Stable Identification of Cliques with Radon Basis Pursuit

Xiaoye Jiang
Stanford University

Yuan Yao
Peking University

Leo Guibas
Stanford University

Abstract

In this paper we study the identification of common interest groups from low order interactive observations. We present a new algebraic approach based on the Radon basis pursuit in homogeneous spaces. We prove that if the common interest groups satisfy a condition that overlaps between them are small, then they can be recovered in a robust way by solving a linear programming problem. We demonstrate the applicability of our approach with examples on identifying social communities in the social network of *Les Miserables* and on identifying coauthorship cliques within large-scale networks.

1 Introduction

In this paper we consider the problem of identifying common interest groups or cliques based on partial information. This problem arises in a variety of situations, from identity management [8], statistical ranking [6, 9], and in particular, social networks. The following three examples provide a glimpse on the typical problems which could be addressed with the techniques discussed in this paper.

Motivating example 1 (Tracking and Identifying Teams) We consider the scenario of multiple targets moving in the environment monitored by sensors. We assume each moving target has an identity and they each belong to some teams or groups. However, we only get partial interaction information due to the problem structure. For example, consider watching a grey-scale video of a basketball game (when it may be hard to tell apart the two teams), we observe ball passes or collaborative offensive/defensive interactions between teammates. The observations are partial due to the fact that players have mostly pairwise interactions in basketball games. It is seldom to observe a single event which involves all team members. Our objective is to infer membership information (which team the players belong to) from partially observed interactions.

In Figure 1-(a), we show a weighted graph (shown on the top) illustrating pairwise interactions among 10 basketball players. Nodes in this graph represent players and weights on edges represent frequencies of cooperative pairwise interactions. Note that we may get noisy data due to observation errors where people from different teams have ambiguous interactions. However, given that the noise is not too large, we hope to be capable to identify the two teams (shown at the bottom) from such partially observed pairwise interaction information.

Motivating example 2 (Detecting Communities in Social Networks) Detecting social communities in social networks is of extraordinary importance. It can be used to understand the organization or collaboration structures within the social network. However, we do not have direct mechanisms to sense what the social communities are. Instead, we have partial, low order interaction information. For example, we observe pairwise or triple-wise co-appearance among people who hang out for some leisure activities together. We hope to detect those social communities in the social network from such partially observed data.

In Figure 2-(a), we show an example as the social network of Victor Hugo's novel *Les Miserables* which was studied in [10]. In the weighted graph, the nodes represent 33 key characters and weights on edges represent frequencies of co-appearance. Several social communities arise in the network, formed by either friendships, street gangs, kinships, student society, or drama conflicts. We wish to detect those social communities from pairwise co-appearance frequencies data. Note that in this example, different social communities may have different sizes and one people may belong to several social communities.

Motivating example 3 (Inferring Partial Rank-

Preliminary work. Under review by AISTATS 2010. Do not distribute.

ings of High Order) The problem of clique identification also arises in the ranking problems. Consider a collection of items are to be ranked by a collection of users. Each user can propose his/her top j, say 3, items in favor but without relative preference within. We wish to infer what are the first tier competitors for the top $k > j$, say 5 items. This problem is the inference of high order partial rankings from low order observations.

In these examples we are typically given a network with some nodes representing players, characters, or items, and with edges summarizing the pairwise interaction observations. Triple-wise and other low order information can be further considered if we consider complete sub-graphs in the networks. *The basic problem here is to determine common interest groups or cliques within the network from observed low order interaction frequencies,* since in reality such low order interactions are often governed by a considerably smaller number of high order communities.

In this paper we assume there are frequency function defined on complete low order subsets and high order susets. Intuitively, the interaction frequency of a particular low order subset should be the sum of frequencies of high order subsets which it belongs to. Hence we consider the following generative model which assume there exists a linear mapping from frequencies on high order subsets(usually sparsely distributed) to low order subsets. One typically can collect data on low order subsets while our task is to find those few dominant high order subsets.

Our problem can be regarded as an extension of the recent work in [9] which studies sparse recovery of *functions on permutation groups*, while we reconstruct *functions on k-subsets* (cliques), often called homogeneous space in literature [6]. In our studies the discrete Radon basis becomes the natural choice instead of the Fourier basis considered in [9]. This leaves us a new challenge on addressing the noiseless exact recovery and stable recovery with noise. Unfortunately the greedy algorithm for exact recovery in [9] can not be applied to noisy settings, and in general the Radon basis does not satisfy the Restricted Isometry Property (RIP) [4] which is crucial for the universal recovery. In this paper, we develop new theories which guarantee the exact sparse recovery and stable recovery under the choice of Radon basis, which has deep roots in Basis Pursuit [5] and its extensions with uniformly bounded noise.

The main content of this paper can be summarized as follows. Section 2 presents the formulation of our problem with a gentle introduction on Radon basis; Section 3 discusses exact recovery conditions without noise; Section 4 addresses stable recovery under uniformly bounded noise, and we generalize our algorithm to handle cliques with mixed sizes; The last section demonstrates three successful applications to some motivating examples discussed above.

2 Problem Formulation

We introduce a graph $G = (V, E)$ to facilitate our discussion. The set of vertices V represents individual identities such as people in the social network, basketball players, or items to be ranked. Each edge in E is associated with some weights which represent interactive frequency information.

We assume there are several common interest groups or communities within the network, represented by cliques or complete sub-graphs in graph G, which are perhaps of different sizes and may have overlaps. We assume every community has certain interaction frequency which can be viewed as a function on cliques. However, we can only receive partial measurements consisting of low order interaction frequency on subsets in a clique. For example, in the smallest case we only observe pairwise interactions represented by edge weights. Our problem is to reconstruct the function on cliques from partially observed data.

However, to resolve this problem, one has to answer two questions: *what is the suitable representation basis, and what is the reconstruction algorithm?* Below we shall provide an answer that the Radon basis will be the appropriate representation for our purpose which allows the sparse recovery by a simple linear programming reconstruction algorithm.

2.1 Basis Construction

We first consider the constuction of a basis so that we can use such a basis to connect functions on j-subsets to functions on k-subsets ($j \leq k$). Our construction of basis is directly related to *Radon Transform* in combinatorics[6].

2.1.1 Common Interest Groups of Equal Size

For simplicity, we restrict ourselves here to the case that all the common interest groups are all of the same size k ($k > j$). The case with mixed sizes will be handled later. There are even some natural scenarios where such a simple case arises, for example the inference of two teams each of size $k = 5$ from pairwise ($j = 2$) interaction frequencies.

