# Commute Times, Concentration, Discrete Green's Functions, and Kernel Regression on Graph'ish Data

Chris Long

21 August, 2014

# A Model Problem

Graph algorithms based on heuristic analogies to physical systems, e.g. heat or electrical flow, have a long history...

Typically motivated by physical diffusion processes involving a Poisson equation, or some variation thereof:

$$\Delta \mathbf{f}(x) = \mathbf{g}(x)$$

where $\Delta$ is a Laplacian and $\mathbf{f}$ and $\mathbf{g}$ are functions defined on a Riemannian manifold.

# A Model Problem

In the discrete setting, this subsumes a number of relative and absolute centrality algorithms, e.g. [7], [8], [20]...

and even personalized PageRank, for an appropriate Dirichlet boundary condition [18]

Given problem-specific weights, take the source function $\mathbf{g}$ to be a weighted characteristic function for the seed set $S$, i.e.

$$\mathbf{g}_i = \begin{cases} w_i & \text{if } i \in S, \\ 0 & \text{if } i \notin S. \end{cases}$$

Solve the model Poisson problem by convolving the source term with the discrete Green's function $\mathcal{G}$ for $\Delta$:

$$\mathbf{f} = \mathcal{G}\chi_S$$

For a graph without boundary the Green's function $\mathcal{G}$ is just the Moore-Penrose pseudoinverse of the graph Laplacian [5]:

$$\mathcal{G} = \mathcal{L}^{\dagger} = \sum_{\lambda_j > 0} \frac{1}{\lambda_j} \mathbf{u}_j \mathbf{u}_j^T .$$

Hence we "solve" the linear system:

$$\mathcal{L}\mathbf{f} = \chi_S$$

# Generalized Inversion & Least Squares Estimation

The pseudoinverse $A^\dagger$ plays an important role in linear least squares estimation

For a matrix $A$ with full column rank, $A^\dagger$ takes the form:

$$A^\dagger = (A^T A)^{-1} A^T.$$

which is just the matrix form of the solution by normal equations in linear least squares estimation

In general, $\mathcal{L}^\dagger$ may be obtained at the strong regularization limit, i.e. as the ridge parameter $\delta \to 0$

$$(\mathcal{L}^T \mathcal{L} + \delta I)^{-1} \mathcal{L}^T \to \mathcal{L}^\dagger$$

The **commute time** between nodes $i$ and $j$ is just the sum of the hitting time, $H(i,j)$, for a random walk starting at $i$ to $j$ and vice-versa

$$C(i,j) = H(i,j) + H(j,i) = C(j,i).$$

So $C(i,j)$ is really just the expected number of hops for a random walk starting at $i$ to reach $j$ and then return to $i$

$C(i,j)$ is proportional to the **effective resistance** between $i$ and $j$, regarded as resistors in an electrical network [4]

*For any two vertices $i$ and $j$, $C(i,j)$ may be obtained as a quadratic form over $\mathcal{L}^\dagger$...*

# Commute Times & the Combinatorial Laplacian

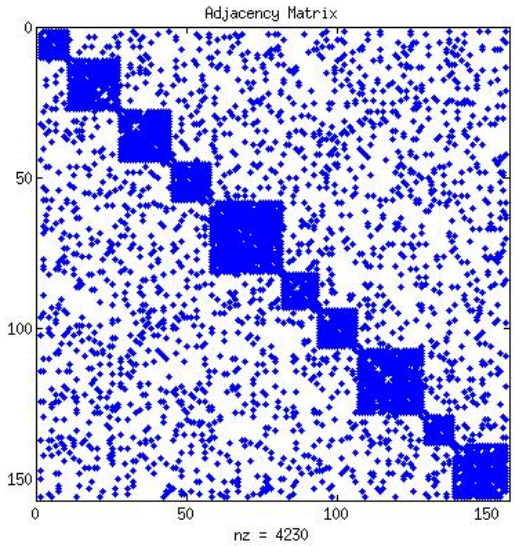
**Big Idea**: Commute time/resistance distance between nodes of the graph may be computed from the eigendecomposition of $\mathcal{L}_C$:

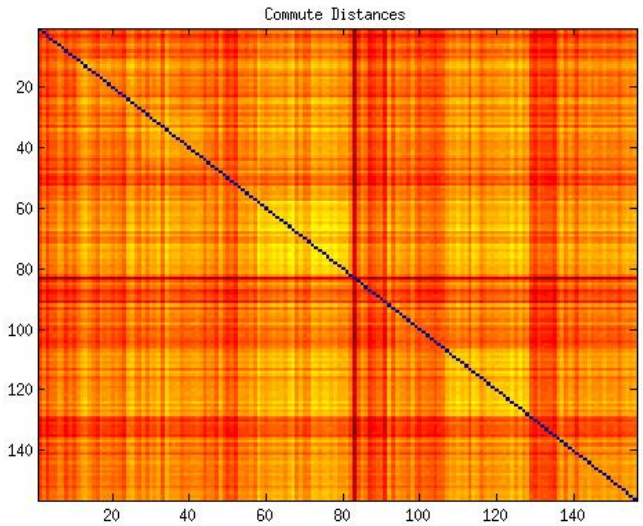
## Theorem

*(Fouss et al. [8]) The commute time $C(i,j)$ between nodes $i$ and $j$ may be written as:*

$$\begin{aligned}\frac{1}{N}C(i,j) &= (\mathbf{e}_i - \mathbf{e}_j)^T \mathcal{L}_C^{\dagger} (\mathbf{e}_i - \mathbf{e}_j) \\ &= \sum_{k=1}^{n-1} \frac{1}{\lambda_k} (\mathbf{u}_k(i) - \mathbf{u}_k(j))^2 \\ &= \mathcal{L}_C^{\dagger}(i,i) - 2\mathcal{L}_C^{\dagger}(i,j) + \mathcal{L}_C^{\dagger}(j,j).\end{aligned}$$

# Neat Math! What Could Possibly Go Wrong?



# Neat Math! What Could Possibly Go Wrong?

Empirically, the raw commute distance can be strongly affected by the node degree.

(von Luxburg et al.) For certain classes of random graphs, known concentration results yield [13]

$$C(i,j) \to \frac{1}{d_i} + \frac{1}{d_j}.$$

This is worthless as a distance measure on a graph, as all nodes then have the same nearest-neighbor, second nearest-neighbor, etc.

Moreover, it was observed that, in all cases, $\mathcal{L}_C^\dagger(ii) \approx \frac{1}{d_i}$, implying that $\mathcal{L}_C^\dagger(i,j) \to 0$.

# Concentration and Commute Distances

This is generally obtained under the following conditions [13]:

- The graph is well-connected
- The graph is free of bottlenecks, in that it has a small isoperimetric number/Cheeger constant
- ...i.e. the Fiedler eigenvalue $\lambda_{n-1}$ is bounded away from 0
- The minimum degree of the graph grows slowly with the order of the graph

The first two conditions are characteristic of expanders

# Concentration and Commute Distances

This is generally obtained under the following conditions [13]:

- The graph is well-connected
- The graph is free of bottlenecks, in that it has a small isoperimetric number/Cheeger constant
- ...i.e. the Fiedler eigenvalue $\lambda_{n-1}$ is bounded away from 0
- The minimum degree of the graph grows slowly with the order of the graph

The first two conditions are characteristic of expanders

So can we correct for the influence of the node degrees?

# Linear Models

Consider the two-dimensional linear model, whereby $x(i,j) \in \mathbb{R}^{m \times n}$ decomposes as:

$$x(i,j) = v + v_1(i) + v_2(j) + v_{12}(i,j)$$

for $v = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} x(i,j)$ termed the **main effect**

$v_1(i) = \frac{1}{n} \sum_{j=1}^{m} x(i,j) - v$ the $i^{th}$ **main row effect**

$v_2(j) = \frac{1}{m} \sum_{i=1}^{n} x(i,j) - v$ the $j^{th}$ **main column effect**

and $v_{12}(i,j)$ the **pairwise linear interaction** between the $i^{th}$ row and $j^{th}$ column.

# A Model Distance?

**Theorem**

*([10] [2]) The commute distance decomposes as a linear model:*

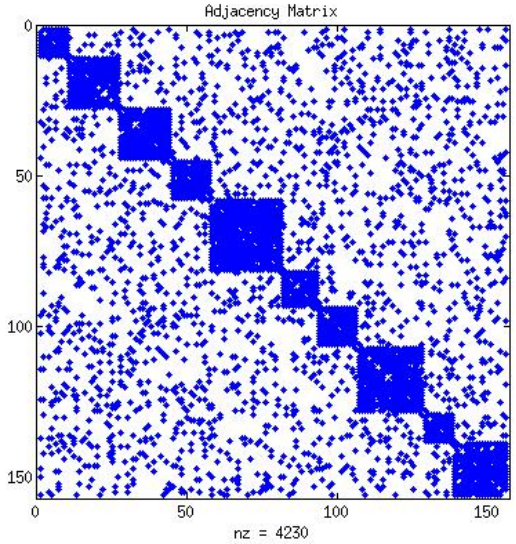
$$C(i,j) = v + v_{1(i)} + v_{2(j)} + v_{12(i,j)}$$

