Communities

...............................

Block model

Optimization

Random

Mining

subspaces

Evamples

Conclusions

Communities, Spectral Clustering, and Random Walks

David Bindel

Department of Computer Science Cornell University

4 Nov 2011

Communities

Introduction

Diagle madel

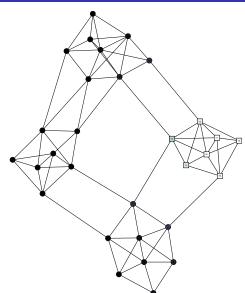
Ontimization

Random

Mining

_. '

Examples



Basic setting

Communities

Introduction

Optimization

Optimizatio

walks

Mining subspace

Ritz vector

Examples

Informal: Community = "unusually tight" node group?

Formal: Graph G = (V, E), seek subgraph G' = (V', E'):

- By model fitting
- By optimization of some metric
- 3 By random walks on G

Unified by linear algebra!

Plan for today

Communities

Introduction

Block mode

Optimization

Random

Mining subspace

Ritz vectors

Examples

- Three routes to an invariant subspace
- How to mine a subspace for information
- From eigenvectors to Ritz vectors
- Some examples

Notation

Communities

Introduction

Adjacency matrix $A \in \{0, 1\}^{n \times n}$ is

$$A_{ij} = \begin{cases} 1, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Pandam

Mining

Ritz vector

Example

Conclusions

Also define

$$e =$$
 vector of n ones $d = Ae =$ degree vector

$$a = 710 = acgree vo$$
 $D = diag(d)$

$$D = \operatorname{diag}(d)$$

$$L = D - A =$$
graph Laplacian

$$B = A - \frac{dd^T}{m} = \text{modularity matrix}$$

Spectrum for a random graph

Communities

Block models

Spectrum of a $G_{n,p}$ graph:

- - One large eigenvalue $\approx np$
 - Other eigs between $\approx \pm \sqrt{np}(1-p)/4$
 - Adjacency matrix = pee^{T} + "noise"



Spectrum for a G_{100,0.2} sample

Communities

ntroduction

Block models

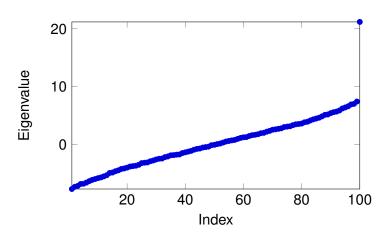
Optimizatio

walks

subspac

Ritz vector

Examples



Perron vector for a $G_{100,0.2}$ sample

Communities

.

Block models

Optimization

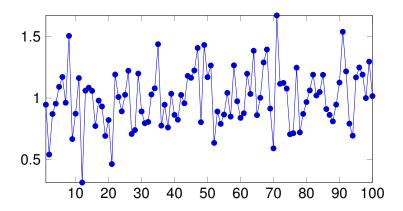
Dandon

walks

subspac

milz vecto

Examples



Block model approach

Communities

Block models

Optimizatio

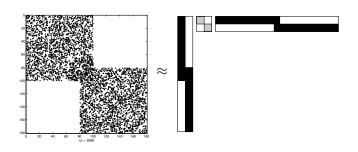
Random

waiks

оавораоо

Lxampies

Conclusions



Composite model: $A \approx S \operatorname{diag}(\beta) S^T$, $S \in \{0, 1\}^{n \times c}$

- Motivation: possibly-overlapping random graphs
- Columns of S are one basis for range space
- Want to go from some general basis back to S



Spectrum for a block model sample

Communities

...............................

Block models

Optimization

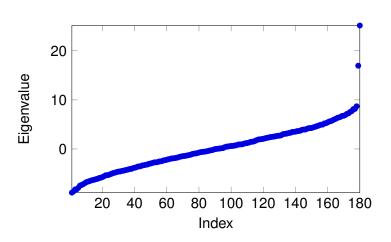
- ----

walks

Mining

Ritz vector

Examples



Dominant vectors for a block model example

Communities

introduction

Block models

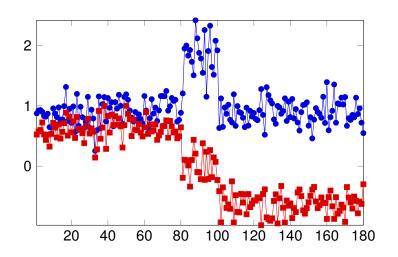
Optimizatio

walks

subspa

Ritz vect

Examples



Same space, different basis

Communities

.

