Correlation Clustering with Noisy Input

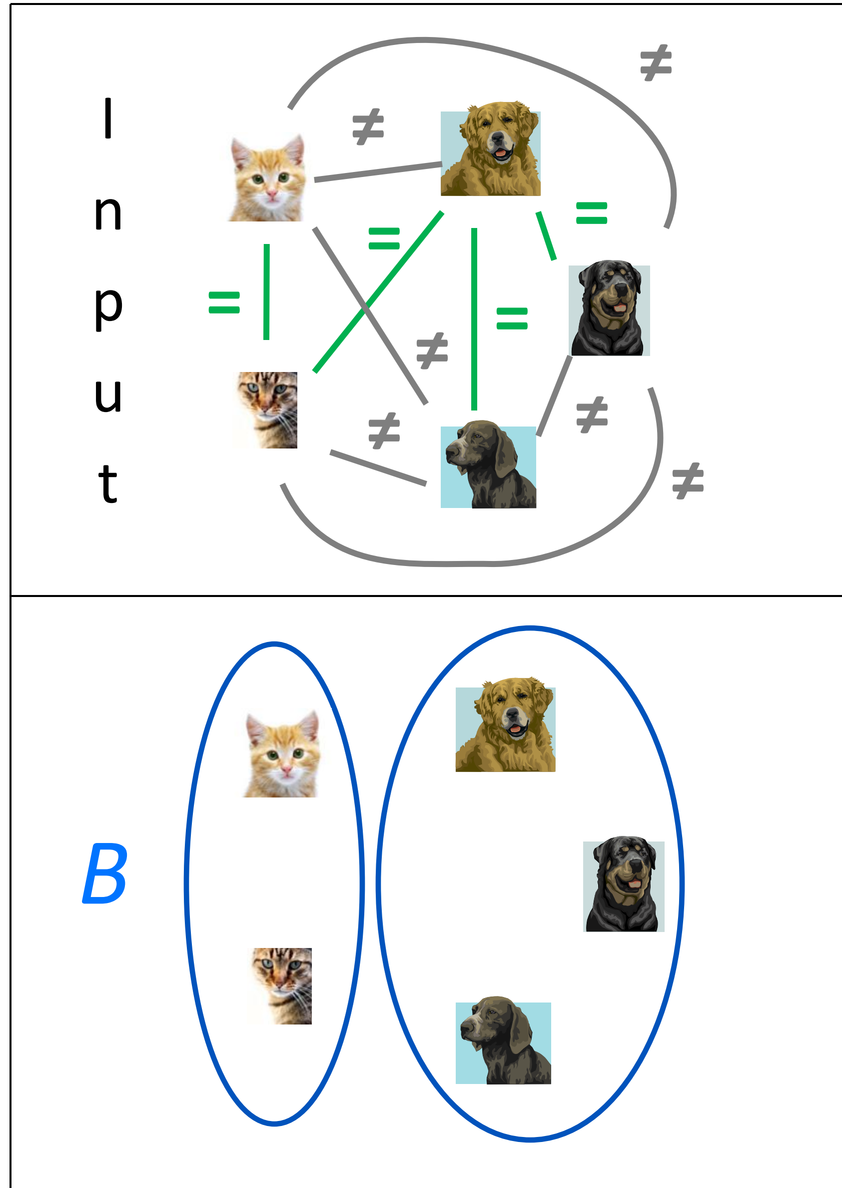
Claire Mathieu
Warren Schudy

Brown University

SODA 2010

Noisy Correlation Clustering Model

- Unknown base clustering *B* of *n* objects
- Noise: each edge is controlled by an adversary with probability *p* and "tells the truth" otherwise
- Problem: reconstruct *B* from the edge labels



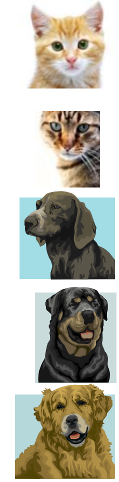
One of our results

- **Theorem:** assume *$p \leq 1/3$*. If all clusters have size at least $\alpha_1 \sqrt{n}$ then the natural semi-definite program (SDP) recovers *B* exactly with high probability.
- **Previous best:** $\alpha_2 \sqrt{n \log n}$ [Bansal, Blum, Chawla '04, Shamir and Tsur '07], combinatorial.
- See paper for other results (including approximation algorithms)

