

# IMPLEMENTING REGULARIZATION IMPLICITLY VIA APPROXIMATE EIGENVECTOR COMPUTATION

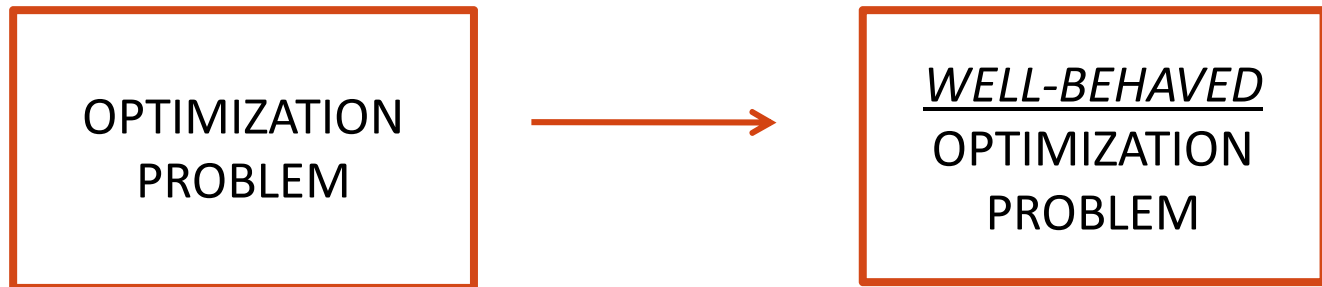
Michael W. Mahoney, Stanford University

**Lorenzo Orecchia**, UC Berkeley

# Regularization

## What is it?

Regularization is a fundamental technique in **optimization**



- Stable optimum
- Unique optimal solution

...

# Regularization

## What is it?

Fundamental technique in **optimization**



## Formal Definition

Initial optimization program

$$\min_{x \in H} L(x)$$

Regularized program

$$\min_{x \in H} L(x) + \lambda \cdot F(x)$$

Parameter  $\lambda > 0$

Regularizer  $F$

# Regularization

## What is it?

Fundamental technique in **optimization**



## Formal Definition

Initial optimization program

$$\min_{x \in H} L(x)$$

Regularized program

$$\min_{x \in H} L(x) + \lambda \cdot F(x)$$

## Benefits of Regularization

In Learning Theory and Statistics:

- Improves level of generalization
- Prevents overfitting
- Decreases sensitivity to random noise

# Regularization

## What is it?

Fundamental technique in **optimization**



## Formal Definition

Initial optimization program

$$\min_{x \in H} L(x)$$

Regularized program

$$\min_{x \in H} L(x) + \lambda \cdot F(x)$$

## Benefits of Regularization

In Learning Theory and Statistics:

- Improves level of generalization
- Prevents overfitting
- Decreases sensitivity to random noise

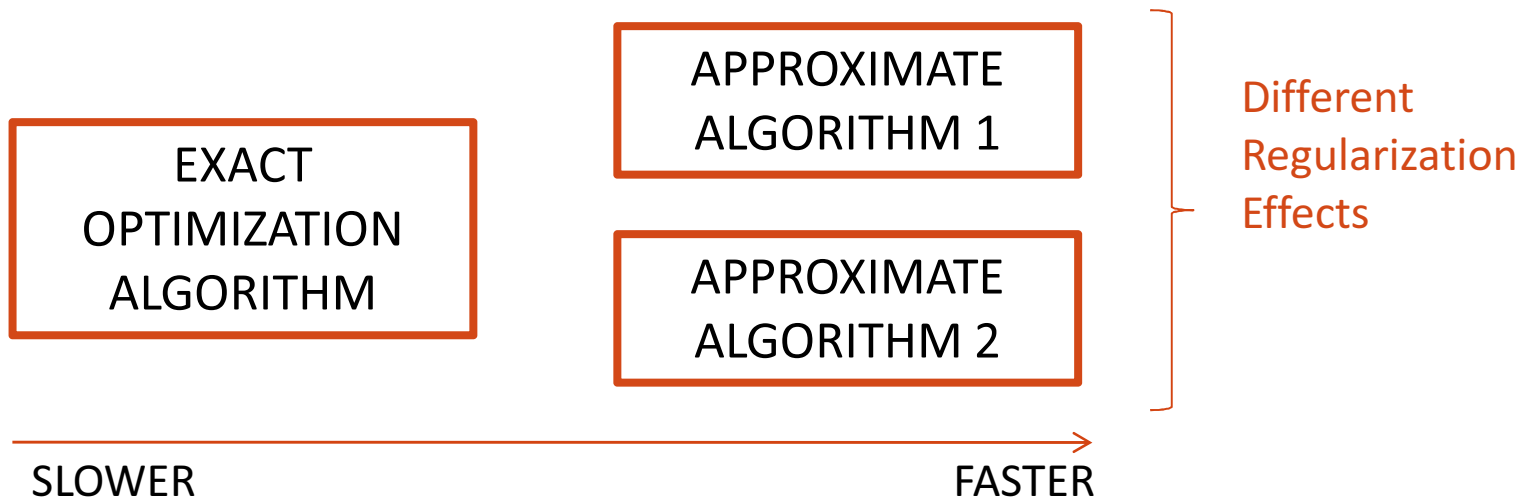
**NB:** Typically, the regularized problem is **explicitly** stated and solved.