Let V_j denote the set of all $j-$subsets of $V = \{1, 2, \cdots, n\}$ and M^j be the set of functions on V_j. The observed partial interaction information, i.e., in-

teraction frequencies on all $j-$subsets, can be viewed as a function on V_j, denoted by $b \in M^j$.

We build a matrix $\tilde{R}^{j,k} : M^k \to M^j$ $(j < k)$ as a mapping from functions on all $k-$subsets of V to functions on all j-subsets of V. For example, $\tilde{R}^{2,5}$ is a $\binom{10}{2}$-by-$\binom{10}{5}$ matrix with rows representing all 2-subsets and columns representing all $5-$subsets. We let entries of $\tilde{R}^{j,k}$ are either 0 or 1 indicating whether the j-subset is a subset of the $k-$subset. Note that every column of $\tilde{R}^{j,k}$ has $\binom{k}{j}$ ones. Lacking *a priori* information, we assume that every $j-$subset has equal probability in interactions, whence choose the same constant 1 for each column. We further normalize $\tilde{R}^{j,k}$ to $R^{j,k}$ so that l_2 norm of each column of $R^{j,k}$ is one. To summarize, we have

$$R^{j,k}_{(\sigma,\tau)} = \begin{cases} \frac{1}{\sqrt{\binom{k}{j}}}, & \text{if } \sigma \subset \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where σ is a $j-$subset and τ is a $k-$subset. As we shall see soon, this construction leads to a canonical basis associated with discrete Radon transform. The size of matrix $R^{j,k}$ clearly depends on the total number of items $n = |V|$, however, we omit n as its meaning will be clear from context.

In the example of basketball games, given information $b \in M^2$ as a function on 2-subsets, we wish to obtain a function $x \in M^5$ on 5-subsets such that $b = Ax$ where $A = R^{2,5}$. Ideally x should be a sparse distribution concentrating on two 5-subsets, representing the two disjoint teams. This is where compressive sensing techniques shall be applied to find such an x.

2.1.2 Relation to Radon Basis

The matrix $R^{j,k}$ constructed above is related to discrete Radon transforms on homogeneous space M^k. In fact, up to a constant, the adjoint or transpose operator $(R^{j,k})^* : M^j \to M^k$ defined by $[(R^{j,k})^* u](\tau) = c\sum_{\sigma\subset\tau} u(\sigma)$, is called in literature [6] the discrete Radon transform from homogeneous spaces M^j to M^k. The collection of all row vectors of $R^{j,k}$ is called as the j-th *Radon basis* for M^k. Our usage here is to exploit the transpose matrix of Radon transform to construct an over-complete dictionary for M^j, such that the observation $b \in M^j$ is represented by a possibly sparse $x \in M^k$ $(k \geq j)$.

The Radon basis was proposed as an efficient way to study partially ranked data in [6], where it was shown that by looking at low order Radon coefficients of function on M^k, we usually get useful and interpretable information. Our approach here adds a reversal of this perspective, *i.e.* the reconstruction of sparse high order functions from low order Radon coefficients. We will discuss this in the following with a connection to the compressive sensing [5, 2].

2.2 Reconstruction Algorithms

Now we give some reconstruction algorithms for detecting high order cliques based on low order information exploiting the basis matrix we talked about in the last section.

Suppose x_0 is a sparse function on common interest groups or cliques. To reconstruct this sparse function based on low order observation data, we consider the following linear programming first known as Basis Pursuit [5], etc.

$$\mathcal{P}_1 : \quad \min \quad \|x\|_1, \quad \text{subject to} \quad Ax = b,$$

where the matrix A is $R^{j,k}$. For robust construction against noise, we also consider the following algorithm

$$\mathcal{P}_{1,\delta} : \quad \min \quad \|x\|_1, \quad \text{subject to} \quad \|Ax - b\|_\infty \leq \delta.$$

It differs from Lasso [12] or BPDN [5] in that a l_∞ norm is used to control the noise instead of the l_2 norm, and also differs from the Dantzig selector [3] which uses $\|A^*(Ax - b)\|_\infty \leq \delta$ in the constraint. The reason for our choice lies in the fact that the typical examples we discussed above often exhibit bounded noise rather than Gaussian-like noise. Our choice will be suitable to incorporate this kind of prior knowledge on noise.

2.3 Failure of Restricted Isometry Property and Universal Recovery

Recently it was shown by [2, 4] that $\mathcal{P}_1$ has a unique sparse solution x_0, if the matrix A satisfies the so called *Restricted Isometry Property* (RIP), i.e. for every set of columns T with $|T| \leq s$, there exists a certain universal constant $\delta_s \in [0, 1)$ (e.g. $\delta_{2s} < \sqrt{2} - 1$ in [4]) such that

$$(1 - \delta_s)\|x\|_{l_2}^2 \leq \|A_T x\|_{l_2}^2 \leq (1 + \delta_s)\|x\|_{l_2}^2, \quad \forall x \in R^s.$$

This exact recovery holds for all $s-$sparse signals x_0, whence called the *universal recovery*.

Unfortunately, in our basis construction of matrix $A = R^{j,k}$, RIP is not satisfied unless $s < \binom{k+j+1}{k}$ which cannot scale up with n. To see this, we extract a set of columns $T = \{\tau : \tau \subset \{1, 2, \cdots, k + j + 1\}\}$ (τ is interpreted as a $k-$subset) and form a submatrix $R^{j,k}_T$. By discarding zero rows, we know the rank of $R^{j,k}_T$ is determined by a small submatrix of $R^{j,k}_T$ of size $\binom{k+j+1}{j}$ by $\binom{k+j+1}{k}$. This matrix has more columns

than rows. This means the extracted columns must be linearly dependent. In other words, there exist an h where $\text{supp}(h) \subset T$ such that $R^{j,k}h = 0$. So in general, we can not expect that the sparse recovery holds universally for all $s-$sparse signals when $s \geq \binom{k+j+1}{k}$.

Therefore, in our case, the correct strategy is to look for the sparsity patterns corresponding to cliques which can be recovered by $\mathcal{P}_1$ or $\mathcal{P}_{1,\delta}$. In general, *we hope to be able to recover a collection of sparse signals x_0, whose sparsity pattern satisfies certain conditions instead of meeting a universal sparse recovery.* Such conditions might naturally occur in reality, which will be shown in the sequel as simply requiring small overlaps between cliques.

3 Exact Recovery Conditions

In this section we present our main results on noiseless exact recovery conditions of x_0 from the given information $b \in M^j$ by solving the linear program $\mathcal{P}_1$.

