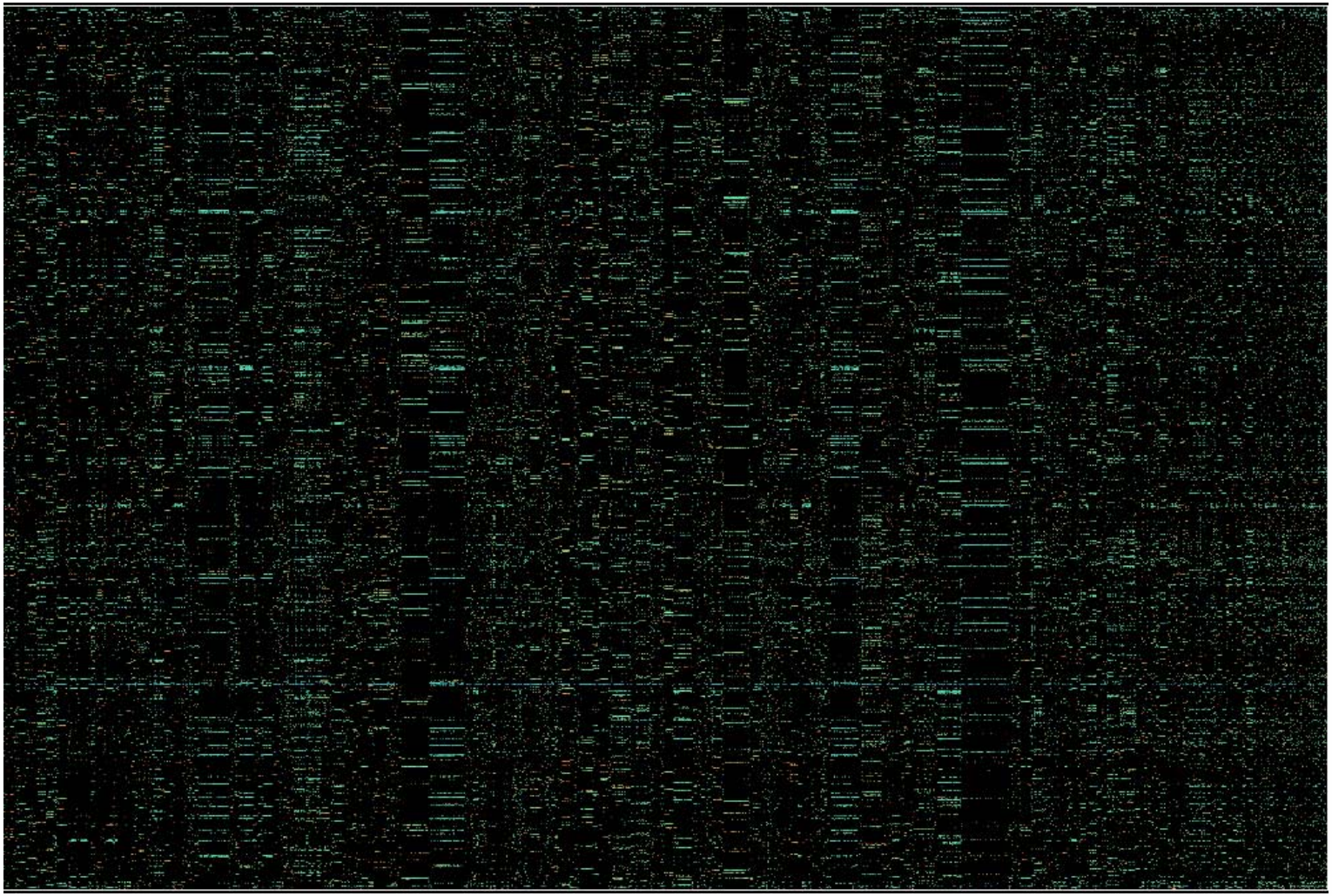


perm = [1 2 6 7 3 8 4 5]

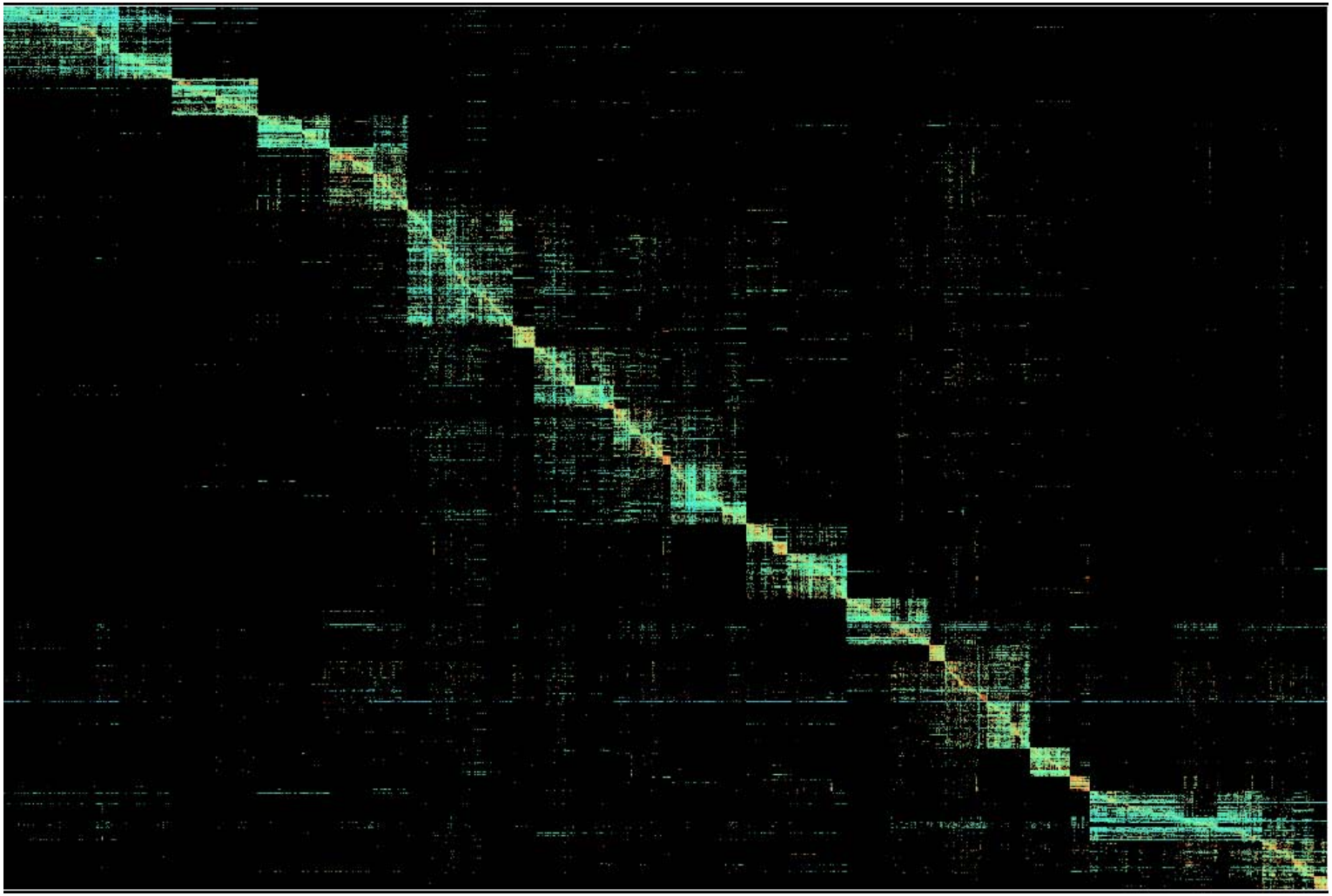
Spectral : additions

- Recursively
- Min cut, ratio cut, normalized cut, quotient cut, …
- Sweep for optimal value
- Flow based refinement
- Optimize partition among several eigenvectors
- Bi-partite graphs
- + other bells & whistles