# Implicit Regularization

## Empirical Observation:

Many heuristics and approximation techniques, designed to speed-up computations, seem to have **regularizing effects**.

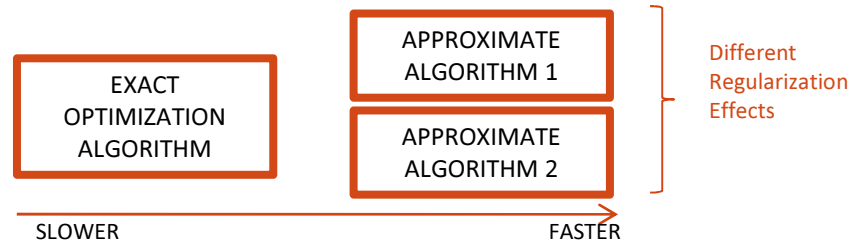
**EXAMPLES:** early stopping, binning, ...



**NB:** This regularization is **implicit**, no regularized optimization is explicitly solved and the **regularizer is unknown**.

# Goal of our paper

Study connection between  
approximate computation and regularization



## QUESTION:

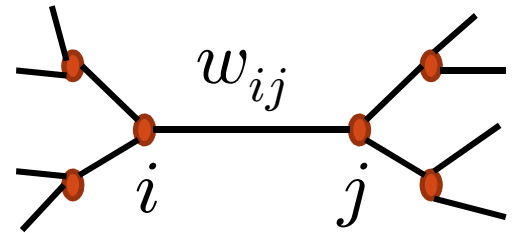
Can we characterize **approximate computation** procedures as computing **optimal solutions to explicit regularized problems**?

## SPECIFIC SETTING:

Computation of first non-trivial eigenvector of a graph

# Spectral Graph Theory

- Undirected Weighted Graph  $G = (V, E, w)$
- Associated Matrices



Degree Matrix  $D$

Adjacency Matrix  $A$

$$D = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_i & 0 & 0 \\ 0 & 0 & d_j & 0 \\ 0 & 0 & 0 & d_n \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & w_{ij} & 0 \\ 0 & w_{ij} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

# Spectral Graph Theory

- Undirected Weighted Graph  $G = (V, E, w)$
- Associated Matrices

Degree Matrix  $D$

Adjacency Matrix  $A$

Laplacian Matrix  $L = D - A$

$$\begin{pmatrix} d_i & -w_{ij} \\ -w_{ij} & d_j \end{pmatrix}$$

# Spectral Graph Theory

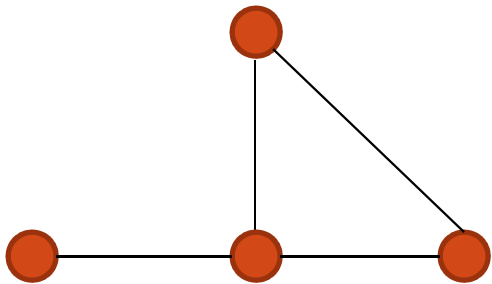
- Undirected Weighted Graph  $G = (V, E, w)$
- Associated Matrices

Degree Matrix  $D$

Adjacency Matrix  $A$

Laplacian Matrix  $L = D - A$

Natural Random Walk Matrix  $W = A D^{-1}$



$$W = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix}$$

# Spectral Graph Theory

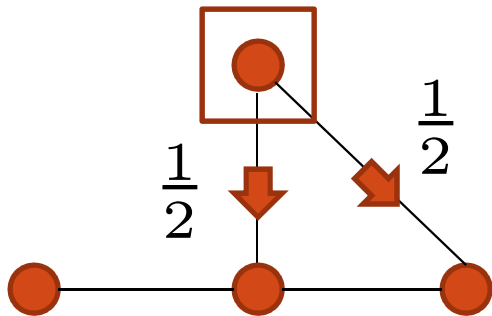
- Undirected Weighted Graph  $G = (V, E, w)$
- Associated Matrices

Degree Matrix  $D$

Adjacency Matrix  $A$

Laplacian Matrix  $L = D - A$

Natural Random Walk Matrix  $W = A D^{-1}$



$$W = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix}$$

# Random Walks and Graph Eigenvectors

The walk  $W$  has a **stationary distribution**  $\pi \propto D\mathbf{1}$ , uniform over the edges.

$$WD\mathbf{1} = D\mathbf{1}$$

The first (smallest) eigenvector of  $L$  is the **constant eigenvector**  $\mathbf{1}$  with eigenvalue 0.

$$L\mathbf{1} = 0$$

# Random Walks and Graph Eigenvectors

The walk  $W$  has a **stationary distribution**  $\pi \propto D\mathbf{1}$ , uniform over the edges.