3.1 Irrepresentable Condition

Suppose A is a M-by-N matrix and x_0 is a sparse signal. Let $T = \text{supp}(x_0)$, T^c be the complement of T, and A_T (or A_{T^c}) be the submatrix of A where we only extract column set T (or T^c, respectively). A regularization path of $\mathcal{P}_{1,\delta}$ refers to the map $\delta \mapsto x_\delta$ where x_δ is a solution of $\mathcal{P}_{1,\delta}$.

Theorem 1 *Assume that $A_T^*A_T$ where $*$ denote matrix transpose is invertible and there exists a vector $w \in R^M$ such that*

(1) $A_T^*w = \iota^* sgn(x_0)$,

(2) $\|A_{T^c}^*w\|_\infty < 1$,

where ι is an imbedding operator $\iota : l_2(T) \to l_2(N)$ extending a vector on T to a vector in R^N by placing zeros outside of T, and ι^* is the dual restriction $\iota^* sgn(x_0) = sgn(x_0)|_T$. Then x_0 is the unique solution for $\mathcal{P}_1$, and it is also a necessary condition that x_0 lies on a unique regularization path of $\mathcal{P}_{1,\delta}$

The sufficiency for the unique solution x_0 of $\mathcal{P}_1$ is shown by [2]. The necessity can also be derived from convex optimization theory. Detailed proofs will be given in Appendix.

However this condition is difficult to check due to the presence of w. However if we further assume that $w \in im(A_T)$, then the condition in Theorem 1 reduces to the following condition.

Irrepresentable condition *$A_T^*A_T$ is invertible and*

$$\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty < 1, \tag{2}$$

where $*$ denote matrix transpose and $\|\cdot\|_\infty$ stands for the matrix ∞-norm, i.e. the maximum absolute row sum of the matrix such that $\|A\|_\infty := \max_j \sum_i |A_{ij}|$.

Note that this condition only depends on A and the true sparsity pattern of x_0, which is easy to check. The restriction $w \in im(A_T)$ does not put a too strong constraint, which is actually the necessary condition that x_0 can be reconstructed by Lasso [12] or Dantzig selector [3], even under some Gaussian-like noise assumptions [15, 14].

Corollary 1 *If the Irrepresentable condition holds, then x_0 is the unique solution of $\mathcal{P}_1$ and lies on a unique regularization path of $\mathcal{P}_{1,\delta}$.*

In the following we will present some further conditions which are easily checkable to satisfy the Irrepresentable condition in (2).

3.2 Common Interest Groups of Equal Size

We consider the case where A is $R^{j,k}$. Given data b defined on all $j-$subsets, we wish to infer common interest groups on all $k-$subsets so that low order interaction data b can be viewed as induced from high order common interest groups. Suppose x_0 is a sparse signal on all $k-$subsets. We have the following theorem:

Theorem 2 *Let $T = supp(x_0)$, if we allow overlaps among common interest groups to be no larger than r, then the maximum r that can guarantee the irrepresentable condition is $j-2$.*

This is a direct conclusion of the following three results.

Lemma 1 *Let* $T = \text{supp}(x_0)$, and $j \geq 2$. Suppose that for any $\sigma_1, \sigma_2 \in T$, there holds $|\sigma_1 \cap \sigma_2| \leq r$.

1. If $r = j-2$, then $\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty < 1$;

2. If $r = j-1$, then $\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty \leq 1$ where equality holds with certain examples;

3. If $r = j$, there are examples such that $\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty > 1$.

The proofs are based on combinatorial arguments and will be given in Appendix. Theorem 2 thus provides us with a theoretical sufficient and necessary condition on how many overlaps we should allow to guarantee the Irrepresentable Condition. Clique overlaps no more than $j-2$ will be suffice to guarantee the exact sparse recovery by $\mathcal{P}_1$, while larger overlaps may violate the Irrepresentable Condition. Note that this theorem is an analysis in the worse case, so in application, one may encounter examples which has larger overlaps than $j-2$ where $\mathcal{P}_1$ still works.

To conclude this section, we note that Irrepresentable condition(*IRR*) is sufficient and almost necessary to guarantee exact recovery. Theorem 2 tells us the intuition behind the *IRR* is that *overlaps between cliques are small* which is also easily verifiable. In the next section, we will see *IRR* is also sufficient to guarantee stable recovery of cliques.

4 Stable Recovery with Bounded Noise and Mixed Sizes

In real applications, one almost always encounters examples with noise such that exact sparse recovery is impossible. In this case, $\mathcal{P}_{1,\delta}$ will be a good replacement of $\mathcal{P}_1$ as a robust reconstruction algorithm. We will also generalize our algorithm so as to deal with identifying cliques with mixed sizes.

4.1 Stable Recovery Theorems

In the previous sections, we have given various sufficient conditions to recover sparse signal x_0 from the convex program $\mathcal{P}_1$, where b exactly equals Ax_0. In reality, one often meets with noisy observations with $b = Ax_0 + z$, where z accounts for noise. Extended algorithms from $\mathcal{P}_1$ to denoising has been studied extensively in the literature, under the names of BPDN [5], LASSO [12], and Dantzig selector [3], etc. These methods differ in their assumptions on the noise. In this paper, we choose $\mathcal{P}_{1,\delta}$ as we found it heuristically useful to assume bounded noise $|z| \leq \epsilon$ in our applications.

The following theorem is about the stable recovery of $\mathcal{P}_{1,\delta}$ under bounded noise assumptions; its proof is given in the Appendix.

Theorem 3 *Assume that $\|z\|_\infty \leq \epsilon$, $|T| = s$, and the Irrepresentable condition $\|A^*_{T^c} A_T (A^*_T A_T)^{-1}\|_\infty \leq \alpha \leq 1/s$. Then the following error bound holds for any solution $\hat{x}_\delta$ of $\mathcal{P}_{1,\delta}$,*

$$\|\hat{x}_\delta - x_0\|_1 \leq \frac{2s(\epsilon+\delta)}{1-\alpha s} \|A_T (A^*_T A_T)^{-1}\|_1, \qquad (1)$$

In the particular case where $k = j + 1$, we have the following corollary.