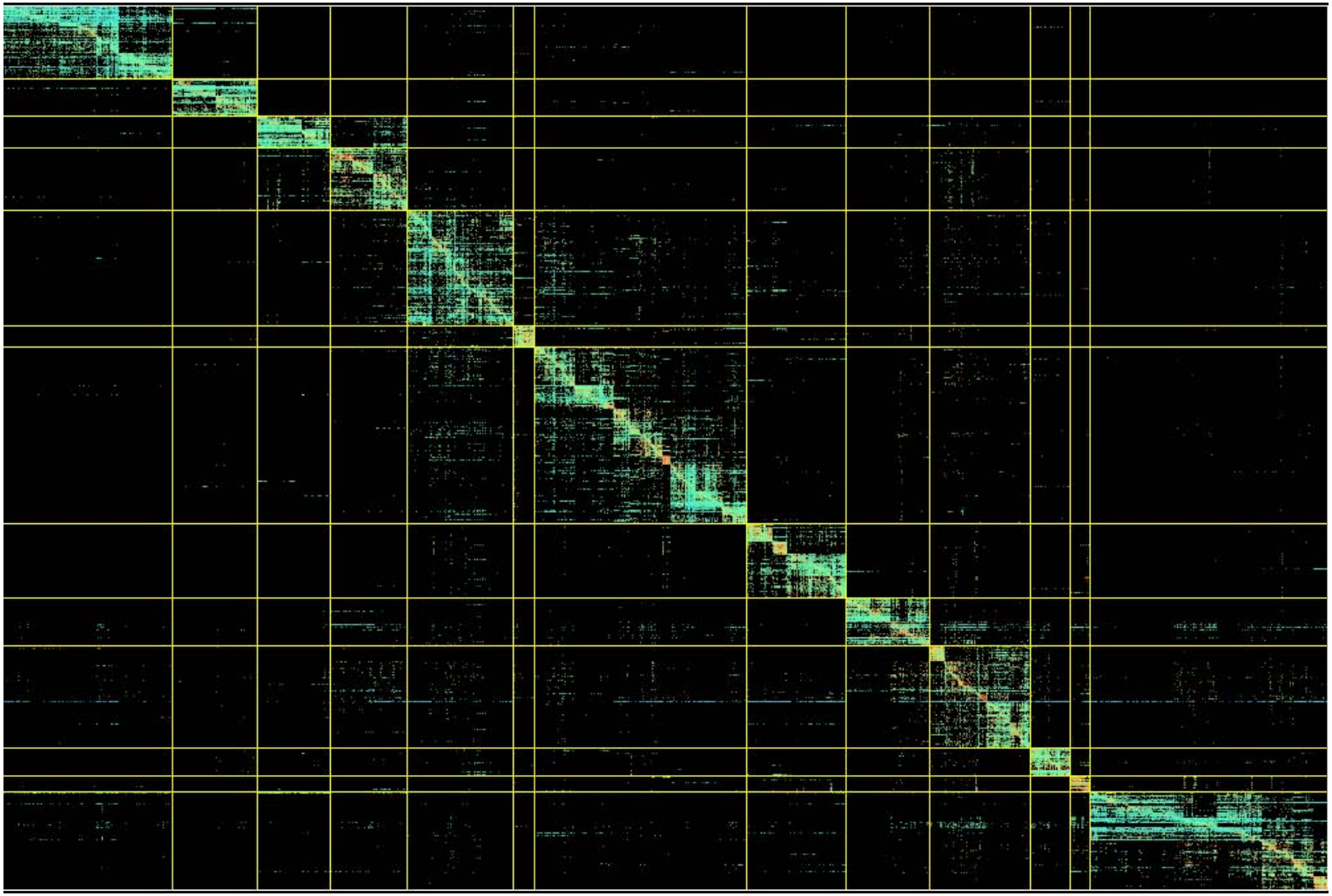
Search Market Place



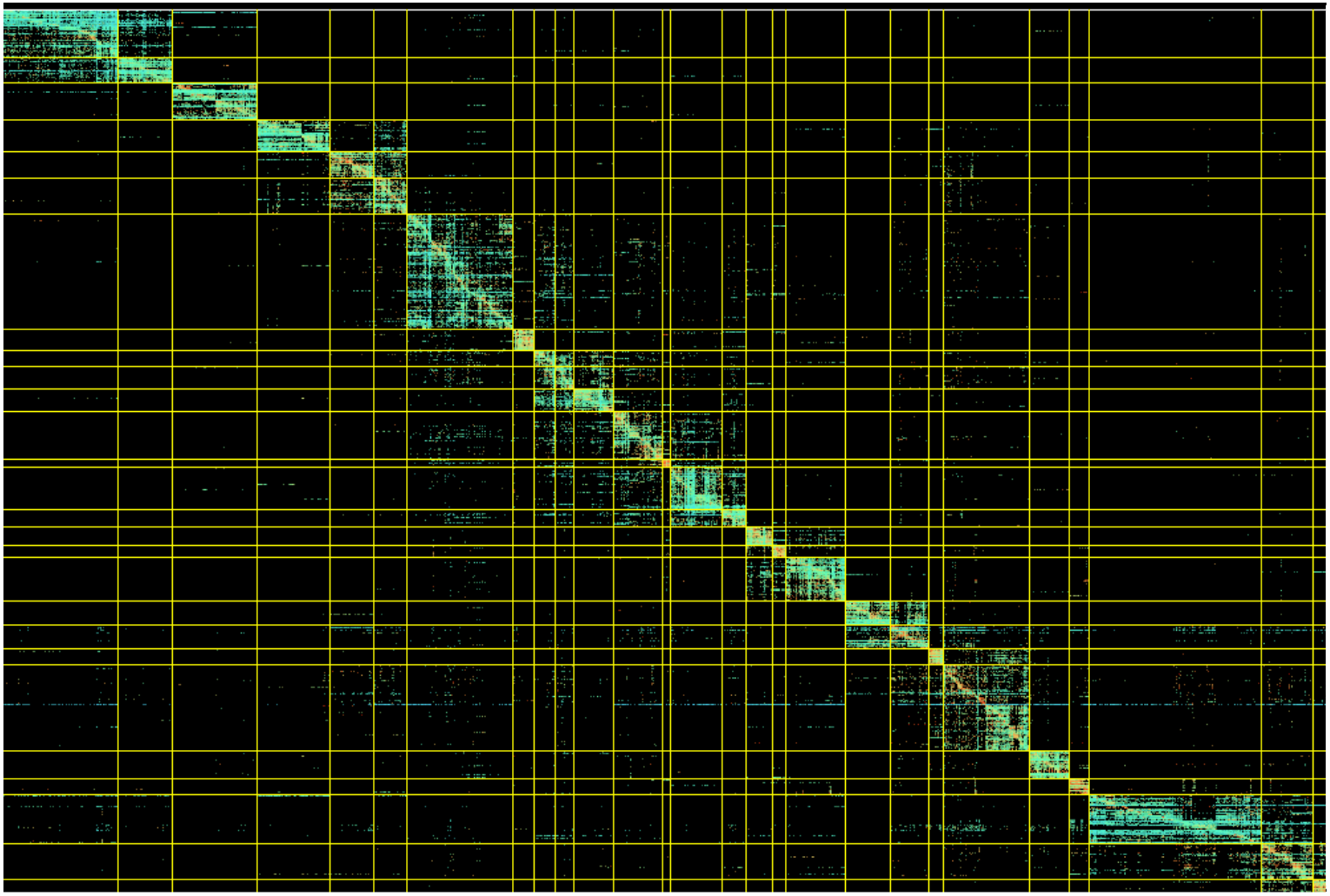
Search Market Place



Search Market Place



Search Market Place



Sample clusters

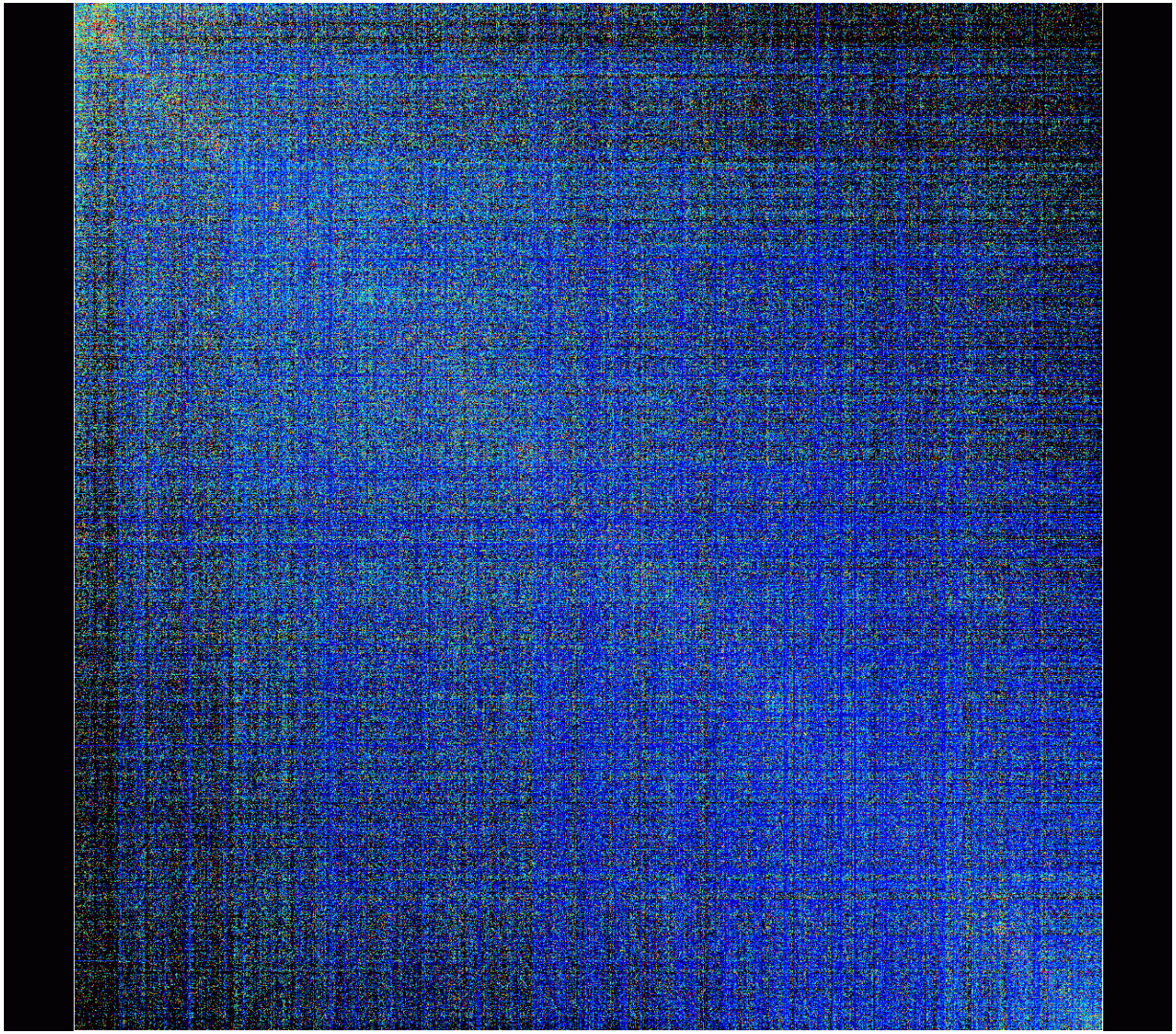
'A.S.C. Incorporated'
'Atlantic Telecom'
'AudioLink'
'Cost Plus Electronics'
'Headset Express'
'Hello Direct'
'PDA Mountain'
'PK TECH INC'
……..

'accessory phone'
'cordless headset'
'cordless headset phone'
'free hands'
'headset'
'headset microphone'
'headset phone'
'headset plantronics'
'headset wireless'
'isdn'
'plantronics'
…….

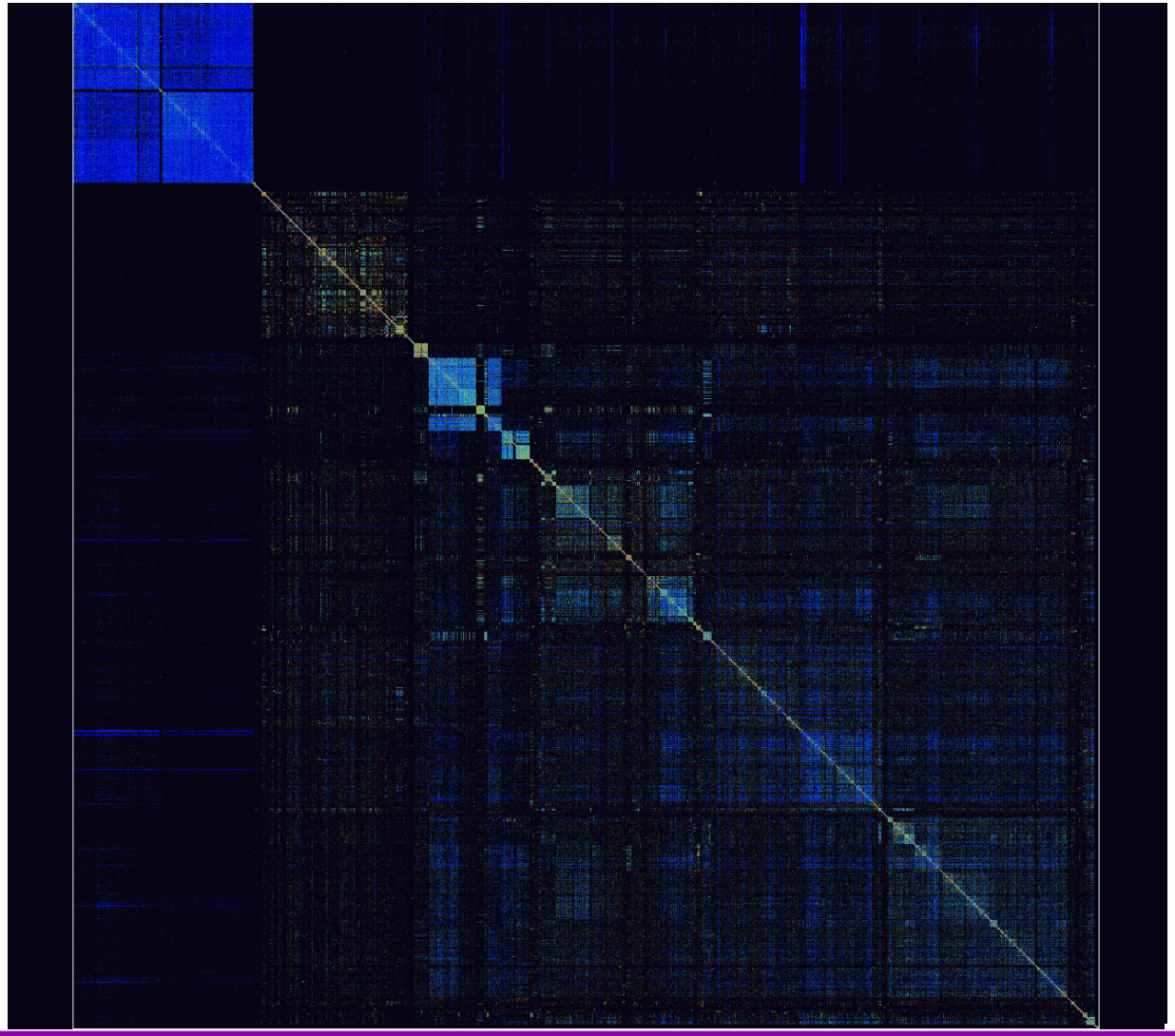
'Berent Associates Center for Shyness and Social Therapy'
'Dr. Puff'
'New York Psychotherapy Collective'
'Ruby Shoes'
'Self-employed psychologist'
'www.kabbalah.com'
…….