Block models

Optimizatio

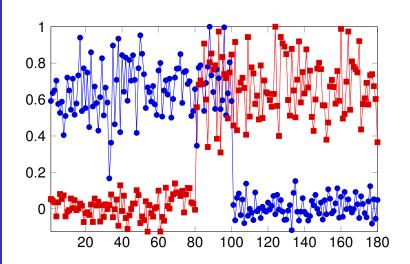
Develope

walks

subspa

Ritz vect

Evamplo



Questions

Communities

Introduction

Block models

Optimizatio

Random

subspace

Ritz vector

. Canalusian

- What about different matrices (e.g. *L*)?
- What about more interesting graph structures?
- How do we find the "right" subspace basis?

Measurement by quadratic forms

Communities

Introduction

Optimization

walks

Mining

Pitz vooto

Examples

Conclusions

Indicate $V' \subseteq V$ by $s \in \{0,1\}^n$. Measure subgraph:

$$s^T A s = |E'| = \text{internal edges}$$

$$s^T D s =$$
edges incident on subgraph

$$s^T L s = \text{edges between } V' \text{ and } \bar{V}'$$

$$s^T B s =$$
 "surprising" internal edges

Graph bisection

Communities

Disabassasist

Optimization

Optimization

walks

subspace:

Ritz vector

Lxamples

Idea: Find $s \in \{0,1\}^n$ such that $e^T s = n/2$ to

- \blacksquare minimize $s^T L s$ (min cut)
- maximize $s^T B s$ (max modularity)

Equivalently: Find $\bar{s} \in \{\pm 1\}^n$ such that $e^T \bar{s} = 0$ to

- \blacksquare minimize $\bar{s}^T L \bar{s} = s^T L s$ or
- \blacksquare maximize $\bar{s}^T B \bar{s} = s^T B s$

Oops - NP hard!

Relax!

Communities

nitioduction

Optimization

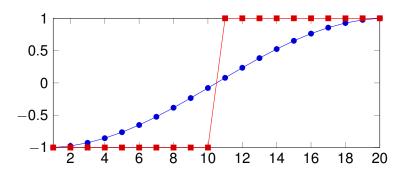
walks

Mining

Ritz vecto

Examples

Canaluaia



Hard: $\min \bar{s}^T L \bar{s}$ s.t. $e^T \bar{s} = 0$, $\bar{s} \in \{\pm 1\}^n$. Easy: $\min v^T L v$ s.t. $e^T v = 0$, $v \in \mathbb{R}^n$, $||v||^2 = n$.

Rayleigh quotients

Communities

Introduction

Block model

Optimization

Random

Mining subspace

HILZ VECIO

Examples

Conclusions

$$\frac{s^T A s}{s^T s} = \text{mean internal degree in subgraph}$$

$$\frac{s^T L s}{s^T s} = \text{edges cut between } V' \text{ and } \bar{V}'$$

$$\frac{s^T A s}{s^T D s} = \text{fraction of incident edges internal to } V'$$

$$\frac{s^T L s}{s^T D s} = \text{fraction of incident edges cut}$$

$$\frac{s^T B s}{s^T s} = \text{mean "surprising" internal degree in subgraph}$$

$$\frac{s^T B s}{s^T s} = \text{mean "surprising" internal degree in subgraph}$$

 $\frac{s^T B s}{s^T D s} = \text{mean fraction of internal degree that is surprising}$

Rayleigh quotients and eigenvalues

Communities

introduction

Optimization

Optimization

Mining

Subspaces

Lxamples

Basic connection (*M* spd):

$$\frac{x^T K x}{x^T M x}$$
 stationary at x \iff $K x = \lambda M x$

Easy despite lack of convexity.

Limits of Rayleigh quotients

Communities

Introduction

.

Optimization

Optimization

walks

subspace

Ritz vecto

Examples

Conclusion:

But small variations kill us:

$$\max_{x\neq 0} \frac{x^T A x}{\|x\|_2^2} = \lambda_{\max}(A), \text{ but}$$
$$\max_{x\neq 0} \frac{x^T A x}{\|x\|_1^2} = 1 - \omega^{-1}$$

where ω is the max clique size (Motzkin-Strauss).