Corollary 2 *Assume that $k = j+1, |T| = s$, and overlap $|\sigma_1 \cap \sigma_2| \leq j-2$ for any $\sigma_1, \sigma_2 \in T$. Then $\|A^*_{T^c} A_T (A^*_T A_T)^{-1}\|_\infty \leq 1/(j+1)$ and the following error bound for solution $\hat{x}_\delta$ of $\mathcal{P}_{1,\delta}$ holds:*

$$\|\hat{x}_\delta - x_0\|_1 \leq \frac{2s(\epsilon+\delta)}{1-\frac{s}{j+1}} \sqrt{j+1}, \qquad s < j+1.$$

4.2 Identifying Cliques with Mixed Sizes

In general, we need to deal with identifying cliques of mixed sizes. Suppose we wish to detect high order cliques of sizes $k_1, k_2, \cdots, k_l (k_1 < k_2 < \cdots < k_l)$ from low order information b on j-subsets. One possible way is to construct basis matrix A by concatenating $R^{j,k}$ with different k's together. We can solve $\mathcal{P}_1$ and $\mathcal{P}_{1,\delta}$ for exact recovery and stable recovery with this new concatenated basis matrix A.

Theorem 4 *Suppose x_0 is a sparse signal on cliques of sizes $k_1, k_2, \cdots, k_l (k_1 < k_2 < \cdots < k_l)$ and $b = Ax_0$. Let $T = supp(x_0)$, if the common interest groups in T have no overlaps, then they can be identified by solving $\mathcal{P}_1$. Moreover, if the data $b = Ax_0 + z$ is noised, then solving $\mathcal{P}_{1,\delta}$ will find the approximating solution of x_0 where inequality (1) still holds.*

The above theorem gives us a sufficient condition to guarantee exact sparse recovery with concatenated basis and stable recovery theory is also established.

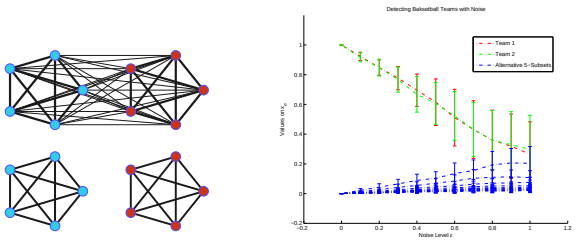
We note that we can also detect cliques with mixed sizes in a stagewise way, i.e., we built different linear programming problems $\mathcal{P}_{1,\delta}$'s with different $A = R^{j,k}$ and b's where k ranges from k_1 to k_l. We can detect cliques of sizes k_i from solving linear programming problem to yield solutions $\hat{x}_i$ which tells us important k_i-cliques. Once a solution $\hat{x}_i$ is obtained, we need to remove its effect by feeding the residue $b_i - A_i \hat{x}_i$ into the next stage as data on $j-$subsets. In our practical experience, this stagewise algorithm works well. Detecting cliques with mixed sizes in increasing order or decreasing order yield similar results.

4.3 Complexity

The basis matrix $R^{j,k}$ is of size $\binom{n}{j}$ by $\binom{n}{k}$ which makes solving the LP program $\mathcal{P}_1$ or $\mathcal{P}_{1,\delta}$ be impossible for all but very small n. However, the following approaches can be considered to improve scalability.

One way is a divide-and-conquer approach utilizing spectral clustering, which can be used to partition, for example, the large social network into subsets(each of size about $20-30$), followed by our LP algorithm in each subset to detect cliques within clusters(interesting cliques typically arise within clusters of this size). This approach is efficient for obtaining good approximation solutions in practice. In theory, it is equivalent to down-sampling columns of the basis matrix A, preserving the support of signals as much as possible, whence the theoretical results above still hold.

The second way to achieve scalability is to solve the LP program iteratively, without writing out A explic-



(a) (b)

Figure 1: Detecting Basketball Teams with Noise. (a) Two teams in a virtual Basketball Game, with intra-team interaction 1 and cross-team interaction noise no more than ϵ; (b) Under a large noise level $\epsilon < 0.9$, the two teams are identifiable. For each noise level, we run 100 simulations repeatedly, whose errorbar plot of weights on cliques are shown.

itly. Note that the Radon basis matrix $R^{j,k}$ has special structure to exploit($R^{j,k}$ is sparse and has a combinatorial structure). Hence, we may not need to form matrix A explicitly if we use iterative algorithms, e.g. Bregman iterations which has a guaranteed convergence rate[1]. If the basis matrix A is a down-sampled matrix of $R^{j,k}$, then a single iteration of the Bregman algorithm has complexity $\mathcal{O}(\binom{k}{j}(\binom{n}{j}+N))$ where N is the number of down-sampling columns. In real cases, we often have data on pairwise interactions which reduce the one-step complexity to be $\mathcal{O}(k^2(n^2+N))$.

Another way to address the scalability is to develop heuristic algorithm. For sparse recovery problem, one may proceeds by matching pursuit, which relies on iteratively find the basis which maximize the correlation between the basis itself and the data residual. We note that computing the correlation is computationally feasible by making use of the special combinatorial structure of $R^{j,k}$. However, to search within all possible basis to select the one which has the largest correlation is a hard problem. Some heuristic techniques, such as branch and bound algorithm, may be considered to address this problem.

5 Application Examples

We demonstrate three application examples, which show the effectiveness of the scheme proposed in this paper. As we will see in this section, our clique-based model can deal with overlaps between cliques which gives us more community structural information compared against using purely clustering methods. Our model can be combined with clustering methods which help improve the scalability of our algorithm.

5.1 Basket-ball team detection

Detecing two basketball teams from pairwise interactions among plays is an ideal scenario. Suppose we have x_0 which is a signal on all 5-subsets of the 10-player set. We assume it is sparsely concentrated on two 5-subsets which correspond to the two teams with magnitudes both equal to one. Assume we have observations b of pairwise interactions which is $b = Ax_0 + z$, where z is uniform random noise distributed in $[-\epsilon, \epsilon]$. We solve $\mathcal{P}_{1,\delta}$, with $\delta = \epsilon$, which is a linear programming searching over $x \in R^{\binom{10}{5}} = R^{252}$ with parameters $A \in R^{\binom{10}{2}\times\binom{10}{5}} = R^{45\times 252}$ and $b \in R^{45}$.

The two $5-$subsets correspond to the two teams have no overlap, hence satisfy the Irrepresentable Condition. In Figure 1-(b), we try to detect the two teams under different noise levels $\epsilon \in [0, 1]$. The two basketball teams can be detected under fairly large noise levels. This example can also be dealt with using spectral clustering techniques where we normalize the pairwise interaction data to get the transition matrix, followed by spectral clustering on eigenspaces. We observed that both our method and spectral clustering works very well under noise level less than 0.8.

