SAND2011-1401C

Spectra of Large Networks

David F. Gleich

Sandia National Laboratories

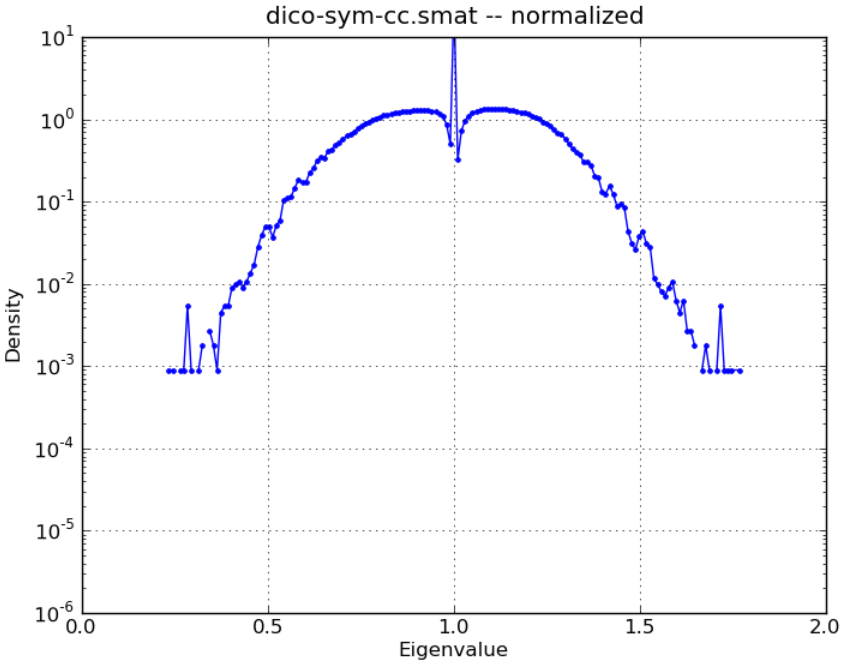
iCME la/opt seminar

24 February 2011

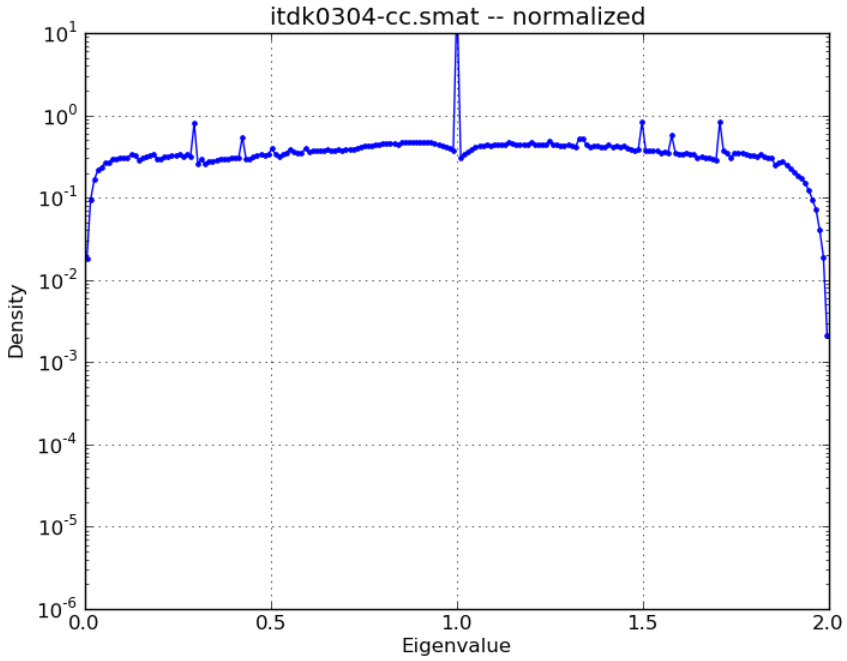
Thanks to Ali Pinar, Jaideep Ray, Tammy Kolda, C. Seshadhri, Rich Lehoucq, and Jure Leskovec for helpful discussions.

Supported by Sandia's John von Neumann postdoctoral fellowship and the DOE Office of Science's ASCR Graphs project.

There's information inside the spectra



Words in dictionary definitions
111,000 vertices, 2.7M edges



Internet router network
192k vetices, 1.2M edges

Also noted in a study of smaller graphs by Banerjee and Jost (2009)

Overview

Graphs and their matrices

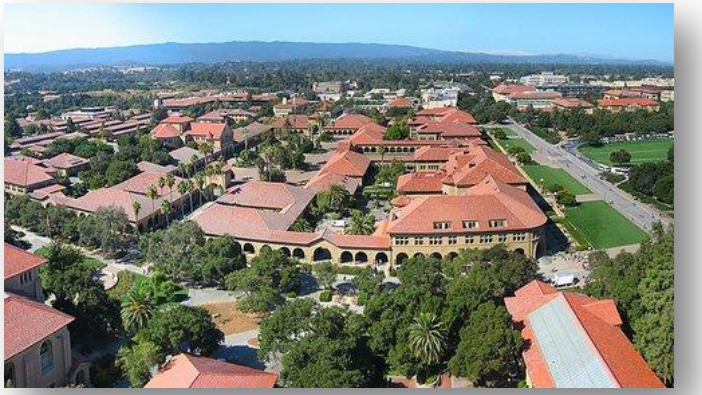
Data for our experiments

Computing spectra for large networks

Issues with computing spectra

Many examples of graph spectra

~~*Conclusion*~~ Future work



Images taken from Stanford, flickr, and Purdue, respectively

GRAPHS AND THEIR MATRICES

As well as things we already know about graph spectra.

