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

FIFA WORLD CUP
Brasil

k-Nearest Neighbor

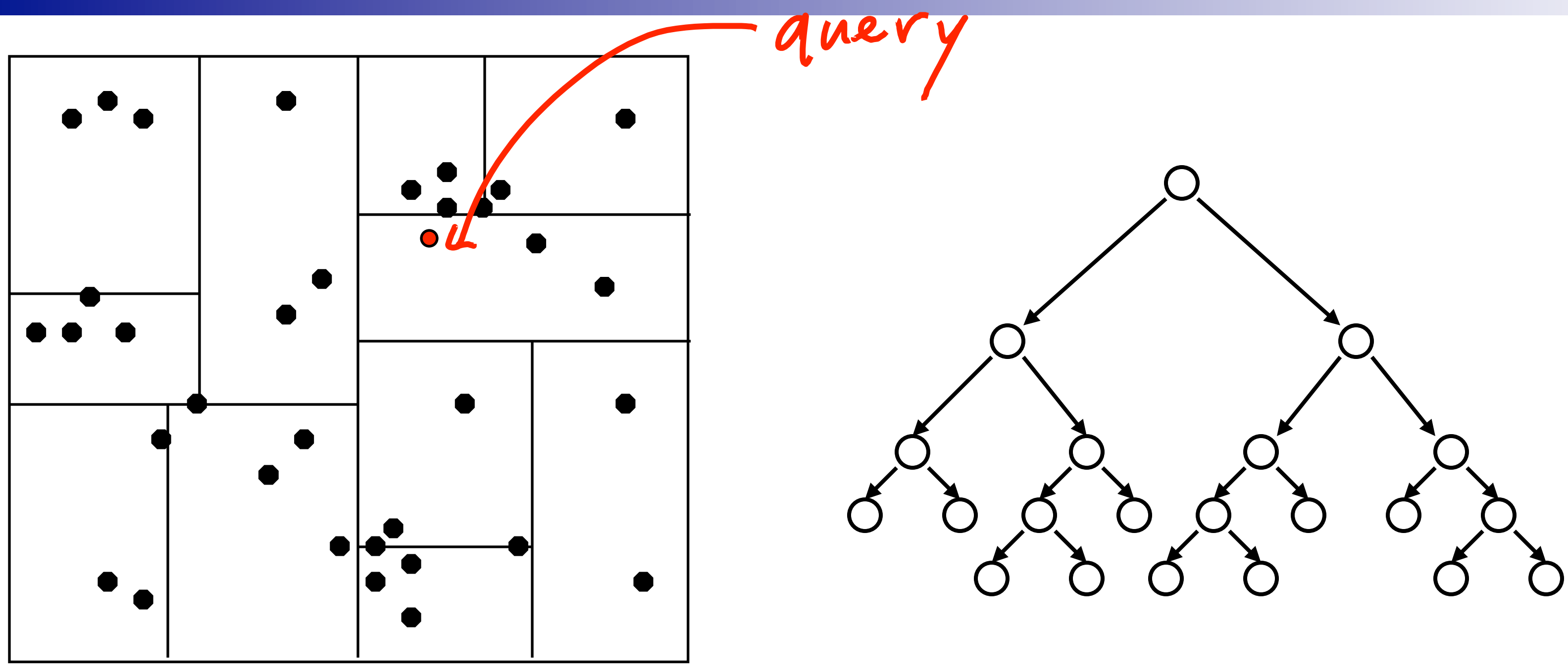
- Articles $X = \{x^1, \ldots, x^N\}, \quad x^i \in \mathbb{R}^d$
- Query: $x \in \mathbb{R}^d$
- k-NN
  - Goal: Find k articles in X closest x
  - Formulation:

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

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

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

Nearest Neighbor with KD Trees



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

Task 2: Cluster Documents

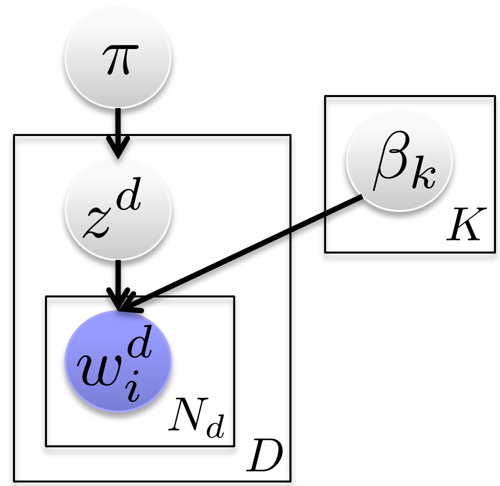
- **Setup**
  - **Input:** Corpus of documents
  - **Output:** Topic assignment per document

sports

world news

A Generative Model

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



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

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

word prob. for cluster/topic z^d

Bayesian approach:

$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$

$\beta_k \sim \mathrm{Dir}(\lambda_1, \ldots, \lambda_V)$

β_k is a V-dim pmf

size of vocab.

Inference

- **Two tasks**
  - **Point estimation:** $\hat{\theta}^{ML}$, or $\hat{\theta}^{MAP}$ ← $p(\theta)$ prior
    - Expectation-Maximization (EM)
  - **Characterize posterior:**
    - Gibbs sampling
    - Variational methods
    - Stochastic variational inference