5.2 Communities in social networks

We consider the social network [10] of Victor Hugo's novel *Les Miserables*, where we extract 33 characters, and represent the social network of those characters in a weighted graph manner (Figure 2-(a)). The weights on edges represent frequencies of co-appearences.

The underlying social community, which is regarded as the groundtruth for the data can be roughly summarized in figure 2-(a) where several social communities arise. We can run spectral clustering on this social network and the result is shown in figure 2-(b) where the first three red cuts are reasonable while the following three blue cuts destroyed a lot of community structures within the network.

We test our algorithm directly on this social network. Our implementation first detects 3-cliques from pairwise interactions. Among $\binom{33}{3} = 5456$ triangles, the top 5 triangles are shown in Table 1. After that, we remove those triangles' effects and detect 4-cliques from the residual. The top 3 tetrahedra from $\binom{33}{4} = 40929$ tetrahedra are shown in Table 1. Figure 2-(c) and (d) depict these 3 and 4 cliques respectively. The sparsity patterns of those cliques satisfy the irrepresentable condition where overlaps between them are generally not large. However, they do not necessarily satisfy the condition in Lemma 1.1 which is based on worst-case considerations.

Table 1: The Social Network of Key Characters in *Les Miserables*

Cliques	Names of Characters	Relationships
$\{1, 2, 3\}$	{Myriel, Mlle Baptistine, Mme Magloire}	Friendship
$\{4, 12, 16\}$	{Valjean, Fantine, Javert}	Dramatic Conflicts
$\{4, 13, 14\}$	{Valjean, Mme Thenardier, Thenardier}	Dramatic Conflicts
$\{4, 15, 22\}$	{Valjean, Cosette, Marius}	Dramatic Conflicts
$\{20, 21, 22\}$	{Gillenormand, Mlle Gillenormand, Marius}	Kinship
$\{5, 6, 7, 8\}$	{Tholomyes, Listolier, Fameuil, Blacheville}	Friendship
$\{9, 10, 11, 12\}$	{Favourite, Dahlia, Zephine, Fantine}	Friendship
$\{14, 31, 32, 33\}$	{Thenardier, Gueulemer, Babet, Claquesous}	Street Gang

Clearly, the result of our algorithm gives more abundant social structure information than using clustering techniques. Our algorithm can return social communities with overlaps which is impossible to happen using clustering methods. However, searching among all $k-$cliques out of n nodes will be intractable for all but very small n. To resolve this issue, we run spectral clustering to pre-process the data and then within each cluster we detect cliques using our method, whose results are shown in figure 2-(e). More important social cliques, such as the student union clique, can be identified in this case.

We finally note that some simple schemes will not work well. For example, one may think of scoring each large clique by the mean scores of the included small cliques. In this example, since two or three key characters appear very often, we will end up with finding that the top high order cliques always contain them. In fact, among the top ten 3-cliques, seven of them contain node 4 and six of them contain node 15, which does not give us good results.

5.3 Coauthorships in Network Science

In this section, we show an example of application of our algorithm to large scale social networks. We use bipartite spectral graph partitioning algorithm to preprocess the data followed by our cliques identification algorithms within each cluster whose sizes can be handled. We look at the persistence of identified cliques in the binary tree decomposition of bipartite spectral clustering of the network in a bottom-up way. Cliques which persist through more levels will give us meaningful community structural information.

In particular, we studied coauthorship relations between scientists working on network theory and experiment. The network contains 379 individuals whose names appear as authors of papers and weights assigned to edges as described in [11](Figure 3-(a)). We run bipartite spectral clustering on the data which returns us a binary tree decomposition of the network with each node in the binary tree represents a cluster.

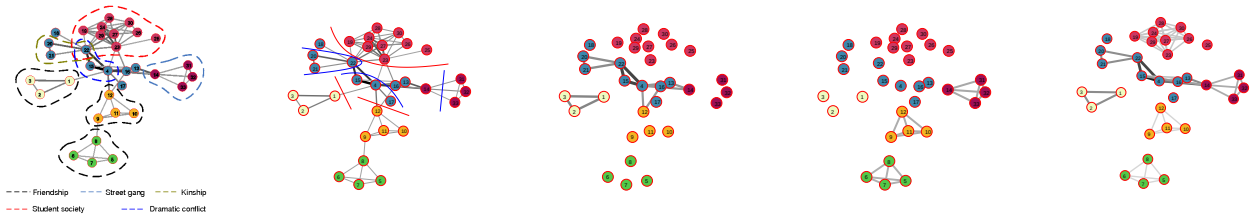
In figure 3-(b), a small fraction of the binary tree decomposition of bipartite spectral clustering is depicted, where child nodes are spectral bipartition of the parent node. We can detect cliques of mixed sizes(up to 10) within the child nodes, e.g., C and D by solving $\mathcal{P}_{1,\delta}$ where basis matrix is concatenation of Radon basis matrix. Once cliques within clusters C, D are identified, we then backtrack to the parent node B and A to see if the identified cliques still persist.

We can identify 3 cliques(c_1={KR, RP, RS, TA}, c_2={KS, RP, RS}, c_3={RP, RS, TA, KS} where KR=Kumar S, RP=Raghavan P, RS=Rajagopalan S, TA=Tomkins A, KS=Kumar S) within C and 3 cliques(d_1={FG, LS, GC, CF}, d_2={FG, LS, GC, PD, GE}, d_3={FG, LS, GC} where FG=Flake G, LS=Lawrence S, GC=Giles C, CF=Coetzee F, PD=Pennock D, GE=Glover E) within D which persist parents to B and A. We can identify papers whose authors are exactly those cliques. Using only clustering will not get this result since there are heavy overlaps between them. In figure 3-(b), for simplicity, we only show two persistent cliques: c_1={KR, RP, RS, TA} and d_1={FG, LS, GS, CF} which are the most important cliques(having the largest weights when solving LP program) in cluster C and D respectively. These two cliques are also the most important two cliques in cluster B, and if we even further back track them to clustering A, they are still ranked as #1 and #3 in terms of weights among all cliques identifiable in A.

This application example shows that our approach can be used to identify cliques in social networks with hundreds or even thousands of nodes, with the help of spectral clustering methods.