Matrices from graphs

Adjacency matrix

$\boldsymbol{A} : n \times n, \boldsymbol{A} = \boldsymbol{A}^T$

$A_{i,j} = 1$ if $(i,j) \in E$

Laplacian matrix

$\boldsymbol{D} = \text{diag}(\boldsymbol{A}\mathbf{e})$

$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$

Normalized Laplacian matrix

$\tilde{\boldsymbol{L}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$

Not covered

Signless Laplacian matrix

Incidence matrix
(It is incidentally discussed)

Seidel matrix

Random walk matrix

$\boldsymbol{L}^{\boldsymbol{W}} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A}$

Modularity matrix

$\mathbf{d} = \boldsymbol{A}\mathbf{e}$

$\boldsymbol{M} = \boldsymbol{A} - 1/(2|E|)\mathbf{d}\mathbf{d}^T$

Everything is undirected.

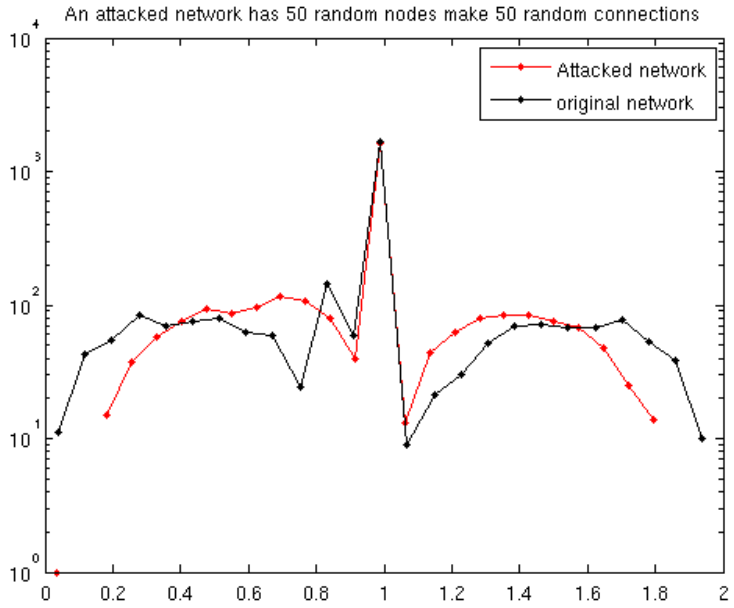
Why are we interested in the spectra?

Modeling

Properties

Moments of the adjacency

Anomalies



Regularities

Network Comparison

Fay et al. 2010 – Weighted Spectral Density

The network is as19971108 from Jure's snap collect (a few thousand nodes) and we insert random connections from 50 nodes

Spectral bounds from Gerschgorin

$$-d_{\max} \leq \lambda(\boldsymbol{A}) \leq d_{\max}$$

$$0 \leq \lambda(\boldsymbol{L}) \leq 2d_{\max}$$

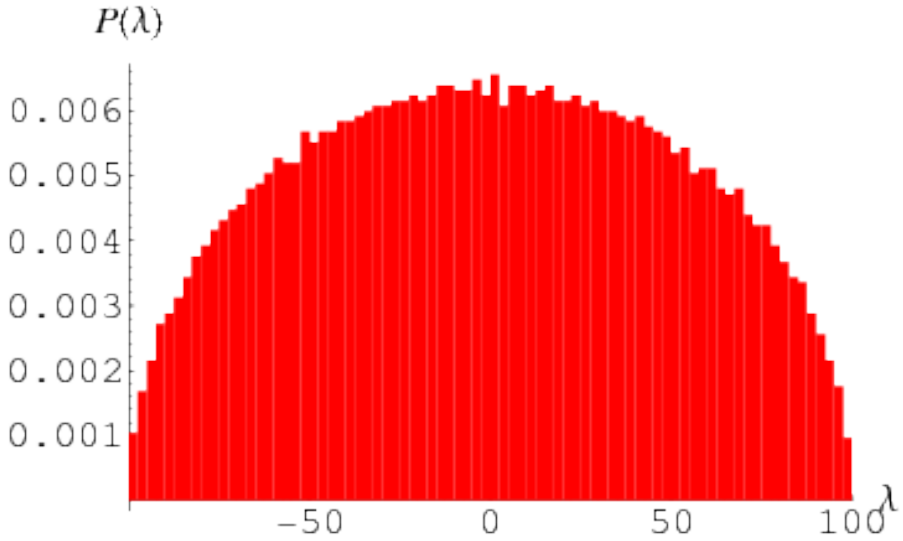
$$0 \leq \lambda(\tilde{\boldsymbol{L}}) \leq 2$$

(from a slightly different approach)

Semi-circle law

Wigner's semi-circle law

Eigenvalues of a random symmetric matrix where each entry is independent have a semi-circle density

