Locally-biased and semi-supervised eigenvectors

Michael W. Mahoney

ICSI and Dept of Statistics, UC Berkeley

(For more info, see:
http:// www.stat.berkeley.edu/~mmahoney/
or Google on "Michael Mahoney")

Locally-biased analytics

You have **BIG** data and want to analyze a small part of it:

Solution 1:

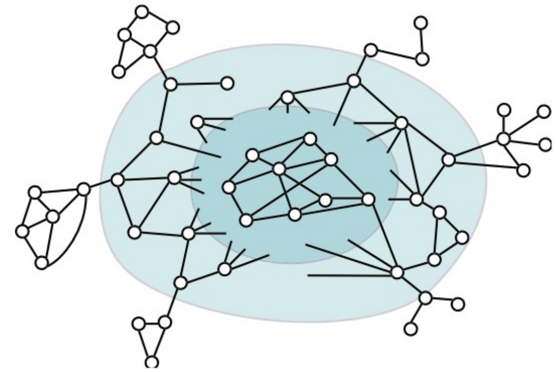
- Cut out small part and use traditional methods
- Challenge: cutting out may be difficult a priori

Solution 2:

- Develop locally-biased methods for data-analysis
- Challenge: Most data analysis tools (implicitly or explicitly) make strong local-global assumptions*

*spectral partitioning "wants" to find 50-50 clusters; recursive partitioning is of interest if recursion depth isn't too deep; eigenvectors optimize global objectives, etc.

Locally-biased analytics



Locally-biased community identification:

- Find a “community” around an exogenously-specified seed node

Locally-biased image segmentation:

- Find a small tiger in the middle of a big picture

Locally-biased neural connectivity analysis:

- Find neurons that are temporally-correlated with local stimulus

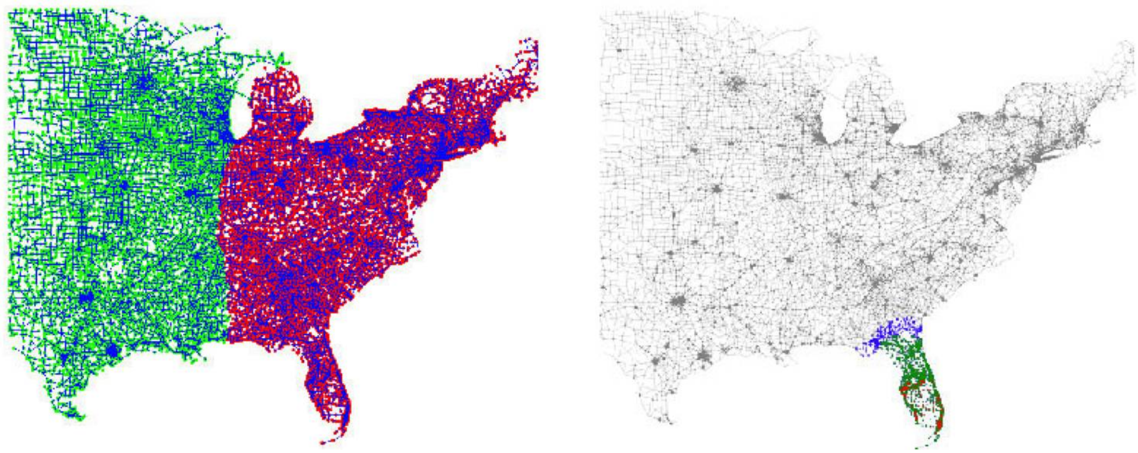
Locally-biased inference, semi-supervised learning, etc.:

- Do machine learning with a “seed set” of ground truth nodes, i.e., make predictions that “draws strength” based on local information

Global spectral methods DO work well

(1) Construct a graph from the data

(2) Use the second eigenvalue/eigenvector of Laplacian: do clustering, community detection, image segmentation, parallel computing, semi-supervised/transductive learning, etc.



Why is it useful?

(*) Connections with random walks and sparse cuts

(*) Isoperimetric structure gives controls on capacity/inference

(*) Relatively easy to compute

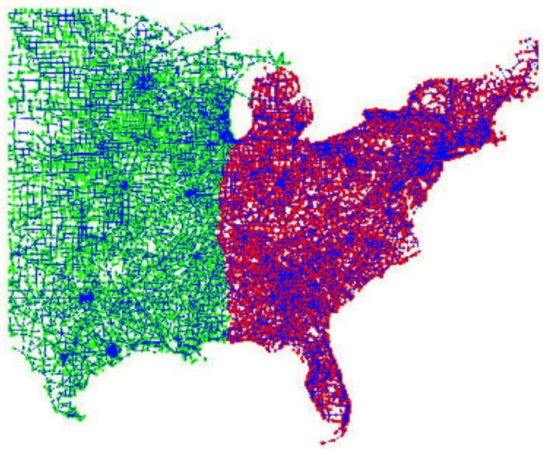
Global spectral methods DON'T work well

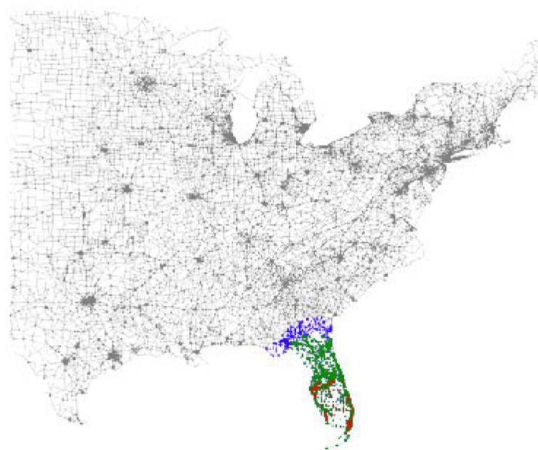
(1) Leading nontrivial eigenvalue/eigenvector are inherently global quantities

(2) May NOT be sensitive to local information:

(*) Sparse cuts may be poorly correlated with second/all eigenvectors

(*) Interesting local region may be hidden to global eigenvectors that are dominated by exact orthogonality constraint.





QUES: Can we find a locally-biased analogue of the usual global eigenvectors that comes with the good properties of the global eigenvectors?

(*) Connections with random walks and sparse cuts

(*) This gives controls on capacity/inference

(*) Relatively easy to compute

Outline

Locally-biased eigenvectors

- A methodology to construct a locally-biased analogue of leading nontrivial eigenvector of graph Laplacian

Implicit regularization …

- … in early-stopped iterations and teleported PageRank computations

Semi-supervised eigenvectors

- Extend locally-biased eigenvectors to compute multiple locally-biased eigenvectors, i.e., locally-biased SPSD kernels

Implicit regularization …

- … in truncated diffusions and push-based approximations to PageRank
- … connections to strongly-local spectral methods and scalable computation

Outline

Locally-biased eigenvectors

- A methodology to construct a locally-biased analogue of leading nontrivial eigenvector of graph Laplacian

Implicit regularization ...

- ... in early-stopped iterations and teleported PageRank computations

Semi-supervised eigenvectors

- Extend locally-biased eigenvectors to compute multiple locally-biased eigenvectors, i.e., locally-biased SPSD kernels

Implicit regularization ...