Plan

- The semi-definite program
- Its dual
- Using the dual

Clusterings



1	1	0	0	0
1	1	0	0	0
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1

- Clusterings are represented by 0/1 matrices:
  X_{ij}=1: *i* and *j* in same cluster
- In general a clustering satisfies:

$$X = \sum_k v_k v_k^T \text{ for some 0/1 orthogonal vectors } v_1, v_2, \ldots, v_m, \text{ one per cluster}$$

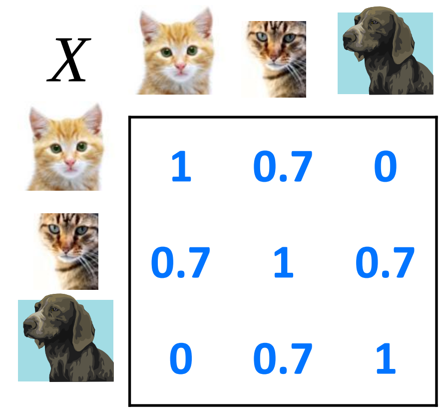
- E.g.

v_{cat} : 1 1 0 0 0

v_{dog} : 0 0 1 1 1

Relaxation of clusterings

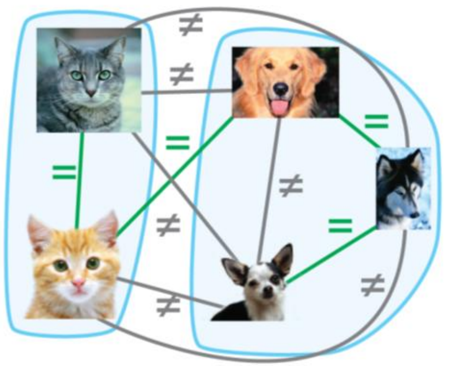
- Relaxation ~~Clustering~~
  - $X = \sum_k v_k v_k^T$ for some $\cancel{0/1}$ vectors $v_1, v_2, \ldots, v_m$
  - $X_{ii} = 1$ for all i
  - $X_{ij} \geq 0$ for all i,j



- The following are equivalent (X symmetric):
  - $X = \sum_k v_k v_k^T$ for some vectors $v_1, v_2, \ldots, v_m$
  - X is *positive semi-definite (p.s.d.)*

Objective

- Maximize number of agreements:



$$\max \sum_{i<j} \begin{cases} X_{ij} & \text{if } i \;\underline{\quad = \quad}\; j \\ \cancel{1} - X_{ij} & \text{if } i \;\underline{\quad \neq \quad}\; j \end{cases}$$

Drop the constant

- I.e. $\max \sum_{i<j} X_{ij} \overline{E}_{ij}$

where

$$\overline{E}_{ij} = \begin{cases} 1 & \text{if } i \;\underline{\quad = \quad}\; j \\ -1 & \text{if } i \;\underline{\quad \neq \quad}\; j \end{cases}$$

Summary of SDP

$$\max \sum_{i<j} X_{ij} \overline{E}_{ij} \text{ s.t.}$$

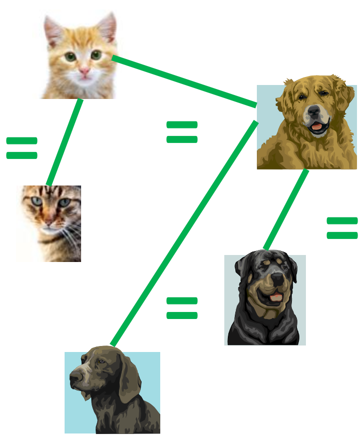
$$X \text{ p.s.d.}$$

$$X_{ii} = 1$$

$$X_{ij} \geq 0$$

This SDP was previously used by:

- [Charikar, Guruswami, Wirth ‘05]
- [Swamy ‘04]