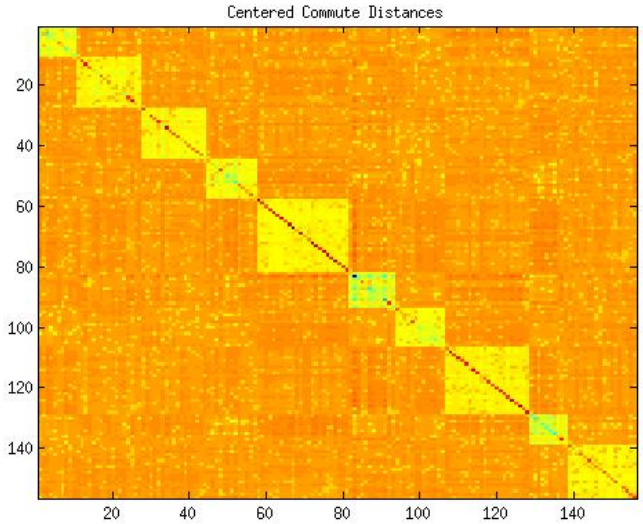
*wherein the parameters of the decomposition may be written in terms of the Moore-Penrose inverse of the Laplacian $\mathcal{L}_C$:*

$$v = \frac{2}{n}\sum_{i=1}^{n} \mathcal{L}_C^{\dagger}(i,i),$$

$$v_{1(i)} = v_{2(i)} = \mathcal{L}_C^{\dagger}(i,i) - \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_C^{\dagger}(j,j).$$

# A Model Distance?



# Concentration in the Wild?

What I am **not** claiming: that graphs corresponding to real-world networks are liable to be subject to the concentration artifacts described by von Luxburg et al.

- Real-world networks often lack sufficient expansion properties
- In particular, eigenvalue separation in can be extremely poor
- ...so the Cheeger constant often can't be bounded well away from the zero eigenvalue
- Moreover, the minimum vertex degree is often constant and does not grow with the order of the graph

# Concentration in the Wild?

What I **am** claiming:

- That using the commute interactions in lieu of the raw commute time is statistically justifiable for all graphs
- ...and appears to handle the currently known cases gracefully
- For a given seed set, computing the linear interactions is **much** easier than computing $C(i, j)$
- i.e. one can efficiently obtain the implicit convolution of the Green's function with a given seed set.

# A Scalable Algorithm

Consider the following decomposition of $\mathcal{L}_C^\dagger$ [16]:

$$\mathcal{L}_C^\dagger = \left(\mathcal{L}_C + P_h\right)^{-1} - P_h$$

We solve for all centered commute times $\tilde{C}_S$ with respect to a set of labeled nodes $S$ by computing:

$$\tilde{C}_S = \mathcal{L}_C^\dagger \chi_S = \left(\mathcal{L}_C + P_h\right)^{-1} \chi_S - P_h \chi_S$$

which requires the solution of a single linear system over the positive definite matrix $\mathcal{L}_C + P_h$ [10].

Actually just a rediscovery (or a re-rediscovery?) of an existing algorithm with a different theoretical justification [7], [8].

# Personalized PageRank

Similarly, personalized PageRank may be obtained by solving the following linear system [18]:

$$\mathbf{u} = K^{-1}\mathbf{v}$$

for $\mathbf{u} \doteq D^{1/2}\mathbf{x}$, $\mathbf{v} \doteq D^{1/2}\chi_S$, and $K \doteq \alpha I + (1-\alpha)\mathcal{L}_S$.

Hence, Personalized PageRank uses a different Laplacian and the following regularization parameter:

$$\delta = \frac{\alpha}{1-\alpha}$$

.