EM Algorithm

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

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

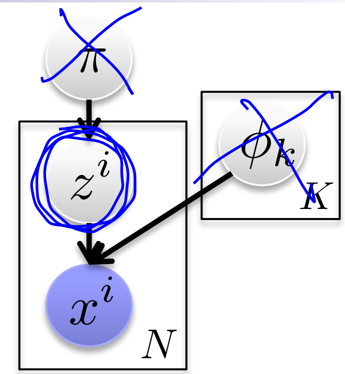
- **M-Step**

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

Collapsed Gibbs Sampling

$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z^i \sim \pi$$

$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$

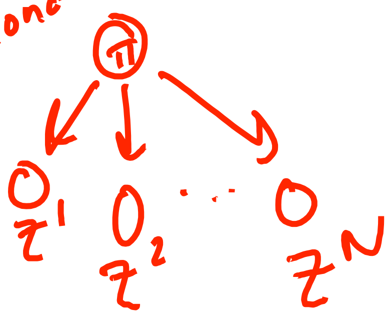


- Collapsed sampler

For $i = 1, \dots, N$

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

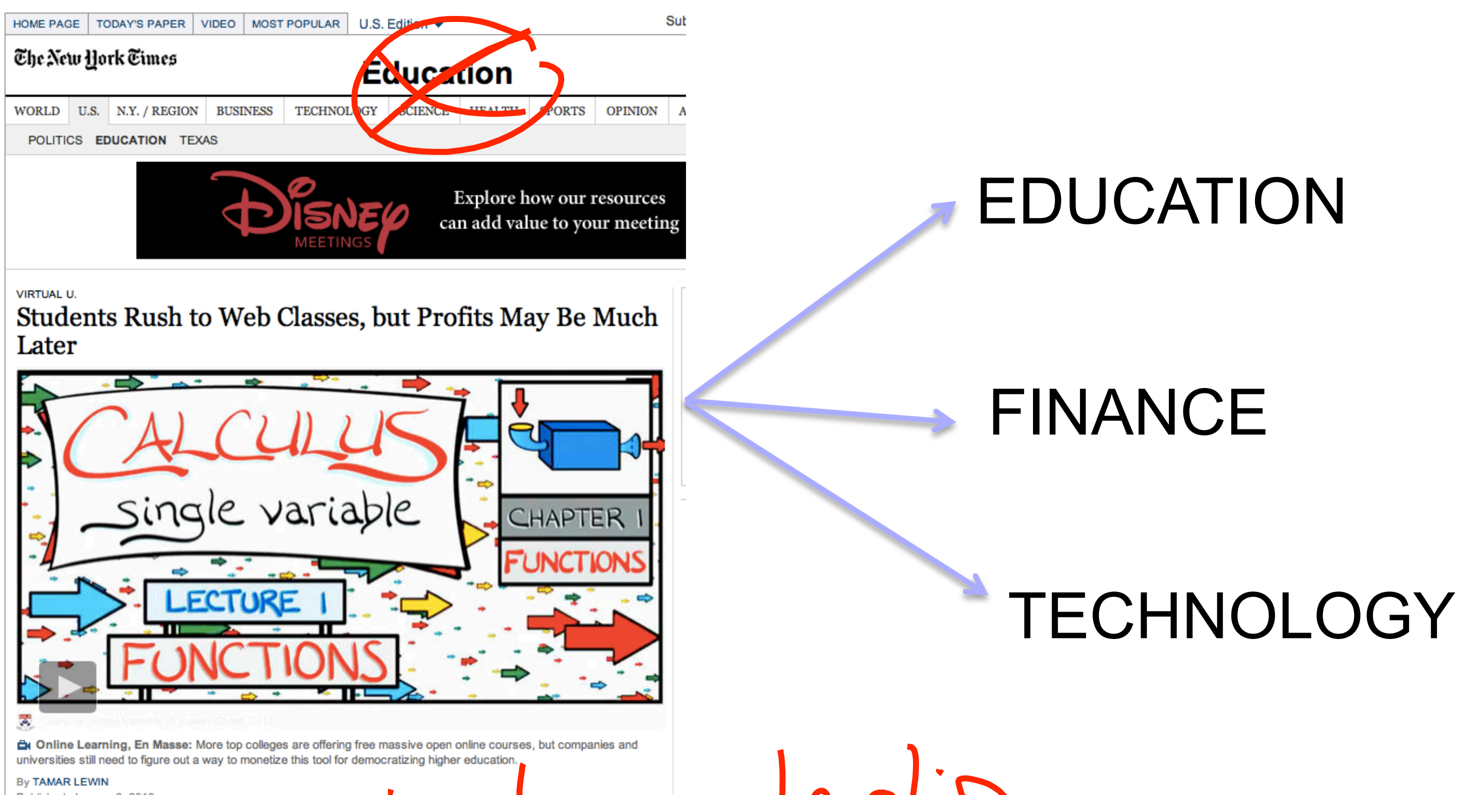
cond. ind.



marg. π

Task 3: Mixed Membership Model

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

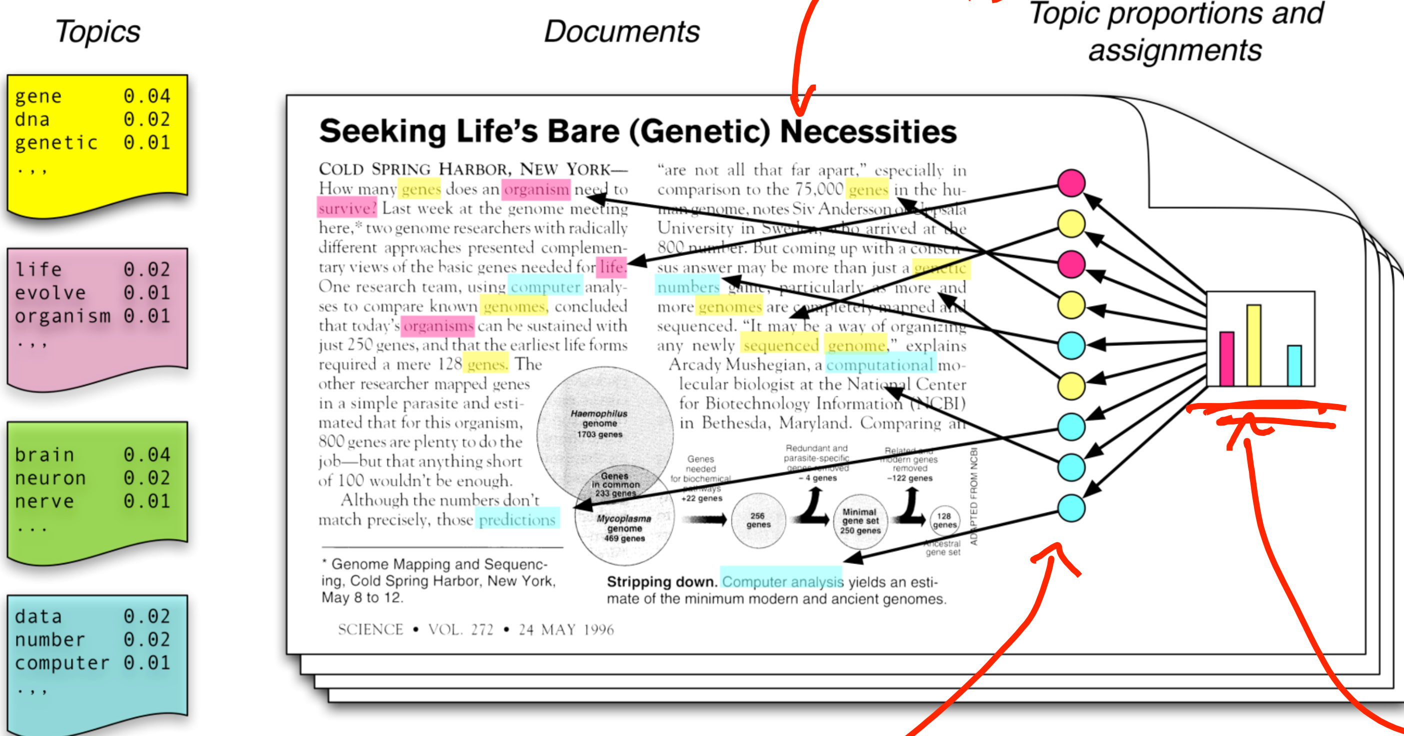


mixed membership

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 $p(\theta, z \mid x)$
  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 Q and want $q \in Q$ 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:

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

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

- Evidence lower bound (ELBO)