Rayleigh quotients and eigenproblems

Communities

Introduction

Optimization

Optimizatio

walks

subspac

niiz vecio

Examples

Conclusion

Decompose:

$$W^TMW = I$$
 and $W^TKW = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$.

For any $x \neq 0$,

$$\frac{x^T K x}{x^T M x} = \sum_{j=1}^n \lambda_j z_j^2$$
, where $z = \frac{W^{-1} x}{\|W^{-1} x\|_2}$.

So

$$rac{s^T \mathit{Ks}}{s^T \mathit{Ms}} pprox \lambda_{\mathsf{max}} \implies s pprox \sum_{\lambda_j pprox \lambda_{\mathsf{max}}} w_j z_j.$$

So look at invariant subspaces for extreme eigenvalues.

The random walker

Communities

Introduction

Dia ali manda

Optimizatio

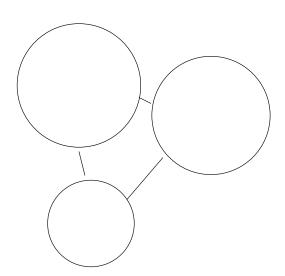
Random

walks

subspace

Ditz vocto

Examples



The random walker

Communities

ntroduction

Optimization

Bandom

walks Mining

Ritz vectors

Lxample

Basic idea: extract structure from random walk.

Old: start at seed and walk forward

Day 1: I came up with a funny joke!

Day 2: I tell everyone in my family

Day 3: My mother tells a friend?

New: look at how quickly source is forgotten

Day 1: David came up with a funny joke!

Day 2: There's a joke going around Cornell CS.

Day 3: I read this bad joke on the web...

The random walker

Communities

Lazy random walk with transition matrix $T = \frac{1}{2}(I + AD^{-1})$.

11 Start at p_0 , take k steps. Distribution:

$$p_k = T^k p_0 \quad (\rightarrow d/m \text{ as } k \rightarrow \infty)$$

2 End at q_0 after k steps. Conditional distribution on start:

$$q_k \propto (T^T)^k q_0 \quad (\rightarrow e/n \text{ as } k \rightarrow \infty)$$

Note: If the graph is undirected, $T^T = D^{-1}TD$.

Random walks

Simon-Ando theory

Communities

Markov chain with loosely-coupled subchains:

Rapid *local* mixing: after a few steps

$$p_k \approx \sum_{j=1}^c \alpha_{j,k} p_{\infty}^{(j)}$$

where $p_{\infty}^{(j)}$ is a local equilibrium for the *i*th subchain

■ Slow equilibration: $\alpha_{i,k} \rightarrow \alpha_{i,\infty}$.

Alternately, rapid local mixing looks like:

$$q_k \approx \sum_{j=1}^c \gamma_{j,k} s_j$$

where s_j is an indicator for nodes in one subchain.

Random walks

Simon-Ando theory

Communities

Introduction

Block mode

Optimization

Random

Mining

- ...

THE VOCION

=xamples

In chemistry: transitions among metastable states.

In network analysis: transitions among communities?

Spectral Simon-Ando picture

Communities

Introduction

Optimization

Random walks

subspaces

Examples

Conclusio

Exactly decoupled case (*c* decoupled chains):

- Eigenvalue one has multiplicity *c*.
- Eigenvectors of *T* are local equilibria.
- Eigenvectors of T^T are indicators for chains.
- Rapid mixing \implies large gap to λ_{c+1} .

Weakly coupled case:

- Cluster of *c* eigenvalues near 1.
- Eigenvectors of *T* are combinations of local equilibria.
- Eigenvectors of T^T are combinations of indicators.
- Large gap between λ_c and λ_{c+1} .

Summary so far

Communities

Introduction

Ontimization

Optimization

Random walks

subspaces

Ritz vector

Examples

Conclusion

Indicator vectors approximately in invariant subspaces

- Several possible motivations
- \blacksquare Several possible matrices (I like T^T)

But how do we go from the subspace to the indicators?

Indicators from subspaces: spectral clustering

Communities

Rlock model

Optimization

- - -

Mining subspaces

Ritz vector

_ .