The first (smallest) eigenvector of  $L$  is the **constant eigenvector**  $\mathbf{1}$  with eigenvalue 0

Given vector  $x$  such that  $x^T D\mathbf{1} = 0$

$$Dx \longrightarrow W(Dx) \longrightarrow W^2(Dx) \longrightarrow \mathbf{MIXING} \longrightarrow W^t(Dx) \longrightarrow 0$$

**What determines the rate of convergence for a specific  $x$ ?**

# Random Walks and Graph Eigenvectors

The walk  $W$  has a **stationary distribution**  $\pi \propto D\mathbf{1}$ , uniform over the edges.

The first (smallest) eigenvector of  $L$  is the **constant eigenvector**  $\mathbf{1}$  with eigenvalue 0

Given vector  $x$  such that  $x^T D\mathbf{1} = 0$

$Dx \longrightarrow W(Dx) \longrightarrow W^2(Dx) \longrightarrow \text{MIXING} \longrightarrow W^t(Dx) \longrightarrow 0$

**What determines the rate of convergence for a specific  $x$ ?**

FAST MIXING

SLOW MIXING

$x^T Lx$  large

$x^T Lx$  small

The first non-trivial eigenvector  $x^*$  describes the **most slowly mixing vector**  $x^*$  and its eigenvalue  $\lambda_2$  determine many important **structural properties**

# Random Walks and Graph Eigenvectors

The walk  $W$  has a **stationary distribution**  $\pi \propto D\mathbf{1}$ , uniform over the edges.

The first (smallest) eigenvector of  $L$  is the **constant eigenvector**  $\mathbf{1}$  with eigenvalue 0

Given vector  $x$  such that  $x^T D\mathbf{1} = 0$

$$Dx \longrightarrow W(Dx) \longrightarrow W^2(Dx) \longrightarrow \text{MIXING} \longrightarrow W^t(Dx) \longrightarrow 0$$

**What determines the rate of convergence for a specific  $x$ ?**

FAST MIXING

SLOW MIXING

$$x^T Lx \text{ large}$$

$$x^T Lx \text{ small}$$

The first non-trivial eigenvector  $x^*$  describes the **most slowly mixing vector**  $x^*$  and its eigenvalue  $\lambda_2$  determine many important **structural properties**

Computation of  $x^*$

$$x^* = \lim_{t \rightarrow \infty} \frac{D^{-1} W^t y_0}{\|W^t y_0\|_{D^{-1}}}$$

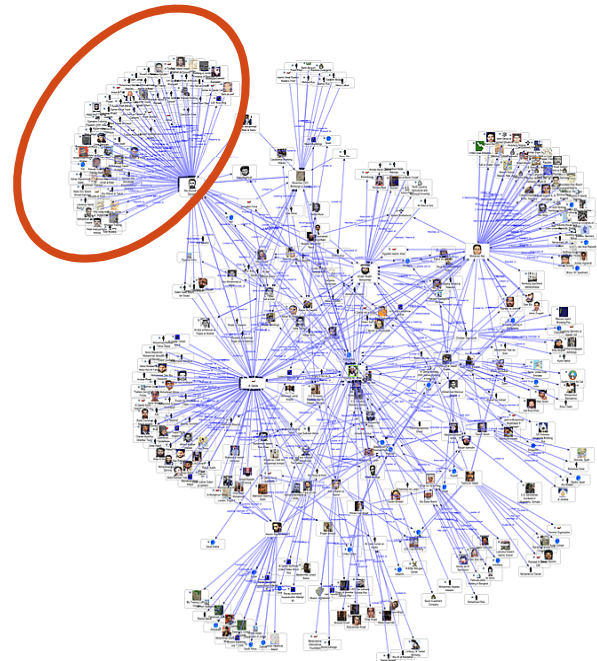
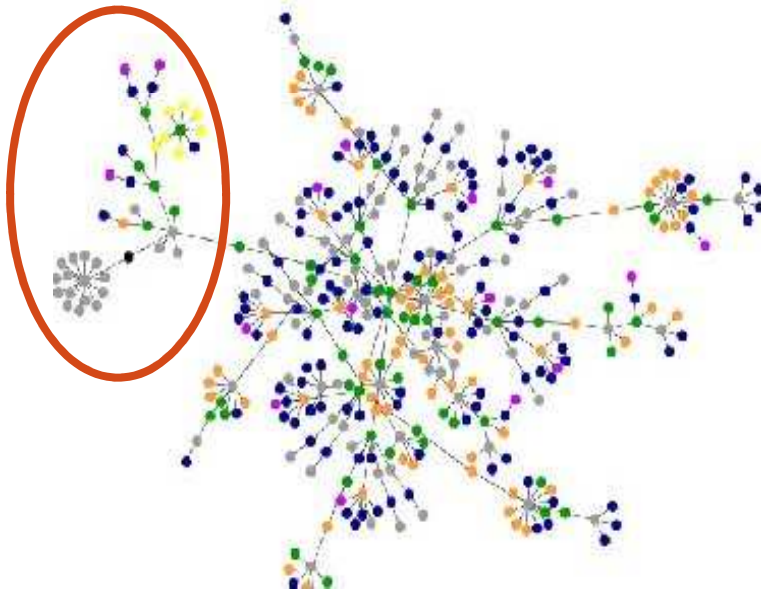
for random  $y_0$  such that  $y_0^T D^{-1} \mathbf{1} = 0$