- ... in truncated diffusions and push-based approximations to PageRank
- ... connections to strongly-local spectral methods and scalable computation

Recall spectral graph partitioning

The basic optimization problem:

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t.} \quad & \langle x, x \rangle_D = 1 \\ & \langle x, 1 \rangle_D = 0 \end{aligned}$$

- Relaxation of:

$$\phi(G) = \min_{S \subset V} \frac{E(S, \bar{S})}{Vol(S)Vol(\bar{S})}$$

- Solvable via the eigenvalue problem:

$$\mathcal{L}_G y = \lambda_2(G) y$$

- Sweep cut of second eigenvector yields:

$$\lambda_2(G)/2 \leq \phi(G) \leq \sqrt{8\lambda_2(G)}$$

Geometric correlation and generalized PageRank vectors

Given a cut T, define the vector:

$$s_T := \sqrt{\frac{\mathrm{vol}(T)\mathrm{vol}(\bar{T})}{2m}} \left(\frac{1_T}{\mathrm{vol}(T)} - \frac{1_{\bar{T}}}{\mathrm{vol}(\bar{T})} \right)$$

Can use this to define a geometric notion of correlation between cuts:

$$< s_T, 1 >_D = 0$$

$$< s_T, s_T >_D = 1$$

$$< s_T, s_U >_D = K(T, U)$$

Defn. Given a graph $G = (V, E)$, a number $\alpha \in (-\infty, \lambda_2(G))$ and any vector $s \in R^n$, $s \perp_D 1$, a ***Generalized Personalized PageRank (GPPR)*** vector is any vector of the form

$$p_{\alpha,s} := (L_G - \alpha L_{K_n})^+ Ds.$$

Local spectral partitioning *ansatz*

Mahoney, Orecchia, and Vishnoi (2010)

Primal program:

$$\begin{array}{ll} \text{minimize} & x^T L_G x \\ \text{s.t.} & < x, x >_D = 1 \\ & < x, s >_D^2 \geq \kappa \end{array}$$

Dual program:

$$\begin{array}{ll} \max & \alpha - \beta(1-\kappa) \\ \text{s.t.} & L_G \succeq \alpha L_{K_n} - \beta \left(\frac{L_{K_T}}{\text{vol}(\bar{T})} + \frac{L_{K_{\bar{T}}}}{\text{vol}(T)} \right) \\ & \beta \geq 0 \end{array}$$

Interpretation:

- Find a cut well-correlated with the seed vector *s*.
- If *s* is a single node, this relaxes:

$$\min_{S \subset V, s \in S, |S| \leq 1/k} \frac{E(S, \bar{S})}{Vol(S) Vol(\bar{S})}$$

Interpretation:

- Embedding a combination of scaled complete graph K_n and complete graphs T and $\underline{T}$ (K_T and $K_{\underline{T}}$) - where the latter encourage cuts near (T,$\underline{T}$).

Main results (1 of 2)

Mahoney, Orecchia, and Vishnoi (2010)

Theorem: If x^* is an optimal solution to LocalSpectral, it is a GPPR vector for parameter α, and it can be computed as the solution to a set of linear equations.

Proof:

(1) Relax non-convex problem to convex SDP

(2) Strong duality holds for this SDP

(3) Solution to SDP is rank one (from comp. slack.)

(4) Rank one solution is GPPR vector.

Main results (2 of 2)

Mahoney, Orecchia, and Vishnoi (2010)

Theorem: If x^* is optimal solution to LocalSpect (G,s,κ), one can find a cut of conductance $\leq 8\lambda(G,s,\kappa)$ in time $O(n \lg n)$ with sweep cut of x^*.

Upper bound, as usual from sweep cut & Cheeger.

Theorem: Let s be seed vector and κ correlation parameter. For all sets of nodes T s.t. $\kappa' := \langle s, s_T \rangle_D^2$, we have: $\phi(T) \geq \lambda(G,s,\kappa)$ if $\kappa \leq \kappa'$, and $\phi(T) \geq (\kappa'/\kappa)\lambda(G,s,\kappa)$ if $\kappa' \leq \kappa$.

Lower bound: Spectral version of flow-improvement algs.

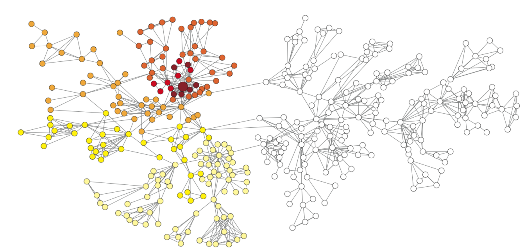
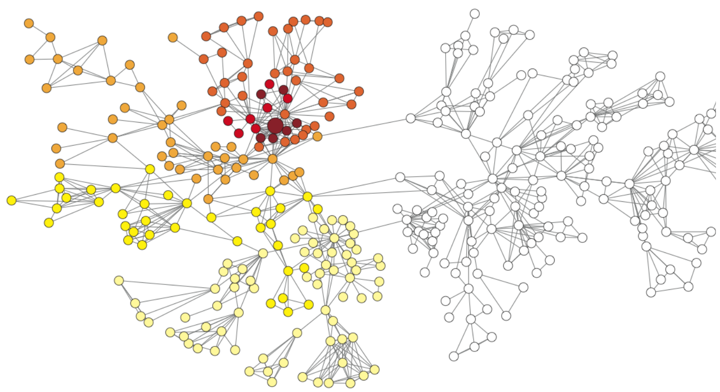
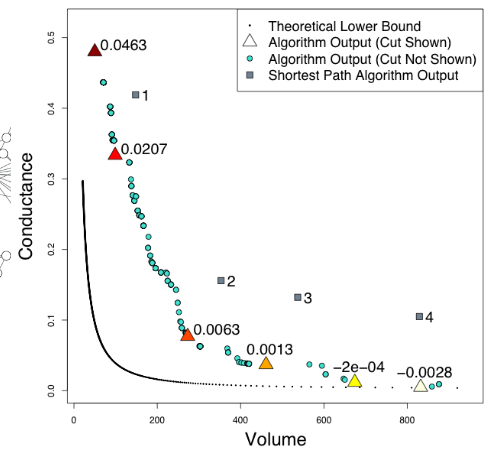
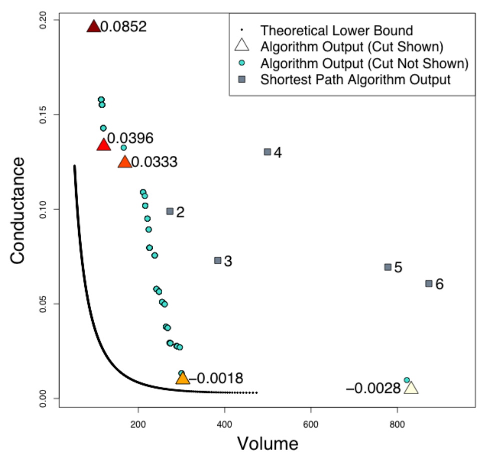


