Exact semidefinite relaxation for the clustering and biclustering problems

Brendan Ames

Institute for Mathematics and its Applications
University of Minnesota

IMA Postdoc seminar
Tuesday, November 8, 2011

Outline

1. SDP relaxation for clustering
2. Biclustering
3. Numerical examples

The clustering problem

- Clustering: Want to partition a given data set so that items in each cluster are similar to each other and items not in the same cluster are dissimilar.
- Intractable in general.
- If data is actually clustered, can we identify the clusters efficiently?

The clustering problem

- Clustering: Want to partition a given data set so that items in each cluster are similar to each other and items not in the same cluster are dissimilar.
- Intractable in general.
- If data is actually clustered, can we identify the clusters efficiently?
    - **Answer:** Yes, under appropriate assumptions on the data (Ostrovsky et al. 2006, Ames-Vavasis 2010, Oymak-Hassibi 2011, Jalali et al 2011, Yu et al 2011).

Graph clustering

- Can model data as weighted complete graph $G_S = (V, E, W)$ called the similarity graph.
- V is the set of items in the data.
- Level of similarity between any two items i, j is indicated by the weight W_{ij} on the edge ij.
- If i and j are highly similar, then W_{ij} is large.
- A cluster in the data will induce a complete subgraph of G_S with large average edge-weight.

The Average Weight k-Disjoint Clique Problem

- A k-disjoint-clique subgraph of a graph G is a subgraph of G whose set of nodes consists of k disjoint cliques.
- Average weight k-disjoint-clique problem (WKDC): find a k-disjoint-clique subgraph such that the sum of the average edge weights of the k complete subgraphs induced by the cliques is maximized.

Nonconvex QP formulation

- Can formulate WKDC as

$$\max_{S=\{\mathbf{v}^1,\ldots,\mathbf{v}^k\}} \quad \sum_{i=1}^{k} \frac{(\mathbf{v}^i)^T W \mathbf{v}^i}{(\mathbf{v}^i)^T \mathbf{v}^i}$$

$$\text{s.t.} \quad (\mathbf{v}^i)^T \mathbf{v}^j = 0 \quad \text{if } i \neq j$$

$$\mathbf{v}^i \in \{0,1\}^V \quad i = 1,\ldots,k.$$

Nonconvex QP formulation

- Can formulate WKDC as

$$\max_{S=\{\mathbf{v}^1,\ldots,\mathbf{v}^k\}} \quad \sum_{i=1}^{k} \frac{(\mathbf{v}^i)^T W \mathbf{v}^i}{(\mathbf{v}^i)^T \mathbf{v}^i}$$

$$\text{s.t.} \quad (\mathbf{v}^i)^T \mathbf{v}^j = 0 \quad \text{if } i \neq j$$

$$\mathbf{v}^i \in \{0,1\}^V \quad i = 1,\ldots,k.$$

- Can relax to the nonconvex quadratic program

$$\max_X \quad Tr(X^T W X)$$

$$\text{s.t.} \quad \text{cols of } X \text{ are orthonormal}$$

$$X \in \mathbf{R}_+^{V \times k}$$

Matrix lifting procedure

- A standard relaxation technique for combinatorial optimization problem:

$$\begin{aligned} \max \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b},\ \mathbf{x} \in \{0,1\}^n \end{aligned}$$

  is to replace the integer variable $\mathbf{x}$ with the matrix variable $X = \mathbf{x}\mathbf{x}^T$.

Matrix lifting procedure

- A standard relaxation technique for combinatorial optimization problem:

$$\begin{aligned} \max \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b},\ \mathbf{x} \in \{0,1\}^n \end{aligned}$$

  is to replace the integer variable $\mathbf{x}$ with the matrix variable $X = \mathbf{x}\mathbf{x}^T$.

- Can relax the original problem as a semidefinite program with additional rank constraints.

$$\begin{aligned} \max \quad & Tr(CX) \\ \text{s.t.} \quad & \mathcal{A}(X) = \tilde{\mathbf{b}} \\ & rank(X) = 1 \\ & X \succeq 0 \end{aligned}$$

- Relax further to a SDP by replacing the rank constraint.

Relaxation to SDP

- We replace $X = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k]$ with $\tilde{X} = \sum_i \mathbf{x}^i (\mathbf{x}^i)^T$. Get rank-constrained SDP

$$\begin{aligned} \max \quad & Tr(WX) \\ \text{s.t.} \quad & X\mathbf{e} \leq \mathbf{e} \\ & rank(X) = k \\ & X_{ij} \geq 0 \quad \forall\, i, j \in V \\ & X \succeq 0. \end{aligned}$$

Relaxation to SDP

- We replace $X = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k]$ with $\tilde{X} = \sum_i \mathbf{x}^i (\mathbf{x}^i)^T$. Get rank-constrained SDP

$$\begin{aligned} \max \quad & Tr(WX) \\ \text{s.t.} \quad & X\mathbf{e} \leq \mathbf{e} \\ & Tr(X) = k \\ & X_{ij} \geq 0 \quad \forall\, i,j \in V \\ & X \succeq 0. \end{aligned}$$