# Motivation: Communities in Data Networks

- Model as undirected weighted graph  $G = (V, E, w)$
- **GOAL:** Find meaningful communities by optimizing “score” over all cuts

**Conductance  
Score**

$$\phi(S) = \frac{w(S, \bar{S})}{\min\{\text{vol}(S), \text{vol}(\bar{S})\}}$$



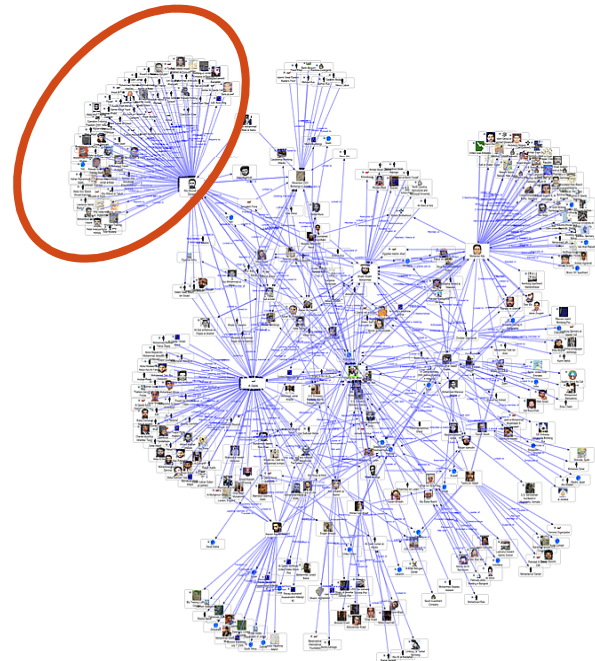
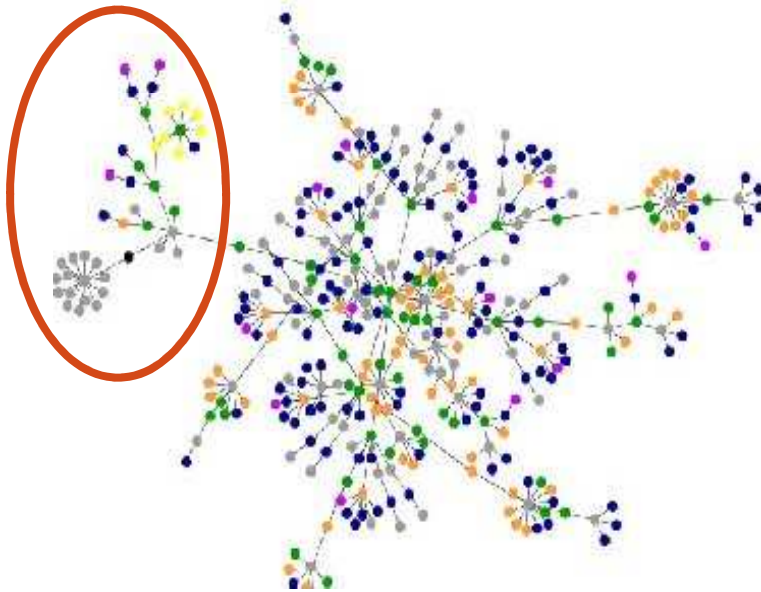
# Motivation: Communities in Data Networks

- Model as undirected weighted graph  $G = (V, E, w)$
- **GOAL:** Find meaningful communities by optimizing “score” over all cuts