Illustration on small graphs



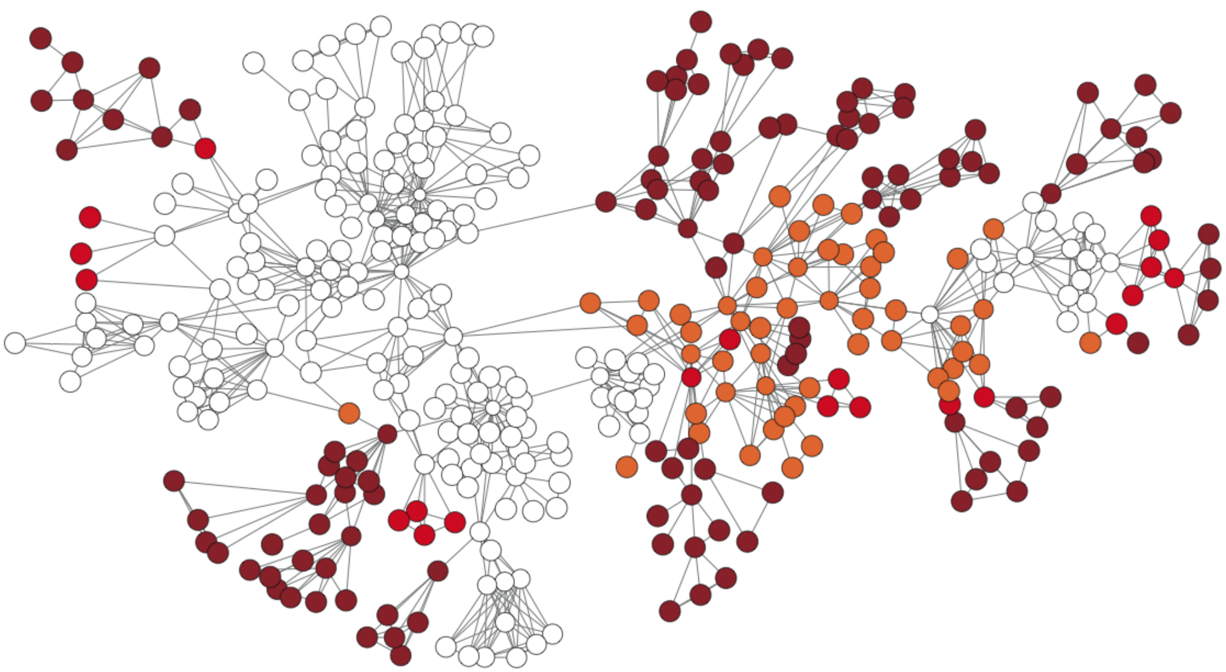




- Similar results if we do local random walks, truncated PageRank, and heat kernel diffusions.
- Linear equation formulation is more "powerful" than diffusions
- I.e., can access all $\alpha \in (-\infty, \lambda_2(G))$ parameter values

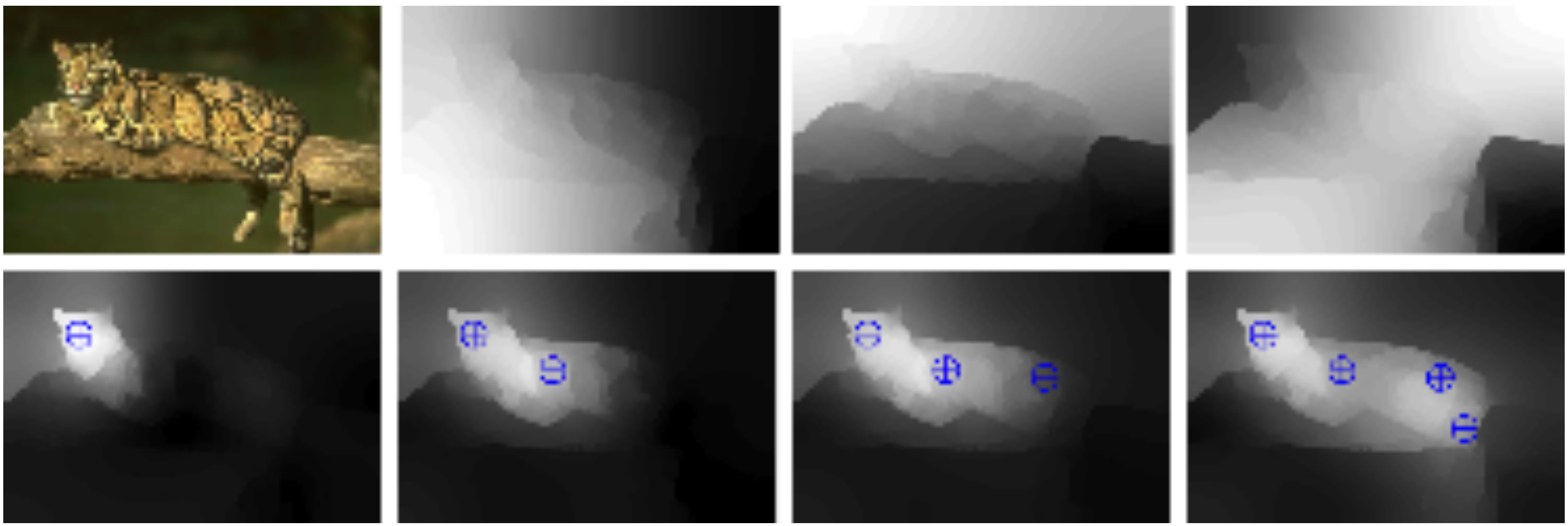
Illustration with general seeds

- Seed vector doesn't need to correspond to cuts.
- It could be any vector on the nodes, e.g., can find a cut "near" low-degree vertices with $s_i = -(d_i - d_{av})$, $i \varepsilon [n]$.



New methods are useful more generally

Maji, Vishnoi,and Malik (2011) applied Mahoney, Orecchia, and Vishnoi (2010)



- Cannot find the tiger with global eigenvectors.
- Can find the tiger with the LocalSpectral method!

Outline

Locally-biased eigenvectors

- *A* methodology to construct a locally-biased analogue of leading nontrivial eigenvector of graph Laplacian

Implicit regularization ...

- ... in early-stopped iterations and teleported PageRank computations

Semi-supervised eigenvectors

- Extend locally-biased eigenvectors to compute multiple locally-biased eigenvectors, i.e., locally-biased SPSD kernels

Implicit regularization ...

- ... in truncated diffusions and push-based approximations to PageRank
- ... connections to strongly-local spectral methods and scalable computation

PageRank and implicit regularization

Recall the usual characterization of PPR:

$$\pi(\gamma, s) = \gamma s + (1-\gamma) M \pi(\gamma, s)$$

$$R_\gamma = \gamma \left(I - (1-\gamma) M\right)^{-1}$$

Compare with our definition of GPPR:

Defn. Given a graph $G = (V, E)$, a number $\alpha \in (-\infty, \lambda_2(G))$ and any vector $s \in R^n$, $s \perp_D 1$, a ***Generalized Personalized PageRank (GPPR)*** vector is any vector of the form

$$p_{\alpha,s} := \left(L_G - \alpha L_{K_n}\right)^+ Ds.$$

Question: Can we formalize that PageRank is a regularized version of leading nontrivial eigenvector of the Laplacian?