'health mental'
'help self'
'improvement self'
'parenting'

Flickr: “10-core” spectral ordering

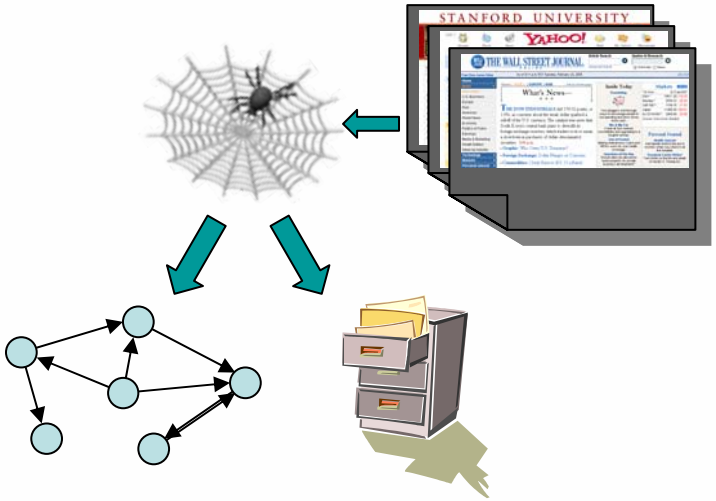


Flickr: “10-core” spectral ordering



Websearch Engines

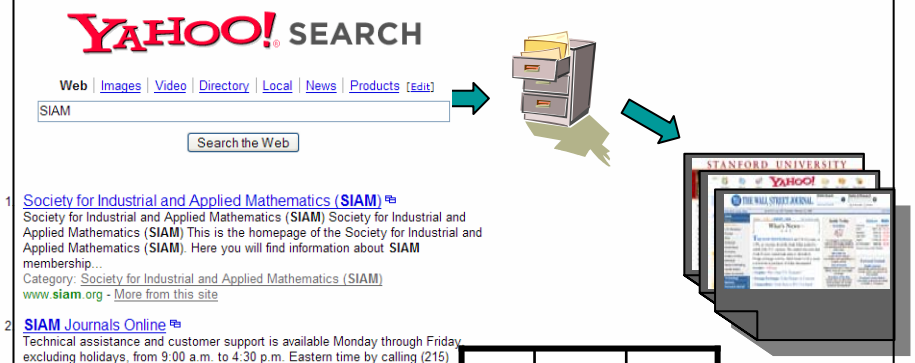
- **Crawler + indexer**



Link Analysis
PageRank

Text Analysis
Inverted Index

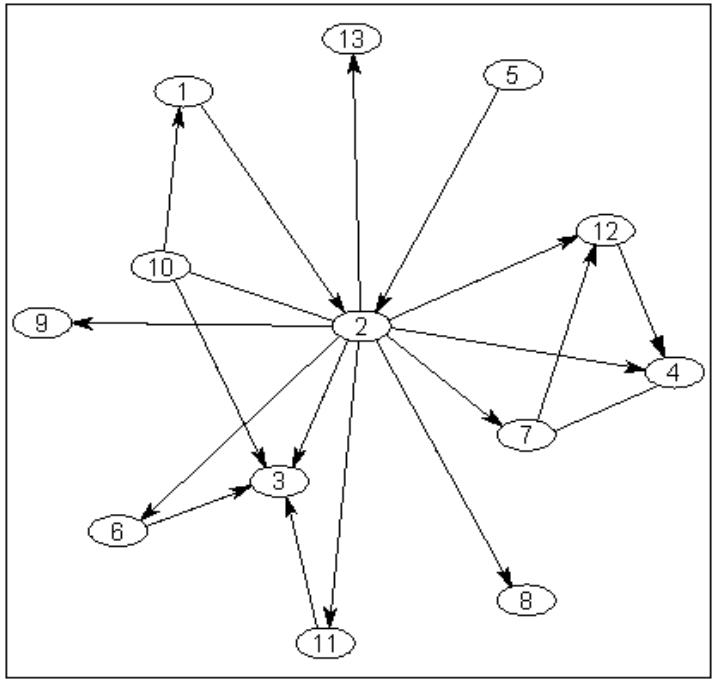
- **Front end**

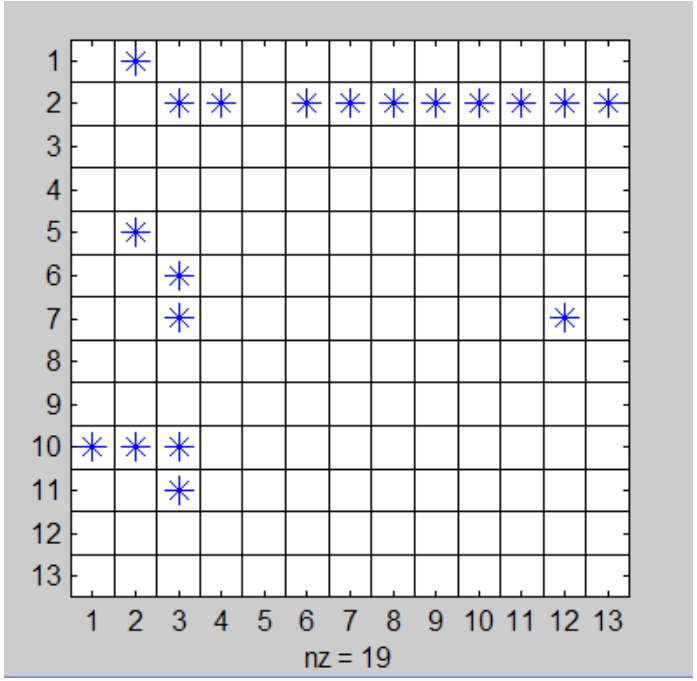


#	Qs	PR
1	0.7	0.4
2	0.8	0.1
3	0.1	0.5

PageRank model

- Random walk on graph
- Markov process: memory-less, homogeneous
- Stationary distribution: existence, uniqueness, convergence





Perron-Frobenius theorem, irreducible, every state is reachable from every other, and aperiodic – no cycles

PageRank

- Construct probability matrix:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$$

- Construct transition row-stochastic matrix for Markov process

$$\mathbf{P}' = \mathbf{P} + (\mathbf{d}\mathbf{v}^T)$$

- Correct reducibility:

$$\mathbf{P}'' = c\mathbf{P}' + (1-c)(\mathbf{e}\mathbf{v}^T)$$

- Find stationary distribution (exist & unique):

$$\mathbf{P}''^T p = \lambda p$$

c – teleportation coefficient, **v** – personalization vector, **d** – dangling nodes (sinks) indicator

PageRank linear system

- Eigensystem:

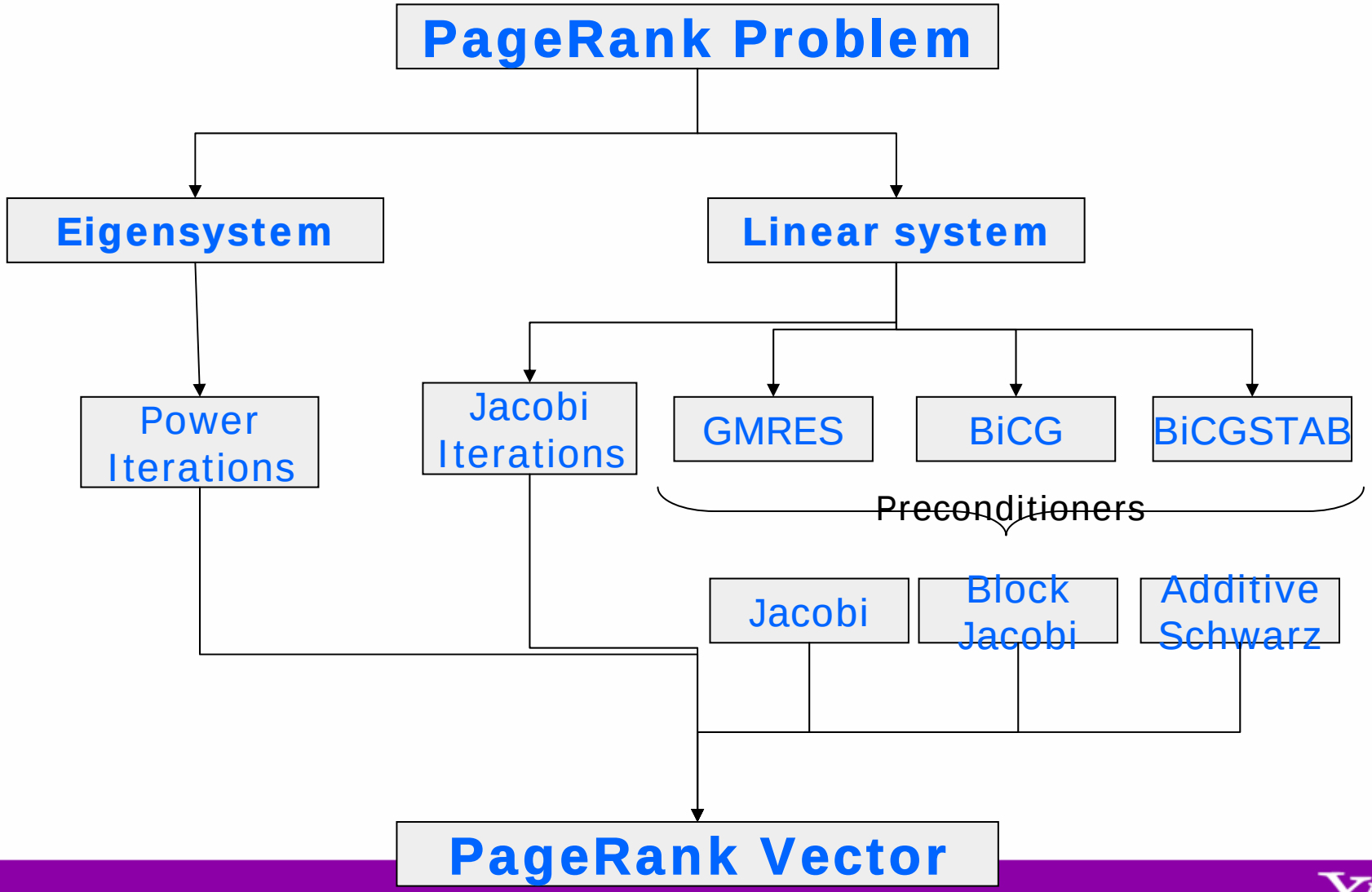
$$[cP + c(d \cdot v^T) + (1-c)(e \cdot v^T)]^T p = \lambda p$$