# And Now For Something Completely Different...

**But these are not what we should focus on going forward...**

# It's Not a Damn Graph

Data often contains rich features of potential predictive value that are thrown away to ~~hammer~~ fit the data into a graph abstraction:

- E-mails are naturally modeled as hyperedges (e.g. $n$-way communications due to multiple recipients)
- Netflow has a variety of attributes (e.g. protocol, port, packet and byte counts, tc/icp flags)

Relational information is useful, but it is not the only (nor always the most useful) property of the data

# It's Not a Damn Graph

We would like to use methods that:

- Make use of topological features (but can handle the corresponding high-dimensional feature spaces)...
- and also extend naturally to heterogeneous data
- Don't deviate "too far" from familiar graph-theoretic techniques, e.g. PageRank, centered commute times, etc., when given strictly 2-way relational data?
- Provide a statistically interpretable - and a statistically defensible - result

# Log-Linear Models

Many graph algorithms based on random-walks are really just instances of kernel regression in graph-theoretic disguise [14]

However, in this setting they:

- Minimize an $\ell_2$ penalty on a binary response
- Incestuously use the nodes as both observations and features
- Are less interpretable than log-linear models

Logistic linear models elegantly address these objections without straying far from current practice

# Bait-and-Switch

But what if only topological information is available?

**Spoilers:** There is a fascinating relationship to existing random-walk inspired algorithms potentially applicable to a wide variety of kernels

Let $D : C^1 \longrightarrow C^0$ be the $n \times m$ edge-incidence matrix

$D$ has columns of the form $\mathbf{x}_e = \delta_{e^+} - \delta_{e^-}$ where $e^+$ and $e^-$ denote the head and tail vertices of edge $e$, respectively

$$D = \begin{array}{c} \\ u \\ \\ v \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} | & | & & \mathbf{e}_i & & | & | \\ | & | & & \vdots & & | & | \\ | & | & & 1 & & | & | \\ \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \vdots & \cdots & \mathbf{e}_{m-1} & \mathbf{e}_m \\ | & | & & -1 & & | & | \\ | & | & & \vdots & & | & | \end{pmatrix}$$

Take the directed edges as observations and the nodes as features (for now). So the design matrix becomes $X \doteq D^T$

A trained model may be applied to an incoming edge $\mathbf{x}_e$, incident on nodes $u$ and $v$, as:

$$\log(\frac{p(1)}{1-p(1)}) = \mathbf{x}_e^T \beta = \beta(u) - \beta(v)$$

But why take the edges as observations?

- In a stream, edges are the observations and the associated nodes constitute a tractable feature set...
- in tandem with other "cheap" features such as protocol, port, duration, timestamps, packet count, etc.
- It is difficult to regard edges as features outside of a batch setting...one rarely sees all of node's adjacencies in a stream
- Obvious extensions to heterogeneous data records like netflow

The Newton-Raphson update formula is now:

$$\beta \leftarrow \beta + (DVD^T)^{-1} D(\mathbf{y} - \mathbf{p}) = \beta + \mathcal{L}^{-1}\mathbf{b}$$

with $\mathbf{b}(v) \doteq \sum_{e^+=v} (\mathbf{y}(e) - \mathbf{p}(e)) - \sum_{e^-=v} (\mathbf{y}(e) - \mathbf{p}(e))$

The Newton-Raphson update formula is now:

$$\beta \leftarrow \beta + (DVD^T)^{-1} D(\mathbf{y} - \mathbf{p}) = \beta + \mathcal{L}^{-1}\mathbf{b}$$

with $\mathbf{b}(v) \doteq \sum_{e^+=v} (\mathbf{y}(e) - \mathbf{p}(e)) - \sum_{e^-=v} (\mathbf{y}(e) - \mathbf{p}(e))$

*This is just a Laplacian system with the estimated binomial variances assigned as edge weights!*

So the dynamics at each iteration are analogous to PageRank, centered commute times, et al.

The update to $\beta$ is a weighted sum of the columns of $\mathcal{L}^{-1}$

A node's contribution to $\beta$ is controlled by the residuals $\mathbf{r}(e) \doteq \mathbf{y}(e) - \mathbf{p}(e)$ of its incident edges

For binary $\mathbf{y}(e)$, a small residual corresponds to rank deficiency in the Hessian for nodes on which $e$ is incident

This is equivalent to neglecting contributions from certain spanning forests on 2 components that contain $e$

# A Shocking Admission

**Disclaimer:** The incidence matrix is probably *not* a realistic choice:

- Undesirable cancellation among the entries of

$$\mathbf{b}(v) = \sum_{e^+=v} \mathbf{r}(e) - \sum_{e^-=v} \mathbf{r}(e)$$

  i.e. perhaps one could "miss" desirable updates corresponding to large residuals?
- Should the log-odds of an edge between two equally relevant nodes be penalized by the sign?