Two versions of spectral partitioning

VP:

$$\begin{aligned} \text{min.} \quad & x^T L_G x \\ \text{s.t.} \quad & x^T L_{K_n} x = 1 \\ & < x, 1 >_D = 0 \end{aligned}$$

↓

R-VP:

$$\begin{aligned} \text{min.} \quad & x^T L_G x + \lambda f(x) \\ \text{s.t.} \quad & constraints \end{aligned}$$

Two versions of spectral partitioning

VP: ⟷ **SDP:**

$$\begin{aligned} \text{min.} \quad & x^T L_G x \\ \text{s.t.} \quad & x^T L_{K_n} x = 1 \\ & < x, 1 >_D = 0 \end{aligned}$$

$$\begin{aligned} \text{min.} \quad & L_G \circ X \\ \text{s.t.} \quad & L_{K_n} \circ X = 1 \\ & X \succeq 0 \end{aligned}$$

↓ ↓

R-VP:

$$\begin{aligned} \text{min.} \quad & x^T L_G x + \lambda f(x) \\ \text{s.t.} \quad & constraints \end{aligned}$$

R-SDP:

$$\begin{aligned} \text{min.} \quad & L_G \circ X + \lambda F(X) \\ \text{s.t.} \quad & constraints \end{aligned}$$

A simple theorem

Mahoney and Orecchia (2010)

$$\text{(F,}\eta\text{)-SDP} \quad \min \quad L \bullet X + \frac{1}{\eta} \cdot F(X)$$
$$\text{s.t.} \quad I \bullet X = 1$$
$$X \succeq 0$$

Modification of the usual SDP form of spectral to have regularization (but, on the matrix X, not the vector x).

Theorem: Let G be a connected, weighted, undirected graph, with normalized Laplacian L. Then, the following conditions are sufficient for $X^\star$ to be an optimal solution to (F,η)-SDP.

- $X^\star = (\nabla F)^{-1}(\eta \cdot (\lambda^* I - L))$, for some $\lambda^* \in R$,
- $I \bullet X^\star = 1$,
- $X^\star \succeq 0$.

Corollary

If $F_D(X)$ = -logdet(X) (i.e., Log-determinant), then this

gives scaled PageRank matrix, with $t \sim \eta$

I.e., PageRank does two things:

- It ***approximately*** computes the Fiedler vector.
- It ***exactly*** computes a regularized version of the Fiedler vector ***implicitly***!

(Similarly, generalized entropy regularization *implicit* in Heat Kernel computations; & matrix p-norm regularization *implicit* in power iteration.)

Outline

Locally-biased eigenvectors

- A methodology to construct a locally-biased analogue of leading nontrivial eigenvector of graph Laplacian

Implicit regularization …

- … in early-stopped iterations and teleported PageRank computations

Semi-supervised eigenvectors

- Extend locally-biased eigenvectors to compute multiple locally-biased eigenvectors, i.e., locally-biased SPSD kernels

Implicit regularization …

- … in truncated diffusions and push-based approximations to PageRank
- … connections to strongly-local spectral methods and scalable computation

Semi-supervised eigenvectors

Hansen and Mahoney (NIPS 2013, JMLR 2014)

Eigenvectors are inherently global quantities, and the leading ones may therefore fail at modeling relevant local structures.

GlobalSpectral

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t} \quad & x^T D_G x = 1 \\ & x^T D_G 1 = 0 \end{aligned}$$

Generalized eigenvalue problem. Solution is given by the second smallest eigenvector, and yields a "Normalized Cut".

LocalSpectral

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t} \quad & x^T D_G x = 1 \\ & x^T D_G 1 = 0 \\ & x^T D_G s \geq \sqrt{\kappa} \end{aligned}$$

Locally-biased analogue of the second smallest eigenvector. Optimal solution is a generalization of Personalized PageRank and can be computed in nearly-linear time [MOV2012].

Generalized LocalSpectral

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t} \quad & x^T D_G x = 1 \\ & x^T D_G X = 0 \\ & x^T D_G s \geq \sqrt{\kappa} \end{aligned}$$

Semi-supervised eigenvector generalization of [HM2013]. This objective incorporates a general orthogonality constraint, allowing us to compute a sequence of "localized eigenvectors".

Semi-supervised eigenvectors are efficient to compute and inherit many of the nice properties that characterizes global eigenvectors of a graph.

Semi-supervised eigenvectors

Hansen and Mahoney (NIPS 2013, JMLR 2014)

This interpolates between very localized solutions and the global eigenvectors of the graph Laplacian.
- For κ=0, this is the usual global generalized eigenvalue problem.
- For κ=1, this returns the local seed set.

GENERALIZED LOCALSPECTRAL

$$\begin{aligned} \text{minimize} \quad & x^T L_G x \\ \text{s.t.} \quad & x^T D_G x = 1 \quad \leftarrow \text{Norm constraint} \\ & x^T D_G X = 0 \quad \leftarrow \text{Orthogonality constraint} \\ & x^T D_G s \ge \sqrt{\kappa} \quad \leftarrow \text{Locality constraint} \end{aligned}$$

For γ<0 , one we can compute the first semi-supervised eigenvectors using local graph diffusions, *i.e.*, personalized PageRank.
- Approximate the solution using the Push algorithm [ACL06].
- Implicit regularization characterization by [M010] & [GM14].