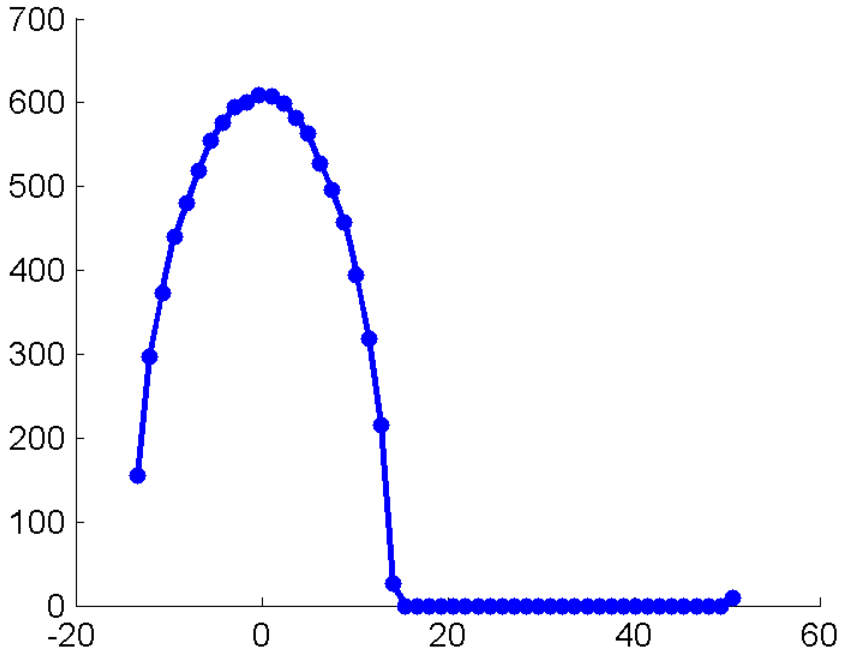


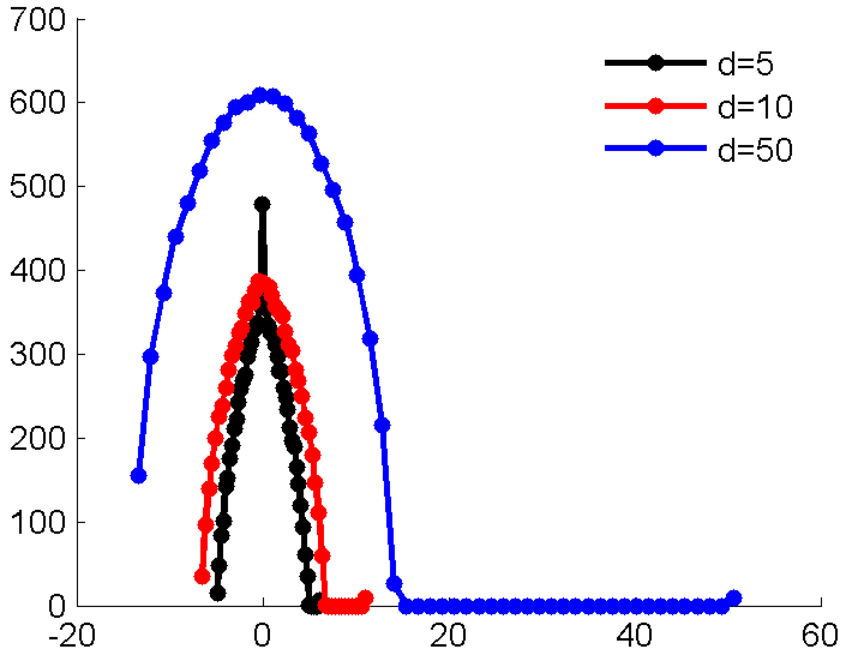
Erdős–Rényi graphs obey a special case

Image from Mathworld

Erdős–Rényi Semi-circles

The eigenvalues of the adjacency matrix for
n=1000, averaged over 10 trials
Semi-circle with outlier if average degree is large enough.





Observed by Farkas and in the book "Network Alignment" edited by Brandes (Chapter 14)

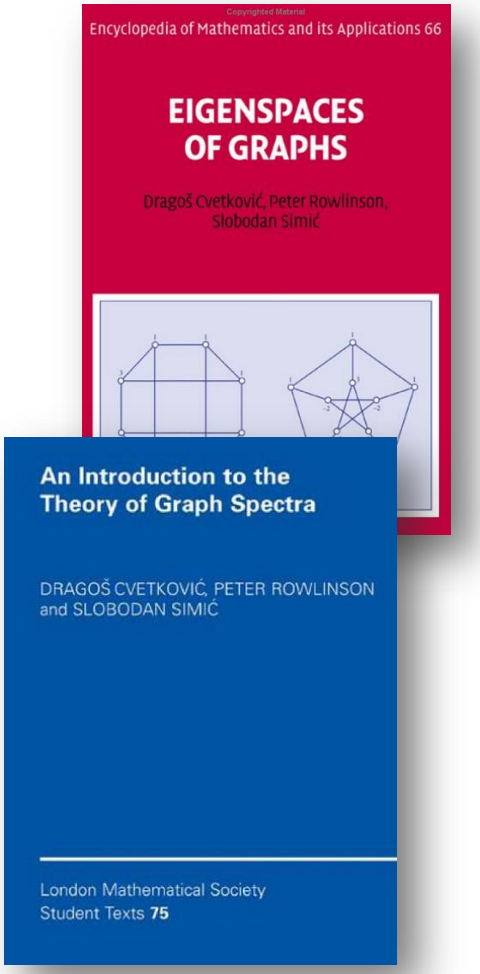
Previous results

Farkas et al. : Significant deviation from the semi-circle law for the adjacency matrix

Mihail and Papadimitriou : Leading eigenvalues of the adjacency matrix obey a power-law based on the degree-sequence

Chung et al. : Normalized Laplacian still obeys a semi-circle law

Banerjee and Jost : Study of types of patterns that emerge in evolving graph models – explain many features of the spectra



In comparison …

We use “exact” computation of spectra, instead of approximation.

We study “all” of the standard matrices over a range of large networks.

Our “large” is bigger.

We look at a few different random graph models.

DATA

Data sources

SNAP	Various	100s-100,000s
SNAP-p2p	Gnutella Network	5-60k, ~30 inst.
SNAP-as-733	Autonomous Sys.	~5,000, 733 inst.
SNAP-caida	Router networks	~20,000, ~125 inst.
Pajek	Various	100s-100,000s
Models	Copying Model	1k-100k 9 inst. 324 gs
	Pref. Attach	1k-100k 9 inst. 164 gs
	Forest Fire	1k-100k 9 inst. 324 gs
Mine	Various	2k-500k
Newman	Various	
Arenas	Various	
Porter	Facebook	100 schools, 5k-60k
IsoRank, Natalie	Protein-Protein	<10k , 4 graphs

Thanks to all who make data available

Big graphs

Arxiv	86376	1035126	Co-authorship
Dblp	9356	356290	Co-authorship
Dictionary(*)	111982	2750576	Word definitions
Internet(*)	124651	414428	Routers
Itdk0304	190914	1215220	Routers
p2p-Gnutella(*)	62561	295756	Peer-to-peer
Patents(*)	230686	1109898	Citations
Roads	126146	323900	Roads
Wordnet(*)	75606	240036	Word relationship
Web-NotreDame	325729	2994268	Web

Models

Preferential Attachment

Start graph with a k-node clique. Add a new node and connect to k random nodes, chosen proportional to degree.