Conclusio

U spans a small subspace (e.g. an invariant subspace)

range(U) \approx range(S), S indicates a partition. Rows of U in the same partition are identical.

Idea: Treat rows of *U* are *latent coordinates*. Cluster.

- 2 Suppose some indicator $s \approx Uy$. Then row U(j, :)
 - forms an acute angle with y when $s_i = 1$
 - is almost normal to y when $s_j = 0$.

Clustering? What if sets overlap?

Clustering and overlap

Communities

_.

Optimization

Optimization

walks

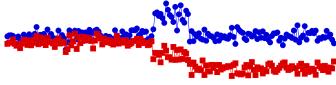
Mining subspaces

Ritz vecto

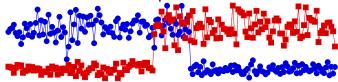
Examples

Conclusion

Dominant eigenvectors for *A*:



Alternate basis for the space:



How do we get the latter basis?



Desiderata

Communities

IIIIIOduction

Optimization

Dandon

Mining subspaces

Ritz vectors

Examples

Conclusion:

Given a basis U, want to extract a vector \tilde{s} s.t.

- \bullet \tilde{s} lies close to the span of U
- \bullet \tilde{s} is almost an indicator for a community
 - Maybe nonnegative?
 - Not too many ones?

Indicators from subspaces: LP version

Communities

Mining

subspaces

Suppose $s \approx Uy$ for some y, $s_i = 1$. Want to find s. Try optimization (a linear program):

```
minimize \|\tilde{s}\|_1 (proxy for sparsity of \tilde{s})

s.t. \tilde{s} = Uy (\tilde{s} in the right space)

\tilde{s}_i \ge 1 ("seed" constraint)

\tilde{s} > 0 (componentwise nonnegativity)
```

Recovers smallest set containing node i if

- $U = SY^{-1}$ exactly.
- Each set contains at least one element only in that set. (Frequently works if there is not "too much" overlap.)

What about noise? Generally need a thresholding strategy.

Indicators from subspaces: QP version

Communities

Minina

subspaces

Alternate optimization (box-constrained quadratic program):

Recover LP with $P = I - UU^T$ and $\tau \to 0$ (for $U^TU = I$).

- Can let P be general semidefinite matrix (e.g. P = L)
- Size of τ controls sparsity (can automate choice)

Summary so far

Communities

Block models

Optimization

Random

Mining

subspaces

Examples

Conclusio

Two pieces to spectral community detection:

- Pull out an invariant subspace
- Mine the subspace for community structure

Motivation: optimization or random walk dynamics.

But...

- What about when *n* and *c* are both large?
- What if there is no clear spectral gap?

Would like an alternative to invariant subspaces!

Eigenvectors to Ritz vectors

Communities

Introduction Block models

Optimiza

Random walks

Mining subspace

Ritz vectors

Examples

Evamples

Eigenvectors are stationary points of Rayleigh quotients. Find stationary points in a subspace \implies *Ritz* vectors.

Usual approach to large-scale eigenproblems:

Generate a basis for a Krylov subspace

$$\mathcal{K}_k(A, x_0) = \text{span}\{x_0, Ax_0, A^2x_0, \dots, A^{k-1}x_0\}$$

- 2 Ritz values rapidly approximate extreme eigenvalues
- 3 Ritz vectors approximate extreme eigenvectors

Idea: Instead of searching invariant subspace, search in a space spanned by a few scaled Ritz vectors. Pulls out dynamics of *short* random walks (vs long).

Current favorite method

Communities

Dia ak madak

Optimization

Pandam

Mining

Ritz vectors

Examples

- 1 Pick "seed" nodes j_1, j_2, \ldots
- 2 Take short random walks (length k) from each seed
- 3 Extract a few Ritz vectors (fewer than k) from span $\{\phi_0, \phi_1, \dots, \phi_{k-1}\}$.
- Use quadratic programming to find approximate indicators in subspace space spanned by all Ritz vectors.
- 5 Possibly add more seeds and return to step 1.
- 6 Threshold to get initial indicator approximation.
- Greedily optimize angle between indicator and space.

