Case Study 2: Document Retrieval

Spectral Clustering

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox

February 12th, 2013

Document Retrieval

- **Goal:** Retrieve documents of interest

Task 1: Find Similar Documents

- **Setup**
  - **Input:** Query article x
  - **Output:** Set of k similar articles

x

©Emily Fox 2013

k-Nearest Neighbor

- Articles $X = \{x^1, \ldots, x^N\}, \quad x^i \in \mathbb{R}^d$
- Query: $x \in \mathbb{R}^d$
- k-NN
  - Goal: Find k articles in X closest x
  - Formulation:

$$X^{NN} = \{x^{NN_1}, \ldots, x^{NN_k}\} \subseteq X$$

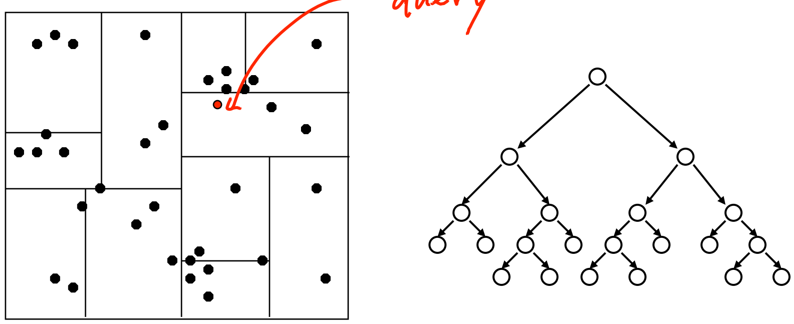
s.t. $\forall x^i \in X \setminus X^{NN}$

$$d(x^i, x) \geq \max_{x^{NN_i} \in X^{NN}} d(x^{NN_i}, x)$$

©Emily Fox 2013

Nearest Neighbor with KD Trees

query



- Traverse the tree looking for the nearest neighbor of the query point.

©Emily Fox 2013

Task 2: Cluster Documents

- **Setup**
  - **Input:** Corpus of documents
  - **Output:** Topic assignment per document

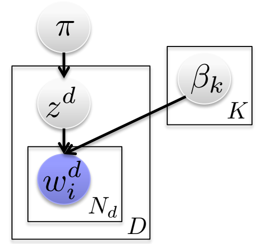
Sports

world news

©Emily Fox 2013

A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:



$z^d \sim \pi \quad d=1,\ldots,D$

$w_i^d | z^d \sim \beta_{z^d} \quad i=1,\ldots,N$

word prob. for cluster/topic z^d

Bayesian approach:

$\pi \sim Dir(\alpha_1,\ldots,\alpha_K)$

$\beta_k \sim Dir(\lambda_1,\ldots,\lambda_V)$

β_k is a V-dim pmf

size of vocab.

©Emily Fox 2013

Inference

- **Two tasks**
  - **Point estimation:**
    - Expectation-Maximization (EM)
  - **Characterize posterior:**
    - Gibbs sampling
    - Variational methods
    - Stochastic variational inference

©Emily Fox 2013

EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration *t*: $\hat{\theta}^{(t)}$

- **E-Step**

  Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$

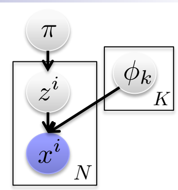
- **M-Step**

  Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

©Emily Fox 2013

Collapsed Gibbs Sampling

$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$



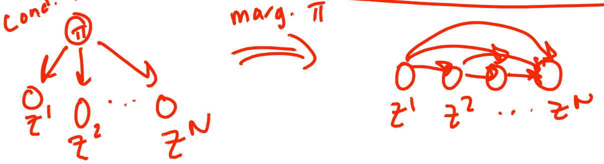
- Collapsed sampler

For $i=1,\dots,N$

$$z^{i(t)} \sim p(z^i \mid \underbrace{z^{1(t)}, \dots, z^{i-1(t)}, z^{i+1(t-1)}, \dots, z^{N(t-1)}}_{z_{\setminus i}^{(t)}}, x_{1:N}, \alpha, \lambda)$$

cond. ind.