Copying model

Start graph with a k-node clique. Add a new node and pick a parent uniformly at random. Copy edges of parent and make an error with probability α

Forest Fire

Start graph with a k-node clique. Add a new node and pick a parent uniformly at random. Do a random "bfs'/"forest fire" and link to all nodes "burned"

COMPUTING SPECTRA OF LARGE NETWORKS

Matlab!

Always a great starting point.

My desktop has 24GB of RAM (less than $2500 now!)

24GB/8 bytes (per double) = 3 billion numbers

~ 50,000-by-50,000 matrix

Possibilities

D = eig(A) – needs twice the memory for A,D

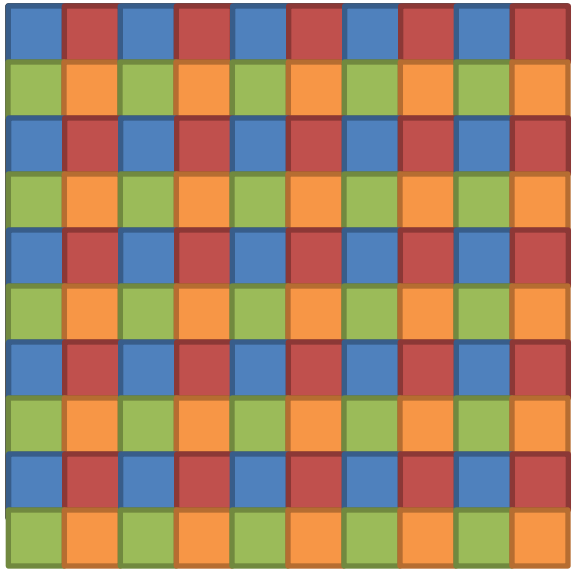
[V,D] = eig(A) – needs three times the memory for A,D,V

These limit us to ~38000 and ~31000 respectively.

ScaLAPACK

LAPACK with distributed memory dense matrices

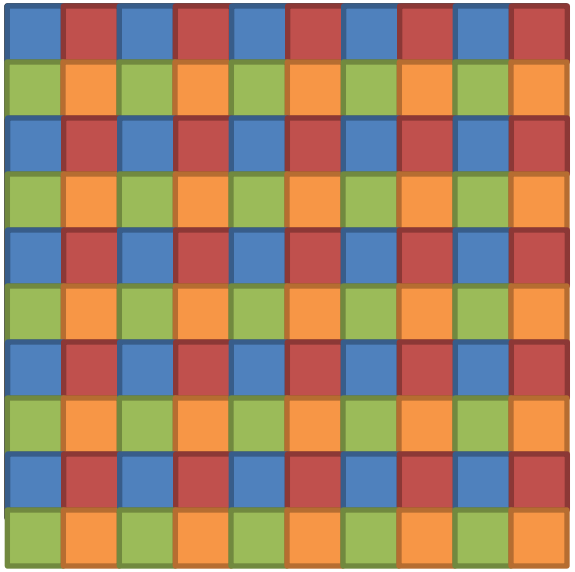
Scalapack uses a 2d block-cyclic dense matrix distribution



Eigenvalues with ScaLAPACK

Mostly the same approach as in LAPACK

1. Reduce to tridiagonal form (most time consuming part)
2. Distribute tridiagonals to all processors
3. Each processor finds all eigenvalues
4. Each processor computes a subset of eigenvectors



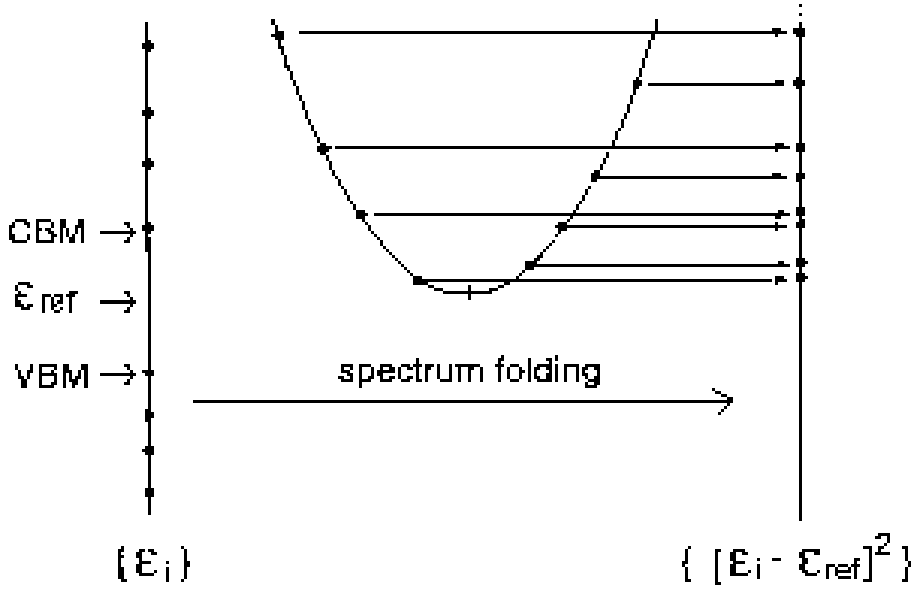
Using the MRRR algorithm steps 3 and 4 are more intricate and faster.

MRRR due to Parlett and Dhillon; implemented in ScaLAPACK by Christof Vomel.

Alternatives

Use ARPACK to get extrema

Use ARPACK to get interior around λ_0 via the folded spectrum $((\boldsymbol{A} - \lambda_0))^k$



Farkas et al. used this approach.

ISSUES WITH COMPUTING SPECTRA

Bugs - Matlab

eig(A)

Returns incorrect eigenvectors

Seems to be the result of a bug in Intel's MKL library.

Bug – ScaLAPACK default

sudo apt-get install scalapack-openmpi
Allocate 36000x36000 local matrix
Run on 4 processors

Code crashes

Bug - LAPACK

Scalapack MRRR

Compare standard lapack/blas to atlas performance

Result: correct output from atlas

Result: incorrect output from lapack

Hypothesis: lapack contains a known bug that's apparently in the default ubuntu lapack

Moral

Always test your software.

Extensively.