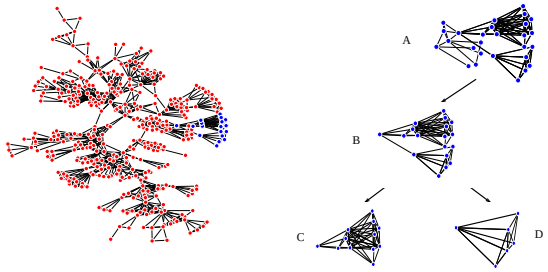
6 Conclusion

In this paper, we have proposed a novel algebraic approach to study the identification of cliques based on low order interaction information. This approach exploits the Radon basis with sparse recovery algorithms rooted in Basis Pursuit. We have shown that noiseless exact recovery and stable recovery with uniformly bounded noise hold under some natural conditions. We have demonstrated successful applications in a simulated model of the basketball team identification, as



(a) (b) (c) (d) (e)

Figure 2: Decomposition of *Les Miserables* social network. (a) Social network of characters in *Les Miserables*; (b) Spectral clustering result; (c) The identified 3-cliques; (d) The identified 4-cliques. (e) The identified cliques after spectral clustering



(a) (b)

Figure 3: (a) Coauthorships in Network Science; (b) Important cliques identified within clusters in a persistent way. Clustering node B is exactly the blue part in (a)

well as two real examples of detecting cliques within medium or large scale social networks. These results show the potential of broad applications of Radon basis pursuit in the studies of identity management, social networks, and statistical ranking.

References

[1] Cai J F and Osher S and Shen Z, Linearized Bregman Iterations for Compressed Sensing, *Math. Comp.*, 78(267): 1515-1536, 2009.

[2] Candès E J and Tao T, Decoding by Linear Programming, *IEEE Transactions on Information Theory*, 51: 4203-15, 2005.

[3] Candès E J and Tao T, The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Statist.*, 35(6): 2313-2351, 2007.

[4] Candès E J, The Restricted Isometry Property and its implications for Compressed Sensing, *Comptes Rendus de l'Académie des Sciences, Paris, Série I*, 346: 589-592, 2008

[5] Chen S, Donoho D L and Saunders M A, Atomic Decomposition by Basis Pursuit, *SIAM J. Scientific Computing*, 20: 33-61, 1999.

[6] Diaconis P, *Group Representations in Probability and Statistics*, IMS Press, 1988.

[7] Goldberg K, T Roeder, D Gupta, and C Perkins, Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2): 133-151, 2001.

[8] Guibas L J, The Identity Management Problem – a short survey, In: 11^{th} *International Conference on Information Fusion*, 2008.

[9] Jagabathula S. and Shah D, Inferring Rankings under Constrained Sensing, In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[10] Knuth D E, *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA, 1993.

[11] Newman M E J, Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev.* E 74, 036104, 2006.

[12] Tibshirani R, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58(1): 267-288, 1996.

[13] Ye Y, *Interior Point Algorithms: Theory and Analysis*, Wiley, 1997.

[14] Yuan M and Y Lin, On the Nonnegative Garrote Estimator, *Journal of the Royal Statistical Society. Series B*, 69 (2), pp. 143-161, 2007.

[15] Zhao P and Yu B, On Model Selection Consistency of Lasso, *Journal of Machine Learning Research*, 7: 2541-2563, 2006.

A. Appendix

A.1 Notations

Given an M by N matrix A, denote by $(v_\tau) \in R^M$ the columns of the matrix A. Let $\bar{x} \in R^N$, $T = \text{supp}(\bar{x})$ and T^c be the complement of T. Denote by A_T the submatrix formed by all columns v_τ where $\tau \in T$ and A_{T^c} the submatrix formed by all columns when $\tau \in T^c$. A^* denote the conjugate of A, which is simply the matrix transpose in this paper.

A.2 Proof of Theorem 1 and Corollary 1

Proof of the following lemma can be found in [2].

Lemma A-1 *The linear programming $\mathcal{P}_1$ has a unique solution $\bar{x}$ if the matrix A_T has full rank and if one can find a vector $w \in R^N$ with the following two properties*

1. $\langle w, v_\tau \rangle = sgn((\bar{x})_\tau)$ *for all* $\tau \in T$,
2. $|\langle w, v_\tau \rangle| < 1$ *for all* $\tau \in T^c$,

where $sgn((\bar{x})_\tau)$ is the sign of $(\bar{x})_\tau$ ($sgn((\bar{x})_\tau) = 0$ for $(\bar{x})_\tau = 0$).

The following lemma is a result by the Karush-Kuhn-Tucker (KKT) condition of $\mathcal{P}_{1,\delta}$.

Lemma A-2 *The two conditions in Lemma A-1 are necessary and sufficient such that the linear programming $\mathcal{P}_{1,\delta}$ has a unique solution.*

Proof. Consider an alternative form of $\mathcal{P}_{1,\delta}$,

$$\begin{array}{rl} \min & 1^T\xi \\ \text{subject to} & -\delta \le Ax - b \le \delta, \quad \delta \ge 0 \\ & -\xi \le x \le \xi, \quad \xi \ge 0 \end{array}$$

whose Lagrangian is

$$\begin{array}{rcl} L_{x,\delta,\xi;\gamma,\lambda,\mu,\nu} & = & 1^T\xi - \mu^T\xi - \nu^T\delta \\ & & -\gamma_+^T(\delta - Ax + b) - \gamma_-^T(Ax - b + \delta) \\ & & -\lambda_+^T(\xi - x) - \lambda_-^T(\xi + x) \end{array}$$

Then the KKT condition gives

1. $0 = \frac{\partial L}{\partial x} = A^*(\gamma_+ - \gamma_-) + (\lambda_+ - \lambda_-)$,
2. $0 = \frac{\partial L}{\partial \xi} = 1 - (\lambda_+ + \lambda_-) - \mu = 0$,

with $\gamma, \lambda, \mu \ge 0$ and $\gamma_+(\tau)\gamma_-(\tau) = \lambda_+(\tau)\lambda_-(\tau) = 0$ for all τ.

Clearly $T = \{\tau : \delta_\tau > 0\}$. Define $w = \gamma_+ - \gamma_-$. Then the first equation leads to

$$\langle w, v_\tau \rangle = -(\lambda_+(\tau) - \lambda_-(\tau)) = -sgn(\bar{x}_\tau), \qquad \tau \in T.$$

On the other hand, by the Strictly Complementary Theorem for linear programming [13], there are $1 > \mu_\tau > 0$ for $\tau \in T^c$ with $\delta_\tau = 0$ such that the second equation leads to

$$|\langle w, v_\tau \rangle| = |\lambda_+(\tau) - \lambda_-(\tau)| = 1 - \mu_\tau < 1,$$

which is the necessary and sufficient condition for the unique solution of $\mathcal{P}_{1,\delta}$. ◇

Theorem 1 is a direct result yielded from the two lemmas above. To see Corollary 1, note that with $M > |T|$ and the injectivity of A_T, if $w \in im(A_T)$, then the first condition in Lemma A-1 leads to

$$w = A_T(A_T^*A_T)^{-1}\iota^* sgn(\bar{x}),$$

where the imbedding operator $\iota : l_2(T) \to l_2(N)$ extends a vector on T to a vector in R^N by placing zeros outside of T and ι^* is the dual restriction $\iota^*\bar{x} = \bar{x}|_T$. With this the second condition in Lemma A-1 can be rewritten as

$$\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\iota^* sgn(\bar{x})\|_\infty < 1,$$

which is exactly the Irrepresentable condition.

