Graph Mining
Guest Lecture

Charalampos E. Tsourakakis

April 1, 2009

1. Prerequisites
   - Matrix-Vector Multiplication
   - Orthogonality
   - Vector and Matrix Norms
2. More on Singular Value Decomposition
   - Image Compression
   - Counting Triangles
3. Graph Patterns and Kronecker Graphs
4. From Christos powerpoint slides:
   - HITS
   - Pagerank
   - Epidemic Threshold
   - More power laws

Matrix Vector Multiplication

Matrix $A \in \mathbb{R}^{mxn}$
Vector $x \in \mathbb{R}^n$
$Ax = b$, $b \in \mathbb{R}^m$

- $b_i = \sum_{j=1}^{n} a_{ij} x_j$ for $i = 1 \ldots m$.
- $b = Ax = [\vec{a_1}| \ldots |\vec{a_n}]x = x_1 \vec{a_1} + \ldots + x_n \vec{a_n}$.

A **linear map**, i.e.,
"Input" in $\mathbb{R}^n$
"Output" in $\mathbb{R}^m$
Why is *A* a linear map?

Orthogonality

Definition (Orthogonal Vectors)

Two vectors v_1 and v_2 are orthogonal if their inner product is 0.

Definition (Linear Independence)

A set of vectors $V = \{v_1, \ldots, v_n\}$, $v_i \in \mathbb{R}^m$, is said to be linearly independent if the equation $\alpha_1 v_1 + \ldots + \alpha_n v_n = 0$, $\alpha_i \in \mathbb{R}$ holds if and only if $\alpha_i = 0$, $i = 1 \ldots n$.

Theorem

Two orthogonal vectors are independent.

Vector Norms

A norm is a function $||.|| : \mathbb{R}^n \rightarrow \mathbb{R}$ with the following three properties:

- $||x|| \geq 0$ and $||x|| =0$ if $x = 0$.
- $||x + y|| \leq ||x|| + ||y||$
- $||\alpha x|| = |\alpha| \, ||x||$

2-norm ($x \in \mathbb{R}^n$)

$$||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{x^T x} \qquad (1)$$

Matrix Norms

$A \in \mathbb{R}^{mxn}$.

Definition (Frobenius norm)

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^2} \quad (2)$$

The 2-norm induces a matrix norm:

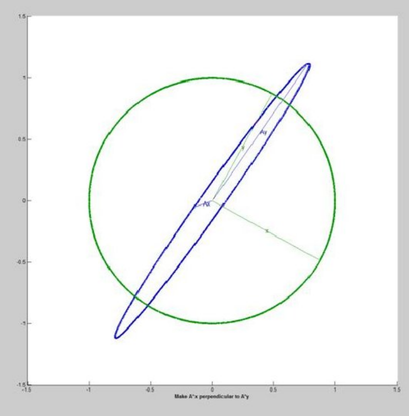
Definition (2-norm)

$$||A||_2 = sup_{x \in \mathbb{R}^n, ||x||=1} ||Ax|| \quad (3)$$

Think! What do they express?

SVD geometric intuition

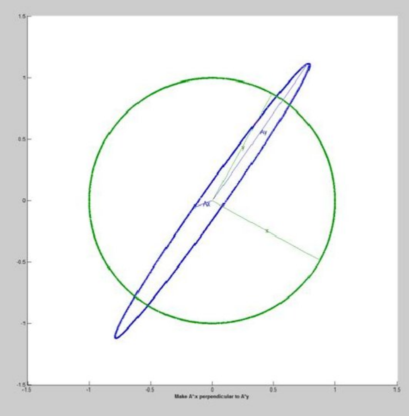
Remember: we view matrix $A \in \mathbb{R}^{mxn}$ as an operator!



The unit circle (sphere) is mapped in an ellipse (hyperellipse).

SVD geometric intuition

Display in this figure $u_1, u_2, v_1, v_2, \sigma_1, \sigma_2$.



Play yourselves with command eigshow in MATLAB!

SVD and dimensionality reduction

Assume rank(A)=r, $A \in \mathbb{R}^{mxn}$, thus $A = \sum_{j=1}^{r} \sigma_j u_j v_j^T$.
The following two theorems show the optimality of SVD with respect to the $L2$ and Frobenious norm as a dimensionality reduction tool.

Theorem

For any $0 \leq k < r$ define $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T$. The following equations hold for any matrix $B \in \mathbb{R}^{mxn}$ whose rank is k or less:

$$||A - A_k||_2 \leq ||A - B||_2 \quad (4)$$

$$||A - A_k||_F \leq ||A - B||_F \quad (5)$$

Image Compression: Original Images

Original image

Original image

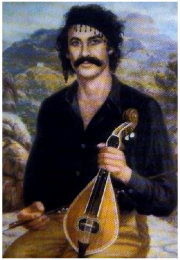


Image Compression: Results from the first image

Figure: Result on first image after applying SVD for k=10,30 and 60.