$\overline{E}$

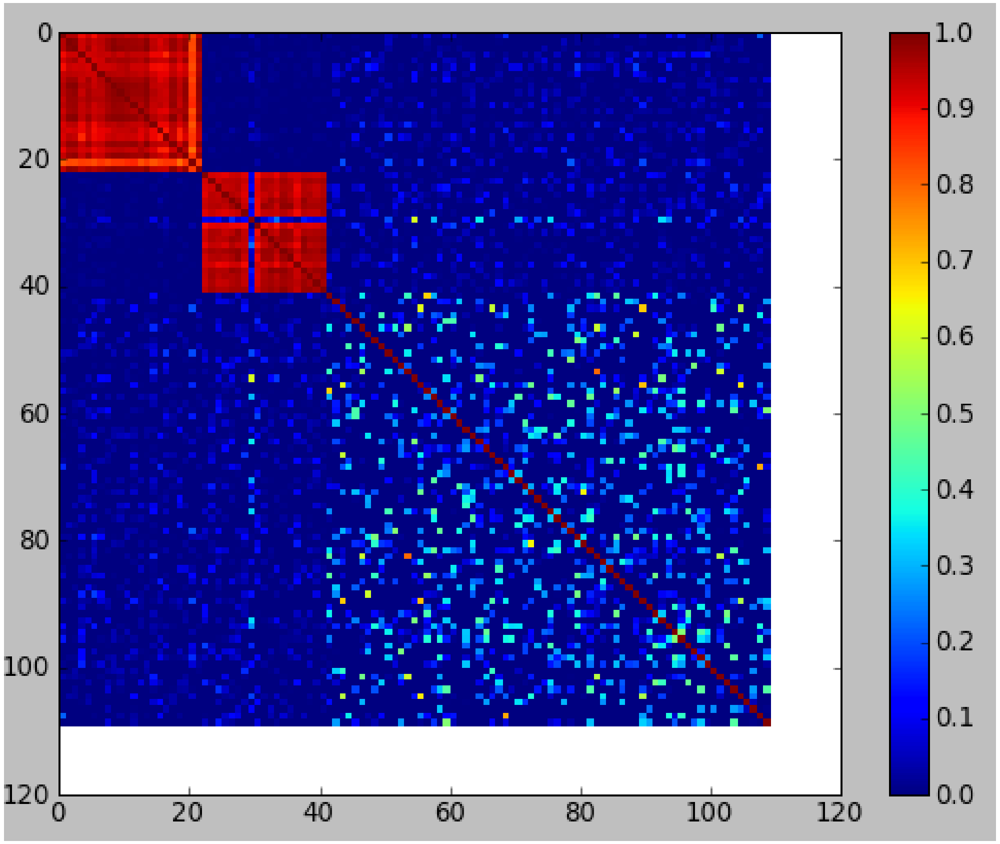
0	1	-1	-1	1
1	0	-1	-1	-1
-1	-1	0	-1	1
-1	-1	-1	0	1
1	-1	1	1	0

X

1	1	0	0	0
1	1	0	0	0
0	0	1	0	0.7
0	0	0	1	0.7
0	0	0.7	0.7	1

Discussion

- Algorithm:
  - Solve SDP
  - If integral, output it. Otherwise fail.
- **Thm:** assume *p≤ 1/3*. If all clusters have size at least $\alpha_1\sqrt{n}$ then the SDP recovers *B* exactly with high probability.



An example *X* matrix from solver in [Elsner and Schudy '09]. That solver scales to a few thousand objects.

Plan

- The semi-definite program
- Its dual
- Using the dual

Translate SDP into LP

The following are equivalent (X symmetric):

- X positive semi-definite
- $u^T X u \geq 0$ for all vectors u ← Linear in X for fixed u

SDP again:

$$\max \sum_{i<j} X_{ij} \overline{E}_{ij} \text{ s.t.}$$

$$X_{ii} = 1$$

$$X_{ij} \geq 0$$

$$X \text{ p.s.d.}$$

LP form:

$$\max \sum_{i<j} X_{ij} \overline{E}_{ij} \text{ s.t.}$$

$$X_{ii} = 1$$

$$X_{ij} \geq 0$$

$$u^T X u \geq 0 \text{ for all vectors } u$$

SDP Dual

Primal:

$$\max \sum_{i<j} X_{ij} \overline{E}_{ij} \text{ s.t.}$$

$$X_{ii} = 1 \text{ for all } i \qquad (d_i)$$

$$-X_{ij} \leq 0 \text{ for all } i, j \qquad (h_{ij})$$

$$-u^T X u \leq 0 \text{ for all } u \qquad (a_u)$$

Dual:

$$\min \sum_i d_i \text{ s.t.}$$

$$-\sum_u a_u u_i u_j + d_i 1(i = j) - h_{ij} = \overline{E}_{ij} \quad \text{for all i} \leq \text{j}$$

$$\mathrm{a}_{\mathrm{u}}, h_{ij} \geq 0$$

Translate dual LP into SDP

The following are equivalent (X symmetric):