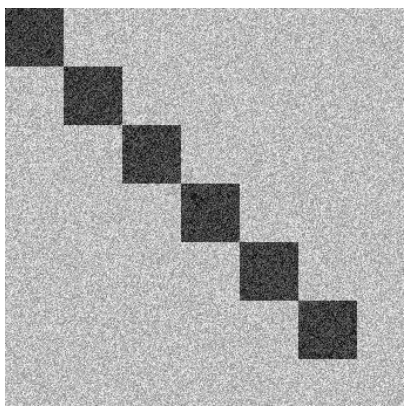
- Convexify by replacing $rank(X)$ with $Tr(X)$.
- Equivalent to underestimating $rank(X)$ with $\|X\|_*$:
  - $X\mathbf{e} \leq \mathbf{e} \Rightarrow \|X\|_* \leq rank(X)$ for all feasible X.
  - $X \succeq 0 \Rightarrow \|X\|_* = Tr(X)$ for all feasible X.

The planted cluster model

- The SDP relaxation may not return a good solution for all program inputs.
- We consider only inputs where a good partition into k exists.
- There is k-disjoint-clique subgraph $\{C_1, \ldots, C_k\}$ such that intraclique edges have higher weight than interclique edges.
- Randomly sample entries of W from two distributions Ω_1, Ω_2:
  - If u, v in same clique, sample from Ω_1: $E[W_{uv}] = \alpha$.
  - Otherwise sample from Ω_2: $E[W_{uv}] = \beta$ for $\alpha > \beta$.

The planted cluster model

- The SDP relaxation may not return a good solution for all program inputs.
- We consider only inputs where a good partition into k exists.
- There is k-disjoint-clique subgraph $\{C_1, \ldots, C_k\}$ such that intraclique edges have higher weight than interclique edges.
- Randomly sample entries of W from two distributions Ω_1, Ω_2:
  - If u, v in same clique, sample from Ω_1: $E[W_{uv}] = \alpha$.
  - Otherwise sample from Ω_2: $E[W_{uv}] = \beta$ for $\alpha > \beta$.



Why is this a good model?

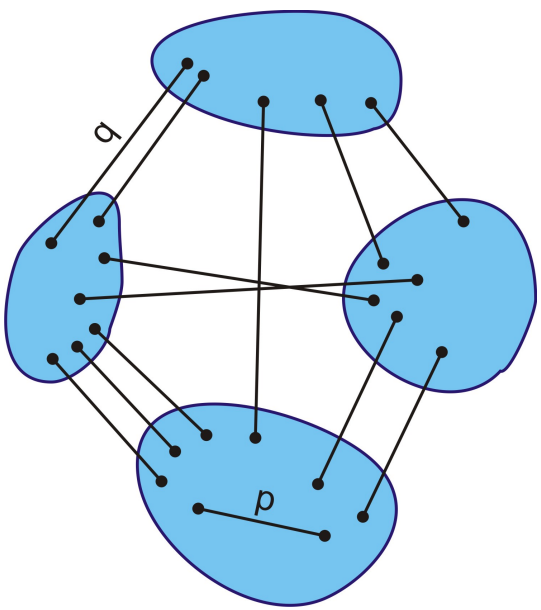
- Under this model there exists a partition such that the resulting clustering exhibits high similarity within cluster and low similarity between clusters.
- Contains several existing models for clustered data as special cases.

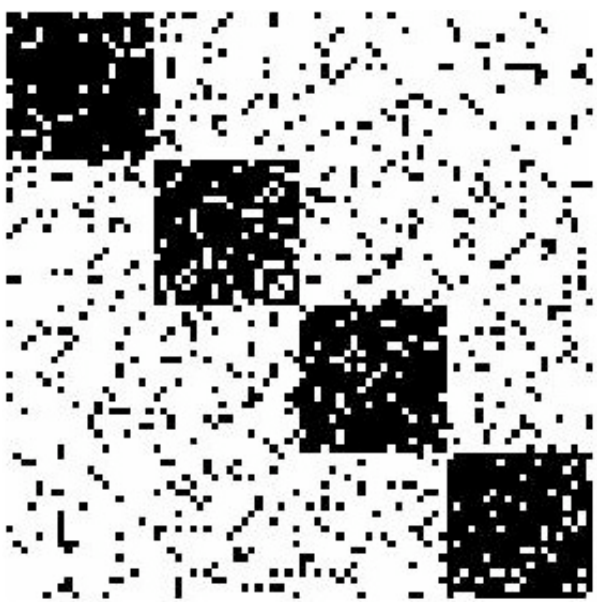
Example: Block stochastic model

- In the block stochastic model have subsets of nodes $\{C_1, \ldots, C_k\}$ such that probability of having an edge uv equals p if u, v in same C_i and equals q otherwise, for some $p >> q$.
- Here Ω_1, Ω_2 are Bernoulli distributions with probability of success equal to p and q respectively ($\alpha = p$, $\beta = q$).

Example: Block stochastic model

- In the block stochastic model have subsets of nodes $\{C_1, \ldots, C_k\}$ such that probability of having an edge uv equals p if u, v in same C_i and equals q otherwise, for some $p >> q$.
- Here Ω_1, Ω_2 are Bernoulli distributions with probability of success equal to p and q respectively ($\alpha = p$, $\beta = q$).