Leading solution

Seed vector

$$x_1^* = c\,(L_G - \gamma_1 D_G)^+ \, D_G s$$

Projection operator

$$x^* \propto (FF^T(L_G - \gamma D_G)FF^T)^+ FF^T D_G s$$

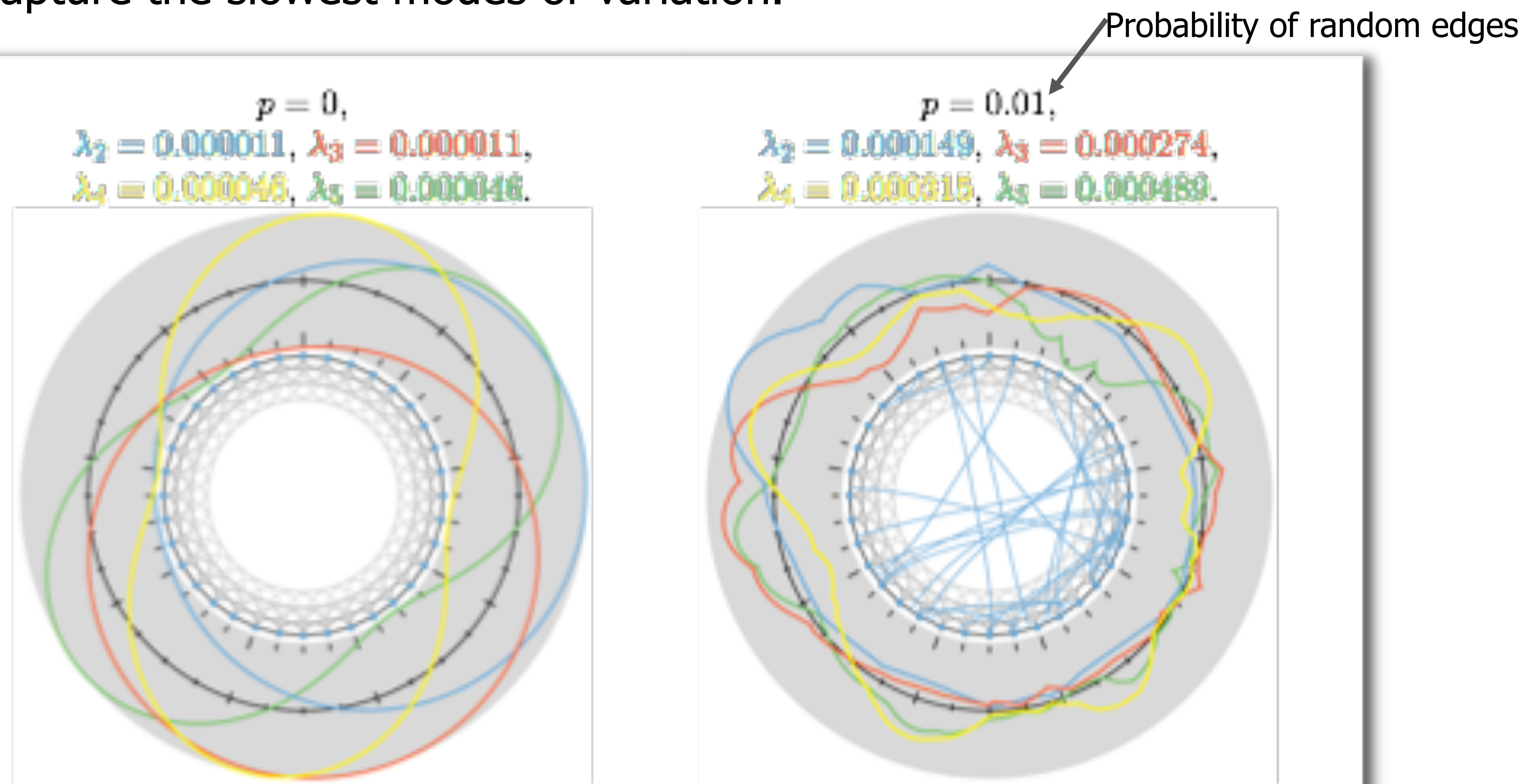
General solution

Determines the locality of the solution.

Convex for $\gamma \in (-\infty, \lambda_2(G))$

Semi-supervised eigenvectors

Small-world example - The eigenvectors having smallest eigenvalues capture the slowest modes of variation.

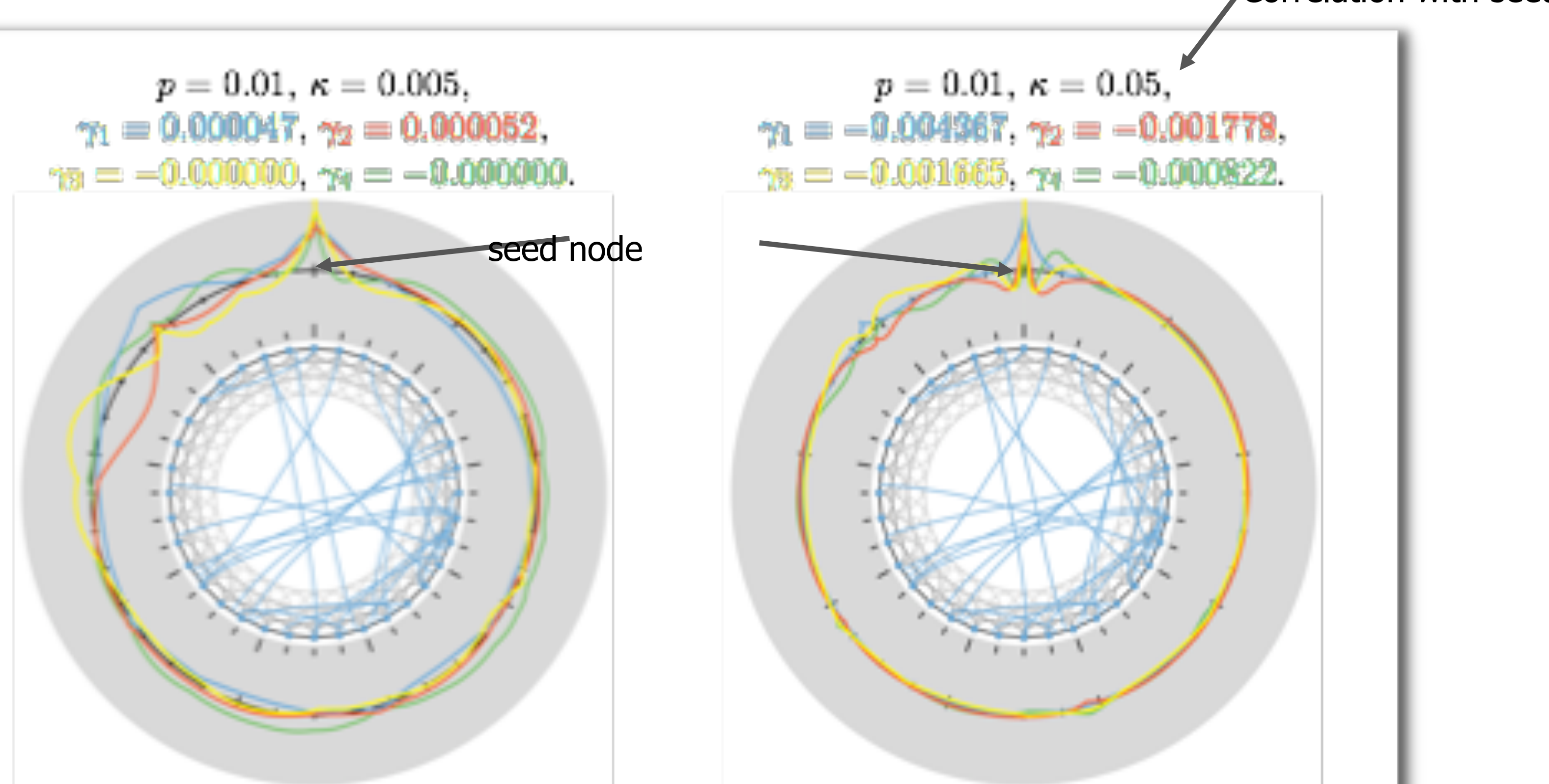


Global eigenvectors

Global eigenvectors

Semi-supervised eigenvectors

Small-world example - The eigenvectors having smallest eigenvalues capture the slowest modes of variation.



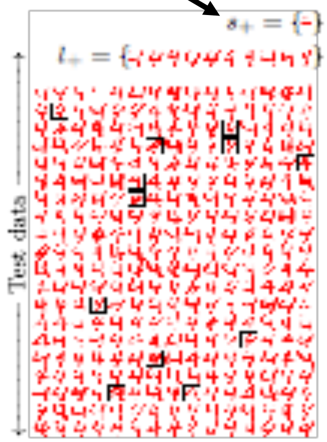
Semi-supervised eigenvectors
Low correlation

Semi-supervised eigenvectors
High correlation

Semi-supervised eigenvectors

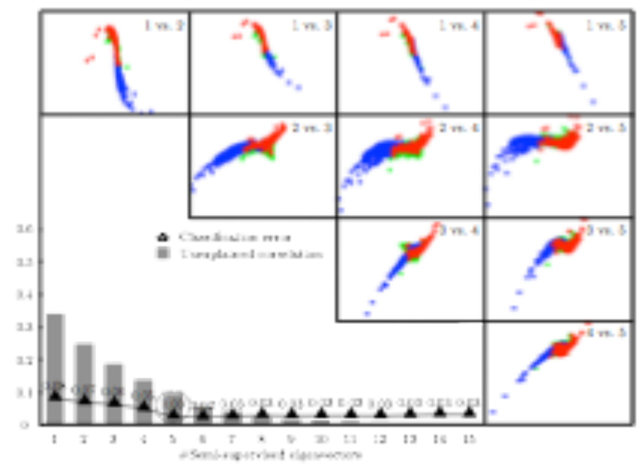
Hansen and Mahoney (NIPS 2013, JMLR 2014)