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

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

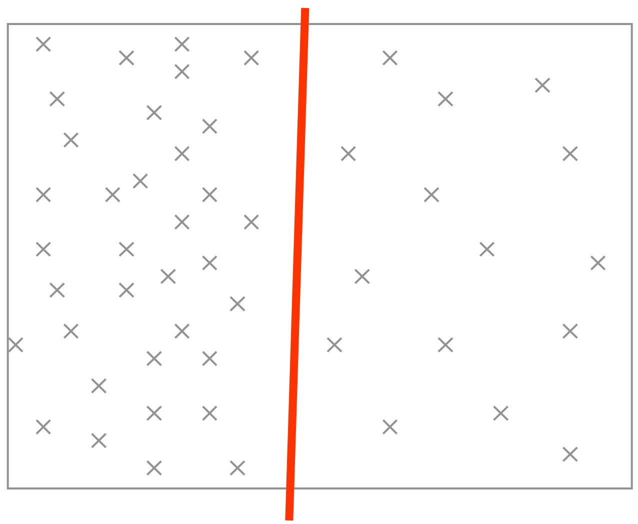
world news

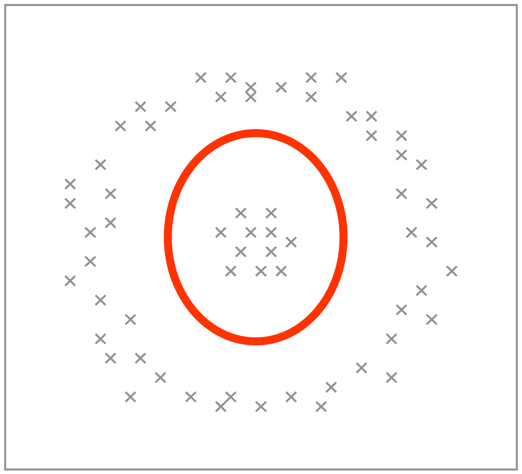
sports

BBC 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

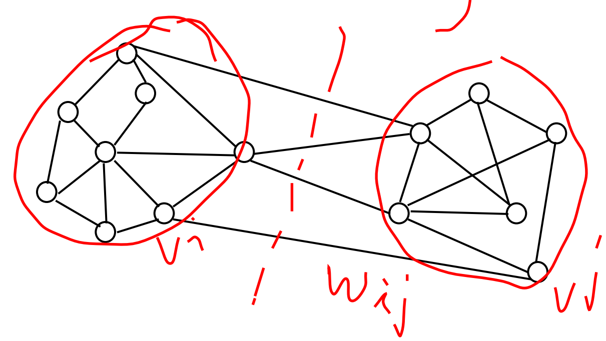
- Data: $x^1, \ldots, x^N$ x^i: doc i … maybe use tf-idf $x^i \in \mathbb{R}^V$
- Similarity metric:

  s_{ij} bt x^i and x^j (eg cosine similarity)

  $s_{ij} = s_{ji}$

- Similarity graph
  - Nodes v^i for each x^i
  - Edge weights

  w_{ij} = fcn of s_{ij}

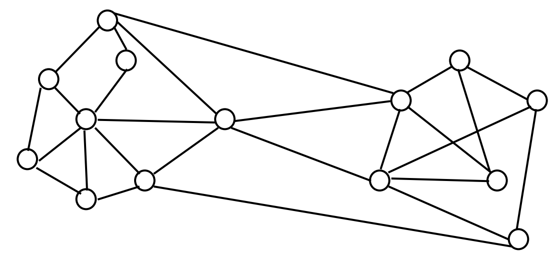


G = {V, E}

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

Types of Graphs

- **ε-neighborhood**:
  - Only include edges with distances < ε
  - Treat as unweighted $w_{ij} = \epsilon$



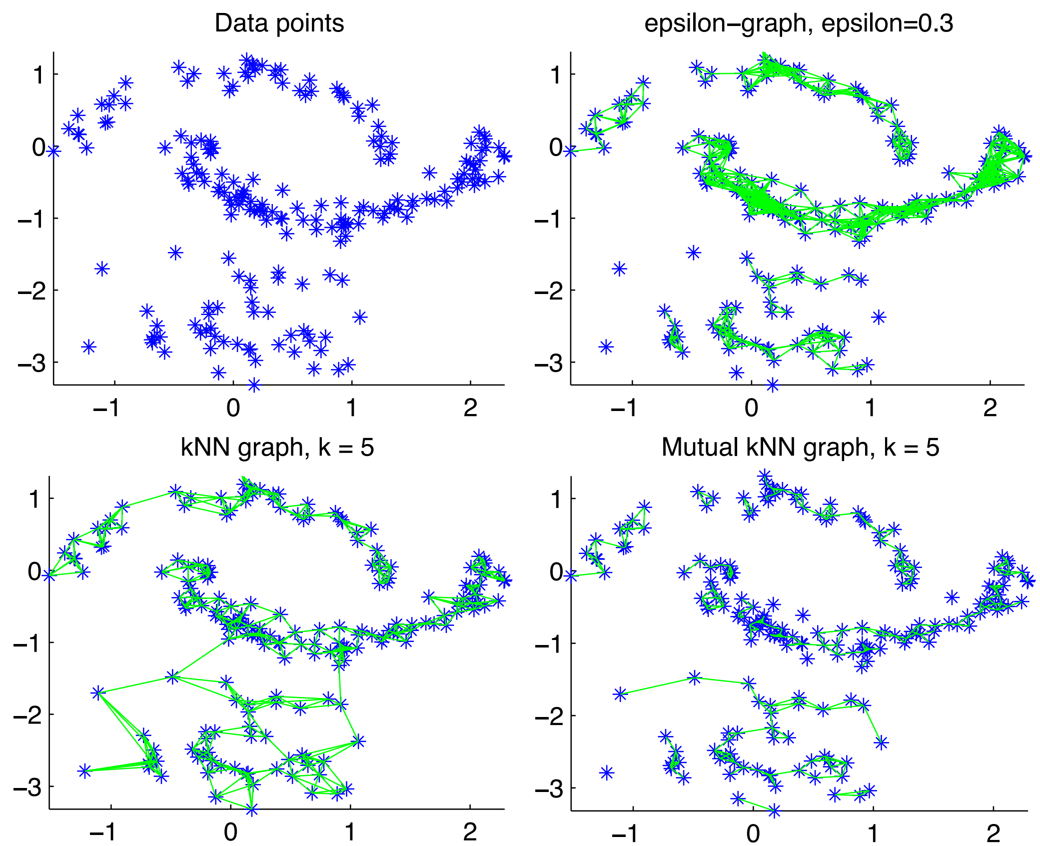
- **k-NN:**
  - Connect v_i and v_j if v_j is a k-NN of v_i
  - Weighted by similarity $s_{ij} = w_{ij}$
  - Directed → undirected

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

Issues with Choosing Graph

- Choosing graph construction techniques and parameters is non-trivial

Choice matters



From
von Luxburg
2007

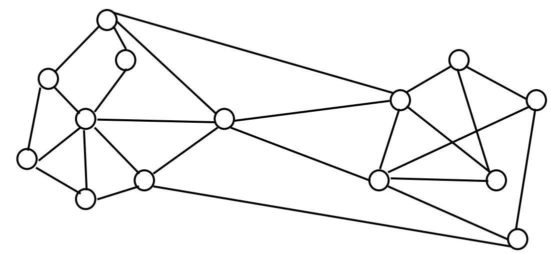
Graph Terminology I

- Weighted adjacency matrix

$W = (w_{ij})_{i,j=1,\dots,N}$

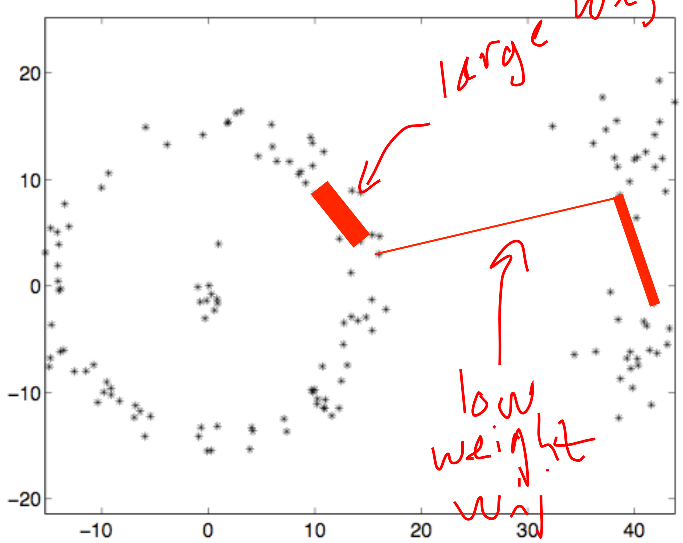
$w_{ij} = 0 \rightarrow$ no edge bt v_i and v_j