(Super)-Computers

Redsky
- 2x Intel Nehalem 2.9 GHz/8 core
- 12 GB/node
- 22TB memory total
- used up to 500 nodes/
  - 6 TB memory

Hopper (I)
- 2x AMD quad-core
  - 2.4 GHz/8 cores
- 16 GB/node

Cielo (testbed) 20 nodes

Adding MPI tasks vs. using threads

Most math libraries have threaded versions (Intel MKL, AMD ACML)

Is it better to use threads or MPI tasks?

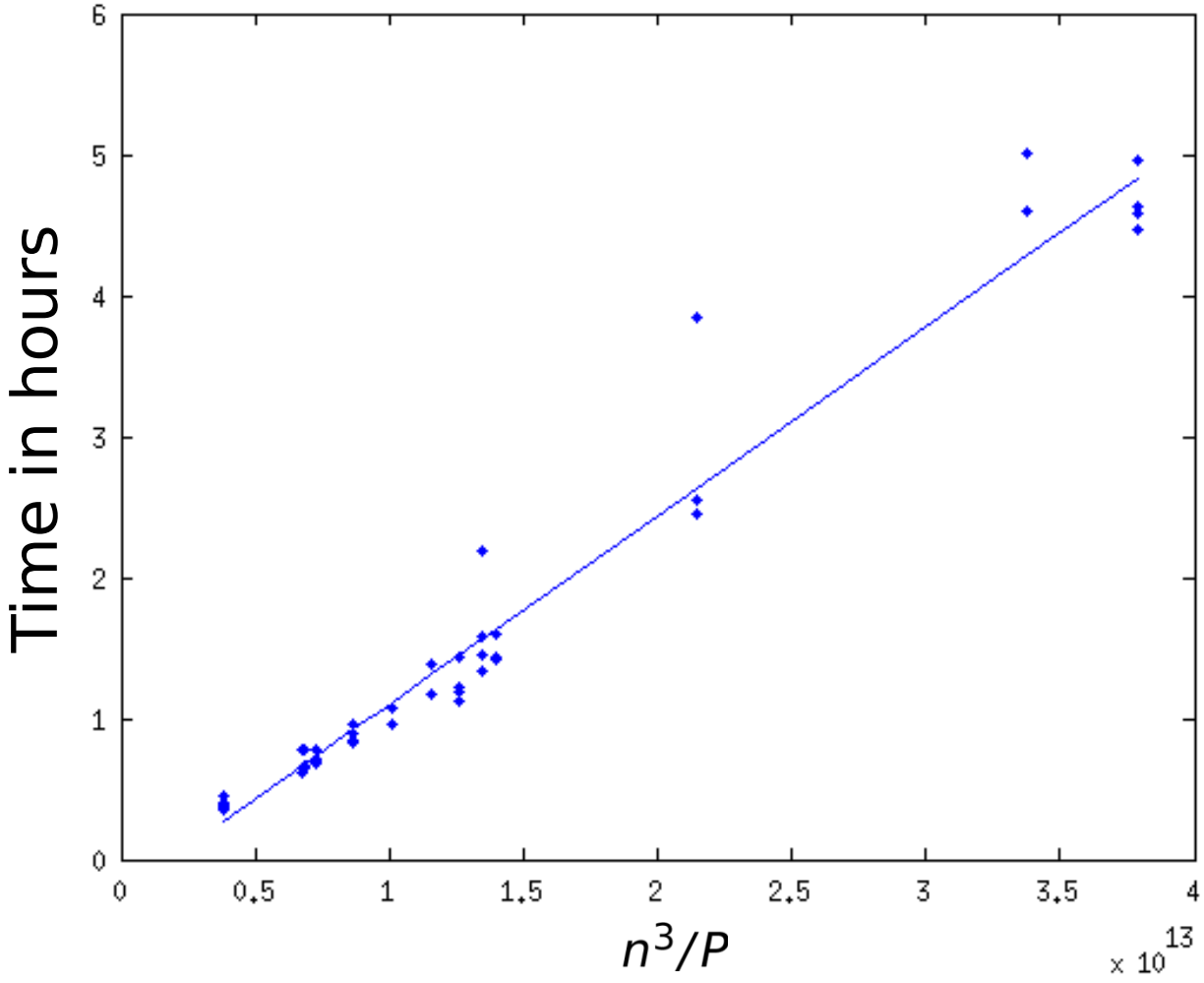
It depends.

Threads	Ranks	Time-T	Time-E
1	36	1271.4	339.0
4	16	1058.1	456.6

Threads	Ranks	Time
1	64	1412.5
4	16	1881.4
16	4	Omitted.