- Linear system: (for $\lambda = 1$ and $||p||_1 = 1$)

$$(I - cP^T)x = kv,$$

$$k = ||x||_1 - c||P^T x||_1 \qquad p = \frac{x}{||x||_1}$$

Computational diagram

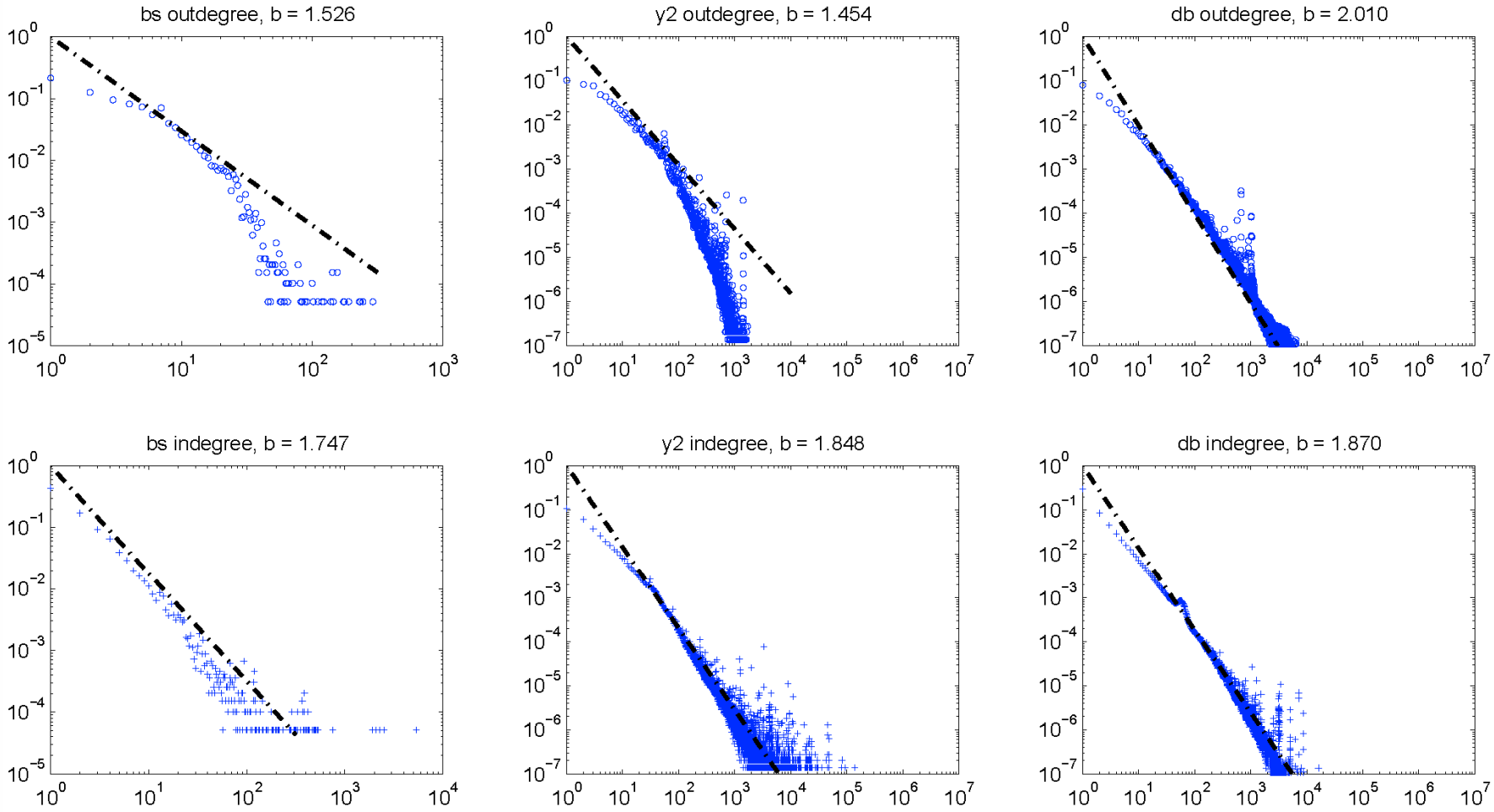


Data: WebGraphs

Name	# Nodes/size	# Links / nnz	Memory
bs	0.3M	1.6M	20.4 MB
edu	2M	14M	176MB
yahoo-r2	14M	266M	3.25GB
uk	18.5M	300M	3.67GB
yahoo-r3	60M	850M	10.4GB
db	70M	1B	12.3GB
av	1.4B	6.6B	80GB

Graph in memory, compressed storage, int, int, double

Power law graphs ($y = x^{-b}$) ?