$w_{ij} \geq 0$



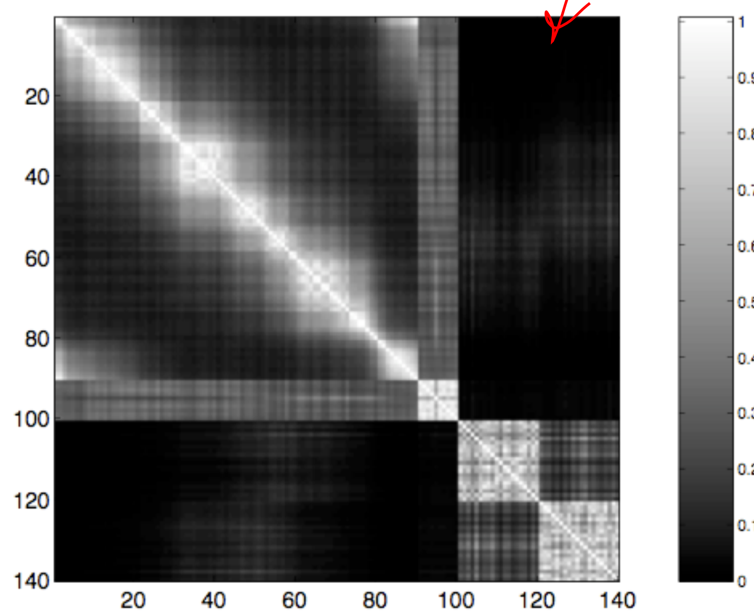
W

Sparse



large w_{ij}

low weight w_{ij}



Graph Cuts

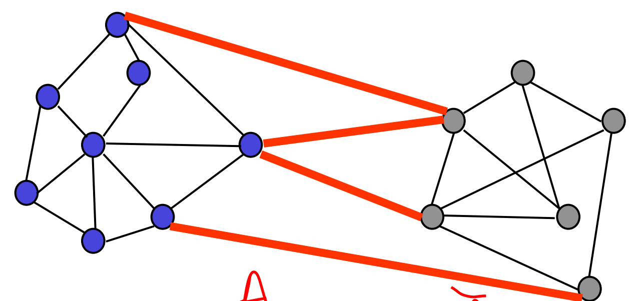
- **Problem:** Partition graph such that edges between groups have low weights
- Define: $\underline{W(A,B)} = \sum_{i \in A, j \in B} w_{ij}$
- MinCut problem:

$$\mathrm{Cut}(A_1,\ldots,A_k) \triangleq \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A}_i)$$

Choose

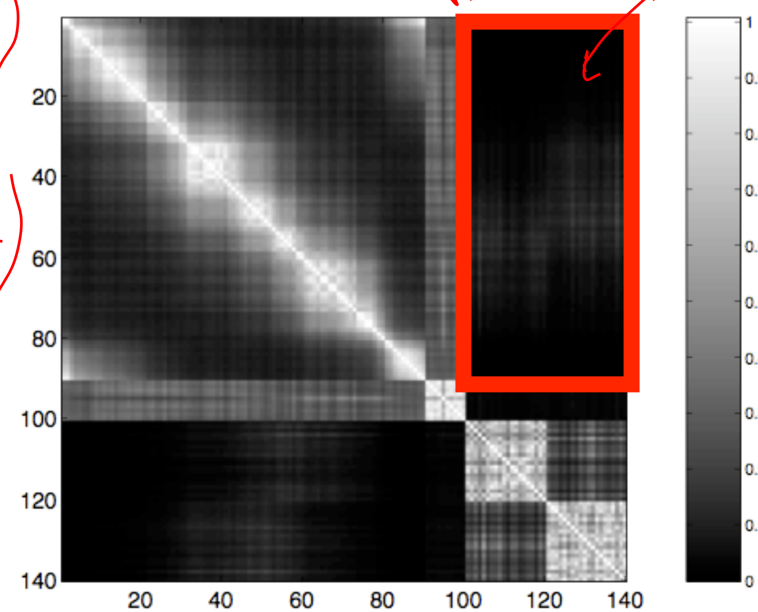
$$A_1,\ldots,A_k = \underset{\substack{A_1,\ldots,A_k \\ \text{disjoint} \subset V}}{\arg\min}\ \mathrm{Cut}(A_1,\ldots,A_k)$$

- Trivial to solve for *k*=2



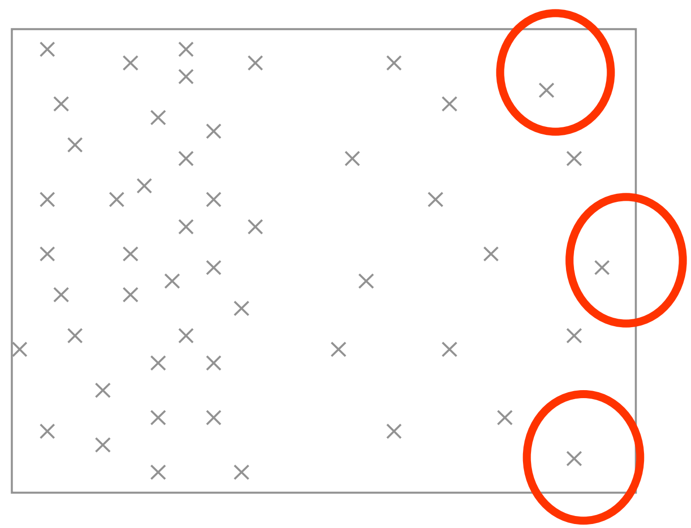
$\underbrace{\qquad\qquad}_{}$ A $\quad$ $\underbrace{\qquad}_{}$ $\bar{A}$