Normalized Laplacian for 36k-by-36k co-author graph of CondMat

Strong Parallel Scaling



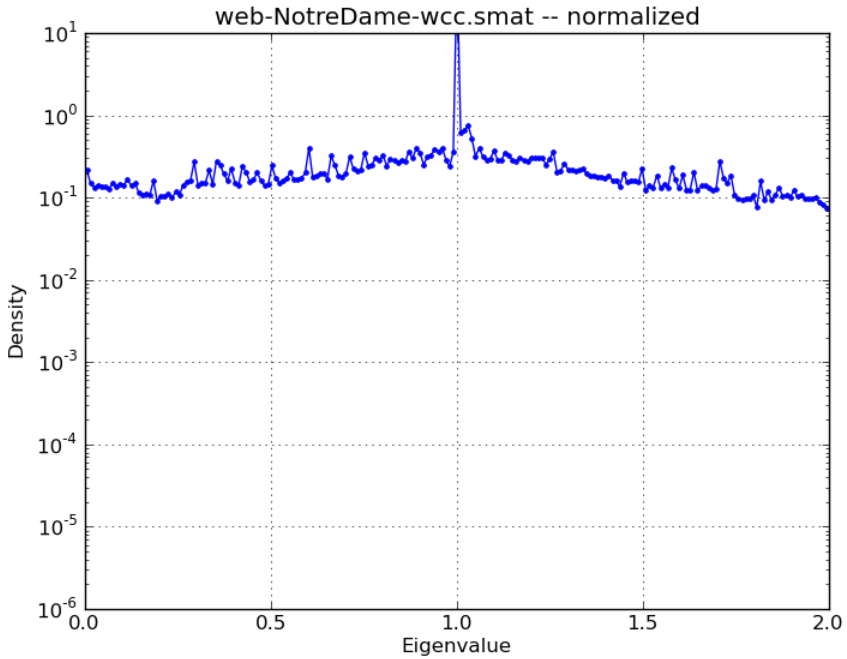
Time $\propto (1.3)n^3/P$

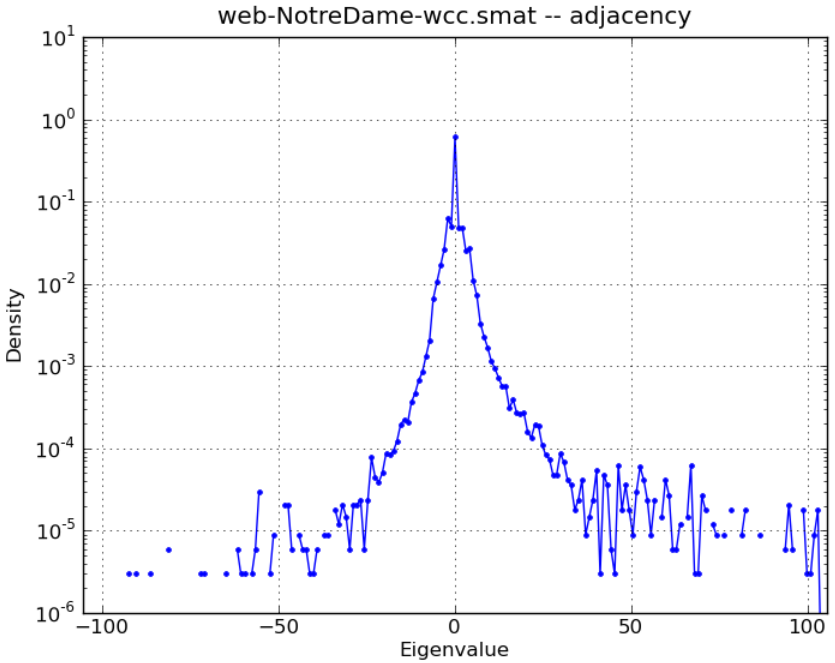
Good strong scaling up to 325,000 vertices

Estimated time for 500,000 nodes **9 hours** with 925 nodes (7400 procs)

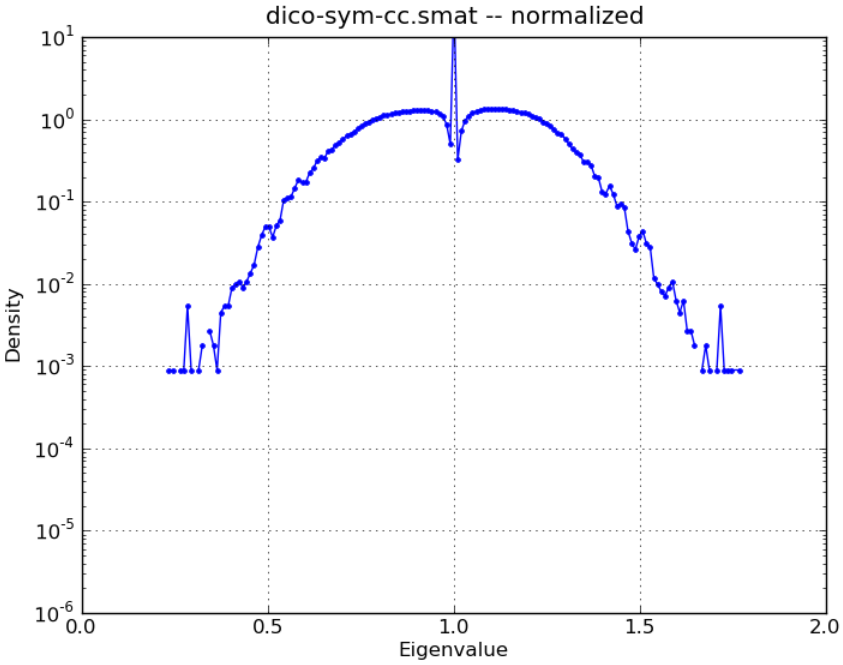
EXAMPLES

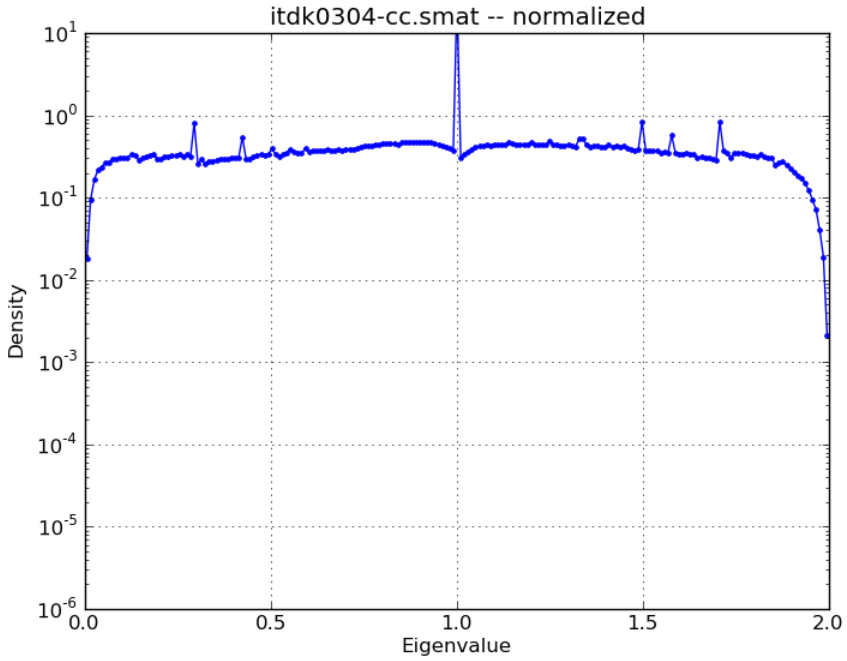
A $40,000 matrix computation





Spikes?