- $X = \sum_k a_u u_k u_k^T$ with $a_u \geq 0$
- X positive semi-definite

Dual again:

$$\min \sum_i d_i \text{ s.t.}$$

$$-\sum_u a_u u_i u_j + d_i 1(i = j) - h_{ij} = \overline{E}_{ij}$$

$$\mathrm{a}_\mathrm{u}, h_{ij} \geq 0$$

Matrix form:

$$\min Trace(D) \text{ s.t.}$$

$$-\overline{E} + D - H = \sum_u a_u u u^T$$

D diagonal

$$\mathrm{a}_\mathrm{u} \geq 0$$

$$H \geq 0$$

Arbitrary positive semi-definite matrix

The Dual SDP

$$\min \mathit{Trace}(D) \text{ s.t.}$$

$$-\overline{E} + D - H \quad \text{positive semi-definite}$$

$$D \quad \text{diagonal}$$

$$H \geq 0$$

Plan

- ✓ The semi-definite program
- ✓ Its dual
- Using the dual

This proof is inspired by a similar result for the planted clique problem [Feige and Krauthgamer ‘00].

Using the dual - overview

- Prove optimality of the base clustering by presenting dual solution *(D,H)* whose value matches value of base clustering *B* (see paper)
- Difficult part: proving that $-\overline{E}+D-H$ is p.s.d.
- The following are equivalent (*Y* symmetric):
  - Y positive semi-definite
  - All eigenvalues of Y are ≥ 0
- We present *b* eigenvectors with eigenvalue *0* (see paper), where *b* is the number of clusters in *B*
- We prove that the $b+1^{th}$ smallest eigenvalue, denoted $\lambda_{b+1}\left(-\overline{E}+D-H\right)$, is positive (sketched next)
- Hence all eigenvalues of $-\overline{E}+D-H$ are ≥ 0

Eigenvalue analysis

$$-\overline{E}+D-H=M_1+M_2+M_3+M_4$$

$$\lambda_{b+1}=\theta(\text{min cluster size})$$

(see paper)

$$\lambda_1 \geq -\theta(\sqrt{n})$$

(next)

We apply the following:

Theorem [Weyl]: If *M* and *N* are symmetric matrices then

$$\lambda_{b+1}(M+N) \geq \lambda_1(M) + \lambda_{b+1}(N)$$

Hence for sufficiently large min cluster size

$$\lambda_{b+1}\left(-\overline{E}+D-H\right) > 0.$$

Random matrices

Theorem [Füredi and Komlós '81]: Let *M* be a random symmetric matrix with independent entries of mean zero. Then with high probability

$$\left|\lambda_i(M)\right| = O\left(\sqrt{n}\right) \text{ for all } i.$$

Application:

$$\lambda_1(M_2) = \lambda_1\left(-\overline{E} - \mathbf{Expectation}\left[-\overline{E}\right]\right) \geq -\theta\left(\sqrt{n}\right)$$

To analyze M_3 we developed a generalization of this theorem that handles limited dependence between the entries.