$\mathrm{cut}(A,\bar{A})$



Issues with MinCut

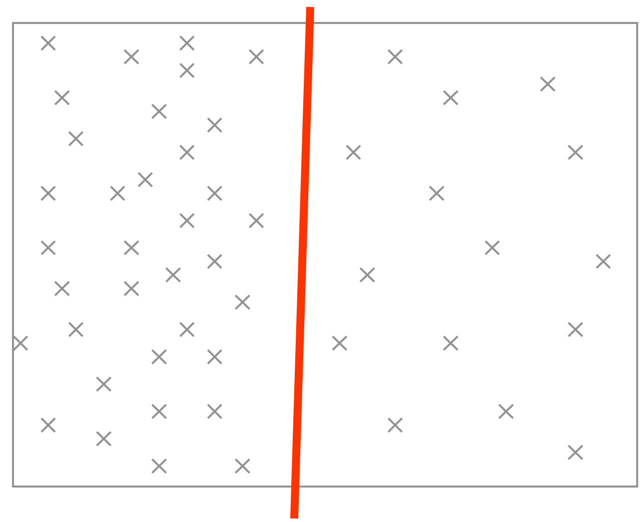
- MinCut favors isolated clusters



nothing working against this

Cuts Accounting for Size

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



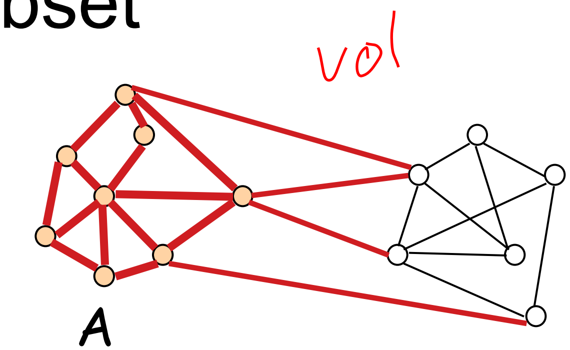
- First need more graph terminology…

to measure "size" of the clusters

Graph Terminology II

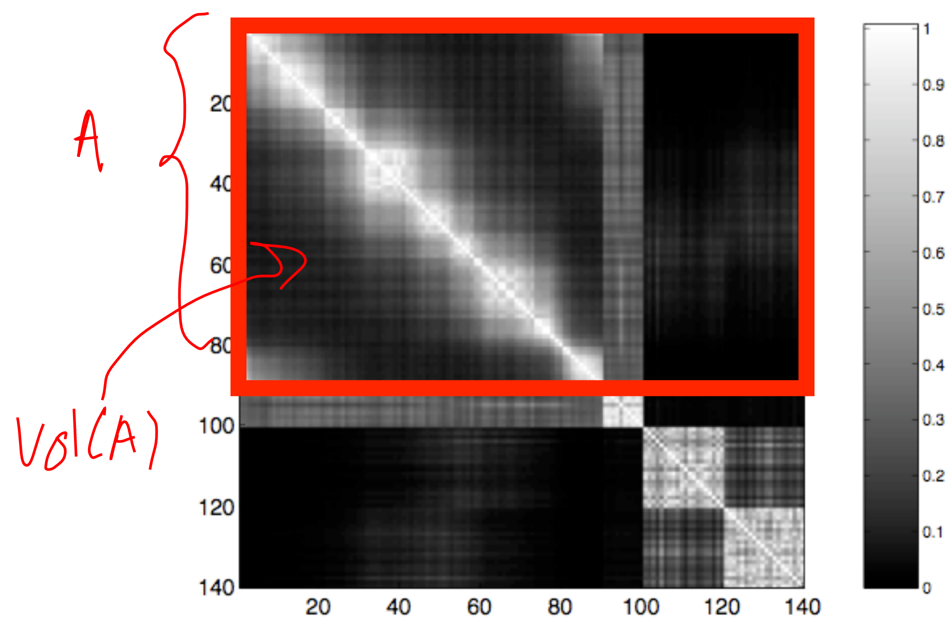
- Two measures of size of a subset
  - Cardinality:

$|A|$ = # of vertices in A



  - Volume:

$\mathrm{vol}(A) = \sum_{i \in A} \sum_{j=1}^{N} w_{ij}$



Cuts Accounting for Size

- Ratio cuts (RatioCut)
  - *k*=2

$$\mathrm{RatioCut}(A, \bar{A}) = \mathrm{cut}(A, \bar{A}) \underbrace{\left(\frac{1}{|A|} + \frac{1}{|\bar{A}|} \right)}$$

min when $|A|$ and $|\bar{A}|$ coincide

  - General *k*

$$\mathrm{RatioCut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_i \frac{W(A_i, \bar{A}_i)}{|A_i|}$$

- Normalized cuts (Ncut)
  - *k*=2

$$\mathrm{Ncut}(A, \bar{A}) = \mathrm{cut}(A, \bar{A}) \left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(\bar{A})} \right)$$

  - General *k*

$$\mathrm{Ncut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_i \frac{W(A_i, \bar{A}_i)}{\mathrm{vol}(A_i)}$$

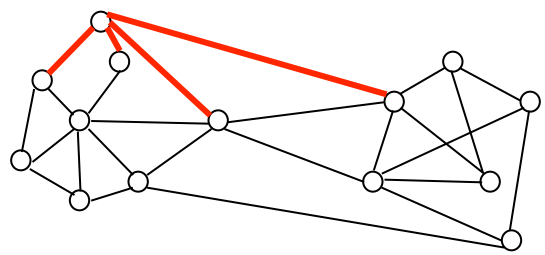
- Problem is NP-hard! Look at relaxation.