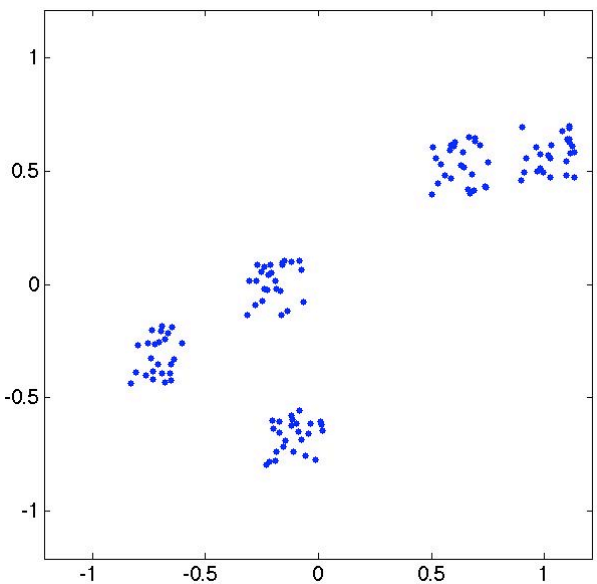


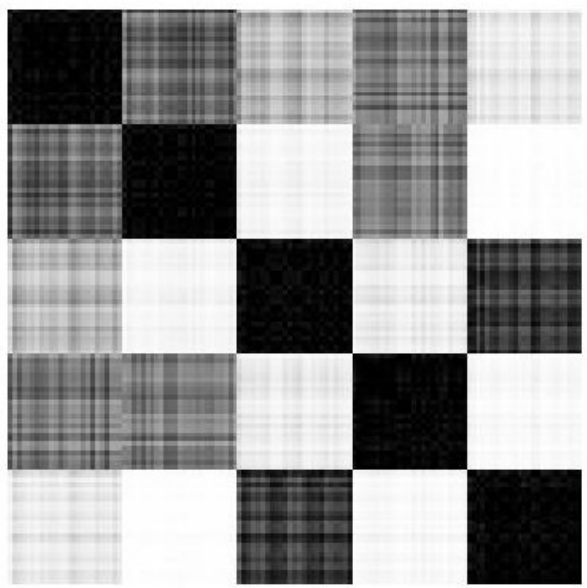
Example: Clustered Euclidean data

- Suppose each data point in the ith cluster C_i is placed uniformly at random in a ball centered at $c_i \in \mathbf{R}^d$. If the centers are sufficiently far apart, then the distance within clusters will be small compared to the distance between clusters.
- Choose $W_{ij} = \exp(-\|\mathbf{x}^i - \mathbf{x}^j\|^2/\sigma)$.

Example: Clustered Euclidean data

- Suppose each data point in the ith cluster C_i is placed uniformly at random in a ball centered at $c_i \in \mathbf{R}^d$. If the centers are sufficiently far apart, then the distance within clusters will be small compared to the distance between clusters.
- Choose $W_{ij} = \exp(-\|\mathbf{x}^i - \mathbf{x}^j\|^2/\sigma)$.





When do we recover the correct partition?

- Let $C_{k+1} := V - (\cup_{i=1}^{k} C_i)$.
- Let $X^* = \sum_{i=1}^{k} (1/r_i)\mathbf{v}^i(\mathbf{v}^i)^T$ where $\mathbf{v}^i$ is the characteristic vector of C_i and $r_i = |C_i|$ for all i.

When do we recover the correct partition?

- Let $C_{k+1} := V - (\cup_{i=1}^{k} C_i)$.
- Let $X^* = \sum_{i=1}^{k} (1/r_i)\mathbf{v}^i(\mathbf{v}^i)^T$ where $\mathbf{v}^i$ is the characteristic vector of C_i and $r_i = |C_i|$ for all i.

Theorem

There exist scalars $c_1, c_2, c_3 > 0$ such that X^* is the unique optimal solution of the SDP relaxation of WKDC (and $\{C_1, \ldots, C_k\}$ is the unique optimal solution for WKDC) if

$$c_1\left(k\sum_{i=1}^{k+1} r_i\right)^{1/2} + c_2\sqrt{N} + c_3\beta r_{k+1} \leq (\alpha - \beta)\hat{r},$$

with probability tending exponentially to 1 as $\hat{r} \to \infty$, where $\hat{r} = \min_{i=1,\ldots,k} r_i$.

When do we recover the correct partition?

- The sufficient condition on last slide cannot be satisfied unless $\hat{r} = \Omega(\sqrt{N})$ and $r_{k+1} = O(\hat{r})$.
- If all clusters are all of size $\hat{r} = N^{\epsilon}$, we can recover $k = O(\sqrt{\hat{r}})$ of them for $\epsilon \in [1/2, 2/3]$.
- Can have clusters of different sizes: suppose we have one large cluster of size $N^{3/4}$. Then we can have $O(N^{1/4})$ smaller clusters of size $N^{1/2}$.