A.3 Proof of Lemma 1

To prove Lemma 1, given any $\tau \in T^c$, we define

$$\mu_\tau := \sum_{\sigma \in T} \frac{\binom{|\tau \cap \sigma|}{j}}{\binom{k}{j}},$$

then $\sup_{\tau \in T^c} \mu_\tau = \|A_{T^c}^*A_T\|_\infty$. As we will see in the following proofs, we essentially try to bound μ_τ for $\tau \in T^c$.

A.3.1 Proof of Lemma 1-1

Under condition 1, since any $\sigma_1, \sigma_2 \in T$ satisfy $|\sigma_1 \cap \sigma_2| \le j-2$, hence any two columns in T are orthogonal. This implies $A_T^*A_T$ is an identity matrix.

Now given $\tau \in T^c$, we will prove $\mu_\tau < 1$ under condition 1. If this is true, then

$$\sup_{\tau \in T^c} \mu_\tau = \|A_{T^c}^*A_T\|_\infty = \|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty < 1$$

Let $T = \{\sigma_1, \sigma_2, \cdots, \sigma_{|T|}\}$ where $\sigma_i (1 \le i \le |T|)$ are $k-$subsets. We need to prove

$$\mu_\tau = \sum_{i=1}^{|T|} \frac{\binom{|\tau \cap \sigma_i|}{j}}{\binom{k}{j}} < 1$$

Let $A_i = \{\rho : |\rho| = j, \rho \subset \tau \cap \sigma_i\}$, so A_i is a collection of $j-$subsets of $\tau \cap \sigma_i$ (Here if $|\tau \cap \sigma_i| < j$, then A_i

is simply an empty set). Obviously, we have $|A_i| = \binom{|\tau\cap\sigma_i|}{j}$. So

$$\sum_{i=1}^{|T|}\binom{|\tau\cap\sigma_i|}{j} = \sum_{i=1}^{|T|}|A_i|.$$

Now we note the fact that for any $1 \le i,l \le |T|$, we have $A_i \cap A_l = \emptyset$. This is true because otherwise suppose $\rho \in A_1 \cap A_2$, then this mean ρ is a $j-$subset of A_1 and A_2. Hence $\rho \subset \tau\cap\sigma_1, \rho \subset \tau\cap\sigma_2$, which implies that

$$|\sigma_1\cap\sigma_2| \ge |(\tau\cap\sigma_1)\cap(\tau\cap\sigma_2)| \ge |\rho| \ge j$$

This contradicts with the condition that σ_i's$(1 \le i \le T)$ have overlaps at most $j-2$. So A_i must be pairwise disjoint. Hence

$$\sum_{i=1}^{|T|}\binom{|\tau\cap\sigma_i|}{j} = \sum_{i=1}^{|T|}|A_i| = |\cup_{i=1}^{|T|}A_i|$$

For any $1 \le i \le |T|$, every $\rho \in A_i$ is a $j-$subset of $\tau\cap\sigma_i$. Hence ρ is of course a $j-$subset of τ. The set τ is of size k. So if we let $A_0 = \{\rho : |\rho| = j, \rho \subset \tau\}$ which is the collection of all $j-$subsets of τ, then we have $\cup_{i=1}^{|T|}A_i \subset A_0$. So $|\cup_{i=1}^{|T|}A_i| \le |A_0| \le \binom{k}{j}$.

Till now, we actually proved $\mu_\tau \le 1$. All the above proof about $\mu_\tau \le 1$ for any $\tau \in T^c$ will remain valid for condition 2. In the next, we prove if any $\sigma_i, \sigma_l \in T$ satisfy $|\sigma_i\cap\sigma_l| \le j-2$, then equality can not hold.

Without loss of generality, we assume $|\sigma_1\cap\tau| \ge j$, otherwise if none of σ_i's satisfies $|\sigma_i\cap\tau| \ge j$, then $\mu_\tau = 0$ which actually finishes the proof. In this case, we can let $\tau = \{1,2,\cdots,k\}$, $\sigma_1 = \{1,2,\cdots,s,k+1,k+2,2k-s\}$ where $j \le s \le k-1 (s \le k-1$ because otherwise $\sigma_1 = \tau$ which contradicts with the fact that $\sigma_1 \in T, \tau \in T^c)$. Now we show that $\rho_0 = \{1,2,\cdots,j-1,s+1\}$ is not a member of $\cup_{i=1}^{|T|}A_i$. Clearly ρ_0 is not a member of A_1 because $s+1 \notin \sigma_1$. Now it remains to show that ρ_0 is not a member of any $A_i (2 \le i \le |T|)$. If this was not true, say $\rho_0 \in A_2$, then $\rho_0 \subset (\tau\cap\sigma_2) \subset \sigma_2$, then $\{1,2,\cdots,j-1\} \subset \sigma_1\cap\sigma_2$, which contradicts with the condition that $|\sigma_1\cap\sigma_2| \le j-2$.

While it is clear that $\rho_0 \in A_0$, so this means $\cup_{i=1}^{|T|}A_i$ is a proper subset of A_0. So $|\cup_{i=1}^{|T|}A_i| < \binom{k}{j}$ which means $\mu_\tau < 1$. ◇

A.3.2 Proof of Lemma 1-2

Under condition 2, then almost the same as proof for lemma 1. We have $A_T^*A_T$ is an identity matrix and $\mu_\tau \le 1$. However, one can not show $\mu_\tau < 1$ in this case. We have the following example where if n is large enough, then μ_τ can happens to be equal to one exactly.

Let $\tau = \{1,2,\cdots,k\} \in T^c$. Denote all the $j-$subsets of τ to be $\rho_1,\rho_2,\cdots,\rho_{\binom{k}{j}}$. For n is large enough, we choose $\binom{k}{j}$ disjoint $(k-j)$-subsets of $\{k+1,k+2,\cdots,n\}$, denoted by $\omega_1,\omega_2,\cdots,\omega_{\binom{k}{j}}$.