Wang test graph

Communities

Introduction

Block mode

Optimizatio

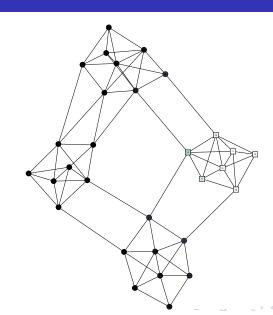
Random

Mining

опрорисс

milz vector

Examples



Spectrum for Wang test graph

Communities

Introduction

Block model

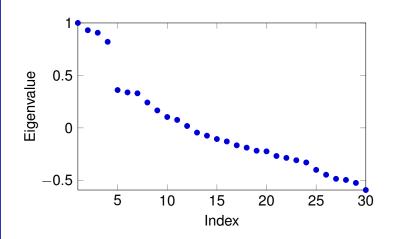
Optimization

walks

subspac

HILZ VECTOR

Examples



Zachary Karate graph

Communities

Introduction

Optimizatio

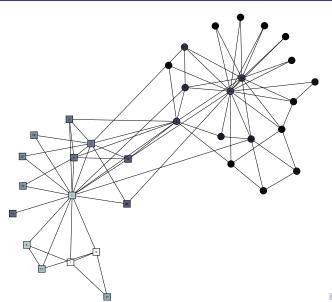
Орини

walks

subspa

HITZ VECTO

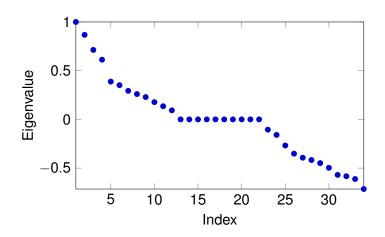
Examples



Spectrum for Karate

Communities

Examples



Football graph

Communities

Slock mode

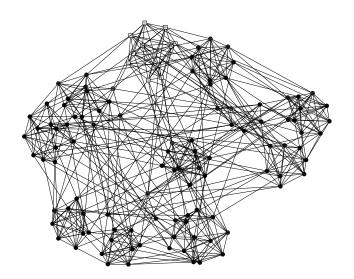
Optimization

Random

Mining

Pitz voote

Examples



Spectrum for Football

Communities

introduction

Dia ale mandal

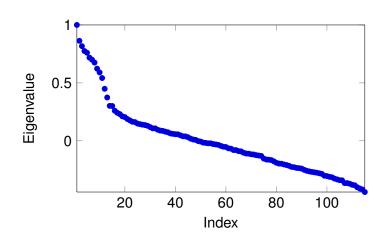
Optimization

waiks

subspace

milz vector

Examples



Dolphin graph

Communities

Introduction

Block mode

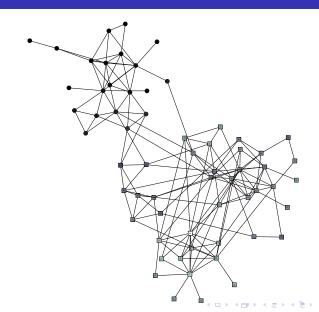
Optimizatio

· · ·

Subspace

Till Vecto

Examples



Spectrum for Dolphin

Communities

Hilloudction

Block model

Optimization

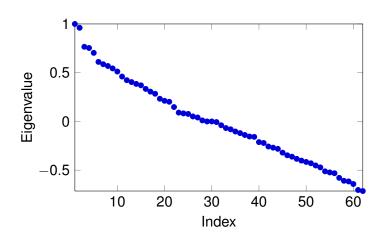
Danden

walks

subspace

Ritz vector

Examples



Non-overlapping synthetic benchmark ($\mu = 0.5$)

Communities

Introduction

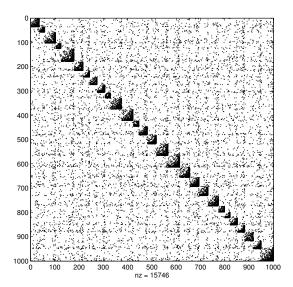
Ontimizatio

Optimizatio

walks

subspace

Examples



Spectrum for synthetic benchmark

Communities

introduction

Plack madal

Optimization

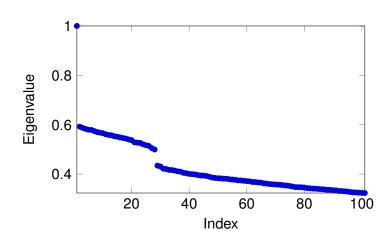
Demoles

waiks

subspace

THE VOOLO

Examples



Communities

Plack models

Optimization

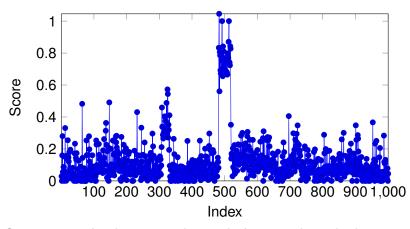
Random

Mining

Ritz vecto

Examples

Capaluai



Score vector for the two-node seed of 492 and 513 in the first LFR benchmark graph. Ten steps, three Ritz vectors.