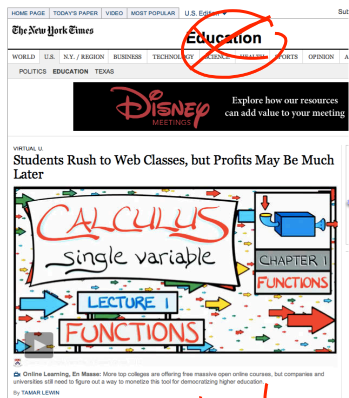
marg. π



©Emily Fox 2013

Task 3: Mixed Membership Model

- **Setup:** Document may belong to multiple clusters



EDUCATION

FINANCE

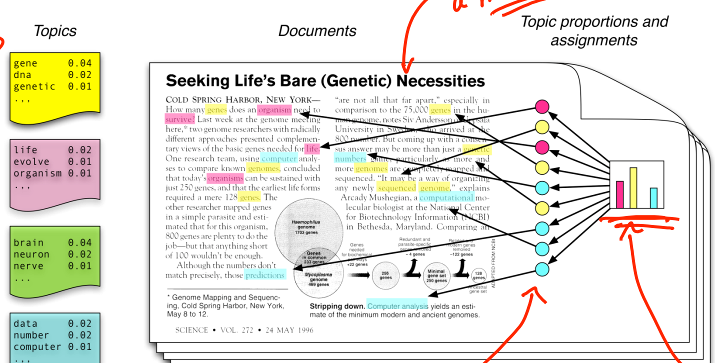
TECHNOLOGY

mixed membership

Latent Dirichlet Allocation (LDA)

each doc is a mixture of these corpus-wide topics

each topic as a dist. over words $\{\beta_k\}$



every word is assigned to a topic

each doc has its own prevelance of topics in doc

Variational Methods

- Recall task: Characterize the posterior $\underset{\text{params}}{\theta}$, ... $p(\theta, z \mid x)$

  params, latent vars, obs

- Turn posterior inference into an optimization task
- Introduce a "tractable" family of distributions over parameters and latent variables
  - Family is indexed by a set of "free parameters"
  - Find member of the family closest to: $p(\theta, z \mid x)$

  Call the family $\mathcal{Q}$ and want $q \in \mathcal{Q}$ that is closest to $p(\theta, z \mid x)$

- Questions:
  - How do we measure "closeness"?
  - If the posterior is intractable, how can we approximate something we do not have to begin with?

Variational Methods

- Similarity measure:

$$D(q(z,\theta) \,||\, p(z,\theta \mid x)) = E_q[\log q(z,\theta)] - E_q[\log p(z,\theta \mid x)]$$

$$= \underbrace{E_q[\log q(z,\theta)] - E_q[\log p(z,\theta,x)]}_{-\mathcal{L}(q)} + \log p(x)$$

- Evidence lower bound (ELBO)

$$\underbrace{\log p(x)}_{\text{const.}} = D(q(z,\theta) \,||\, p(z,\theta \mid x)) + \mathcal{L}(q) \geq \underline{\underline{\mathcal{L}(q)}}$$

add to a const

- Therefore, minimizing KL is equivalent to maximizing a lower bound on the marginal likelihood:
  - Max $\mathcal{L}(q)$ = min $D(q||p)$ = max lower bound of $\log p(x)$

$$\mathcal{L}(q) = E_q[\log p(\theta, z, x)] \underbrace{- E_q[\log q(\theta, z)]}_{\text{entropy of } q}$$

Task 2: Cluster Documents

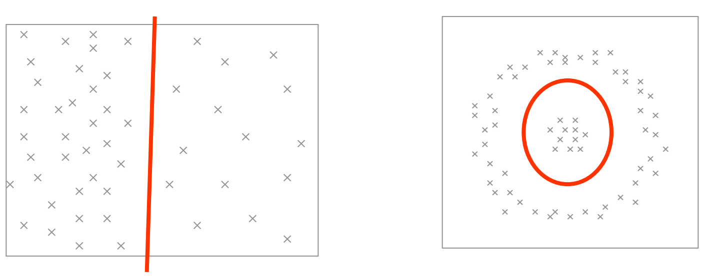
- **Setup**
  - **Input:** Corpus of documents
  - **Output:** Topic assignment per document

sports

world news

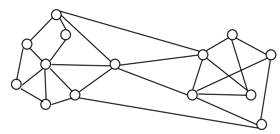
New Approach: Spectral Clustering

- **Goal:** Cluster observations
- **Method:**
  - Use similarity metric between observations
  - Form a similarity graph
  - Use standard linear algebra and optimization techniques to cut graph into connected components (clusters)