Proof outline, pt 1

- The SDP relaxation is strictly feasible (satisfies Slater's constraint qualification).
- To show that X^* is optimal, it suffices to show that there exists $\mu \in \mathbf{R}, \lambda \in \mathbf{R}_+^N, \eta \in \mathbf{R}_+^{N \times N}, S \in \Sigma_+^N$ such that

$$\begin{aligned} -W + \lambda \mathbf{e}^T + \mathbf{e}\lambda^T + \mu I - \eta &= S \\ \lambda^T (X^* \mathbf{e} - \mathbf{e}) = 0, \quad Tr(X^* \eta) = 0, \quad Tr(X^* S) &= 0. \end{aligned}$$

- Complementary slackness condition on S is equivalent to the columns of X^* being in $Null(S)$.
- By the block structure of X^*, this is the same as having the rows and columns of $S(C_i, C_j)$ sum to 0 for all $i, j \in \{1, \ldots, k+1\}$.

Proof outline, pt 2

- The equations $S(C_i, C_i)\mathbf{e} = 0$ give an explicit formula for λ.
- To have $\lambda, \eta \geq 0$, need $\mu = O((\alpha - \beta)\hat{r})$.
- Using standard probabilistic bounds on tail distributions and norms of random matrices can show that there exists η such that $\|S - \mu I\| = O(\text{LHS})$.
- By the triangle inequality:

$$\|S - \mu I\| \leq \mu \Rightarrow \lambda_{min}(S) \geq \mu - \|S - \mu I\| \geq 0.$$

□

Biclustering

Biclustering

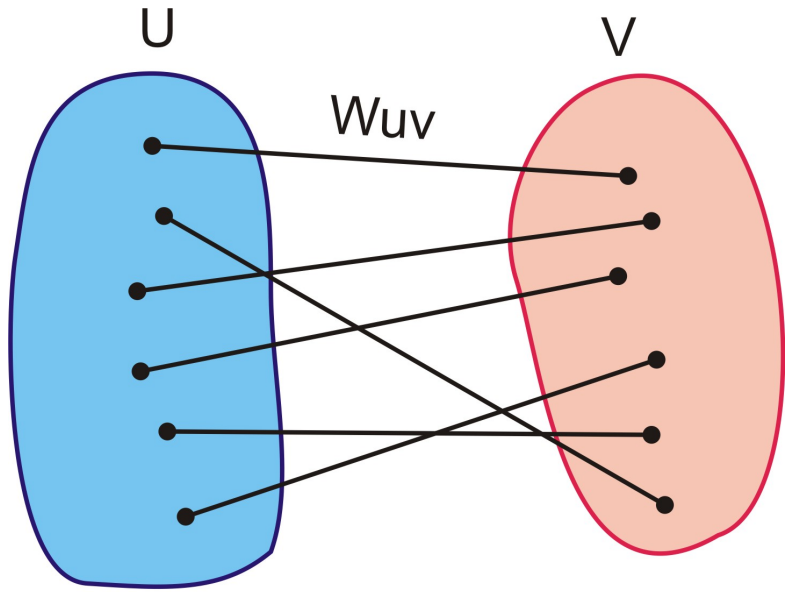
- Given a set of objects and features.
- Want to simultaneously partition objects and features so that each cluster of objects exhibit common features and each cluster of features is shared by a set of similar objects.
- Also known as *co-clustering*, *two-mode clustering*.

Applications

- Biclustering gene expression data:
  - Run a series of experiments where expression level of genes is measured under varying conditions.
  - Obtain matrix where rows are indexed by genes, and each column is the expression level of the genes in a single experiment.
  - Want to identify groups of genes similarly expressed in subsets of experimental conditions.
- Identifying topics in text-database: objects are articles in database, features are keywords, want to identify sets of articles that contain many instances of the same keywords (likely about same topic).
- Recommender systems: want to identify groups of customers and groups of items in catalogue that they prefer.

Graph partitioning model for biclustering

- Data defines bipartite similarity graph $G_S = ((U, V), E, W)$.
- U is the set of objects, V is the set of features, assign weight W_{uv} according to the expression level of object u of feature v.



- Want to simultaneously partition U and V into bicliques $(U_1, V_1), \ldots, (U_k, V_k)$ so that weight within bicliques (U_i, V_i) is higher than weight on edges from U_i to V_j for $i \neq j$.

The WKDB problem

- Want to partition (U, V) so that the sum over of the "average" edge weights from U_i to V_i is maximized:

$$\max \sum_{i=1}^{k} \frac{\mathbf{u}_i \mathbf{v}_i^T}{\sqrt{|U_i||V_i|}},$$

  here $\mathbf{u}_i, \mathbf{v}_i$ are the characteristic vectors of U_i, V_i.

- Can relax as nonconvex QP:

$$\begin{aligned} \max \quad & Tr(X^T W Y) \\ \text{s.t.} \quad & \text{cols of } X \text{ are orthonormal} \\ & \text{cols of } Y \text{ are orthonormal} \\ & X \in \mathbf{R}_+^{U \times k}, \quad Y \in \mathbf{R}_+^{V \times k}. \end{aligned}$$

Relaxation to SDP

- Can symmetrize this QP:

$$\begin{aligned} \max \quad & Tr\left(\begin{pmatrix} X \\ Y \end{pmatrix}^T \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \right) \\ \text{s.t.} \quad & \text{cols of } X \text{ are ON, cols of } Y \text{ are ON} \\ & X \in \mathbf{R}_+^{U \times k}, \quad Y \in \mathbf{R}_+^{V \times k}. \end{aligned}$$