**Conductance  
Score**

$$\phi(S) = \frac{w(S, \bar{S})}{\min\{\text{vol}(S), \text{vol}(\bar{S})\}}$$

**NP-HARD**

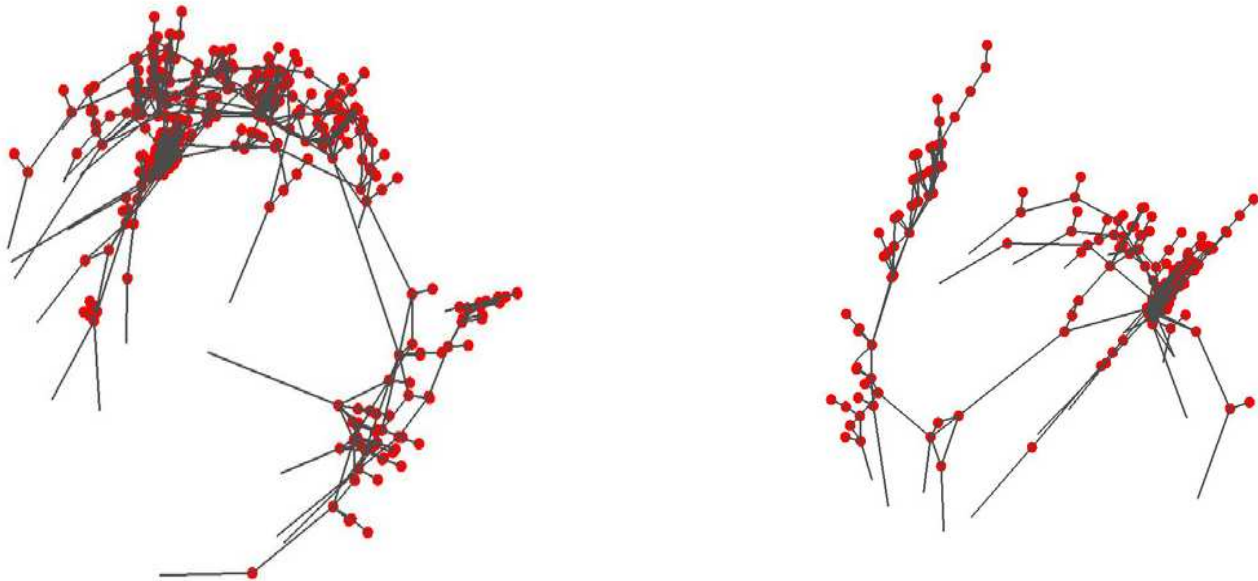


**USE FIRST NON-TRIVIAL EIGENVECTOR TO OPTIMIZE CONDUCTANCE**

# Motivation: Communities in Data Networks

- Eigenvector method and other algorithms optimizing conductance

$$\lim_{t \rightarrow \infty} \frac{D^{-1} W^t y_0}{\|W^t y_0\|_{D^{-1}}}$$



LOW CONDUCTANCE,

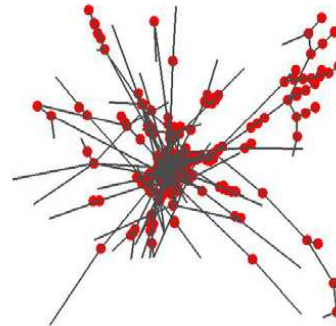
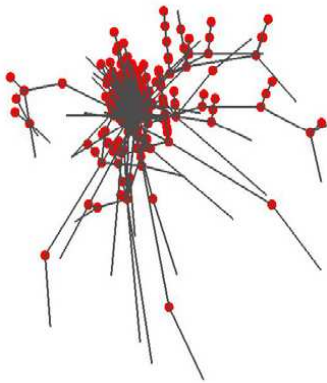
**BUT** OFTEN DISJOINT, ELONGATED, SENSITIVE TO NOISE

# Motivation: Communities in Data Networks

- Random walk approximation to eigenvector

$$\frac{D^{-1}W^t y}{\|W^t y\|_{D^{-1}}}$$

Not too large  $t$   
Analogous to early stopping



HIGHER CONDUCTANCE,  
BUT SMOOTHER, STABLER CLUSTERS

## CONCLUSION

RANDOM WALKS IMPLICITLY REGULARIZE EIGENVECTOR COMPUTATION

# Our Paper

## RESULT:

We show that a number of classic **random walks** arise as optimal solutions to a **regularized version of the graph eigenvector problem**.

## OUTLINE:

1. Definition of **random walks**
2. Construction of the **regularized problem**
3. Definition of the **regularizers**
4. Main result
5. Discussion

# Random Walks

Different random walks approximate eigenvector differently yielding **different regularization properties**.

Random walks considered in our work and their transition matrices:

DISTRIBUTION  
OF NUMBER OF STEPS

- Heat Kernel random walk with parameter  $t$

$$H_t = e^{-tL} = e^{-t} \sum_{i=1}^{\infty} \frac{t^i}{i!} W^i$$

**Poisson ( $t$ )**

- Personalized PageRank random walk

$$R_\alpha = \alpha \sum_{i=0}^{\infty} (1 - \alpha)^i W^i$$

**Geometric ( $\alpha$ )**

- Truncated random walk

$$T_{p,t} = (pI + (1 - p)W)^t$$

**Binomial ( $t, p$ )**

# Regularized Spectral Optimization

- Assume graph  $G$  is  $d$ -regular.