Image Compression: Results from second image

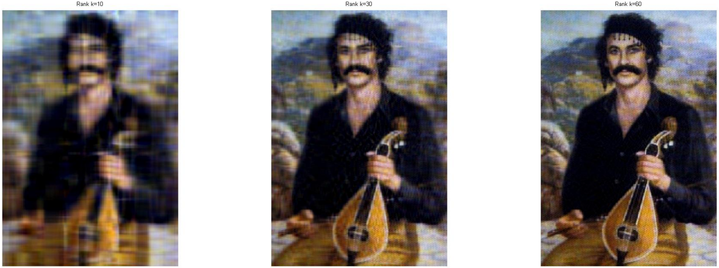


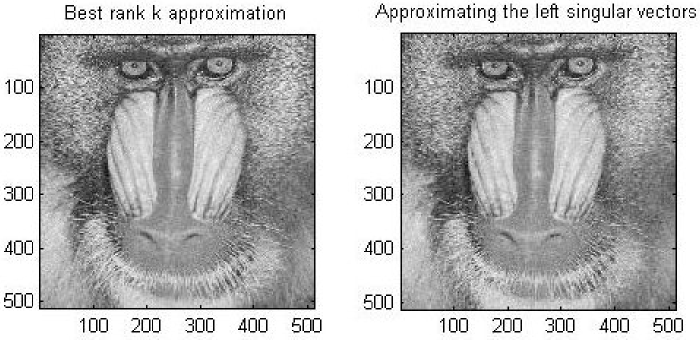
Figure: Result on second image after applying SVD for k=10, 30 and 60.

Randomized SVD

Computing the SVD of a matrix is expensive! Lose little accuracy for speedup. READINGS:

- *Fast Monte Carlo Algorithms for Finding Low Rank Approximations* by **Alan Frieze, Ravi Kannan, Santosh Vempala**.
- *Fast Monte Carlo Algorithms for Matrices II: Computing a Low Rank Approximation to a Matrix* by **P. Drineas, R. Kannan, and M.W. Mahoney**.
- *Improved Approximation Algorithms for Large Matrices via Random Projections*, **T. Sarlós**.

Randomized SVD for image compression



Triangle Counting

Theorem (EigenTriangle)

The total number of triangles in a graph is equal to the sum of cubes of its adjacency matrix eigenvalues divided by 6, namely:

$$\Delta(G) = \frac{1}{6}\sum_{i=1}^{n}\lambda_i^3 \tag{6}$$

Theorem (EigenTriangleLocal)

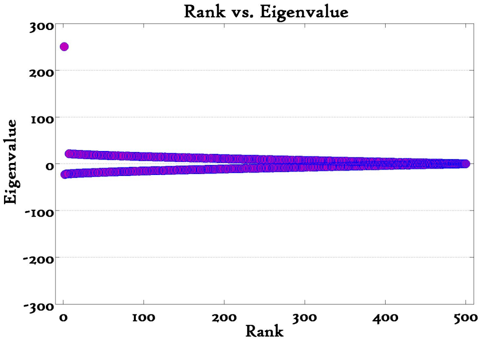
The number of triangles Δ_i that node i participates in, can be computed from the cubes of the eigenvalues of the adjacency matrix

$$\Delta_i = \frac{\sum_j \lambda_j^3 u_{i,j}^2}{2} \tag{7}$$

where $u_{i,j}$ is the j-th entry of the i-th eigenvector.

Triangle Counting

Spectrum of random graph $G_{n,\frac{1}{2}}$:



Triangle Counting

Therefore it suffices to consider the first eigenvalue of the adjacency matrix of $G_{n,\frac{1}{2}}$ to get a good estimate of the number of triangles. What about real world networks?

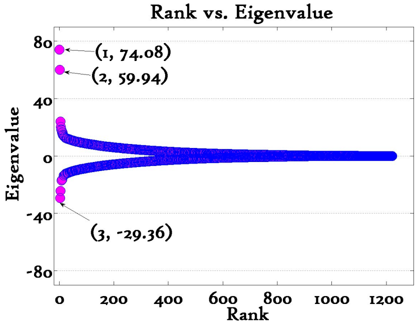


Figure: This figure plots the value of the eigenvalue vs. its rank for a network with $\approx$ 1,2K nodes, 17K edges.