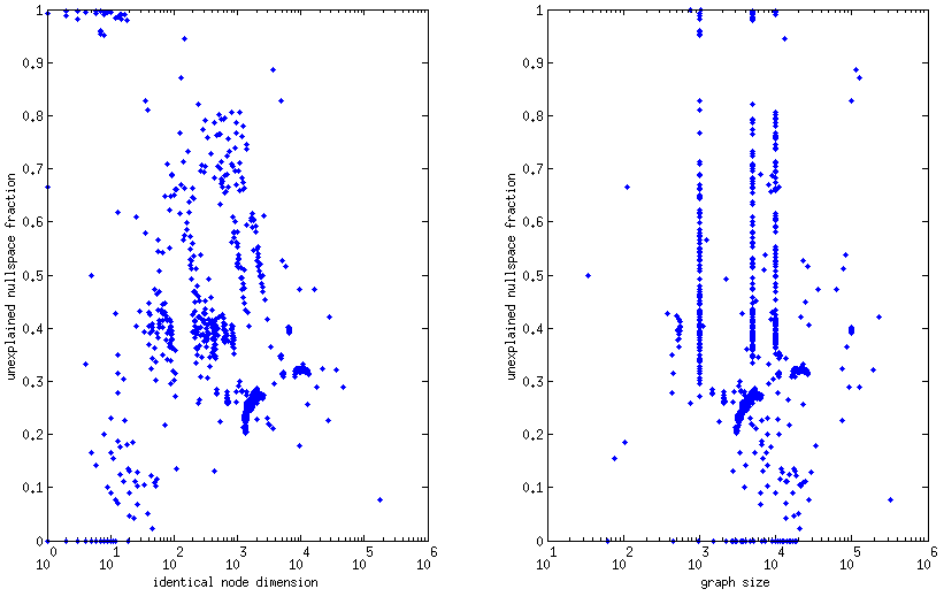

dico-sym-cc.smat -- normalized
Density
Eigenvalue
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0.0
0.5
1.0
1.5
2.0


itdk0304-cc.smat -- normalized
Density
Eigenvalue
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0.0
0.5
1.0
1.5
2.0

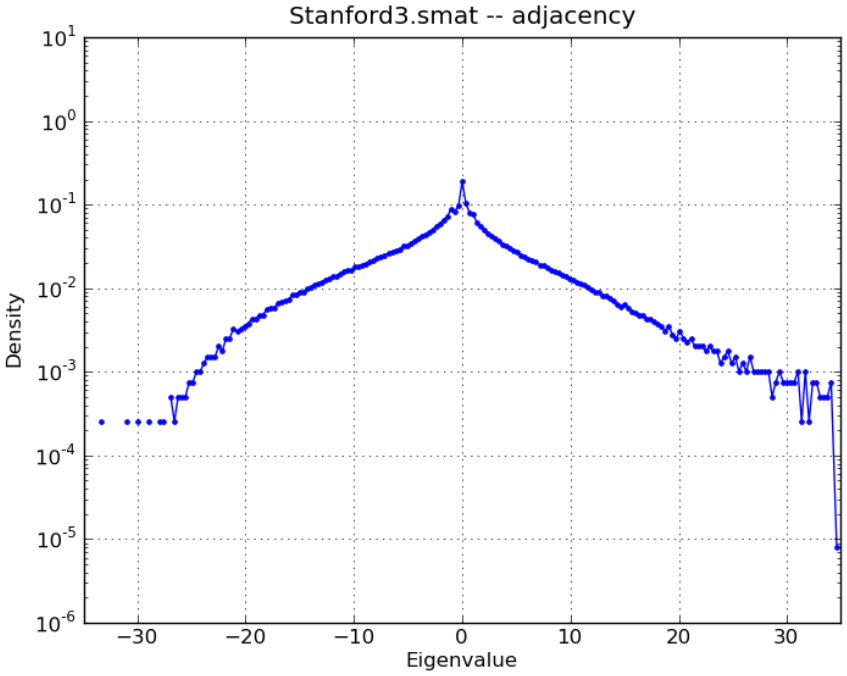
Nullspaces of the adjacency matrix

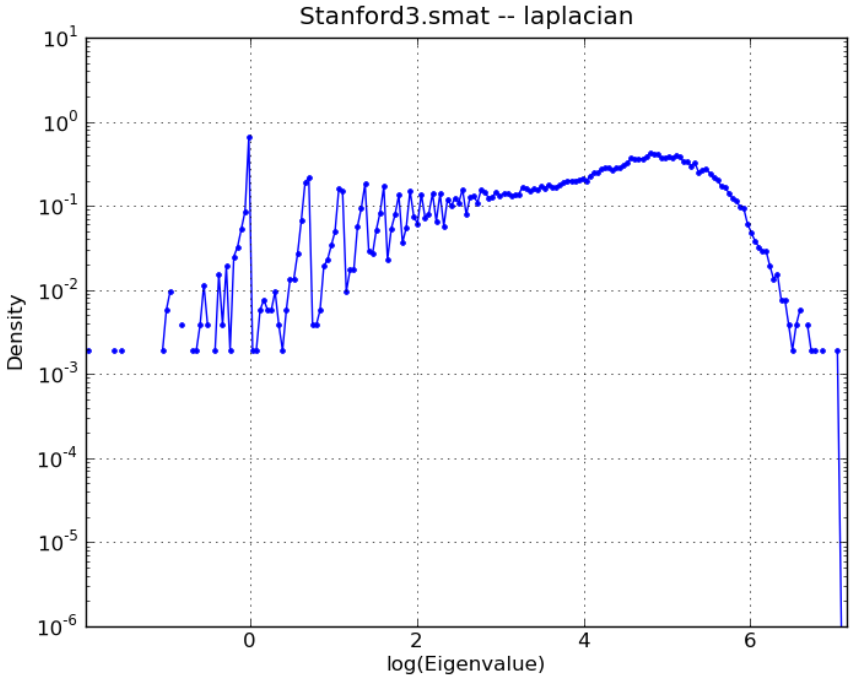
$$(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A})\mathbf{x} = \mathbf{x} \Rightarrow \boldsymbol{A}\mathbf{x} = 0$$

So unit eigenvalues of the normalized Laplacian are null-vectors of the adjacency matrix.