Relaxation to SDP

- Can symmetrize this QP:

$$\begin{aligned} \max \quad & Tr\left(\begin{pmatrix} X \\ Y \end{pmatrix}^T \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \right) \\ \text{s.t.} \quad & \text{cols of } X \text{ are ON, cols of } Y \text{ are ON} \\ & X \in \mathbf{R}_+^{U \times k}, \quad Y \in \mathbf{R}_+^{V \times k}. \end{aligned}$$

- Apply matrix lifting argument from before:

$$\begin{aligned} \max \quad & Tr\left(\begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix} \overbrace{\begin{pmatrix} X & M \\ M^T & Y \end{pmatrix}}^{Z} \right) \\ \text{s.t.} \quad & X\mathbf{e} = \mathbf{e}, \quad Y\mathbf{e} = \mathbf{e} \\ & Tr(X) = k, \quad Tr(Y) = k \\ & Z \geq 0, \quad Z \succeq 0. \end{aligned}$$

The Planted Bicluster Model

- We assume that a good biclustering exists.
- i.e. there exists bicliques $(U_1, V_1), \cdots (U_k, V_k)$ such that the edges within the bicliques have higher weight than edges between bicliques.
- Probabilistic model: weights are chosen independently at random so that for all $u \in U_i, v \in V_j$

$$E[W_{uv}] = \begin{cases} \alpha & \text{if } i = j, \\ \beta & \text{if } i \neq j. \end{cases}$$

Proposed solution

- Note that

$$Z^* = \begin{pmatrix} X^* & M^* \\ (M^*)^T & Y^* \end{pmatrix}$$

is a feasible solution where

$$\begin{aligned} X^* &= \sum_{i=1}^{k} \frac{1}{m_i} \mathbf{u}^i (\mathbf{u}^i)^T, \\ Y^* &= \sum_{i=1}^{k} \frac{1}{n_i} \mathbf{v}^i (\mathbf{v}^i)^T, \\ M^* &= \sum_{i=1}^{k} \frac{1}{\sqrt{m_i n_i}} \mathbf{u}^i (\mathbf{v}^i)^T. \end{aligned}$$

When is the proposed solution optimal?

Theorem

- *Suppose that* $m_i = n_i$ *for all* $i = 1, \ldots, k$ *and let* $\hat{n} = \min_i n_i$.
- *Then there exists scalars* $c_1, c_2 > 0$ *such that* Z^* *is the unique optimal solution of the SDP relaxation of WKDB (and* $(U_1, V_1), \ldots, (U_k, V_k)$ *is the optimal set of bicliques) if*

$$c_1 \Big(k \sum_{i=1}^{k} n_i \Big)^{1/2} + c_2 \sqrt{N} \leq (\alpha - \beta)\hat{n}$$

with probability tending exponentially to 1 *as* $\hat{n} \to \infty$.

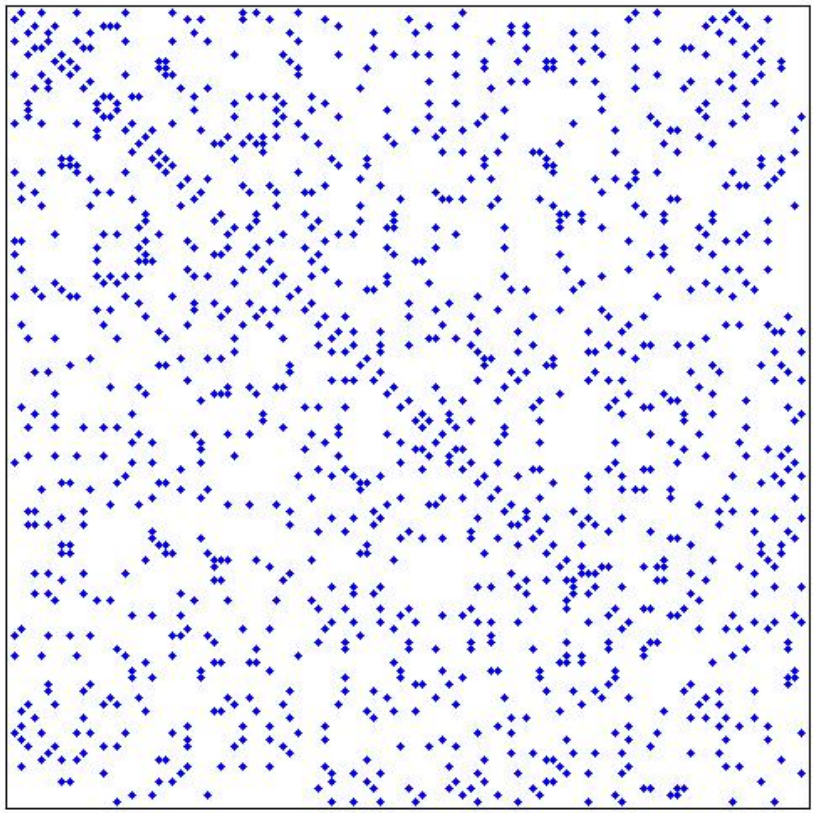
Numerical results

The NCAA Football Network

- First considered by Girvan and Newman 2001.
- Nodes are the 115 Division I football teams.
- Teams are deemed adjacent if they played at least one game against each other in the 2000 season.
- Games are typically scheduled based on membership within athletic conferences and geography.
- Should display community structure based on membership in conferences.