# A Shocking Admission

**Disclaimer:** I'm going to talk about it anyway...

The combinatorial interpretation can be recovered by the *unsigned* incidence matrix, in a larger ambient graph

In fact, it will apply to any choice of symmetric, weakly-diagonally dominant kernel

# Generalized Dirichlet Problems

Let $K$ be a diagonal matrix of conductances on the edges. For $c \in C^1$ on the edges, a *generalized Dirichlet problem* is:

$$DKD^T\phi = Dc$$

We borrow terminology from the theory of electrical networks and call $\phi$ the *potential induced by c*

Define a *source of magnitude j with input p and output q* to be the vector $j(\delta_p - \delta_q)$

# The Tree Solution of a Dirichlet Problem

Define a *2-tree* to be a spanning forest on two components

Finally, denote by $(pv||q)$ the set of 2-trees for which $p$ and $v$ lie in the same component, but $q$ lies in the other component

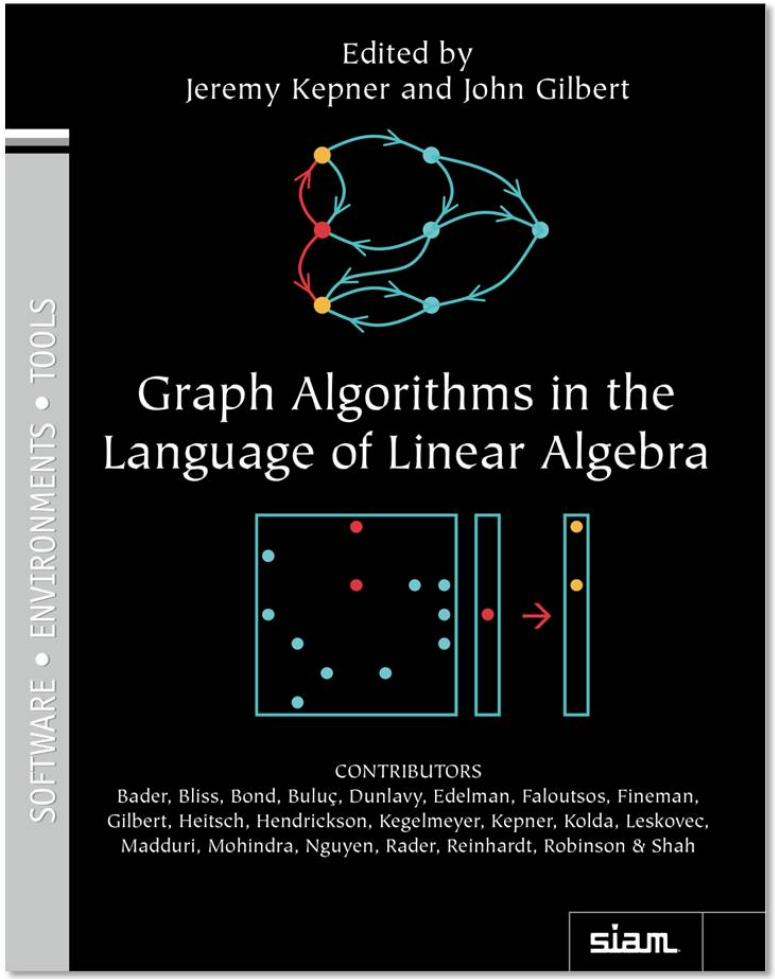
**Theorem**

*[1] Let $c$ be a source of magnitude $j$ with input $p$ and output $q$. Then the potential $\phi$ induced by $c$ for which $\phi(q) = 0$ is given component-wise by:*

$$\phi(v) = (j/\kappa)|(pv||q)|$$

*where $\kappa$ is the tree number of the graph.*

# Mandatory Shameless Plug



# Linear Algebra in the Language of Graphs?

**Corollary**

*Taking the directed edge-incidence matrix as the design matrix, the logistic regression update formula is then:*

$$\beta(v) \leftarrow \beta(v) + \frac{1}{\omega} \sum_{e \in E} [\mathbf{r}(e) \sum_{F \in (e^+ v || e^-)} \prod_{e' \in F} p(e')(1 - p(e'))].$$

i.e. just multiply the binomial variances assigned to the edges among all 2-trees and sum each contribution, weighted by the residuals of the cut edges.