Setup

- Data: $x^1, \dots, x^N$
- Similarity metric:
- Similarity graph
  - Nodes
  - Edge weights



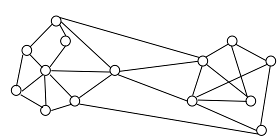
G = {V, E}

- Problem: Want to partition graph such that edges between groups have low weights

©Emily Fox 2013

Types of Graphs

- **ε-neighborhood**:
  - Only include edges with distances < ε
  - Treat as unweighted
- **k-NN:**
  - Connect v_i and v_j if v_j is a k-NN of v_i
  - Weighted by similarity s_{ij}
  - Directed → undirected

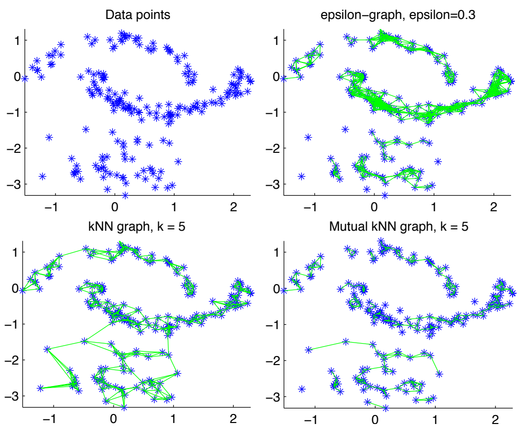


- **Mutual k-NN:**
  - Same as k-NN, but only include mutual k-NN

©Emily Fox 2013

Issues with Choosing Graph

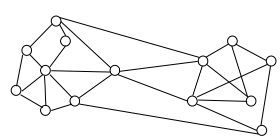
- Choosing graph construction techniques and parameters is non-trivial

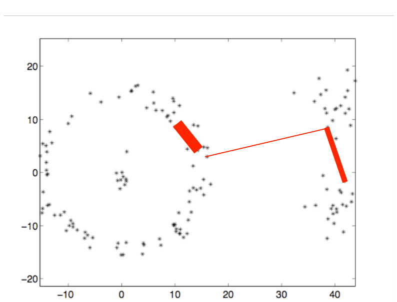


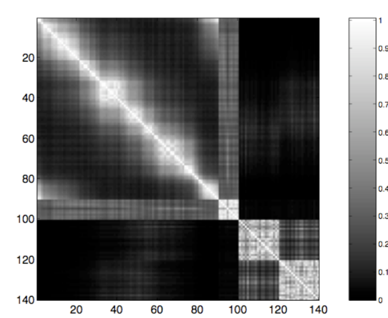
From
von Luxburg
2007

Graph Terminology I

- Weighted adjacency matrix

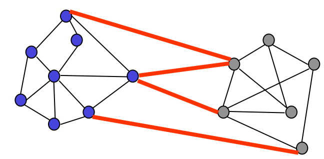




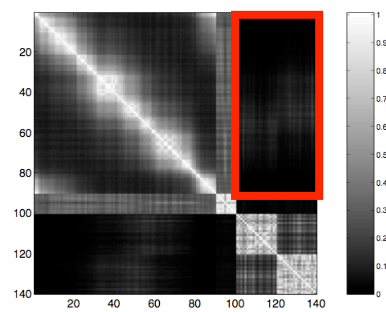


Graph Cuts

- **Problem:** Partition graph such that edges between groups have low weights



- Define: $W(A,B) = \sum_{i \in A, j \in B} w_{ij}$
- MinCut problem:

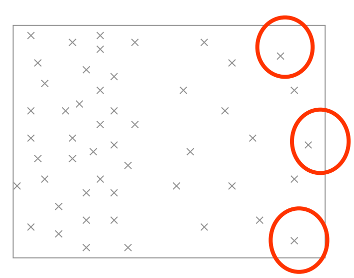


- Trivial to solve for *k*=2

©Emily Fox 2013

Issues with MinCut

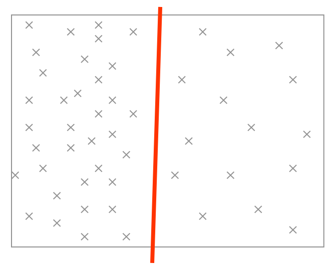
- MinCut favors isolated clusters



©Emily Fox 2013

Cuts Accounting for Size

- Ratio cuts (RatioCut)
- Normalized cuts (Ncut)
- Lead to "balanced" clusters



- First need more graph terminology…

©Emily Fox 2013

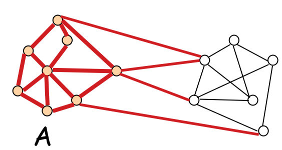
Graph Terminology II

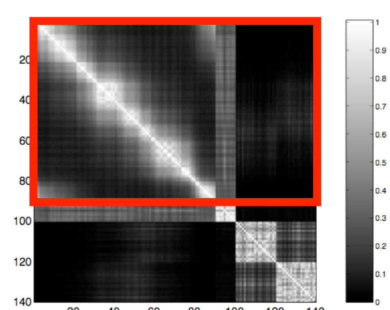
- Two measures of size of a subset
  - Cardinality:

    $|A|$

  - Volume:

    $\mathrm{vol}(A)$





©Emily Fox 2013

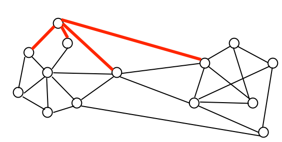
Cuts Accounting for Size

- Ratio cuts (RatioCut)
  - *k*=2
  - General *k*
- Normalized cuts (Ncut)
  - *k*=2
  - General *k*
- Problem is NP-hard! Look at relaxation.