Graph Terminology III

- Degree

$d_i = \sum_{j=1}^{N} w_{ij}$

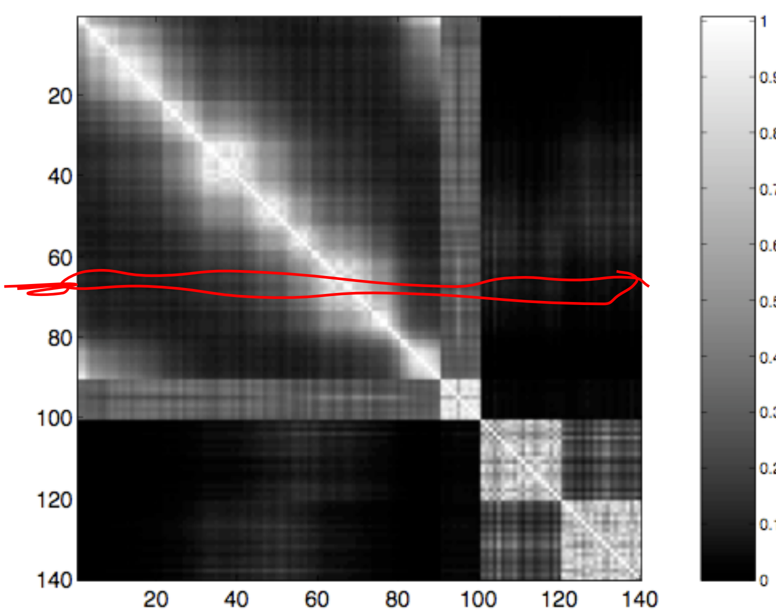
only counts neighbors



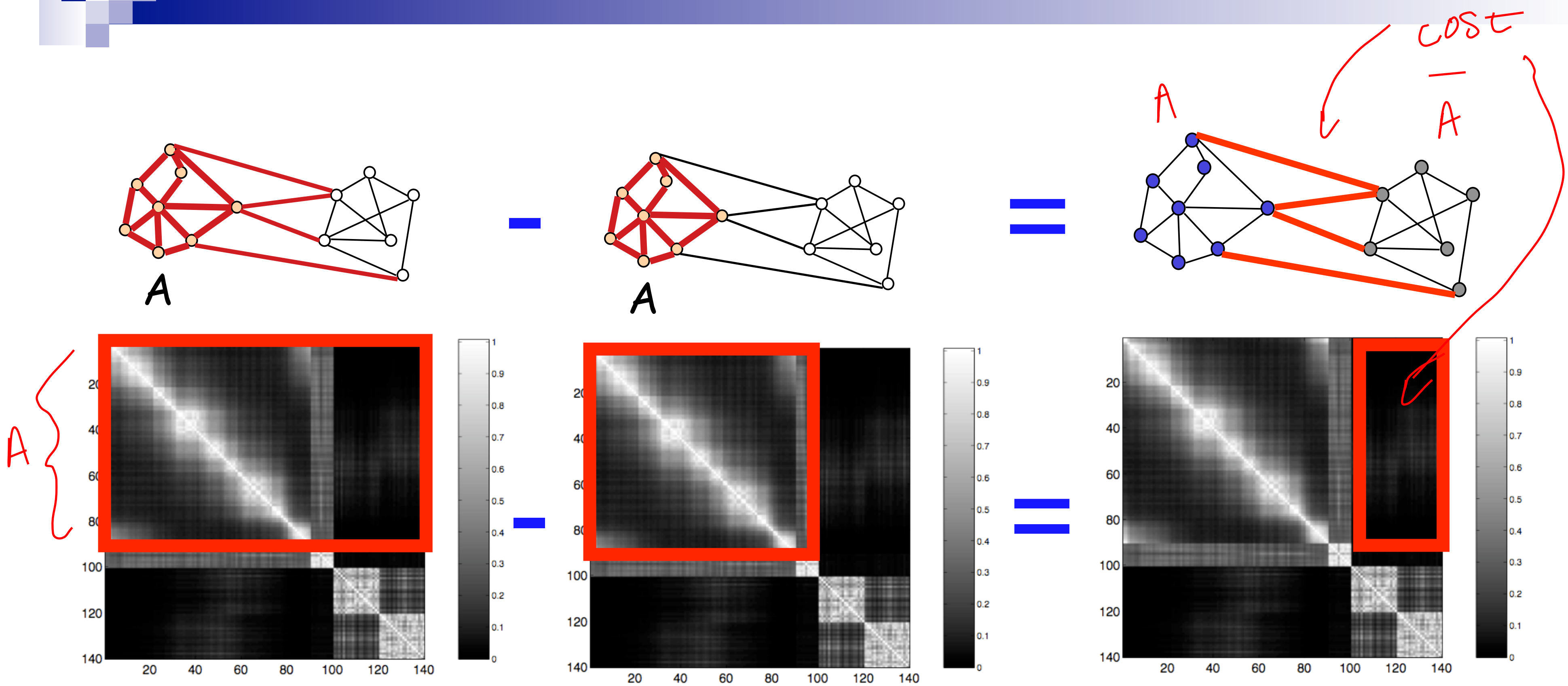
- Degree matrix

$$D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_N \end{bmatrix}$$

i

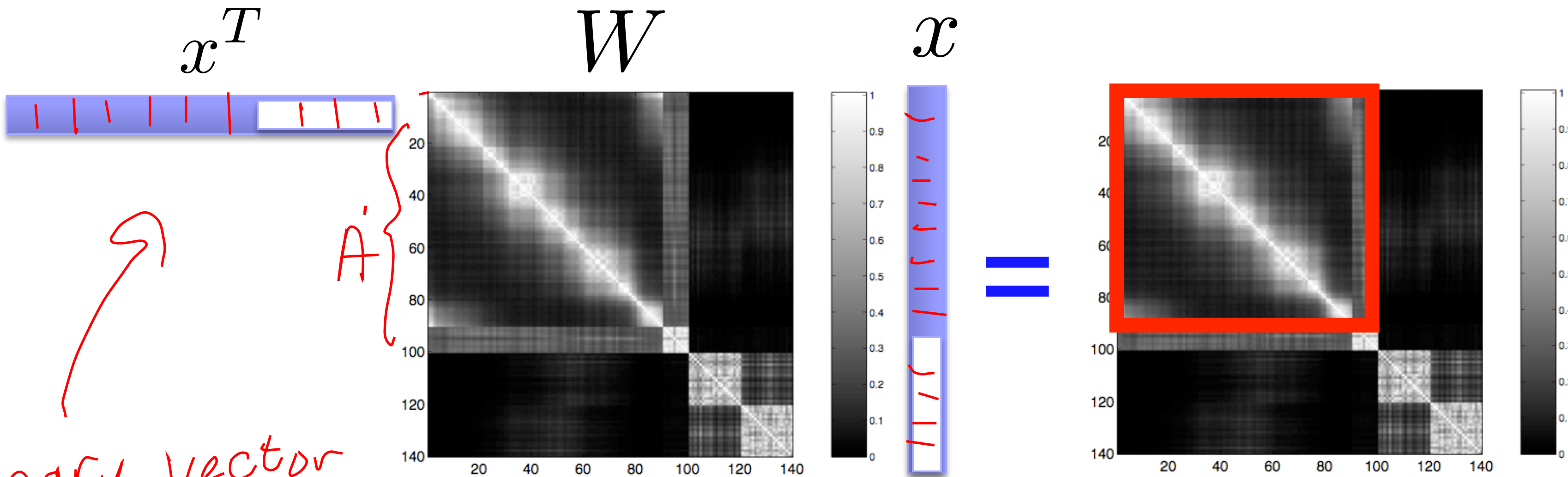


Restating Cut Metric



Volume(A) − "association(A)" = cut cost

Restating Cut Metric – Assoc.