Original Program

$$\frac{1}{d} \min x^T L x$$

$$\text{s.t. } \|x\|_2 = 1$$

$$x^T \mathbf{1} = 0$$



SDP Formulation

$$\frac{1}{d} \min L \bullet X$$

$$\text{s.t. } I \bullet X = 1$$

$$J \bullet X = 0$$

$$X \succeq 0$$

Programs have **same optimum**. Take optimal solution  $X^* = x^*(x^*)^T$

# Regularized Spectral Optimization

SDP Formulation

$$\frac{1}{d} \min L \bullet X$$

s.t. 
$$\left. \begin{aligned} I \bullet X &= 1 \\ J \bullet X &= 0 \\ X &\succeq 0 \end{aligned} \right\} \text{Density Matrix}$$

Eigenvector decomposition of  $X$ :

$$X = \sum p_i v_i v_i^T \left\{ \begin{aligned} \forall i, p_i &\geq 0, \\ \sum p_i &= 1, \\ \forall i, v_i^T \mathbf{1} &= 0. \end{aligned} \right.$$

Eigenvalues of  $X$  define **probability distribution**

# Regularized Spectral Optimization

SDP Formulation

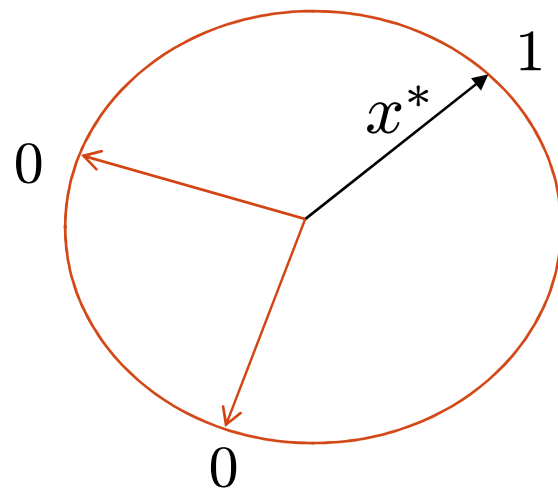
$$\frac{1}{d} \min L \bullet X$$

s.t. 
$$\left. \begin{aligned} I \bullet X &= 1 \\ J \bullet X &= 0 \\ X &\succeq 0 \end{aligned} \right\} \text{Density Matrix}$$

Eigenvalues of  $X$  define **probability distribution**

$$X^* = x^*(x^*)^T$$

TRIVIAL DISTRIBUTION



# Regularized Spectral Optimization

SDP Formulation

$$\frac{1}{d} \min L \bullet X$$

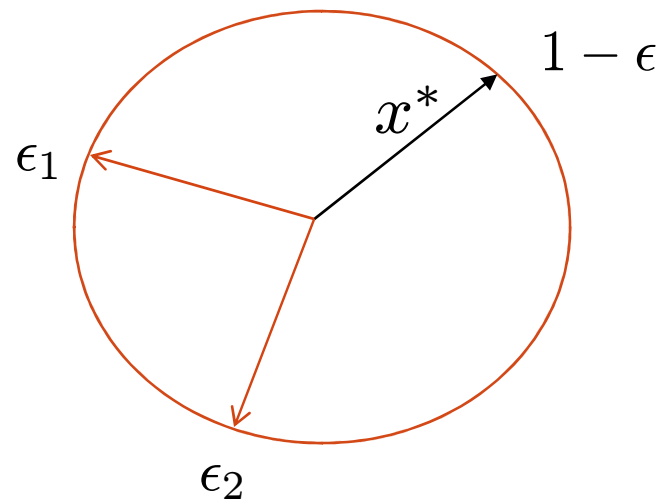
s.t. 
$$\left. \begin{aligned} I \bullet X &= 1 \\ J \bullet X &= 0 \\ X &\succeq 0 \end{aligned} \right\} \text{Density Matrix}$$

Eigenvalues of  $X$  define **probability distribution**

$$X^* = x^*(x^*)^T$$

REGULARIZATION

$$X^* = \sum p_i v_i v_i^T$$



# Regularized Spectral Optimization

$$\begin{aligned} & \text{Regularized SDP} \\ & \frac{1}{d} \min L \bullet X + \eta \cdot F(X) \quad \text{Regularizer } F \\ & \text{s.t.} \quad I \bullet X = 1 \quad \text{Parameter } \eta \\ & \quad \quad J \bullet X = 0 \\ & \quad \quad X \succeq 0 \end{aligned}$$

The regularizer  $F$  forces the distribution of eigenvalues of  $X$  to be non-trivial