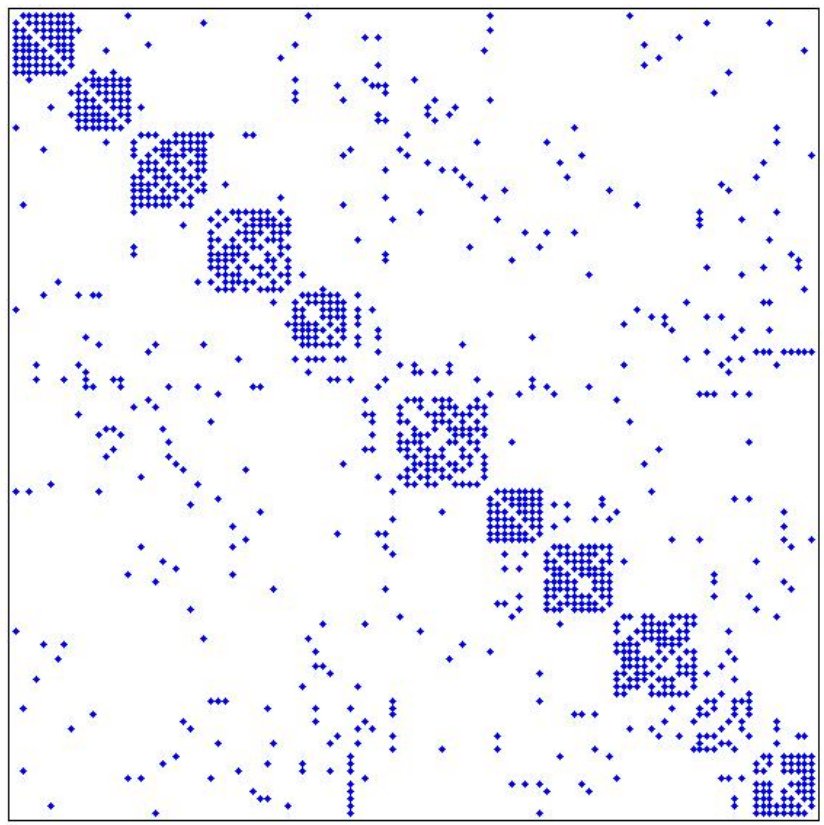
The NCAA Football Network, pt 2

- The node-node adjacency matrix for the football network.



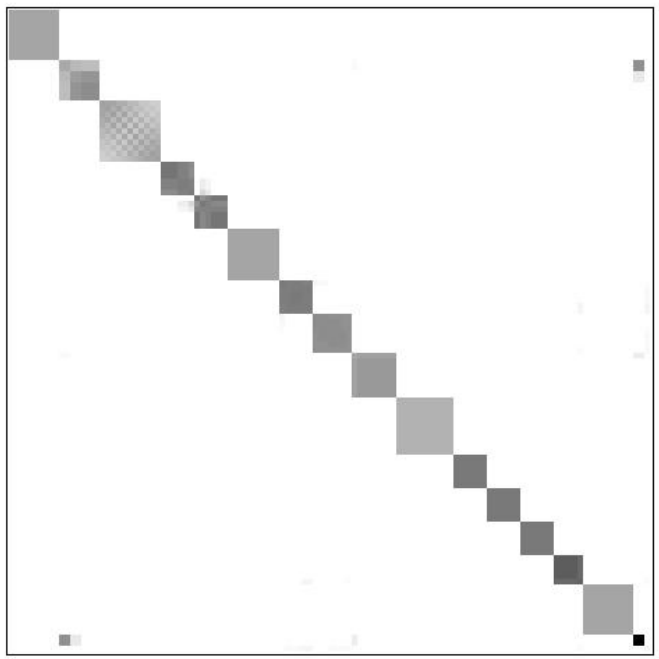
The NCAA Football Network, pt 3

- The adjacency matrix with rows/cols reordered according to conferences.



The NCAA Football Network, pt 4

- Solve instance of WKDC with $W = A$, $k = 16$.
  - Use SDPNAL to solve the SDP (~ 13225 constraints/variables).
- Obtain optimal solution X^*:



The NCAA Football Network, pt 5

Atlantic Coast
Clemson
Duke
Florida State
Georgia Tech
Maryland
NC State
North Carolina
Virginia
Wake Forest

Big East
Boston College
Miami
Pittsburgh
Rutgers
Syracuse
Temple
Virginia Tech
West Virginia

Big Ten
Illinois
Indiana
Iowa
Michigan
Michigan State
Minnesota
Northwestern
Ohio State
Penn State
Purdue
Wisconsin

Big Twelve 1
Colorado
Kansas
Kansas State
Iowa State
Missouri
Nebraska

Big Twelve 2
Oklahoma State
Oklahoma
Texas Tech
Baylor
Texas A&M
Texas

Conference USA
Alabama-Birm
Army
Cincinnati
East Carolina
Houston
Louisville
Memphis
So. Miss.
Tulane