Non-overlapping synthetic benchmark ($\mu = 0.6$)

Communities

minoduction

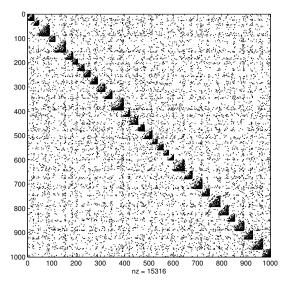
Optimization

Random

Walks

subspace

Examples



Spectrum for synthetic benchmark

Communities

introduction

DI - -I- --- - -I-I

Ontimization

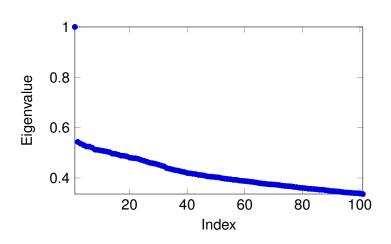
Optimization

walks

subspac

Ritz vector

Examples



Communities

Dississes

Optimization

Оринигаци

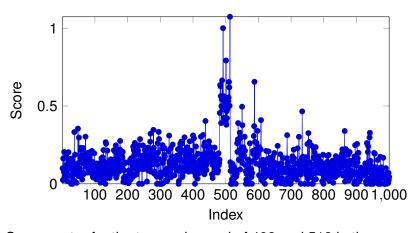
Random walks

Mining subspace

Ritz vecto

Examples

Conclusi



Score vector for the two-node seed of 492 and 513 in the first LFR benchmark graph. Ten steps, three Ritz vectors.

Overlapping synthetic benchmark ($\mu = 0.3$)

Communities

Introduction

Block mode

Optimizatio

Mining subspace

Ritz vector

Examples

- 1000 nodes
- 47 communities
- 500 nodes belong to two communities

Spectrum for synthetic benchmark

Communities

introduction

District and state

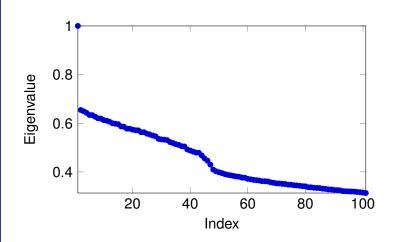
Optimization

waiks

subspace

milz vector

Examples



Communities

Introduction

Optimization

D I

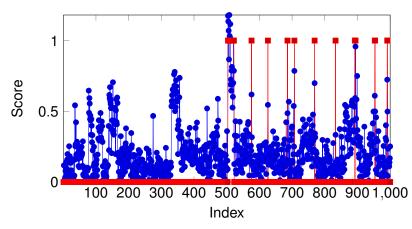
walks

subspac

Ritz vecto

Examples

Canaluai



Score vector for the two-node seed of 521 and 892. The desired indicator is in red.

Communities

ntroduction

Optimizatio

Optimization

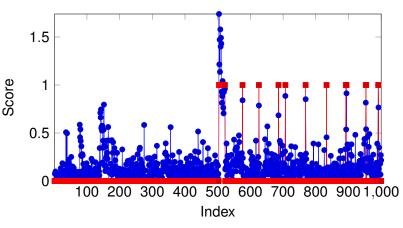
walks

Mining

Ritz vecto

Examples

Capaluai



Score vector for the two-node seed of 521 and 892 + twelve reseeds. The desired indicator is in red.

Conclusions

Communities

ntroduction

Optimization

Ор.....

Mining

Ritz vector

Conclusions

Classic spectral methods use eigenvectors to find communities, but:

- We don't need to stop at partitioning!
 - Overlap is okay
 - Key is how we mine the subspace
- We don't need to stop at eigenvectors!
 - Can also use Ritz vectors
 - Computation is cheap: short random walks