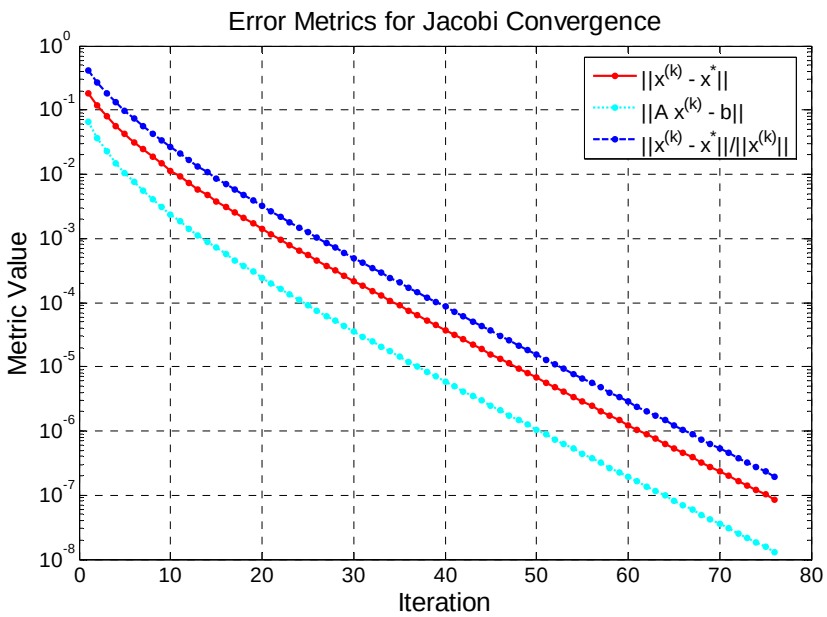
bs: 0.3M nodes

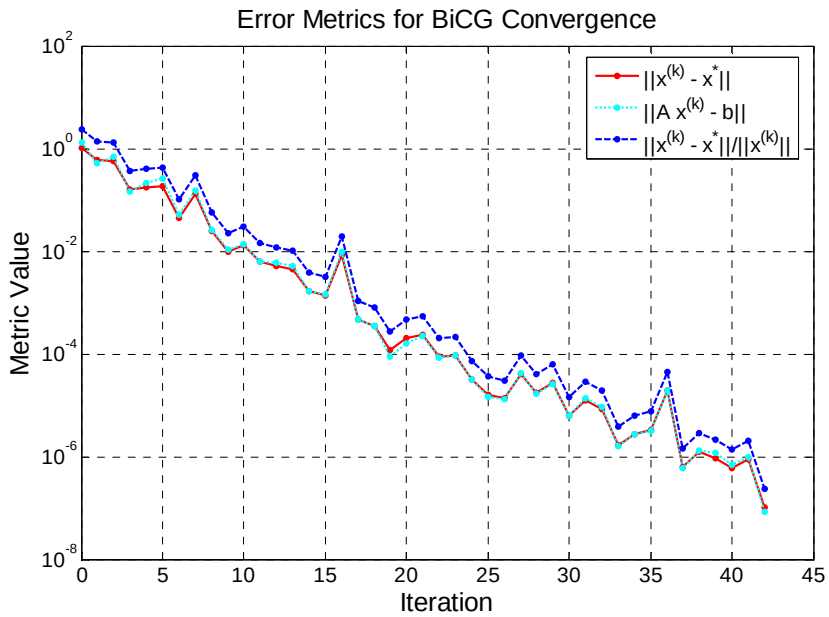
y2: 266 M nodes

db: 1B nodes

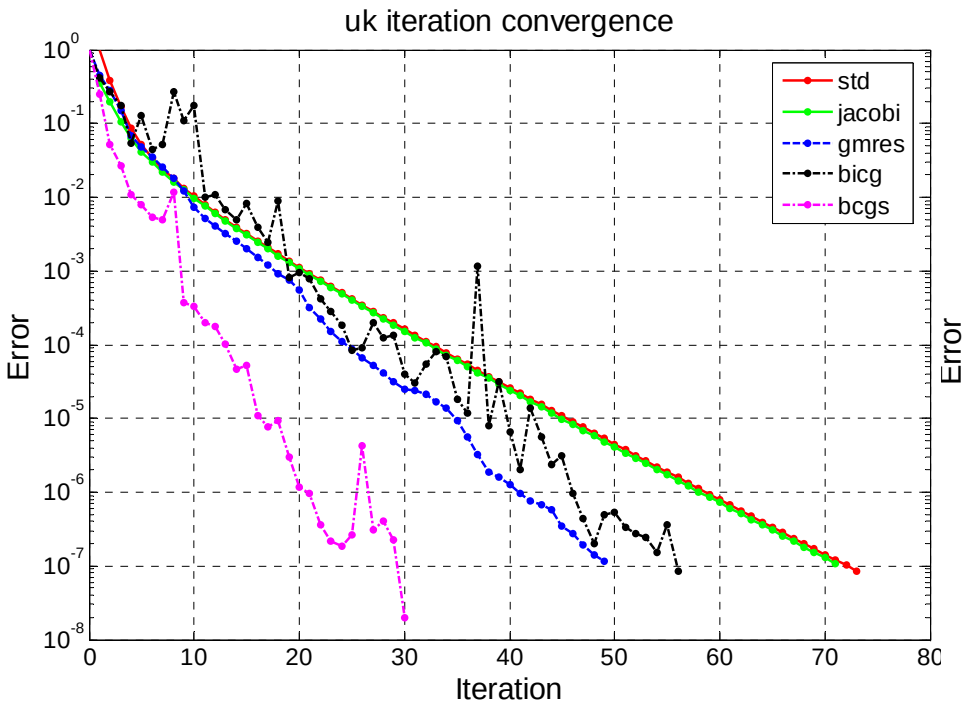
Computational Methods

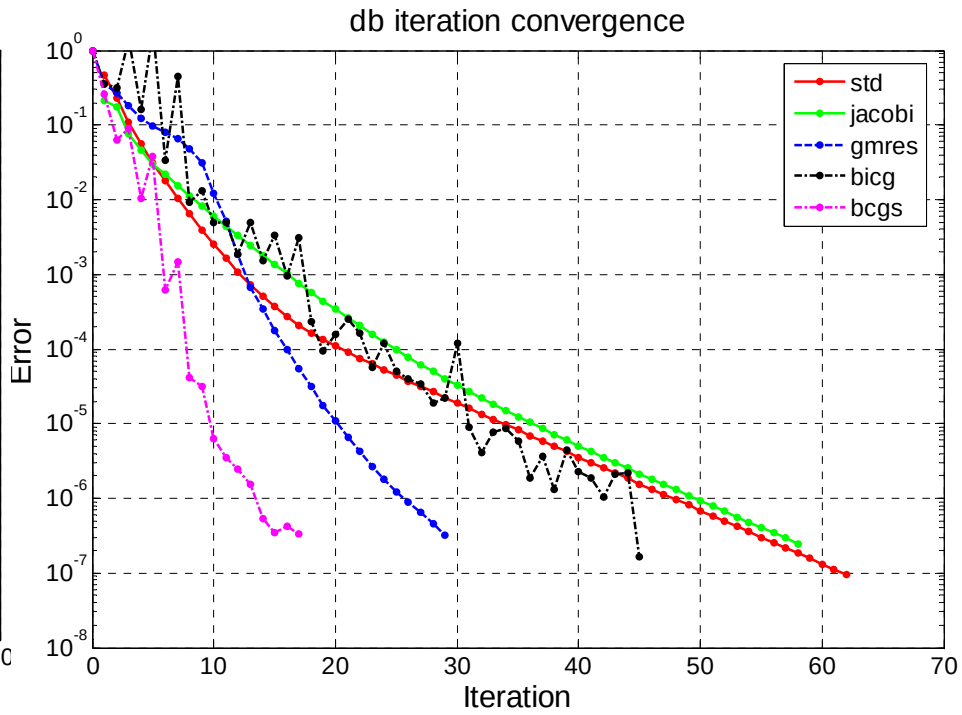
Method	Inner Products	SAXPY	Matrix-Vector	Storage
PAGERANK		1	1	$M + 3v$
JACOBI		1	1	$M + 3v$
GMRES	$i + 1$	$i + 1$	1	$M + (i + 5)v$
BiCG	2	5	2	$M + 10v$
BiCGSTAB	4	6	2	$M + 10v$





Convergence results

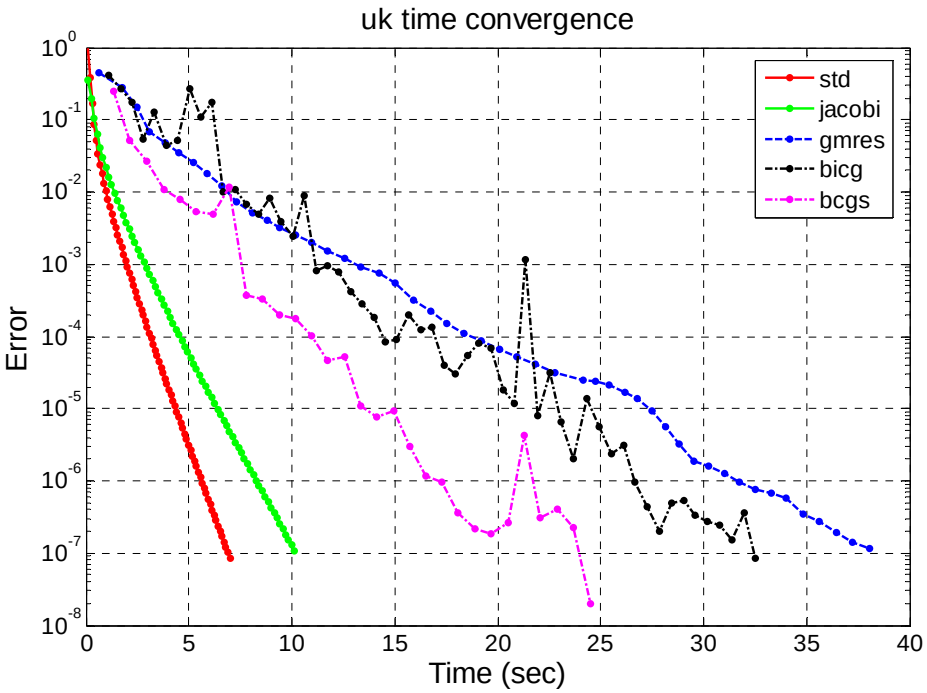


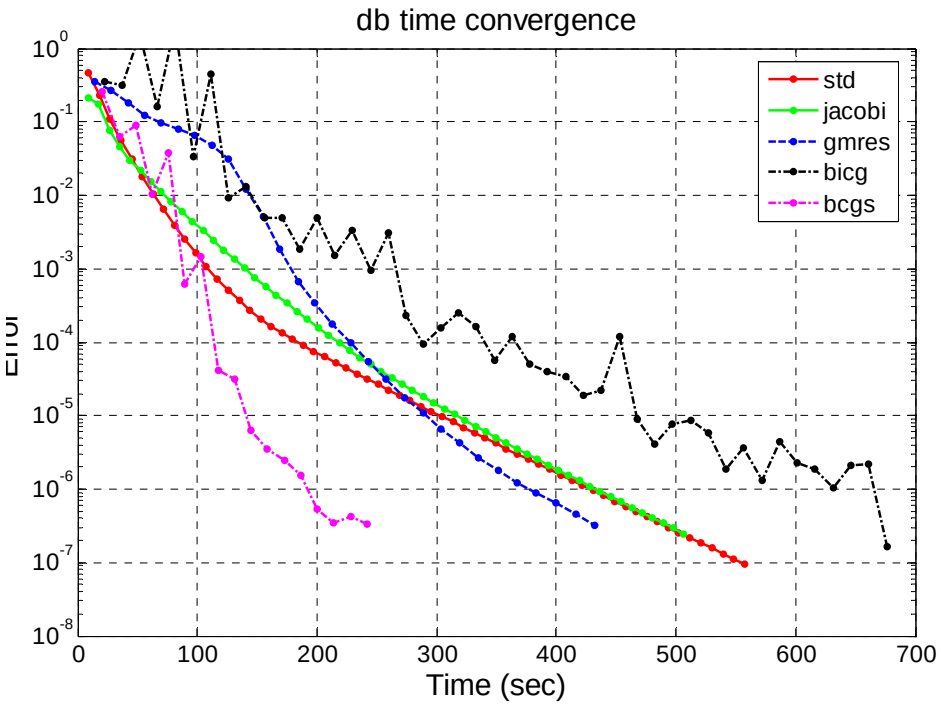


uk: 18.5 M pages, 300 M links

db: 70 M pages, 1 B links.

Convergence results





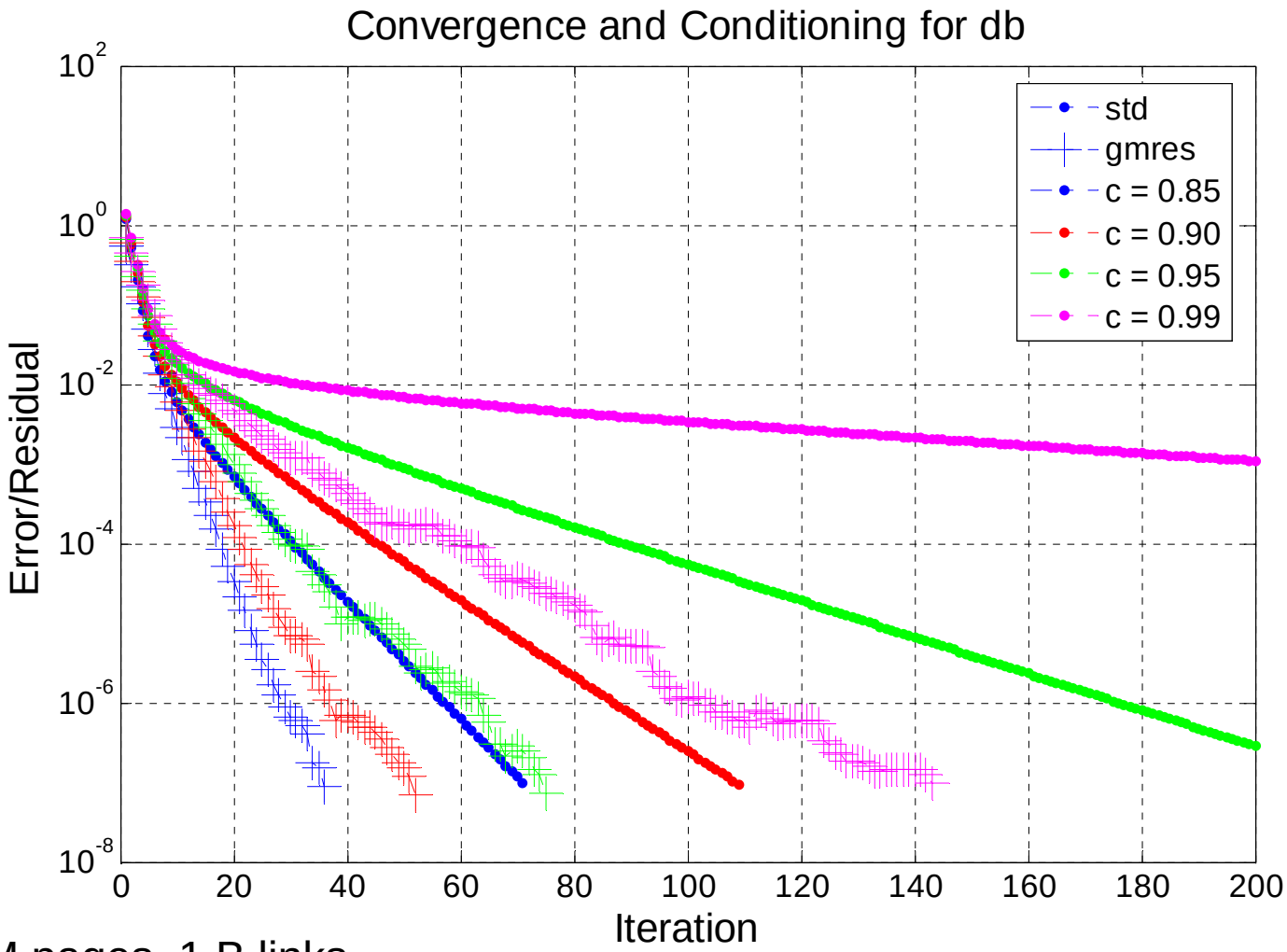
uk: 18.5 M pages, 300 M links

db: 70 M pages, 1 B links.