# Regularized Spectral Optimization

$$\begin{aligned} & \text{Regularized SDP} \\ & \frac{1}{d} \min L \bullet X + \boxed{\eta \cdot F(X)} \quad \text{Regularizer } F \\ & \text{s.t.} \quad I \bullet X = 1 \quad \text{Parameter } \eta \\ & \quad \quad J \bullet X = 0 \\ & \quad \quad X \succeq 0 \end{aligned}$$

The regularizer  $F$  forces the distribution of eigenvalues of  $X$  to be non-trivial

Properties of  $F$ :

- **Strictly convex**, infinitely differentiable
- **Unitarily invariant**: given  $X = \sum p_i v_i v_i^T$ ,  $F(X)$  only depends on  $\{p_i\}$
- Minimized at  $X \propto I$ , i.e. **uniform distribution**

# Regularizers

Regularizers are **SDP-versions** of common regularizers

- von Neumann Entropy

$$F_H(X) = \text{Tr}(X \log X) = \sum p_i \log p_i$$

- Log Determinant

$$F_D(X) = -\log \det(X) = -\sum \log p_i$$

- p-Norm,  $p > 1$

$$F_p(X) = \frac{1}{p} \|X\|_p^p = \frac{1}{p} \text{Tr}(X^p) = \frac{1}{p} \sum p_i^p$$

# Our Main Result

Regularized SDP

$$\frac{1}{d} \min L \bullet X + \eta \cdot F(X)$$

$$\text{s.t.} \quad I \bullet X = 1$$

$$J \bullet X = 0$$

$$X \succeq 0$$

## RESULT:

Explicit correspondence between regularizers and random walks

REGULARIZER

OPTIMAL SOLUTION OF  
REGULARIZED PROGRAM

$$F = F_H \xrightarrow{\text{Entropy}} X^* \propto H_t \quad \text{where } t \text{ depends on } \eta$$

$$F = F_D \xrightarrow{\text{LogDet}} X^* \propto R_\alpha \quad \text{where } \alpha \text{ depends on } \eta$$

$$F = F_p \xrightarrow{p\text{-Norm}} X^* \propto T_{q, \frac{1}{p-1}} \quad \text{where } q \text{ depends on } \eta$$

# Discussion: Vector vs Density Matrix

Variable is **density matrix**, not vector

Q: Can we produce a single vector?

A: Density Matrix  $X$  describes distributions over vectors.

Assuming distribution is Gaussian, sample a vector

$$x = X^{\frac{1}{2}} u$$

where  $u \sim N(0, I_{n-1})$

For example, the vector

$$x = H_{t/2} u \quad \text{Random walk on random seed}$$

is a sample from the solution of an **entropy-regularized problem**

# Discussion: Vector vs Density Matrix

- **OPEN QUESTION:**

Is there an optimization formulation characterizing the vector

$$x = P s$$

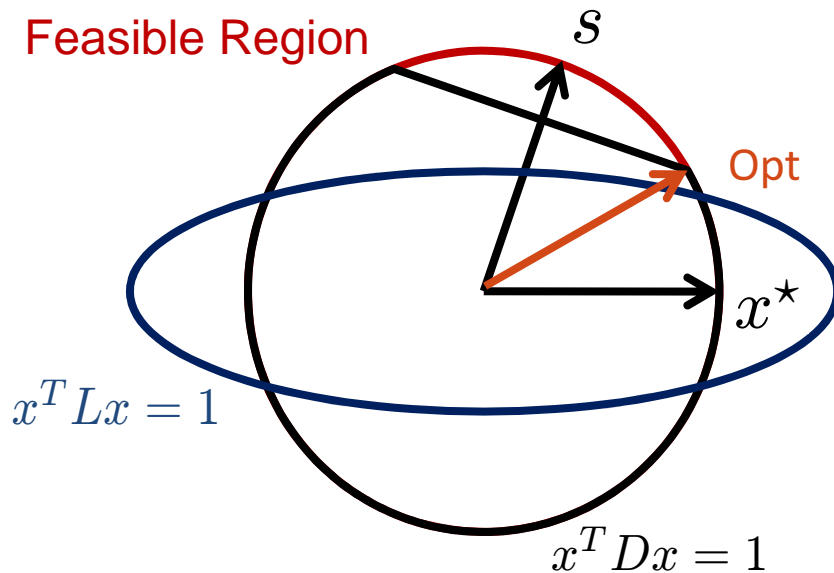
for a random walk  $P$  and a **fixed** seed point  $s$ ?

Unknown for Heat Kernel and Truncated Random Walk.