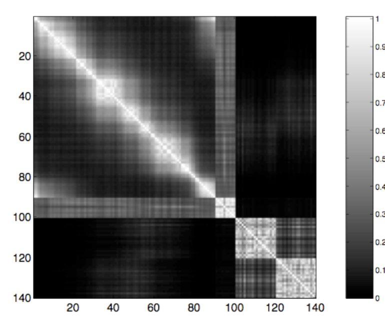
©Emily Fox 2013

Graph Terminology III

- Degree

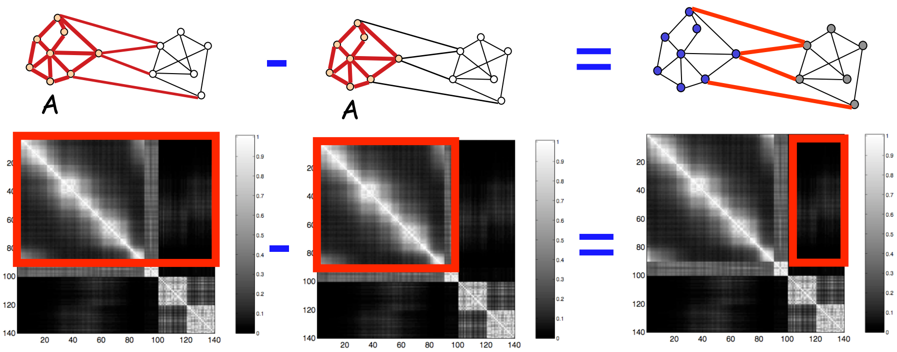


- Degree matrix



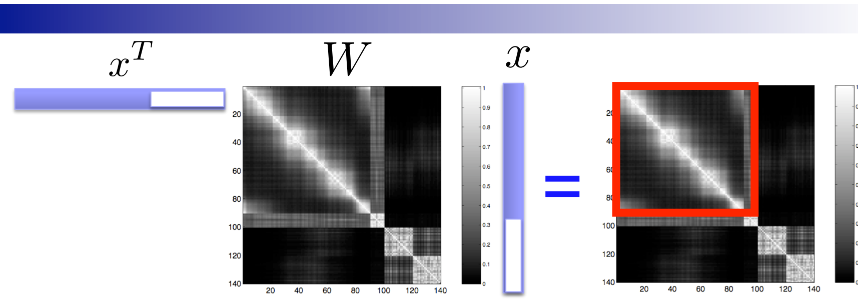
©Emily Fox 2013

Restating Cut Metric



©Emily Fox 2013

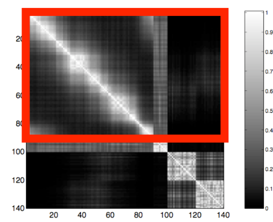
Restating Cut Metric



©Emily Fox 2013

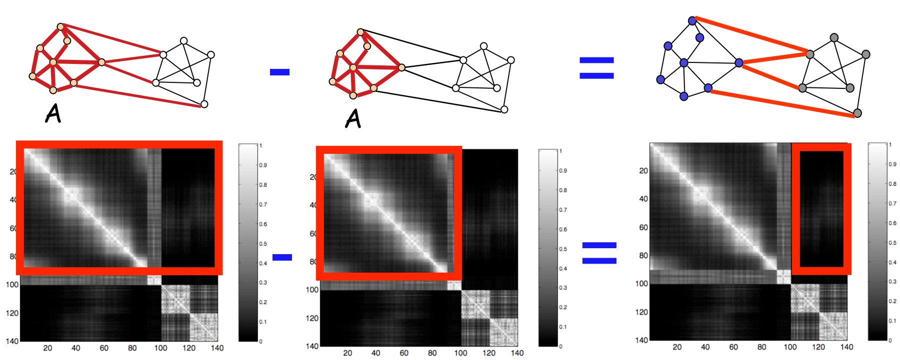
Restating Cut Metric

$$x^T \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & d_N \end{bmatrix} x =$$



©Emily Fox 2013

Restating Cut Metric



$$x^T D x \quad - \quad x^T W x \quad = \quad x^T (D - W) x$$

©Emily Fox 2013

Graph Laplacian

- Definition:
- Facts:
  - Symmetric, positive semi-definite
  - Eigenvalues
  - Invariance to self-edges
  - Inner product in *L* space

Relationship to Identifying Connected Components

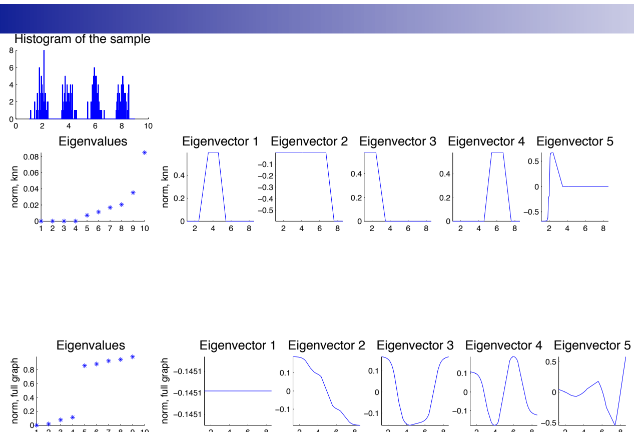
- Proposition:
  - The multiplicity *k* of eigenvalue 0 of *L* is equal to the number of connected components
- Proof: Assume graph is connected (*k*=1)

Relationship to Identifying Connected Components

- Proposition:
  - The multiplicity *k* of eigenvalue 0 of *L* is equal to the number of connected components

- Proof: Assume *k* connected components

Example – Mixture of Gaussians



From
von Luxburg
2007

Graph Laplacians and Ratio Cuts

- Ratio cuts for *k*=2
- Define cluster indicator variables:
- Properties:
- RatioCut
- Reformulating RatioCut problem

©Emily Fox 2013

Relaxation to Formulation

- Let *f* be arbitrary continuous vector
- Rayleigh-Ritz Theorem
  - Which vector maximizes objective subject to constraint that the vector is orthogonal to the first eigenvector and has bounded norm?