Summary

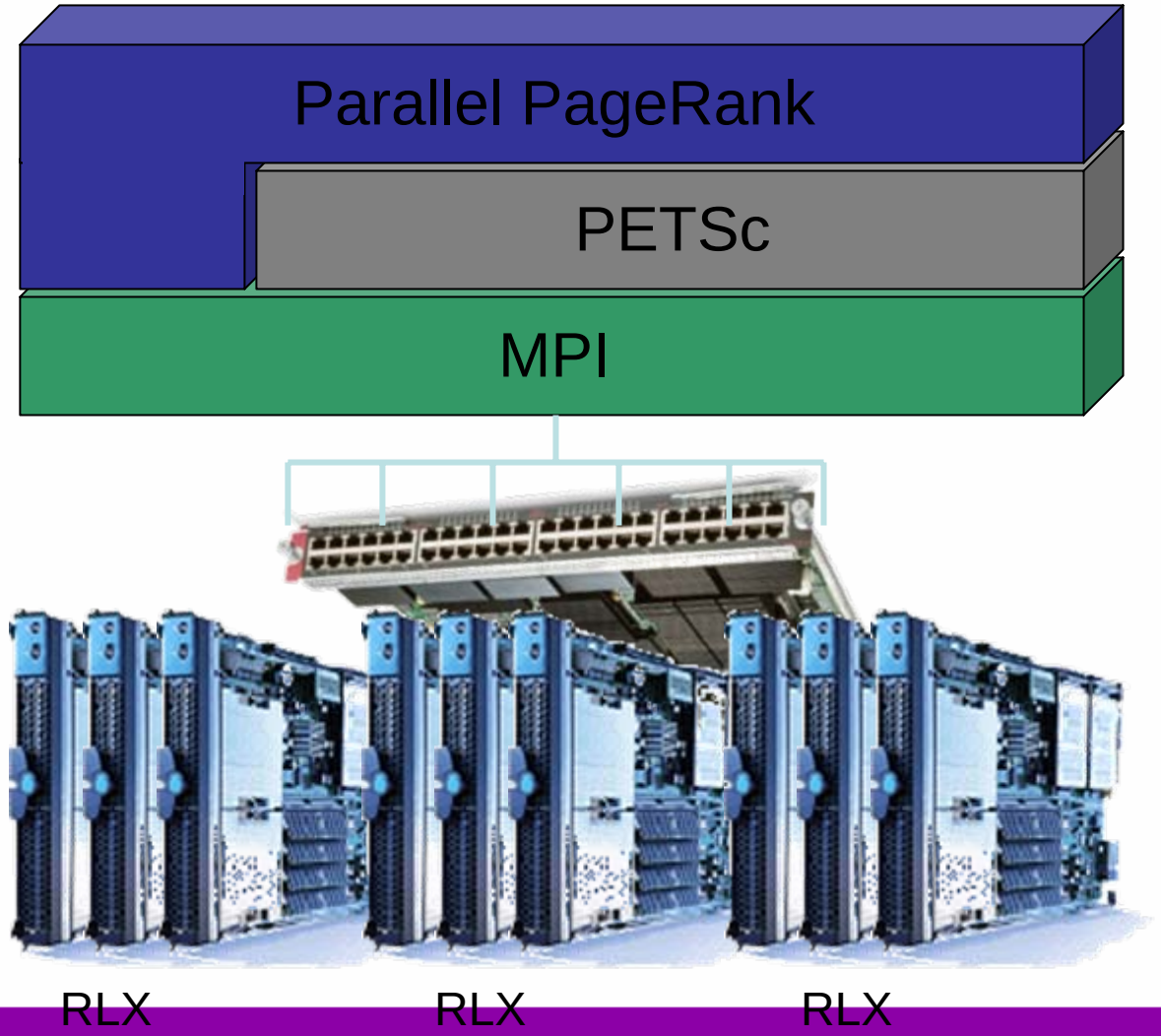
Name	Size	Power	Jacobi	GMRES	BiCG	BCGS
edu 20 procs	2M 14M	84 0.09/7.5s	84 0.07/6.5s	21* 0.6/13.2s	44* 0.4/17.7s	21* 0.4/8.7s
yahoo-r2 20 procs	14M 266M	71 1.8/129s	65 1.9/126s	12* 16/194s	35* 8.6/300s	17* 9.9/168s
uk 60 procs	18.5M 300M	73 0.09/7s	71 0.1/10s	22* 0.8/17.6s	25* 0.8/19.4s	11* 1.0/10.8s
yahoo-r3 60 procs	60M 850M	76 1.6/119s	75 1.5/112s			
db 60 procs	70M 1B	62 9.0/557s	58 8.7/506s	29 15/432s	45 15/676s	15* 15/220s
av 140 procs	1.4B 6.6B	72 4.6/333s				26 15/391s

Reduced teleportation



db: 70 M pages, 1 B links.

YRL research cluster



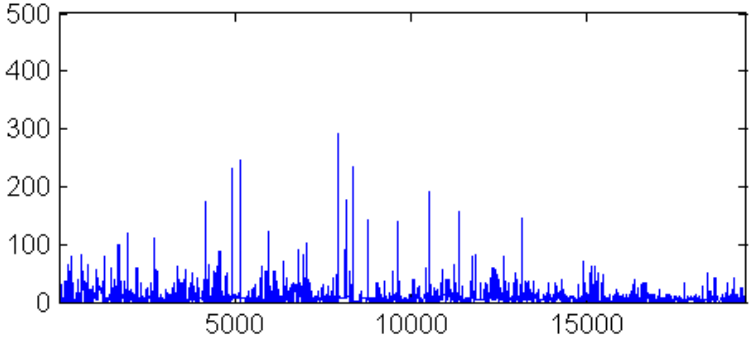
Custom

Off the shelf

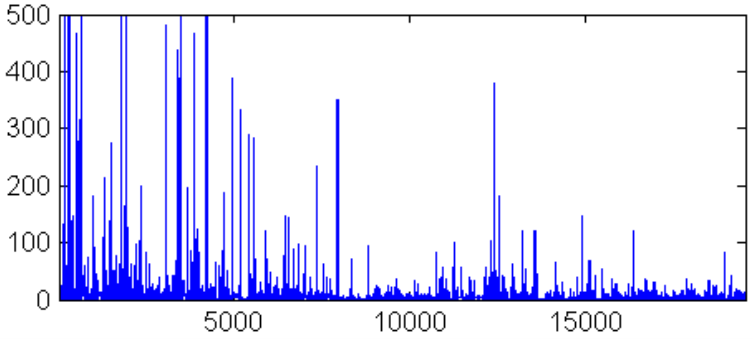
Gigabit Switch

RLX Blades
Dual 2.8 GHz Xeon
4 GB RAM
Gigabit Ethernet
120 Total

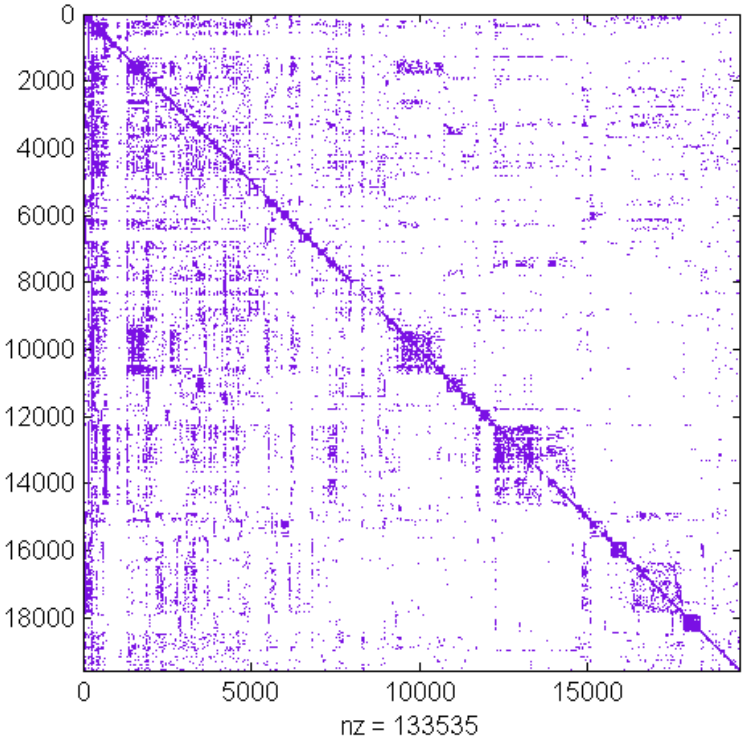
Typical graph



out-degrees.