- **PARTIAL SOLUTION:**

**YES** for Personalized PageRank

# Localization vs Regularization



## Localized Program

$$\begin{array}{l} \min x^T L x \\ \text{s.t. } x^T D x = 1 \\ x^T D \mathbf{1} = 0 \end{array} \left. \vphantom{\begin{array}{l} \min \\ \text{s.t.} \end{array}} \right\} \begin{array}{l} \text{Global} \\ \text{spectral} \\ \text{problem} \end{array}$$

$$x^T D s \geq \kappa \left. \vphantom{x^T D s} \right\} \begin{array}{l} \text{Additional} \\ \text{local} \\ \text{constraint} \end{array}$$

## THEOREM:

For every  $\alpha \in (0,1)$ , there exists a  $\kappa$  such that the optimal solution to the Localized Program is a scaling of the Personalized PageRank  $R_{\alpha} s$

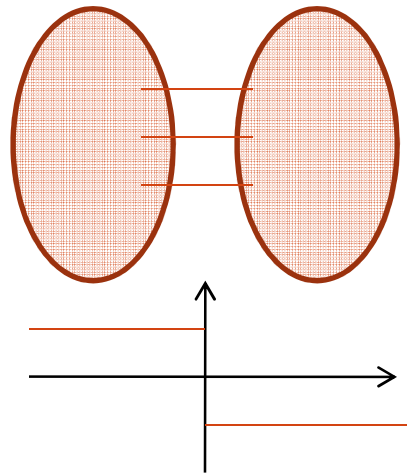
# Applications in Combinatorial Optimization

Regularization by random walk computation is particularly useful in the design of **approximation algorithms for graph partitioning**

**EXAMPLE:** Minimum-conductance **balanced cut**

$$\min_{\text{vol}(S) \geq b \cdot \text{vol}(V)} \phi(S)$$

**IDEA:** **Sensitivity** of the first non-trivial eigenvector makes it a poor tool to detect low-conductance balanced cut



EIGENVECTOR  $x^*$

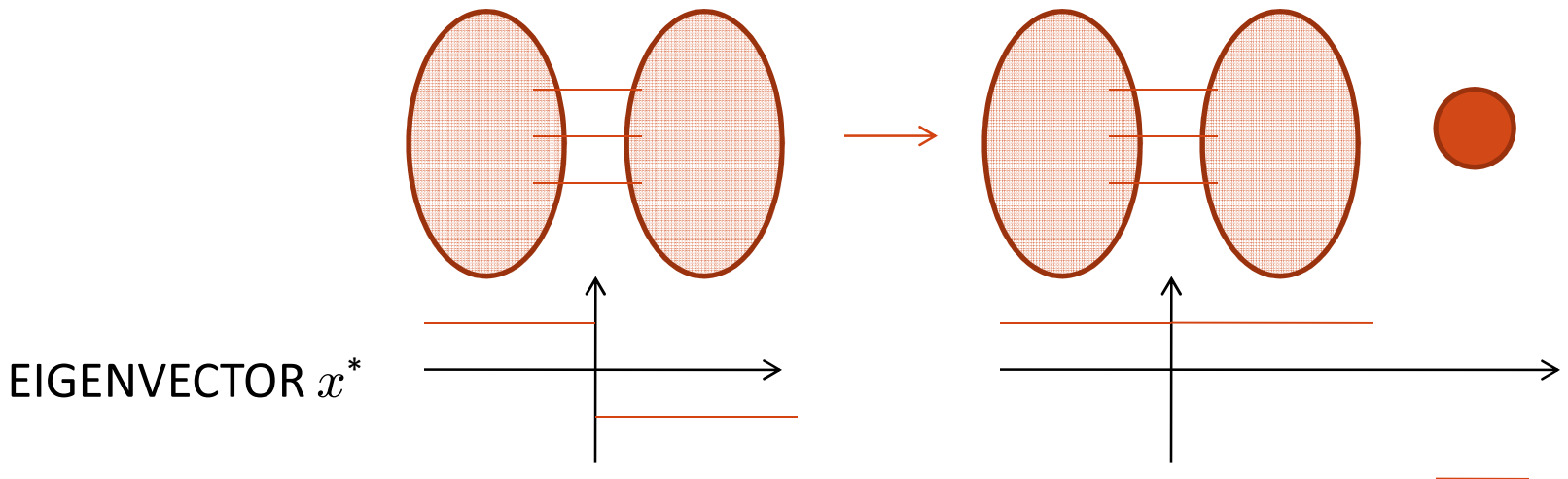
# Applications in Combinatorial Optimization

Regularization by random walk computation is particularly useful in the design of **approximation algorithms for graph partitioning**

**EXAMPLE:** Minimum-conductance **balanced cut**

$$\min_{\text{vol}(S) \geq b \cdot \text{vol}(V)} \phi(S)$$

**IDEA:** **Sensitivity** of the first non-trivial eigenvector makes it a poor tool to detect low-conductance balanced cut



**CUT IS INVISIBLE TO EIGENVECTOR, BUT DETECTABLE BY RANDOM WALKS**