Triangle Counting

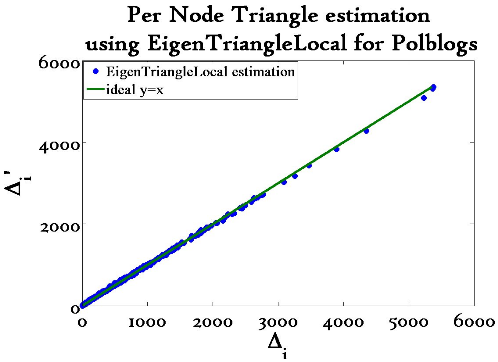


Figure: Local Triangle Reconstruction using a 10-rank approximation for the Political Blogs network

Triangle Counting

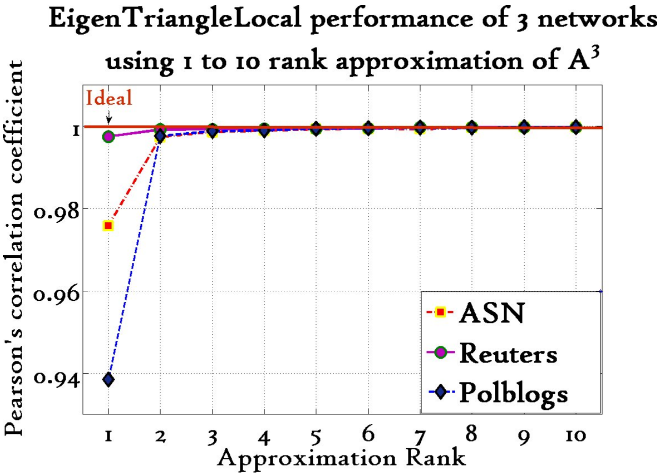


Figure: Local Triangle Reconstruction for three networks, Flickr, Pol Blogs and Reuters.

Triangle Counting

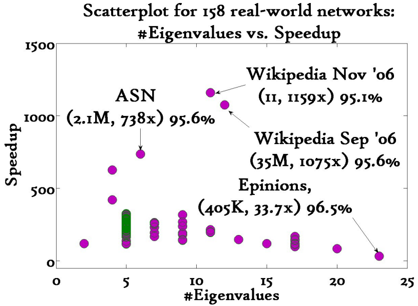


Figure: Scatterplots of the results for 158 networks. Speedup vs. Eigenvalues: The mean required approximation rank for 95% accuracy is 6.2.Speedups are between 33.7x and 1159x, with mean 250.

Triangle Counting

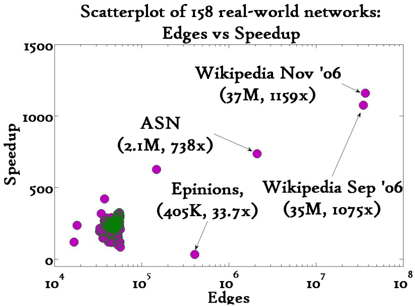


Figure: Speedup vs. Edges: Notice the trend of increasing speedup as the network size grows

Degree Distribution

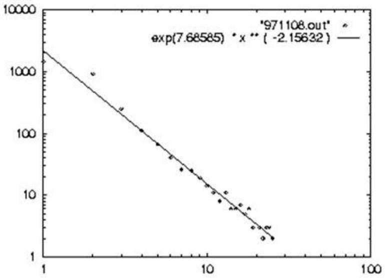


Figure: Outdegree Plot from the *Faloutsos*, *Faloutsos*, *Faloutsos* paper. Observe that the plot is linear in log-log scale. The least squares fitting gives that: $freq = degree^{-2.15}$

Top-Eigenvalues

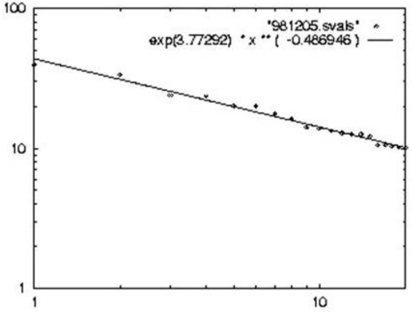


Figure: Top Eigenvalue Plot from the *Faloutsos*, *Faloutsos*, *Faloutsos* paper

Spectrum: Eigenvalue vs. Rank

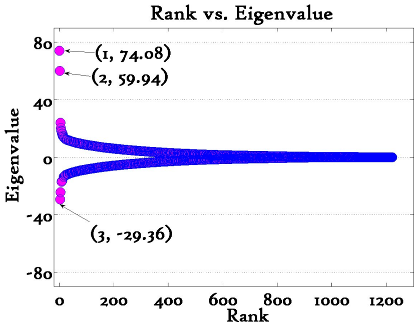


Figure: This figure plots the value of the eigenvalue vs. its rank for a network with $\approx$ 1,2K nodes, 17K edges.

Hop-plot power law

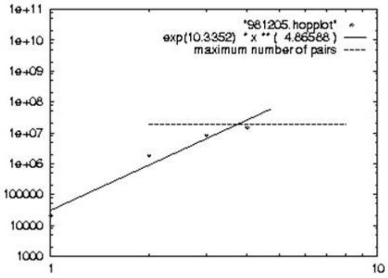


Figure: Hop Plot from the *Faloutsos*, *Faloutsos*, *Faloutsos* paper. Pairs of nodes as a function of hops $N(h) = h^H$ The least squares fitting gives that the exponent is $H = 4.86$.