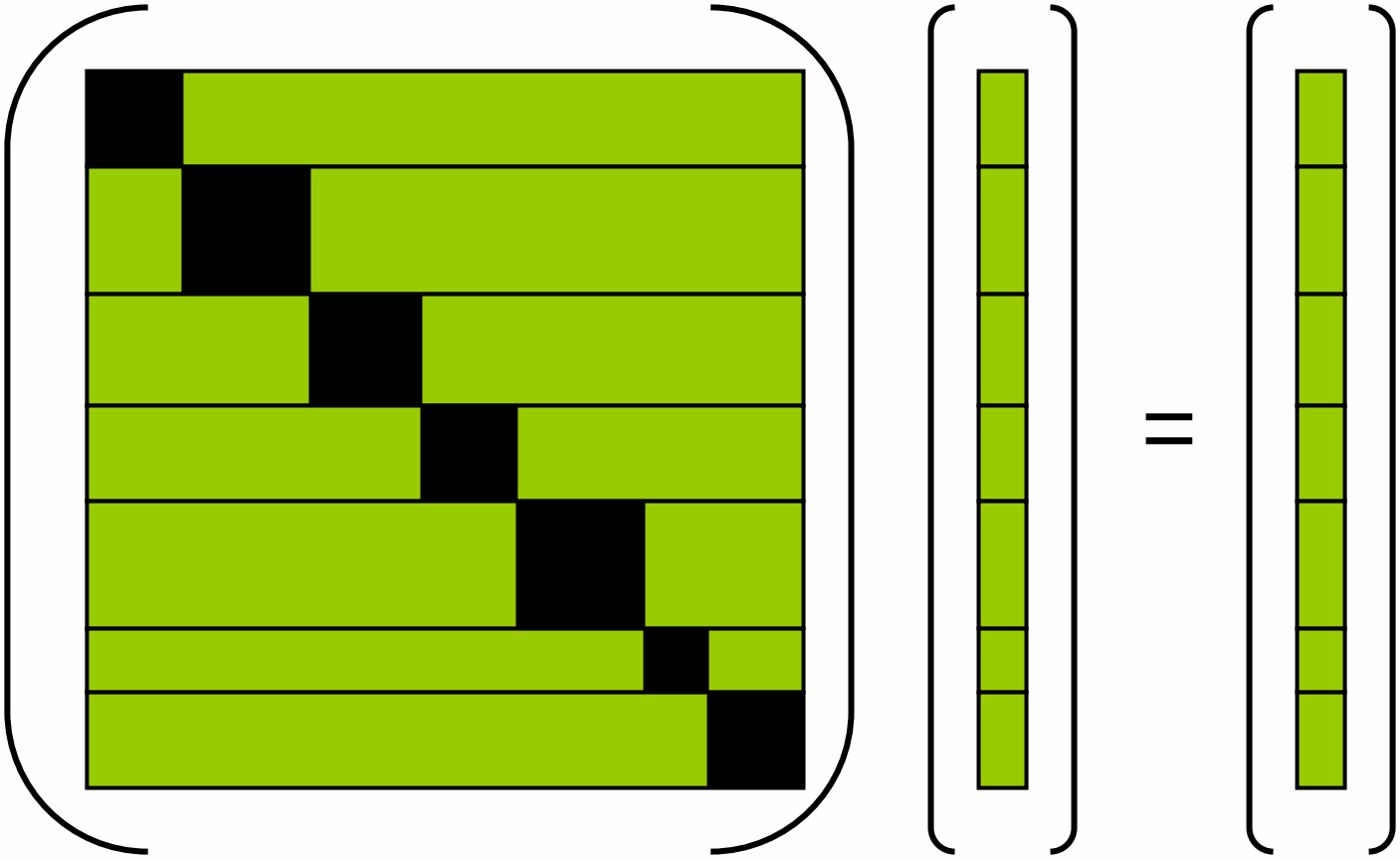


in-degrees.



bs: 70 M pages, 1 B links.

Parallel Matrix-Vector Multiply



Graph distribution methods

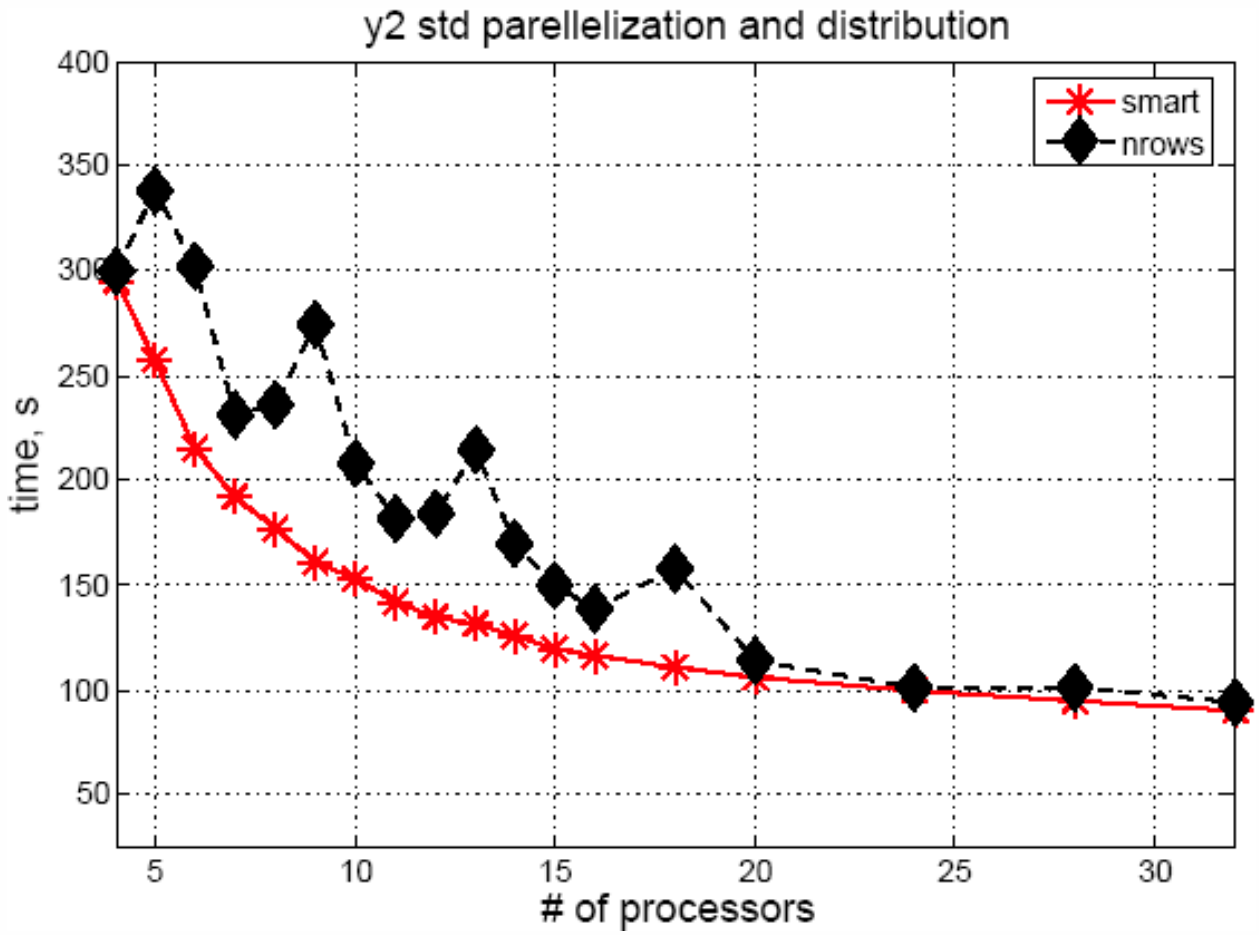
- Balance rows
- Balance nnz
- Random destribution
- Balance rows and nnz (heuristics)