# Linear Algebra in the Language of Graphs?

This is clearly much more expensive than the algebraic Newton update - I do not advocate it as a practical algorithm

However, note that for $p(e)(1 - p(e)) \approx 0$ the contribution from any 2-tree containing $e$ becomes negligble

Could we design an intelligent sampling procedure that would neglect 2-trees containing edges with near zero binomial variance?

Clearly, rank-deficiency due to small binomial variances arises from 2 trees that contribute little to $\beta$

# Gremban's Reduction

The remarks above are specific to the case wherein the Hessian is a graph Laplacian

However, recall that any symmetric, diagonally dominant linear system can be equivalently solved as a Laplacian system

**Theorem**

*(Gremban [9]) Let $M$ be a real, symmetric, weakly-diagonally dominant $n \times n$ matrix. If $M\mathbf{x} = \mathbf{b}$, then there exists a $2n \times 2n$ Laplacian $\mathcal{L}$ such that $\mathcal{L}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ where*

$$\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ -\mathbf{x} \end{bmatrix} \text{ and } \tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix}$$

# Gremban's Reduction

Hence, any linear system over a real, symmetric, diagonally dominant matrix may be equivalently solved as a Laplacian system in a larger ambient graph

Recall that solving the normal equations squares the condition number of the design matrix, and ill-conditioning due to small binomial variances is likely

Given suitably effective preconditioners, it may be advantageous to work with a Laplacian in the ambient space

# References


- N. Biggs. *Algebraic Potential Theory on Graphs*, Bull. London Math. Soc. 29 (1997) 641-682.
- M. Brand. *A Random Walks Perspective on Maximizing Satisfaction and Profit*, Proceedings of the 2005 SIAM International Conference on Data Mining, 2005.
- A. Buluç and J. R. Gilbert. *The Combinatorial BLAS: Design, Implementation, and Applications*, The International Journal of High Performance Computing Applications, 2011.
- A.K. Chandra, P. Raghavan, W.L. Ruzzo, R. Smolensky, and P. Tiwari. *The Electrical Resistance of a Graph Captures Its Commute and Cover Times*, ACM STOC, 1989. p. 574-586.

# References


- F. Chung and S.T. Yau. *Discrete Green's Functions*, Journal of Combinatorial Theory, Series A **91**, 191-214 (2000).
- F. Chung and W. Zhao *PageRank and Random Walks on Graphs*
- C. Ding, R. Jin, T. Li, and H.D. Simon. *A Learning Framework Using Green's Function and Kernel Regularization with Application to Recommender System*. ACM, 2007.
- F. Fouss, J.M. Renders, A. Pirotte, and M. Saerens. *Random-Walk Computations of Similarities between Nodes of a Graph with Applications to Collaborative Filtering*, IEEE Transactions on Knowledge and Data Engineering, Volume 19, 2007.

# References


- K. Gremban. *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*, PhD Thesis, Carnegie Mellon University, CMU-CS-96-123, 1996.
- V.E. Henson, G. Sanders, and J. McCloskey. Private communication.
- P. Komarek. *Logistic Regression for Data Mining and High-Dimensional Classification*, TR-04-34, Department of Mathematical Sciences, Carnegie Mellon University (dissertation).
- L. Lovász. *Random Walks on Graphs: a Survey*. In *Combinatorics, Paul Erdös Is Eighty*, Bolyai Soc. Math. Stud., pp. 353-397. János Bolyai Math. Soc., Budapest, 1993.

# References


- U. von Luxburg and A. Radl and M. Hein: *Getting Lost in Space: Large Sample Analysis of the Commute Distance*, Neural Information Processing Systems (NIPS), 2010.
- M. O'Hara. Private communication.
- P.O. Perry and M.W. Mahoney. *Regularized Laplacian Estimation and Fast Eigenvector Approximation*, in Proceedings of CoRR, 2011.
- C.R. Rao, S.K. Mitra. *Generalized Inverses of Matrices and its Applications*, John Wiley and Sons, 1971.

# References


- M. Saerens, F. Fouss, L. Yen, and P. DuPont. *The Principal Components Analysis of a Graph, and Its Relationships to Spectral Clustering*, Proc. 15th European Conf. Machine Learning (ECML '04), pp.371-383, 2004.
- H. Tong. *Random Walk With Restart and Its Applications*, Proceedings of the ICDM 2006.
- L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2004.
- P. Cheboratev and E. Shamis. *The Matrix-Forest Theorem and Measuring Relations in Small Social Groups*, Automation and Remote Control, 58(9): 1505-1514, 1997.