©Emily Fox 2013

Mapping Back to Partition

- To obtain partition, transform continuous *f* to a discrete indicator
- Cluster coordinates
- Return

Ratio Cuts for General k

- Define cluster indicator variables:

$$F_{ij} = \begin{cases} 1/\sqrt{|A_j|} \\ 0 \end{cases} \qquad F'_{\mathcal{A}} F_{\mathcal{A}} = I$$

- RatioCut

$$\mathrm{RatioCut}(A_1, \dots, A_k) = \sum_{i=1}^{k} f'_{\mathcal{A}i} L f_{\mathcal{A}i} = \mathrm{Tr}(F'_{\mathcal{A}} L F_{\mathcal{A}})$$

- Reformulating RatioCut problem

$$\min_{A_1, \dots, A_k} \mathrm{Tr}(F'_{\mathcal{A}} L F_{\mathcal{A}})$$

- Relaxation

$$\min_{F \in R^{N \times k}} \mathrm{Tr}(F' L F)$$

Ratio Cuts for General k

- Relaxation:

$$\min_{F \in R^{N \times k}} \mathrm{Tr}(F'LF) \quad \text{s.t. } F'F = I$$

- Solution:

- To obtain partition:

©Emily Fox 2013

Graph Laplacians and Norm. Cuts

- Normalized cuts for *k*=2
- Define cluster indicator variables:

- Properties:

- Ncut

- Reformulating Ncut problem

©Emily Fox 2013

Relaxation to Formulation

- Let *f* be arbitrary continuous vector
- Rayleigh-Ritz Theorem

©Emily Fox 2013

Normalized Cuts for General k

- Define cluster indicator variables:

$$F_{ij} = \begin{cases} 1/\sqrt{\mathrm{vol}(A_j)} & v_i \in A_j \\ 0 & ow \end{cases} \qquad \begin{aligned} F'_{\mathcal{A}} F_{\mathcal{A}} &= I \\ F'_{\mathcal{A}} D F_{\mathcal{A}} &= I \end{aligned}$$

- Reformulating RatioCut problem

$$\min_{A_1,\dots,A_k} \mathrm{Tr}(F'_{\mathcal{A}} L F_{\mathcal{A}}) \ \text{ s.t. } \ F'_{\mathcal{A}} D F_{\mathcal{A}} = I$$

- Relaxation

$$\min_{H \in R^{N\times k}} \mathrm{Tr}(H' D^{-1/2} L D^{-1/2} H) \quad \text{s.t. } H'H = I$$

- Solution:
  - H is matrix of first *k* eigenvectors of L_{sym}, which is equivalent to the approximate F being the first *k* eigenvectors of L_{rw}

©Emily Fox 2013

Random Walks on Graphs

- Stochastic process with random jumps from v_i to v_j wp:
- Transition matrix:
- Connection to graph Laplacian:
- Intuitively, want to partition graph s.t. random walk stays in cluster for a while and rarely jumps between clusters

©Emily Fox 2013

Random Walks on Graphs

- Assume that stationary distribution exists and is unique. Then,
- Proposition: $\mathrm{Ncut}(A, \bar{A}) = P(A \mid \bar{A}) + P(\bar{A} \mid A)$
- Proof:
- Minimizing normalized cuts is equivalent to minimizing the probability of transitioning between clusters

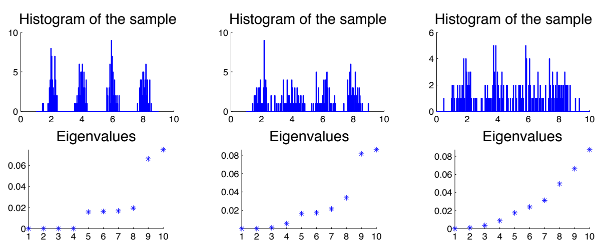
©Emily Fox 2013

Notes

- No guarantee to quality of approximation
- Sensitive to choice of similarity graph (see earlier)
- Which graph Laplacian to use?
  - If degrees in graph vary significantly, then Laplacians are quite different
  - In general, L_{rw} behaves the best
  - Volume gives better measure of within-cluster similarity than cardinality
  - Normalized cuts has consistency results, Ratio cuts does not

Notes

- Choosing the number of clusters *k* can be hard
  - Easy when clusters are well-separated



From
von Luxburg
2007

- k-means to return partition from solution to relaxation is *an* approach, but not the only