$$w_r \cdot n_p + w_n \cdot nnz_p > (w_r \cdot n + w_n \cdot nnz)/p$$

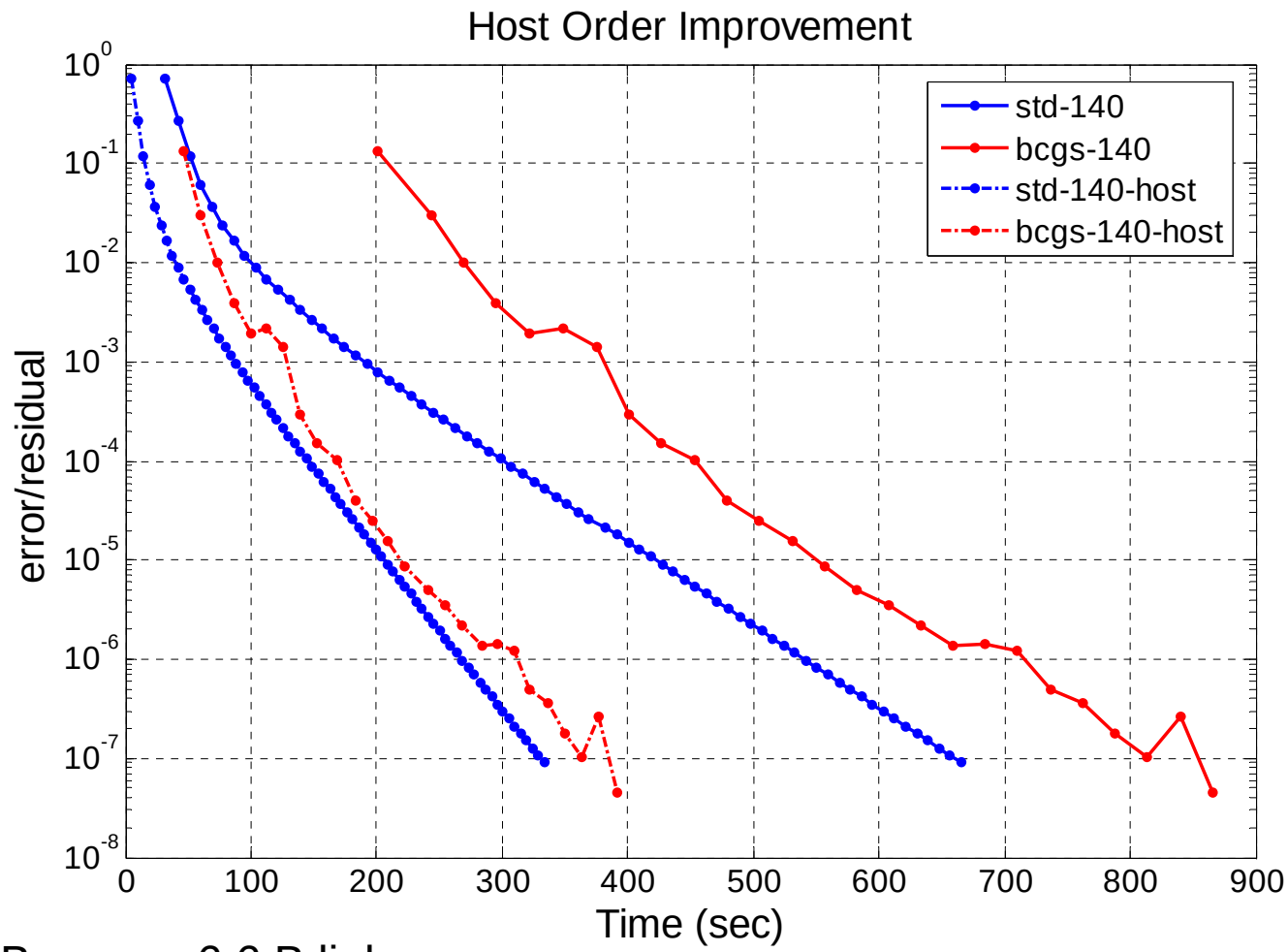
$$w_r : w_n = 1/1, 2/1, 5/1$$

- 2D distributions
- Graph partitioning
- Hypergraph partitioning

Some parallel experiments

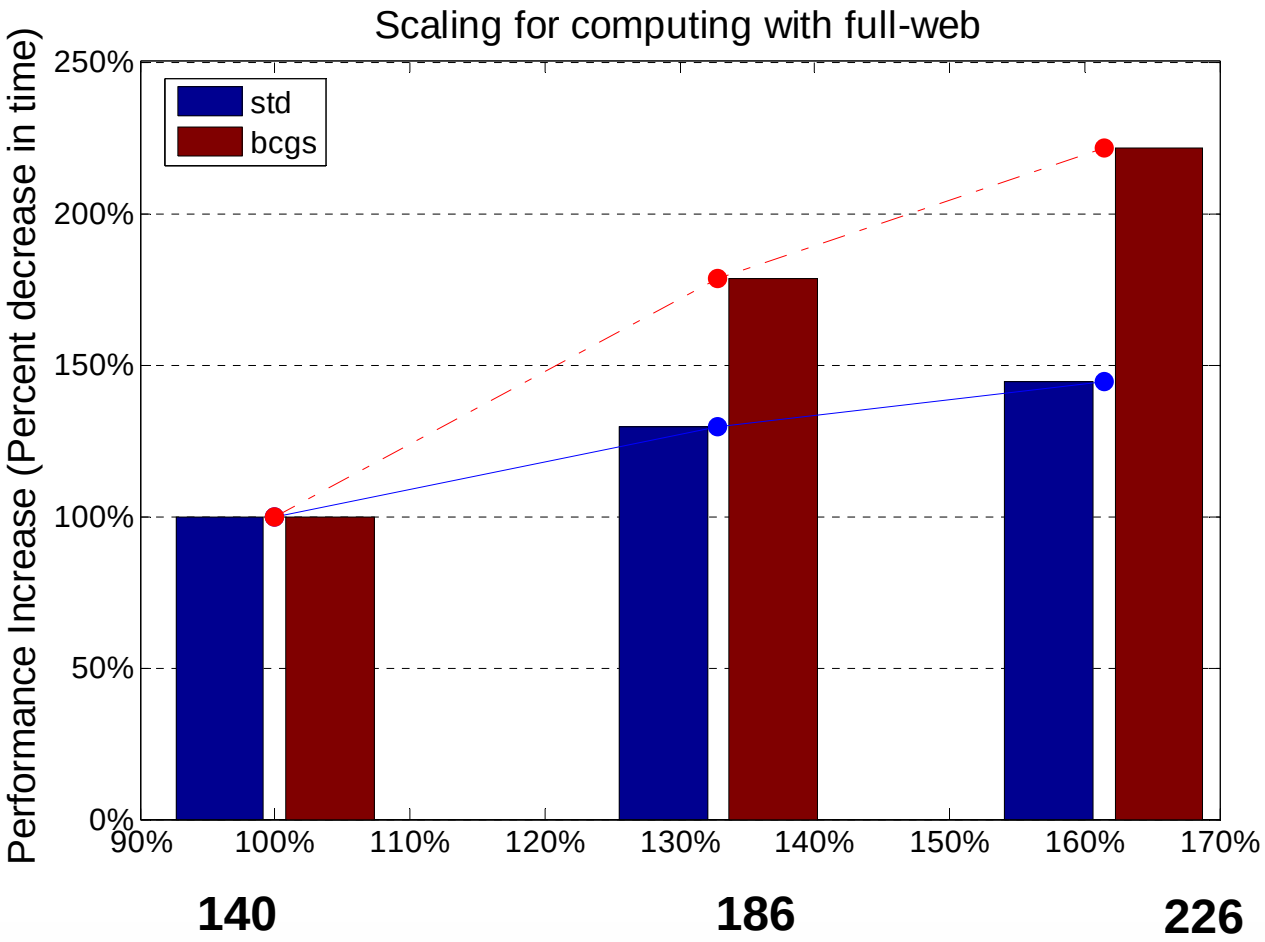


Domain ordering



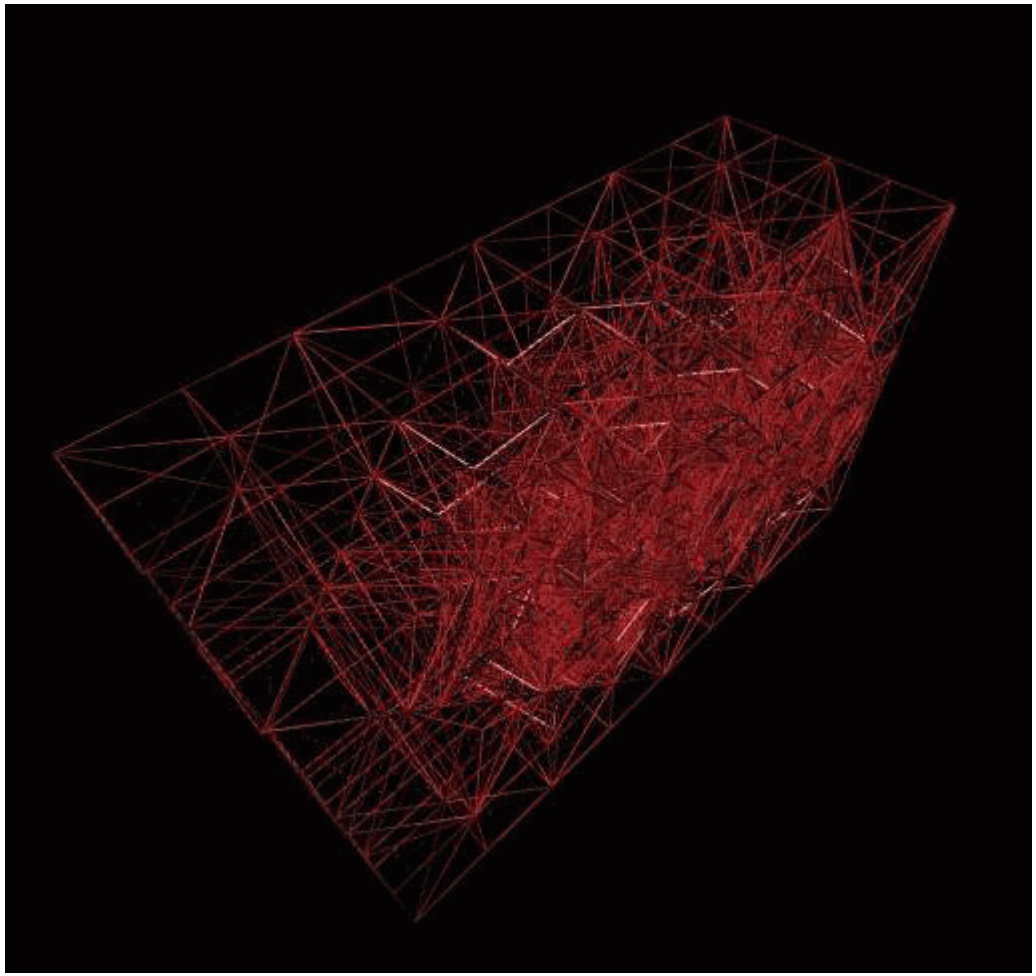
av: 1.4 B pages, 6.6 B links.

Full web graph parallelization

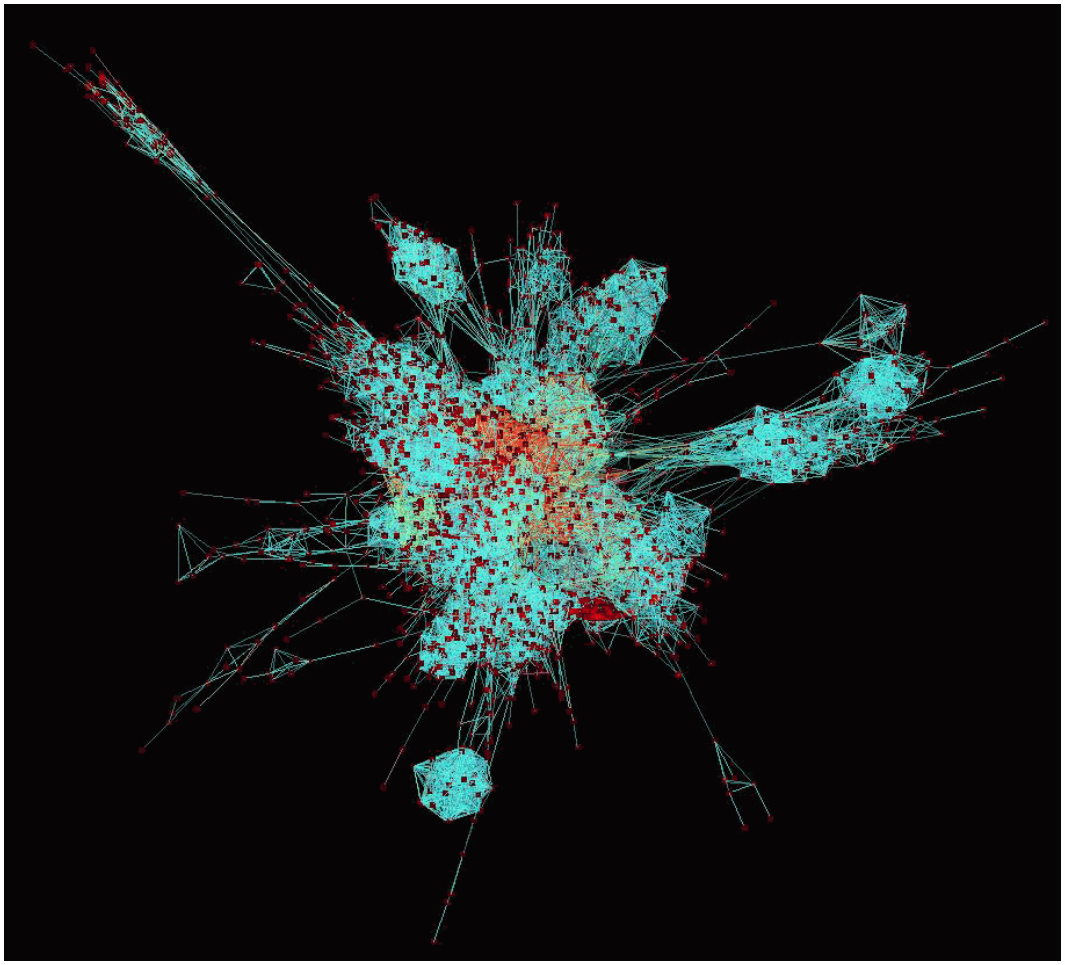


av: 1.4 B pages, 6.6 B links.

FEM mesh for CFD computations



Power-law graph embedding in 3D



Final thoughts…

- Power-law graphs are different
- Have central “core”, long chains, clusters, singletons
- Not all power-law graphs are the same
- Difficult to distribute / parallelize
- Noise: its is experimental data after all
- Lots of cool applications
- Let’s talk …..

References

- **Spectral graph partitioning:**

- M. Fiedler (1973), A. Pothen (1990), H. Simon (1991), B. Mohar (1992), B. Hendrickson (1995), D. Spielman (1996), F. Chang (1996), S. Guattery (1998), R. Kannan (1999), J. Shi (2000), I. Dhillon (2001), A. Ng (2001), H. Zha (2001), C. Ding (2001)

- **PageRank computing:**

- S.Brin (1998), L. Page (1998), J. Kleinberg (1999), A. Arasu (2002), T. Haveliwala (2002-03), A. Langville (2002), G. Jeh (2003), S. Kamvar (2003), A. Broder (2004)