Triangle power laws: Participation Law

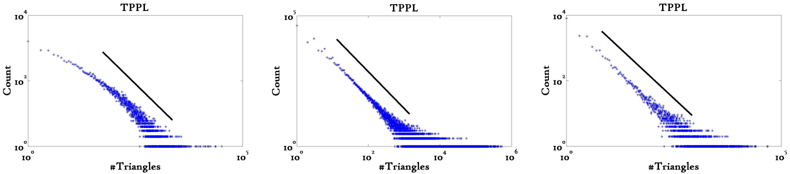


Figure: Triangle Participation Law for three networks HEP-TH, Flickr and Epinions. Observe the emerging power law or the power law tail.

Triangle power laws: Degree Triangle Power Law

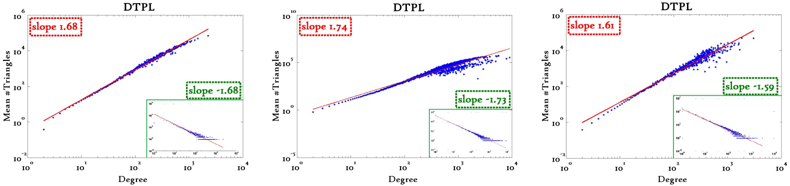


Figure: Degree Triangle Power Law for three networks Reuters, Flickr and Epinions.

Communities

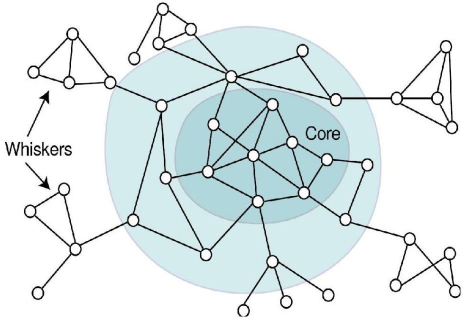
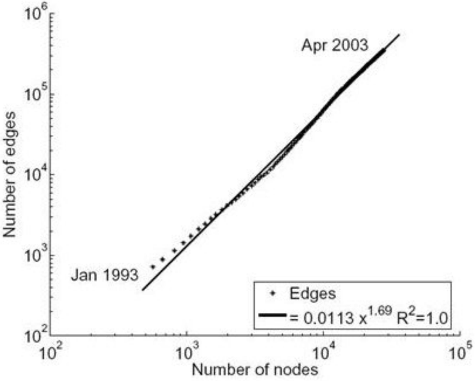


Figure: Figure from the Leskovec et al. paper. Caricature of how a real world network looks like. (courtesy of J. Leskovec).

Densification Law

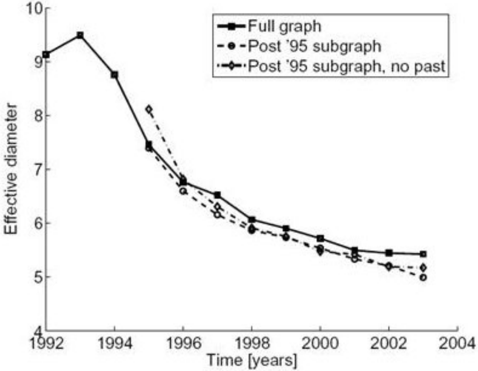
$E(t) \propto N(t)^{\alpha}$, $\alpha = 1.69$.



(a) arXiv

Shrinking Diameter

.... and the shrinking diameter phenomenon.



(a) arXiv citation graph

Figure: Shrinking diameter phenomenon for the Arxiv citation graph.

Kronecker Graphs in a picture

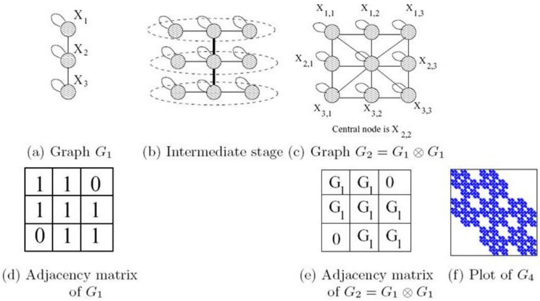


Figure: Deterministic Kronecker Graphs

Kronecker Graphs properties

The following properties hold for deterministic Kronecker graphs:

- Power-law-tail in- and out-degrees
- Power-law-tail scree plots
- constant diameter
- perfect Densification Power Law
- communities-within-communities

Chazelle's Talk, 3:15 pm Wean Hall 7500. SCS Distinguished Lecture