Outlier
Connecticut

MAC West
Ball State
C. Michigan
E. Michigan
N. Illinois
Toledo
W. Michigan

MAC East
Akron
BGSU
Buffalo
Kent
Marshall
Miami Ohio
Ohio

Mountain West
Air Force
BYU
Colorado St
New Mexico
SDSU
UNLV
Utah
Wyoming

Pacific Ten
Arizona
Arizona State
California
Oregon
Oregon State
Stanford
UCLA
USC
Washington
WSU

SEC West
Alabama
Arkansas
Auburn
LSU
Miss St
Mississippi

SEC East
Florida
Georgia
Kentucky
South Carolina
Tennessee
Vanderbilt

Sun Belt 1
Arkansas State
Boise State
Idaho
NMSU
North Texas
Utah State

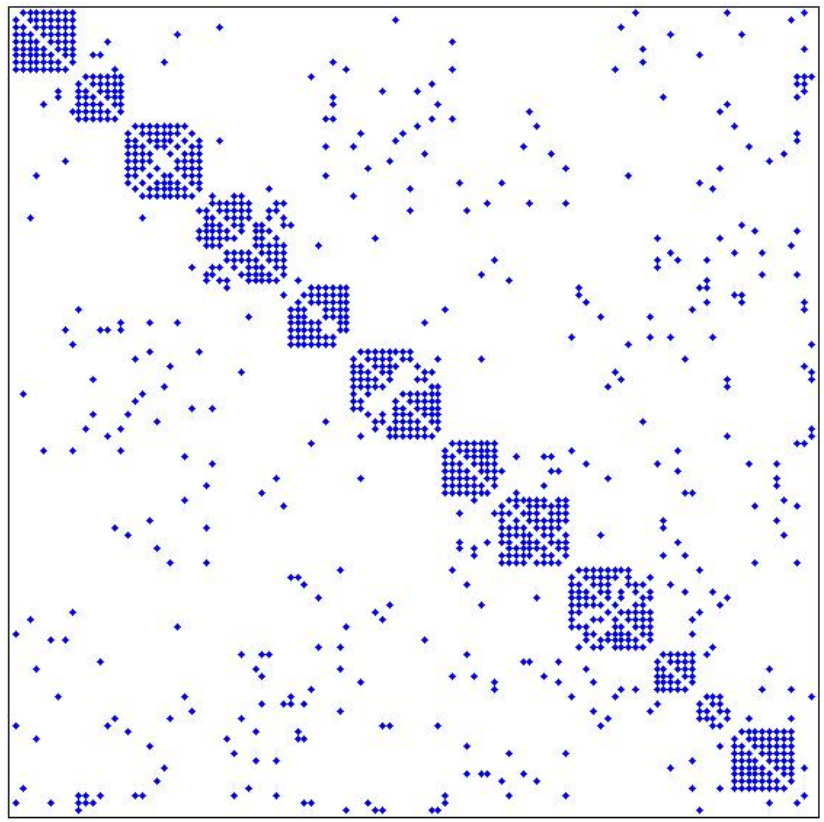
Sun Belt 2
*Central Florida**
L-Lafayette
L-Monroe
*Louisiana Tech**
MTSU

Western Athletic
Fresno St
Hawaii
Nevada
Rice
SJSU
SMU
*Texas Christian**
Texas El Paso
Tulsa

Independents
Navy
Notre Dame

The NCAA Football Network, pt 6

- Reordering of A according to the identified clusters:



Rehnquist Supreme Court

- Data set is the set of U.S. Supreme Court Justices (serving from 1994-95 to 2003-04).
- Assign edge-weights corresponding to fraction of decisions on which Justices agreed:

	St	Br	Gi	So	Oc	Ke	Re	Sc	Th
1 St	1	0.62	0.66	0.63	0.33	0.36	0.25	0.14	0.15
2 Br	0.62	1	0.72	0.71	0.55	0.47	0.43	0.25	0.24
3 Gi	0.66	0.72	1	0.78	0.47	0.49	0.43	0.28	0.26
4 So	0.63	0.71	0.78	1	0.55	0.5	0.44	0.31	0.29
5 Oc	0.33	0.55	0.47	0.55	1	0.67	0.71	0.54	0.54
6 Ke	0.36	0.47	0.49	0.5	0.67	1	0.77	0.58	0.59
7 Re	0.25	0.43	0.43	0.44	0.71	0.77	1	0.66	0.68
8 Sc	0.14	0.25	0.28	0.31	0.54	0.58	0.66	1	0.79
9 Th	0.15	0.24	0.26	0.29	0.54	0.59	0.68	0.79	1

Rehnquist Supreme Court: results, pt 1

- Solving weighted KDC with $k = 2$ yields the following partition of the Supreme court:

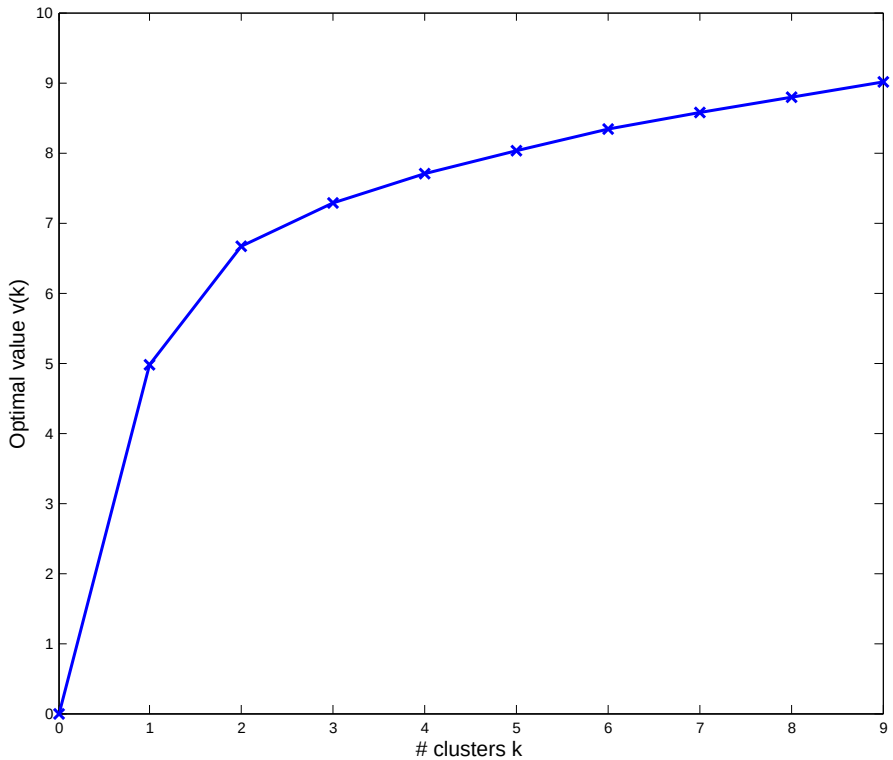
1: "Liberal"	2: "Conservative"
Stevens (St)	O'Connor (Oc)
Breyer (Br)	Kennedy (Ke)
Ginsberg (Gi)	Rehnquist (Re)
Souter (So)	Scalia (Sc)
	Thomas (Th)

Rehnquist Supreme Court: results, pt 2

- Algorithm is sensitive to choice of k.
- Solve with $k = 3$:

1: "Most Conservative"	2: "Moderate Conservative"	3: "Liberal"
Thomas (Th) Scalia (Sc)	O'Connor (Oc) Kennedy (Ke) Rehnquist (Re)	Stevens (St) Breyer (Br) Ginsberg (Gi) Souter (So)

Rehnquist Supreme Court: results, pt 3



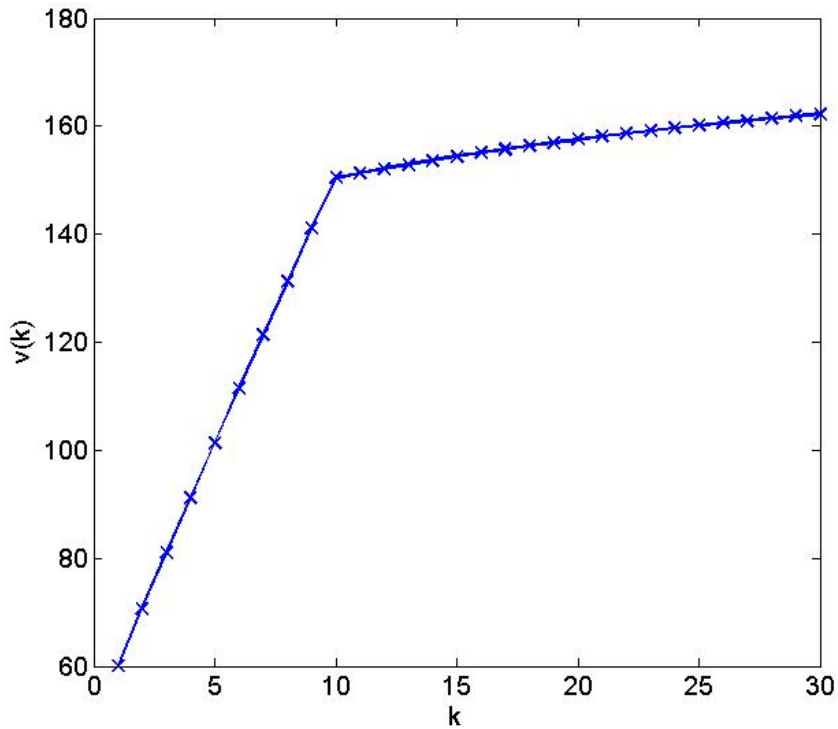
How to choose k?

- Value function $v(k)$ is concave and increasing in k.

- **Conjecture:** If W is generated according to the planted cluster model with k^* clusters then $v(k)$ is approximately piecewise linear with break point at $k = k^*$.

Experimental results

- Generated W from planted cluster model with $k^* = 10$ clusters of size $\hat{r} = ?$, and no diversionary nodes ($N = 10\hat{r}$).
- Ω_1, Ω_2 uniform over $[0.5, 1]$, $[0, 0.5]$
- Solved for choices of $k = 1, ...30$.



Final remarks

- Have presented new semidefinite programming based heuristics for the clustering and biclustering problems.
- If data is sufficiently clusterable (i.e. highly similar within clusters, dissimilar across clusters) then these heuristics successfully recover the correct partition of the data.
- Open problems:
  - Behaviour of $v(k)$.
  - Unbalanced biclusters: $m_i \neq n_i$ for some i.
  - Diversionary nodes in biclustering model.