One seed per class



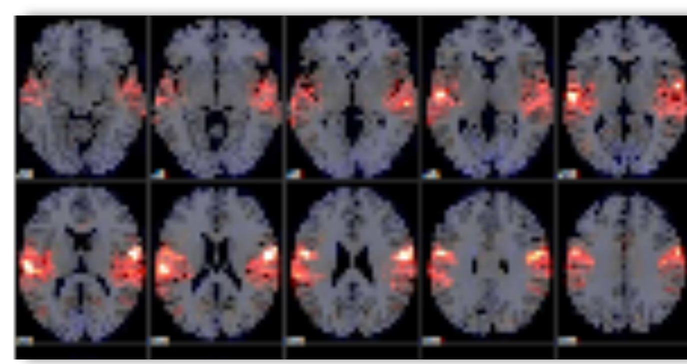
Ten labeled samples per class used in a downstream classifier

Semi-supervised eigenvector scatter plots



Many "real" applications:

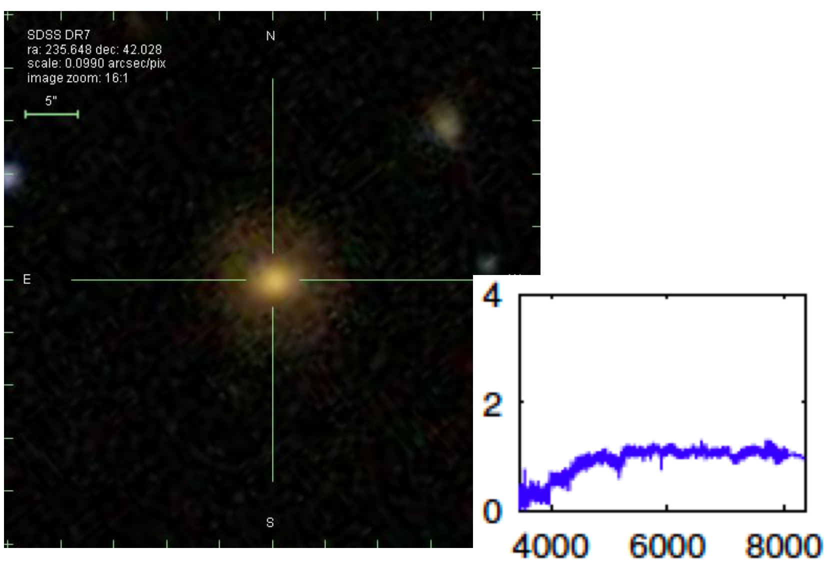
- A spatially guided "searchlight" technique that compared to [Kriegeskorte2006] account for spatially distributed signal representations.
- Large/small-scale structure in DNA SNP data in population genetics
- Local structure in astronomical data
- Code is available at: https://sites.google.com/site/tokejansenhansen/



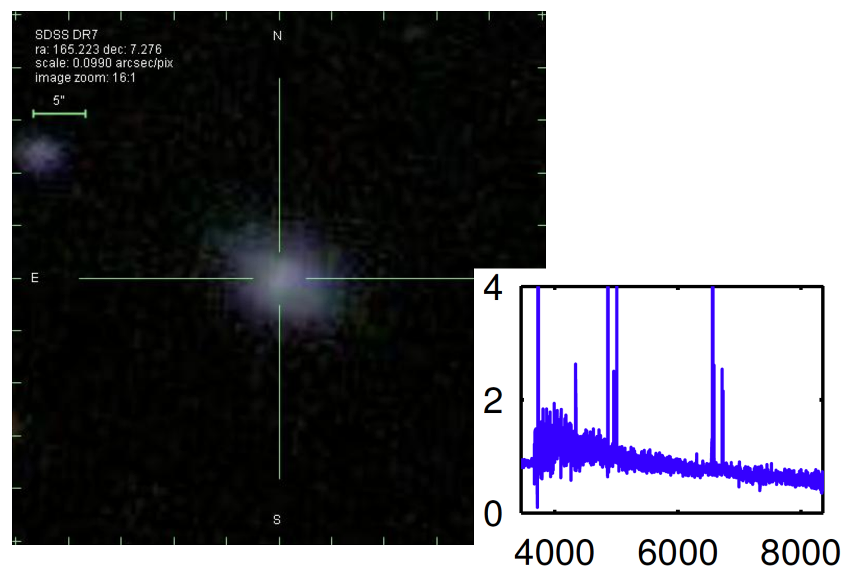
Local structure in SDSS spectra

Lawlor, Budavari, and Mahoney (2014)

- Data: x ε R^{3841}, N ≈ 500k are photon fluxes in ≈ 10 Å bins
- preprocessing corrects for redshift, gappy regions
- normalized by median flux at certain wavelengths



Red galaxy

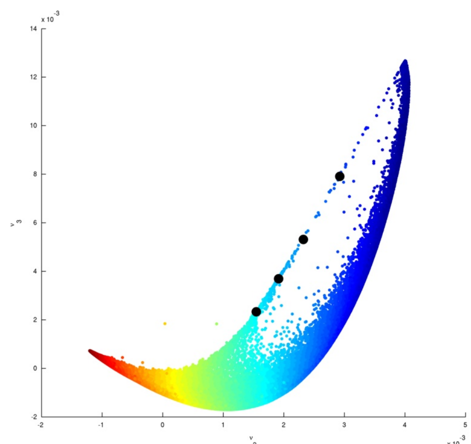


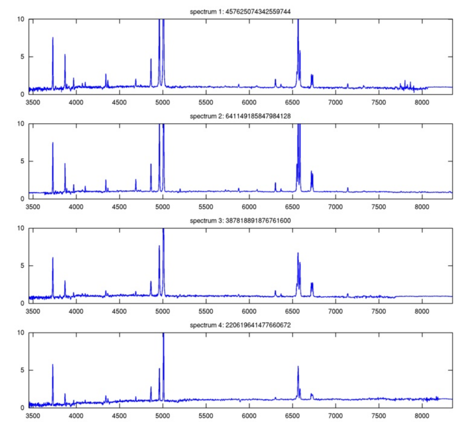
Blue galaxy

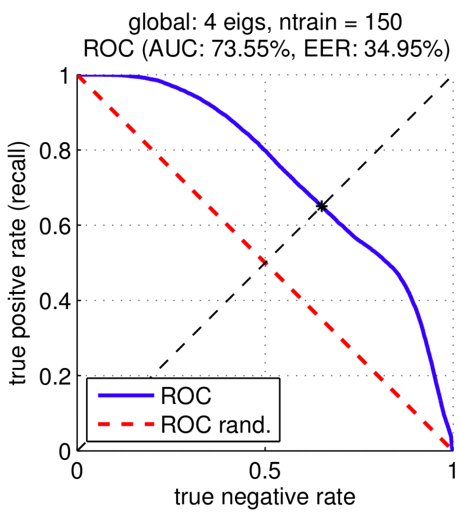
Local structure in SDSS spectra

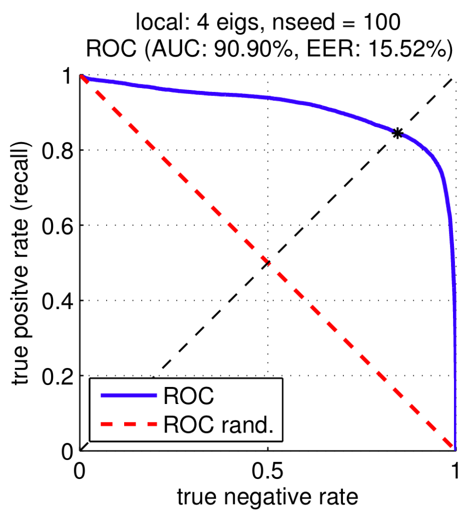
Lawlor, Budavari, and Mahoney (2014)

Galaxies along bridge & bridge spectra









ROC curves for classifying AGN spectra on top four global eigenvectors (left) ; and (right) top four semi-supervised eigenvectors.