binary vector
that's an
indicator on set A

$\mathbb{1}_A$

Restating Cut Metric - Volume

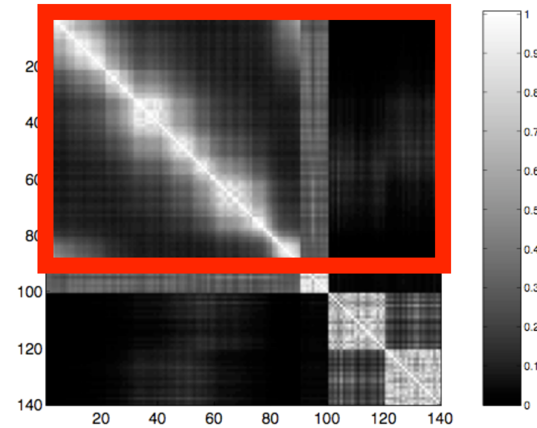
x^T

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & d_N \end{bmatrix}$$

x

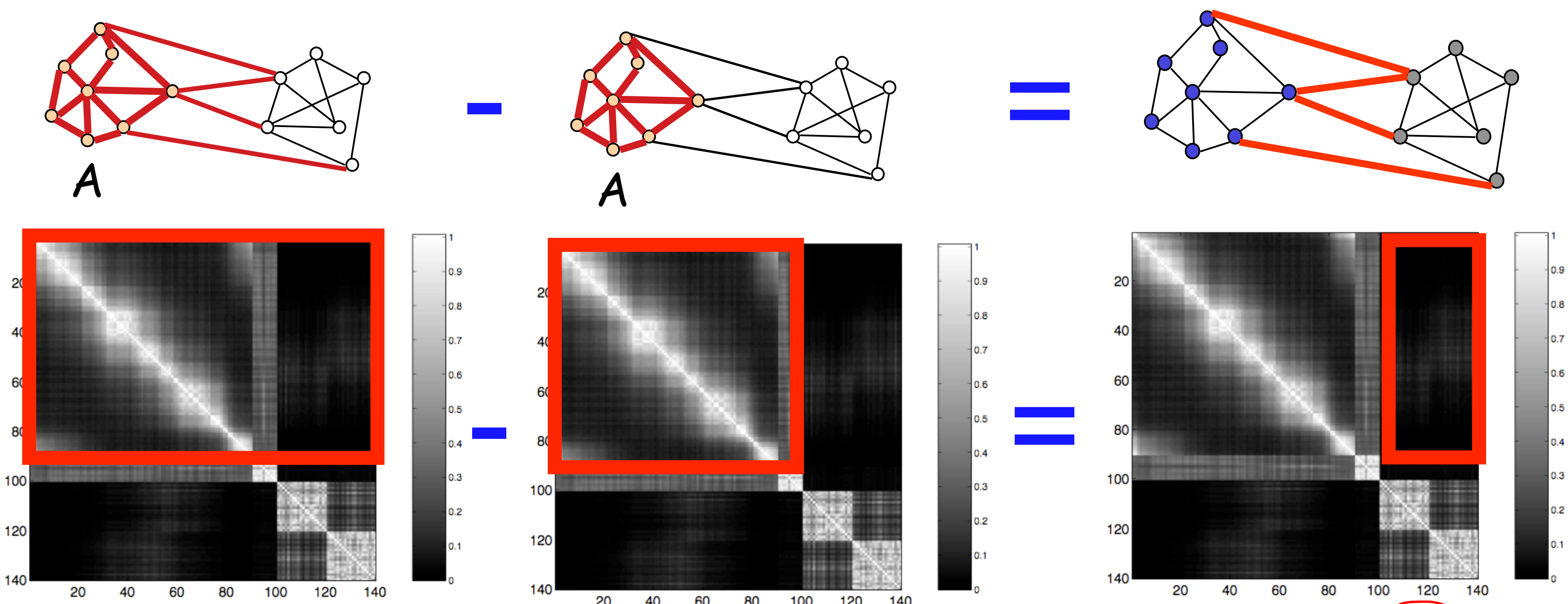
sum of weights in row 2

=



Vol(A)

Restating Cut Metric



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

$\mathbb{1}_A$

cut(A, $\bar{A}$)

key

Graph Laplacian

- Definition: $L = D - W$

- Facts:
  - Symmetric, positive semi-definite
  - Eigenvalues

    $$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$$

    eigvec $u_1 = \mathbb{1}$
  - Invariance to self-edges

    $$\left.\begin{aligned} L_{ii} &= d_i - w_{ii} \\ L_{ij} &= -w_{ij} \end{aligned}\right\} \text{ don't depend on } w_{ii}$$
  - ~~Inner product~~ norm in L space

    $$\forall f \in \mathbb{R}^N \quad f' L f = \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2$$

    useful later

Relationship to Identifying Connected Components

- Proposition:
  - The multiplicity *k* of eigenvalue 0 of *L* is equal to the number of connected components $A_1, \ldots, A_k$

    Furthermore, $u_1, \ldots, u_k = \mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$
- Proof: Assume graph is connected (*k*=1)

$$0 = u_1' L u_1 = \sum_{i,j} w_{ij}(u_{1i} - u_{1j})^2$$

If $w_{ij} > 0 \Rightarrow u_{1i} = u_{1j}$

Since $\exists$ a path bt all i, j, then

$$u_1 = \text{constant} = \mathbb{1}$$

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 $A_1, \ldots, A_k$

Assume WLOG that they're ordered

$$L = \begin{pmatrix} L_1 & & 0 \\ & \ddots & \\ 0 & & L_k \end{pmatrix}$$

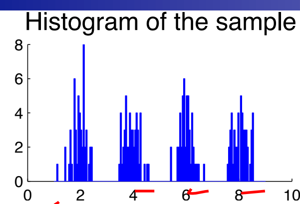
graph Laplacian for subgraph A_1

$\text{eigval}(L) = \bigcup_i \text{eigval}(L_i)$

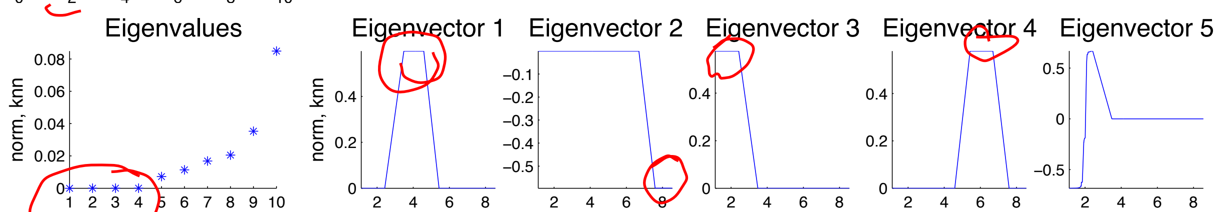
each has 1 eigval equal to 0 + corr. eigvec $\mathbb{1}_{A_i}$

$\Rightarrow$ eigvecs are indicators on the partition

Example – Mixture of Gaussians



← MoG w/ 4 clusters

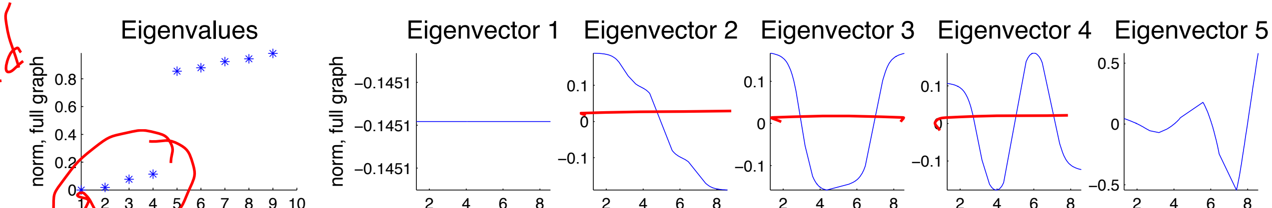


4 disc. comp. in graph

4 eigenval = 0

indicators on the partition

From
~~von Luxburg 2007~~



connected graph

1 eigval = 0

small

still info here

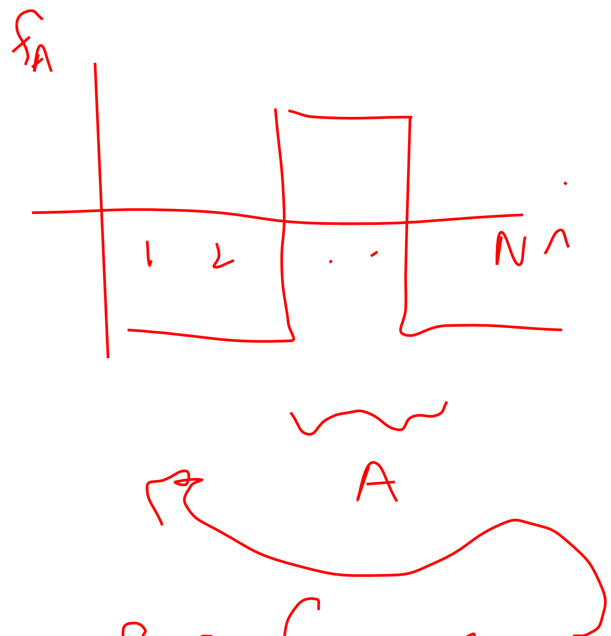
Graph Laplacians and Ratio Cuts

- Ratio cuts for *k*=2

$V = (A, \bar{A})$

- Define cluster indicator variables:

$$f_{A_i} = \begin{cases} \sqrt{|\bar{A}|/|A|} & v_i \in A \\ -\sqrt{|A|/|\bar{A}|} & v_i \in \bar{A} \end{cases}$$



- Properties:

$$\sum_i f_{A_i} = |A|\sqrt{|\bar{A}|/|A|} - |\bar{A}|\sqrt{|A|/|\bar{A}|} = 0$$

$$\|f_A\|^2 = N$$

- RatioCut

$$\text{RatioCut}(A, \bar{A}) = \frac{f_A' L f_A}{|V|} \quad \text{for } f_A \text{ as}$$

- Reformulating RatioCut problem

$$\min_{A \subset V} f_A' L f_A \quad \text{s.t. } f_A \text{ defined as above, } f_A \perp \mathbb{1},\ \|f_A\| = \sqrt{N}$$

f_{A_i} are in a discrete set

Relaxation to Formulation

- Let *f* be arbitrary continuous vector

$$\min_{f \in \mathbb{R}^N} f' L f \quad \text{s.t.} \quad f \perp \mathbb{1} \quad \|f\| = \sqrt{N}$$

graph Laplacian

1st eigvec of L

const

- Rayleigh-Ritz Theorem
  - Which vector maximizes objective subject to constraint that the vector is orthogonal to the first eigenvector and has bounded norm?