Recap

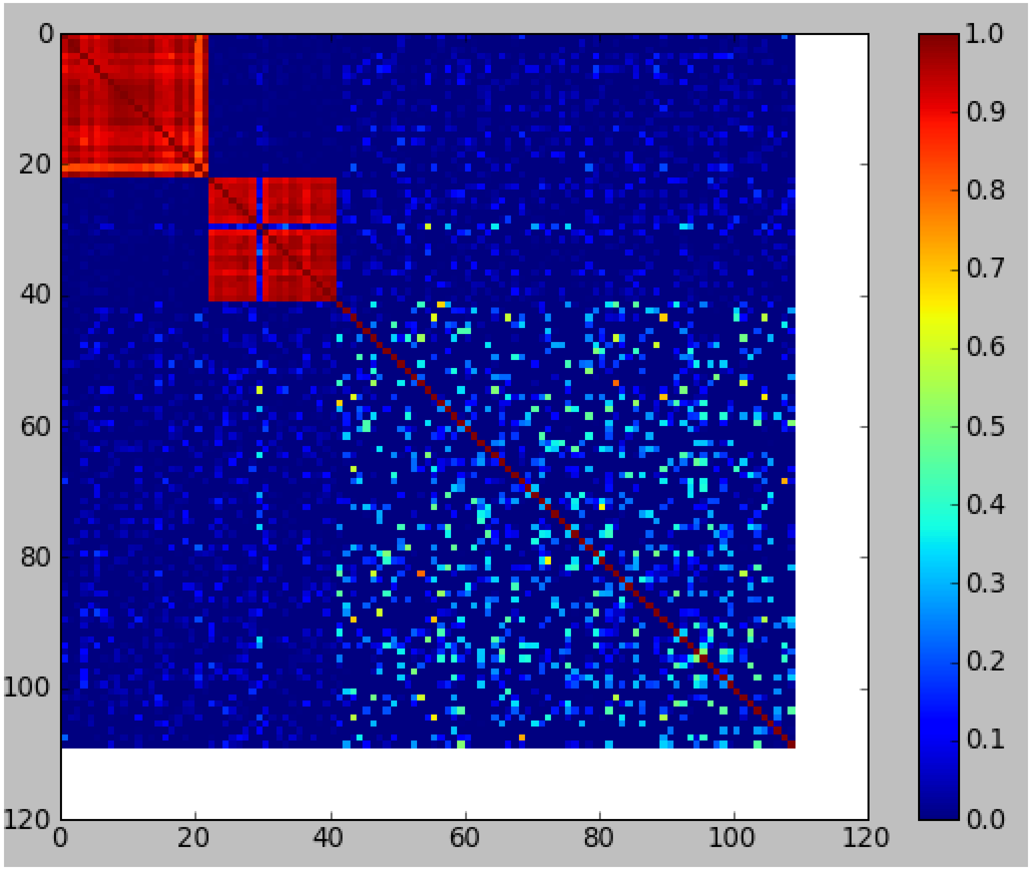
Theorem: assume *p≤ 1/3*. If all clusters have size at least $\alpha_1\sqrt{n}$ then the SDP recovers *B* exactly with high probability.

Proof:

- We wrote a dual solution matrix as a sum of 4 random matrices, used Füredi-Komlós variants to bound their eigenvalues, used Weyl to infer bound on eigenvalues of the matrix, hence p.s.d., hence solution is feasible.
- That solution has value equal to the value of *B*, hence by duality *B* is primal optimal
- *B* is the *unique* primal optimum (see paper), hence SDP will exactly return *B*
- Hence algorithm reconstructs *B* exactly when all clusters have size at least $\alpha_1\sqrt{n}$.

Open Question 1

- Suppose some clusters are size $c_3\sqrt{n}$ and others are size 1. Can the SDP be used to reconstruct the large clusters?



Software: [Elsner and Schudy '09].

Open Question 2

- Planted clique problem = correlation clustering with only one non-singleton and no corruption of within-cluster edges
- Exist polynomial-time algorithm when clique size = $c_1\sqrt{n}$
- Exists $n^{O(\log n)}$ -time algorithm when clique size = $c_1 \log n$
- Can polynomial-time algorithms beat the $c_1\sqrt{n}$ barrier?

Clustering References

- Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. In *STOC '05, pages* 684–693, 2005.
- Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Mach. Learn., 56(1-3):89–113, 2004.*
- Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. *J. Comput. Syst. Sci., 71(3):360–383, 2005.*
- M. Elsner and W. Schudy. Bounding and Comparing Methods for Correlation Clustering Beyond ILP. In *ILP-NLP '09: Proc. NAACL/HLT 2009 Workshop on Integer Linear Programming for Natural Language Processing, pages 19–27, 2009.*
- Ron Shamir and Dekel Tsur. Improved algorithms for the random cluster graph model. *Random Structures and Algorithms, 31(4):418–449, 2007.*

Other References

- F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization,* 5(1):13–51, 1995
- Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Struct. Algorithms, 16(2):195–208,* 2000
- Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica, 1(3):233–241, 1981*