Outline

Locally-biased eigenvectors

- *A methodology to construct a locally-biased analogue of leading nontrivial eigenvector of graph Laplacian*

Implicit regularization ...

- ... in early-stopped iterations and teleported PageRank computations

Semi-supervised eigenvectors

- Extend locally-biased eigenvectors to compute multiple locally-biased eigenvectors, i.e., locally-biased SPSD kernels

Implicit regularization ...

- ... in truncated diffusions and push-based approximations to PageRank
- ... connections to strongly-local spectral methods and scalable computation

Push Algorithm for PageRank

The Push Method
τ, ρ

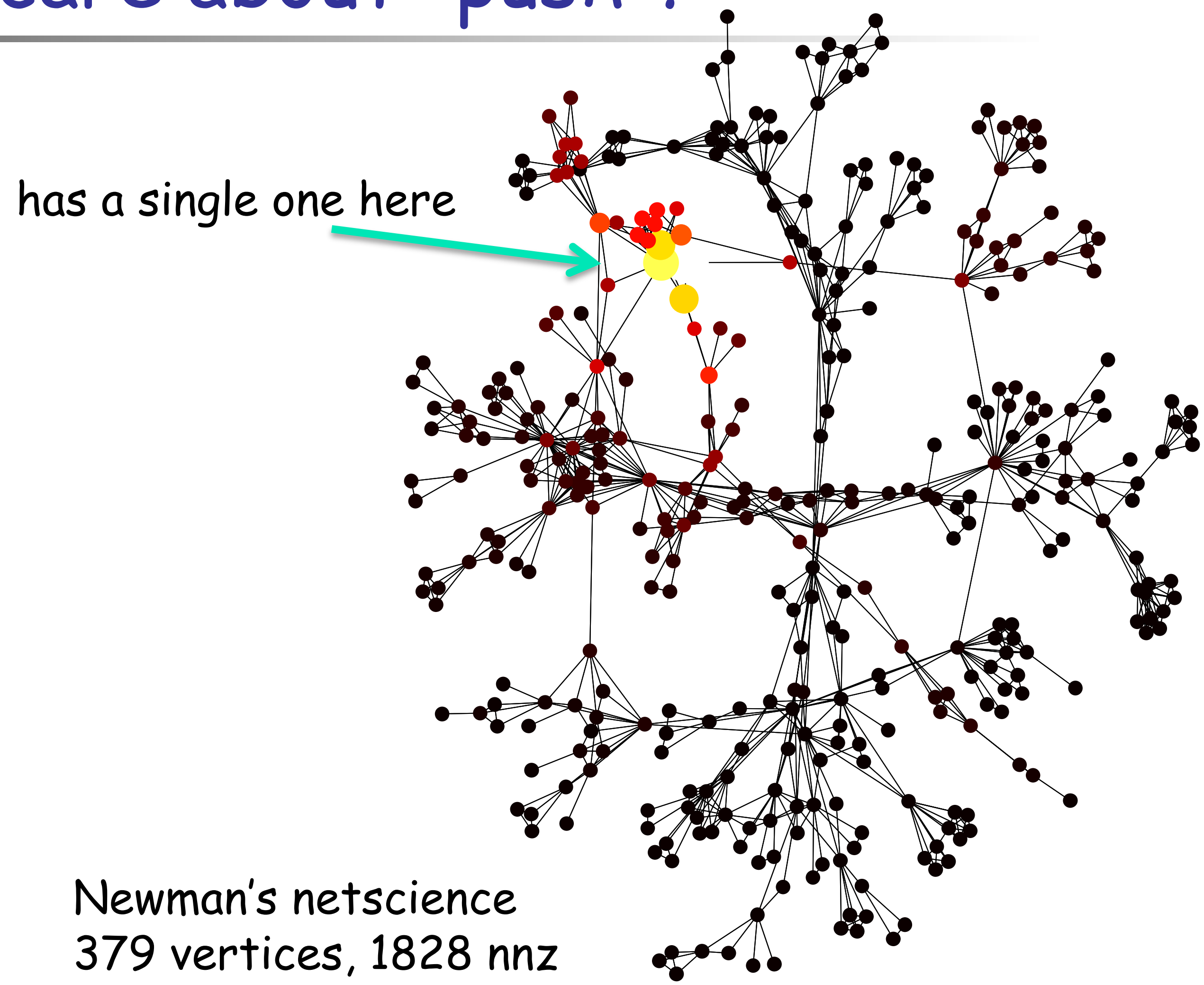
1. $\mathbf{x}^{(1)} = 0, \mathbf{r}^{(1)} = (1-\beta)\mathbf{e}_i, k = 1$
2. *while* any $r_j > \tau d_j$ (d_j *is the degree of node* j)
3. $\quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + (r_j - \tau d_j \rho)\mathbf{e}_j$
4. $$\mathbf{r}_i^{(k+1)} = \begin{cases} \tau d_j \rho & i = j \\ r_i^{(k)} + \beta(r_j - \tau d_j \rho)/d_j & i \sim j \\ r_i^{(k)} & \text{otherwise} \end{cases}$$
5. $\quad k \leftarrow k + 1$

- Proposed (a variant) in ACL06 (also M0x, JW03) for *Personalized PageRank*
- Strongly related to Gauss-Seidel (see Gleich's talk at Simons for this)
- Derived to show improved runtime for balanced solvers
- Applied to graphs with 10M+nodes and 1B+edges

Why do we care about “push”?

1. Widely-used for empirical studies of “communities”
2. Used for “fast PageRank” approximation

- Produces *sparse* approximations to PageRank!
- Why does the “push method” have such empirical utility?



Newman’s netscience
379 vertices, 1828 nnz
“zero” on most of the nodes

How might an algorithm be good?

Two ways this algorithm might be good.

- Theorem 1. [ACL06] The ACL push procedure returns a vector that is ε-worst than the exact PPR and much faster.
- Theorem 2. [GM14] The ACL push procedure returns a vector that exactly solves an L1-regulairzed version of the PPR objective.

I.e., the Push Method does two things:

- It ***approximately*** computes the PPR vector.
- It ***exactly*** computes a regularized version of the PPR vector ***implicitly***!