$f = u_2(L)$ = eigvec assoc. w/ 2nd smallest eigval

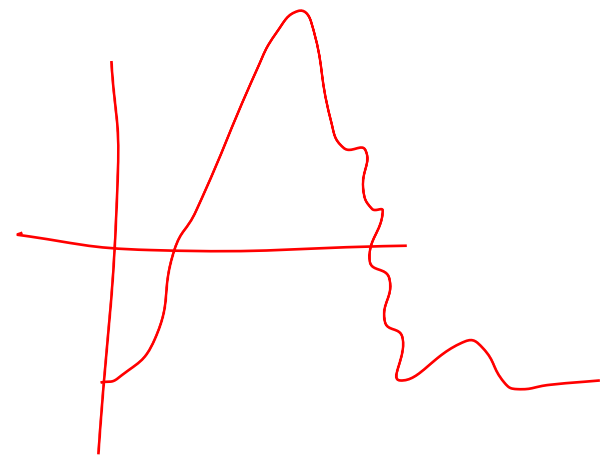
Mapping Back to Partition

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

- Cluster coordinates

$f_i \in \mathbb{R}$ into $C, \bar{C}$

using k-means



- Return

$$\begin{cases} v_i \in A & \text{if } f_i \in C \\ v_i \in \bar{A} & \text{if } f_i \in \bar{C} \end{cases}$$

Ratio Cuts for General k

- Define cluster indicator variables:

$$F_{ij} = \begin{cases} 1/\sqrt{|A_j|} \\ 0 \end{cases}$$

$\forall i \in A_j$

ow

$F_{\mathcal{A}} \in R^{N \times k}$

$$F'_{\mathcal{A}} F_{\mathcal{A}} = I$$

- RatioCut

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

- Reformulating RatioCut problem

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

and F_A w/ $F_A' F_A = I$

- Relaxation

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

s.t. $F'F = I$

Ratio Cuts for General k

- Relaxation:

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

standard trace min problem

- Solution:

$\Rightarrow$ choose F containing first k eig vec(L)

$$\begin{bmatrix} | & & | \\ u_1, & \cdots & u_k \\ | & & | \end{bmatrix}$$

- To obtain partition:

cluster rows of F using k-means

$$F = \begin{array}{c|c} & 1 \quad 2 \quad \cdots \quad k \\ \hline \begin{array}{c} \text{obs } 1 \\ \vdots \\ \\ \text{obs } N \end{array} & \left[\begin{array}{c} \\ \text{———} \\ \\ \\ \end{array} \right] \end{array}$$

if obs i is in cluster w/ obs j, then the rows are the same

Graph Laplacians and Norm. Cuts

- Normalized cuts for *k*=2

$A, \bar{A}$

- Define cluster indicator variables:

$$f_{A_i} = \begin{cases} \sqrt{\mathrm{vol}(\bar{A})/\mathrm{vol}(A)} & v_i \in A \\ -\sqrt{\mathrm{vol}(A)/\mathrm{vol}(\bar{A})} & v_i \in \bar{A} \end{cases}$$

- Properties:

$$(Df_A)'\mathbb{1} = 0 \quad \text{and} \quad f_A' D f_A = \mathrm{vol}(V)$$

- Ncut

$$\mathrm{Ncut}(A, \bar{A}) = \frac{f_A' L f_A}{\mathrm{vol}(V)}$$

- Reformulating Ncut problem

$$\min_{A \subset V} f_A' L f_A \quad \text{s.t.} \quad Df_A \perp \mathbb{1} \quad \text{and} \quad f_A' D f_A = \mathrm{vol}(V)$$

Relaxation to Formulation

- Let *f* be arbitrary continuous vector

$$\min_{f \in \mathbb{R}^N} f' L f \quad \text{s.t.} \quad Df \perp \mathbb{1} \quad f' D f = \mathrm{vol}(V)$$

$$\Updownarrow \quad f = D^{-1/2} g$$

$$\min_{g \in \mathbb{R}^n} g' \underbrace{D^{-1/2} L D^{-1/2}}_{\triangleq L_{sym}} g \quad \text{s.t.} \quad g \perp \underbrace{D^{1/2} \mathbb{1}}_{\text{1st eigvec of } L_{sym}} \quad \underbrace{\|g\|^2 = \mathrm{vol}(V)}_{\text{const}}$$

- Rayleigh-Ritz Theorem

$$g = u_2(L_{sym})$$

$$\Rightarrow f = D^{-1/2} u_2(L_{sym}) = u_2(L_{rw}) \quad \leftarrow I - D^{-1}W$$

$$\boxed{\text{equiv to } f \text{ soln of } \quad Lu = \lambda D u}$$

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}

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

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:

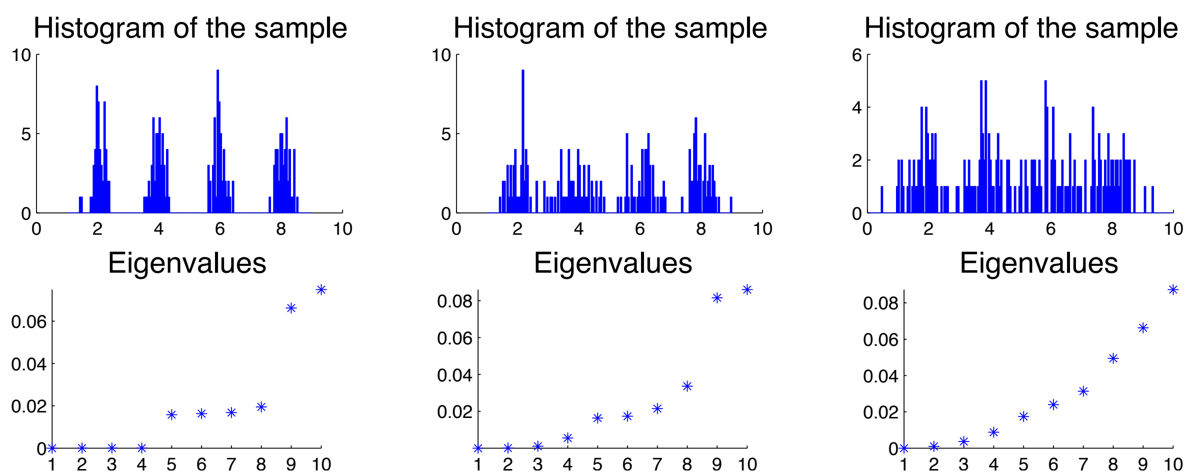
- Minimizing normalized cuts is equivalent to minimizing the probability of transitioning between clusters

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