Stanford's Facebook Network





Data from Mason Porter

Movies of spectra…

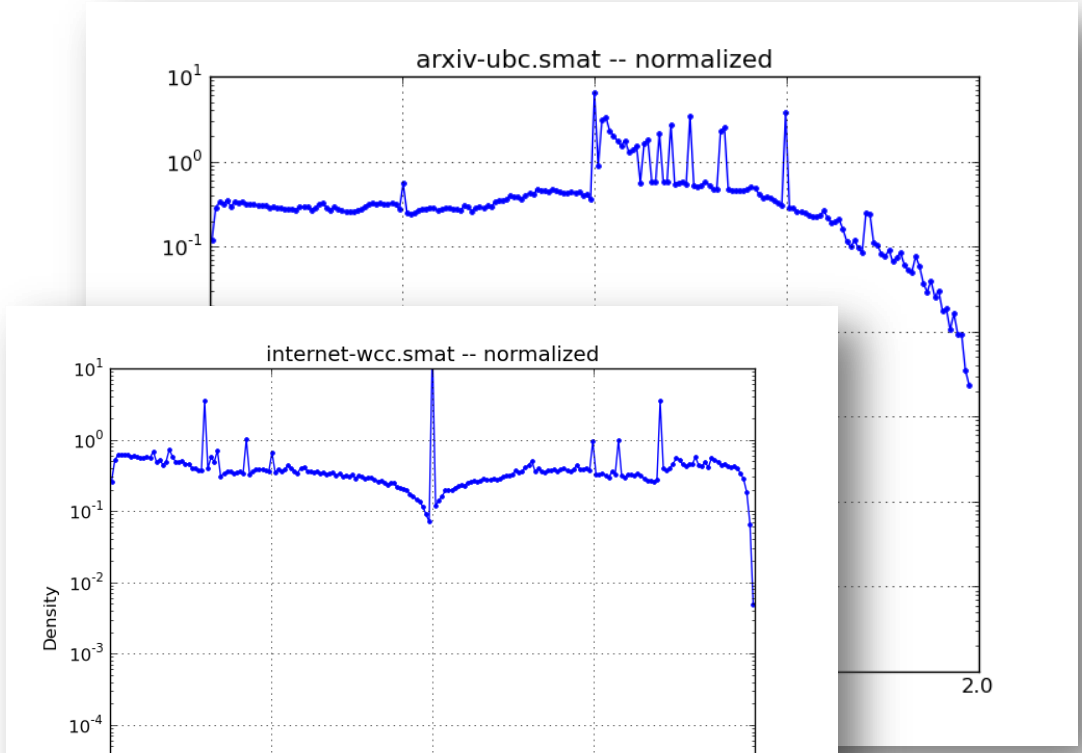
As these models evolve, what do the spectra look like?

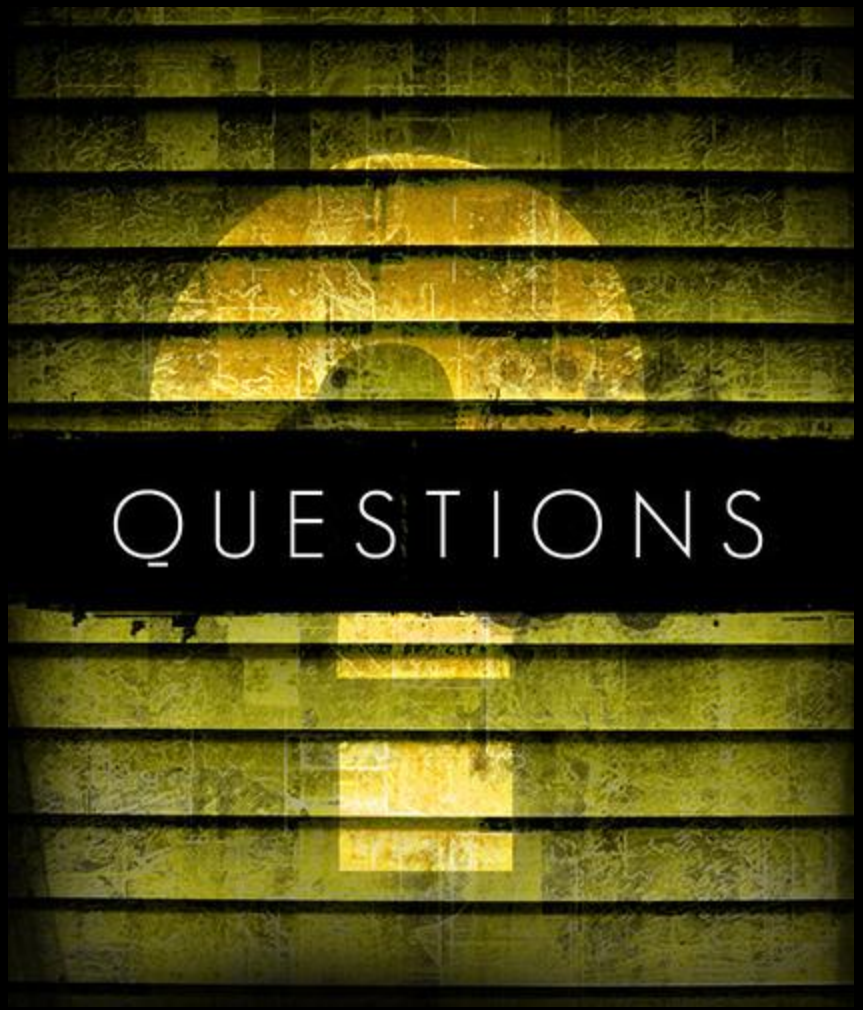
Where is this going?

Keep asking questions

Keep reading

Keep finding answers?





Code will be available eventually. Image from good financial cents.