The s-t min-cut problem

Unweighted incidence matrix

Diagonal capacity matrix

minimize $\|\mathbf{Bx}\|_{C,1} = \sum_{ij \in E} C_{i,j} |x_i - x_j|$

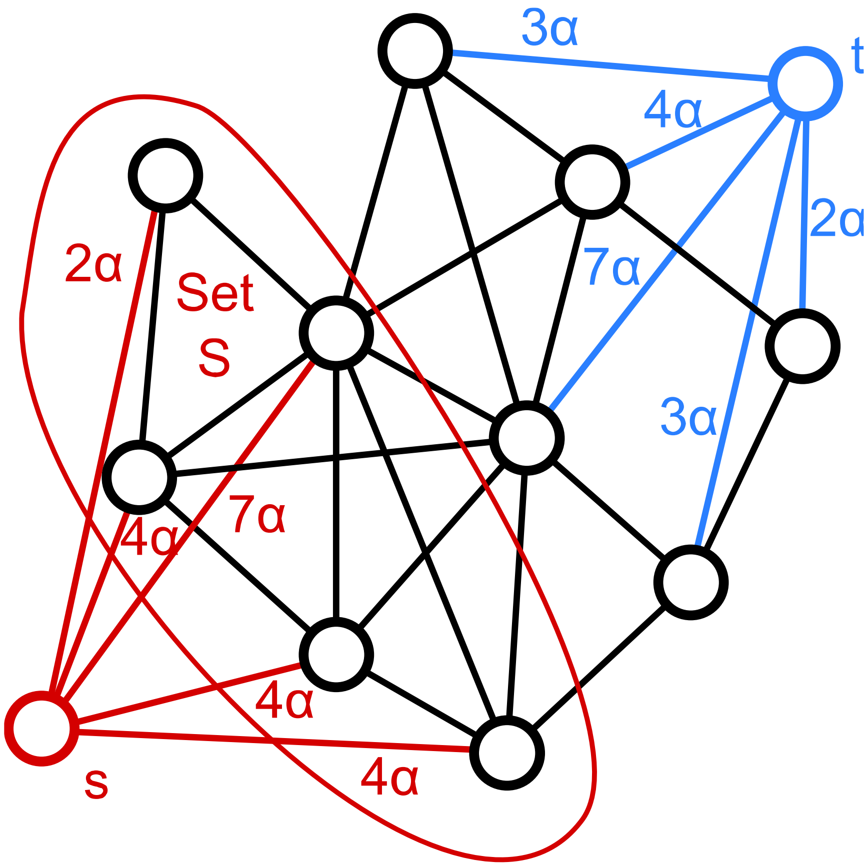
subject to $x_s = 1, x_t = 0, \mathbf{x} \geq 0.$

- Consider L2 variants of this objective & show how the Push Method and other diffusion-based ML algorithms implicitly regularize.

$$A_S = \begin{bmatrix} 0 & \alpha \mathbf{d}_S^T & 0 \\ \alpha \mathbf{d}_S & \mathbf{A} & \alpha \mathbf{d}_{\bar{S}} \\ 0 & \alpha \mathbf{d}_{\bar{S}}^T & 0 \end{bmatrix}$$

The localized cut graph

Gleich and Mahoney (2014)



Connect s to vertices in S with weight $\alpha \cdot$ degree
Connect t to vertices in $\bar{S}$ with weight $\alpha \cdot$ degree

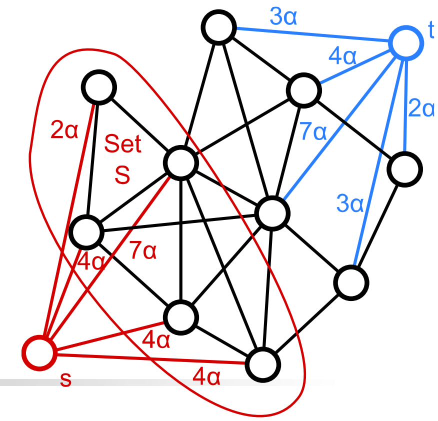
$$\mathbf{B}_S = \begin{bmatrix} \mathbf{e} & -\mathbf{I}_S & 0 \\ 0 & \mathbf{B} & 0 \\ 0 & -\mathbf{I}_{\bar{S}} & \mathbf{e} \end{bmatrix}$$

Solve the s-t min-cut

$$\text{minimize} \quad \|\mathbf{B}_S \mathbf{x}\|_{C(\alpha),1}$$
$$\text{subject to} \quad x_s = 1, x_t = 0$$
$$\mathbf{x} \geq 0.$$

s-t min-cut -> PageRank

Gleich and Mahoney (2014)



The PageRank vector **z** that solves

$$(\alpha \mathbf{D} + \mathbf{L})\mathbf{z} = \alpha \mathbf{v}$$

with $\mathbf{v} = \mathbf{d}_S / \mathrm{vol}(S)$ is a renormalized solution of the electrical cut computation:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{B}_S \mathbf{x}\|_{C(\alpha),2} \\ \text{subject to} & x_s = 1, x_t = 0. \end{array}$$

L1->L2 changes s-t min-cut to "electrical flow" s-t min-cut

Specifically, if **x** is the solution, then

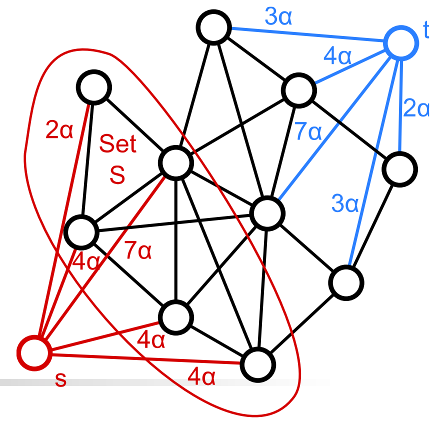
$$\mathbf{x} = \begin{bmatrix} 1 \\ \mathrm{vol}(S)\mathbf{z} \\ 0 \end{bmatrix}$$

Proof

Square and expand the objective into a Laplacian, then apply constraints.

Back to the push method

Gleich and Mahoney (2014)



Let $\mathbf{x}$ be the output from the push method with $0 < \beta < 1$, $\mathbf{v} = \mathbf{d}_S / \mathrm{vol}(S)$, $\rho = 1$, and $\tau > 0$.

Set $\alpha = \frac{1-\beta}{\beta}$, $\kappa = \tau \mathrm{vol}(S)/\beta$, and let $\mathbf{z}_G$ solve:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\|\mathbf{B}_S \mathbf{z}\|^2_{C(\alpha),2} + \kappa \|\mathbf{D}\mathbf{z}\|_1 \\ \text{subject to} \quad & z_s = 1, z_t = 0, \mathbf{z} \geq 0 \end{aligned}$$

Need for normalization

L1 regularization for sparsity

where $\mathbf{z} = \begin{bmatrix} 1 \\ \mathbf{z}_G \\ 0 \end{bmatrix}$.

Then $\mathbf{x} = \mathbf{D}\mathbf{z}_G / \mathrm{vol}(S)$.

Proof Write out KKT conditions
Show that the push method solves them. Slackness was "tricky"

Conclusions

Locally-biased and semi-supervised eigenvectors

- Local versions of the usual global eigenvectors that come with the good properties of global eigenvectors
- Strong algorithmic and statistical theory & good initial results in several applications

Novel connections between approximate computation and implicit regularization

Special cases already scaled up to LARGE data