Let $T = \{\sigma_1,\sigma_2,\cdots,\sigma_{|T|}\}$, where $\sigma_i = \rho_i \cup \omega_i$. Hence $|T| = \binom{k}{j}$ and σ_i's satisfy $|\sigma_i\cap\sigma_j| \le j-1$. But

$$\sum_{i=1}^{|T|}\frac{\binom{|\tau\cap\sigma_i|}{j}}{\binom{k}{j}} = \sum_{i=1}^{|T|}\frac{1}{\binom{k}{j}} = 1$$

◇

A.3.3 Proof of Lemma 1-3

Under condition 3, we can construct examples where $\|A_{T^c}^*A_T(A_T^*A_T)^{-1}\|_\infty > 1$. Let $\rho_1,\rho_2,\cdots,\rho_{\binom{k}{j}}$ be all j-subsets of $\{1,2,\cdots,k\}$. For large enough n, it is possible to choose $\binom{k}{j}+1$ disjoint $(k-j)$-subsets of $\{k+1,k+2,\cdots,n\}$, say $\omega_0,\omega_1,\omega_2,\cdots,\omega_{\binom{k}{j}}$. Let $\sigma_i = \rho_i\cup\omega_i$ for $1 \le i \le \binom{k}{j}$ and $\sigma_0 = \rho_1\cup\omega_0$. Define $T = \{\sigma_0,\sigma_1,\sigma_2,\cdots,\sigma_{\binom{k}{j}}\}$ which is of size $|T| = \binom{k}{j}+1$.

In this case, $|\sigma_i\cap\sigma_l| = j-1$ for any $1 \le i,l \le \binom{k}{j}$ and $|\sigma_0\cap\sigma_1| = j$, $|\sigma_0\cap\sigma_i| \le j-1$ for any $2 \le i \le \binom{k}{j}$. Then $A_T^*A_T$ is a $\binom{k}{j}+1$ by $\binom{k}{j}+1$ matrix shown belog with rows and columns corresponds to $\{\sigma_0,\sigma_1,\cdots,\sigma_{\binom{k}{j}}\}$

$$A_T^*A_T = \left[\begin{array}{cc|ccc} 1 & \epsilon & 0 & 0 & \cdots & 0 \\ \epsilon & 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{array}\right]$$

Here $\epsilon = \frac{1}{\binom{k}{j}}$. The inverse of the matrix is

$$(A_T^*A_T)^{-1} = \left[\begin{array}{cc|ccc} \frac{1}{1-\epsilon^2} & -\frac{\epsilon}{1-\epsilon^2} & 0 & 0 & \cdots & 0 \\ -\frac{\epsilon}{1-\epsilon^2} & \frac{1}{1-\epsilon^2} & 0 & 0 & \cdots & 0 \\ \hline 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{array}\right]$$

Consider $\tau = \{1,2,\cdots,k\} \in T^c$, then the row corresponds to τ for $A_{T^c}^*A_T$ is a vector of length $|T| = \binom{k}{j}+1$ with each entry being $\epsilon = \frac{1}{\binom{k}{j}}$. So the row vector corresponds to τ in $A_{T^c}^*A_T(A_T^*A_T)^{-1}$ is a vector

of length $\binom{k}{j}+1$, $[\frac{\epsilon}{1+\epsilon}, \frac{\epsilon}{1+\epsilon}, \epsilon, \epsilon, \cdots, \epsilon]$. This vector has row sum

$$\begin{aligned} \frac{2\epsilon}{1+\epsilon} + (\binom{k}{j} - 1)\epsilon &= \frac{2\epsilon}{1+\epsilon} + (\frac{1}{\epsilon} - 1)\epsilon \\ &= \frac{1+2\epsilon-\epsilon^2}{1+\epsilon} \\ &> \frac{1+2\epsilon-\epsilon}{1+\epsilon} = 1 \end{aligned}$$

Hence in this example $\|A_{T^c}^* A_T (A_T^* A_T)^{-1}\|_\infty > 1$. ⋄

A.4 Proof of Theorem 3 and Corollary 2

Lemma A-3 *Assume that $\|z\|_\infty \le \epsilon$, $|T| = s$, and the Irrepresentable condition*

$$\|A_{T^c}^* A_T (A_T^* A_T)^{-1}\|_\infty \le \alpha < 1.$$

Then the following error bound holds for any solution $\hat{x}_\delta$ of $\mathcal{P}_{1,\delta}$,

$$\|\hat{x}_\delta - \bar{x}\|_1 \le \frac{2s(\epsilon+\delta)}{1-\alpha s} \|A_T (A_T^* A_T)^{-1}\|_1.$$

Proof of Lemma A-3. Note that $\|A\hat{x}_\delta - b\|_\infty \le \delta$ and $z = A\bar{x} - b$ with $\|z\|_\infty \le \epsilon$. Then

$$\begin{aligned} \|Ah\|_\infty &= \|A\hat{x}_\delta - A\bar{x}\|_\infty = \|A\hat{x}_\delta - b + b - A\bar{x}\|_\infty \\ &\le \|A\bar{x}_\delta - b\|_\infty + \|z\|_\infty \le \delta + \epsilon. \end{aligned} \tag{3}$$

Let $h = \hat{x}_\delta - \bar{x}$. By $\|\bar{x}\|_1 \ge \|\hat{x}\|_1$,

$$\begin{aligned} \|h_T\|_1 &= \|\bar{x} - \hat{x}_\delta|_T\|_1 \ge \|\bar{x}\|_1 - \|\hat{x}_\delta|_T\|_1 \\ &\ge \|\hat{x}_\delta\|_1 - \|\hat{x}_\delta|_T\|_1 = \|\hat{x}_\delta|_{T^c}\|_1 = \|h_{T^c}\|_1. \end{aligned} \tag{4}$$

Therefore,

$$\begin{aligned} & |\langle Ah, A_T (A_T^* A_T)^{-1} h_T \rangle| \\ = \; & |\langle A_T h_T, A_T (A_T^* A_T)^{-1} h_T \rangle + \langle A_{T^c} h_{T^c}, A_T (A_T^* A_T)^{-1} h_T \rangle| \\ \ge \; & \|h_T\|_2^2 - |\langle h_{T^c}, A_{T^c}^* A_T (A_T^* A_T)^{-1} h_T \rangle| \\ \ge \; & \|h_T\|_2^2 - \|h_{T^c}\|_1 \|A_{T^c}^* A_T (A_T^* A_T)^{-1} h_T\|_\infty \\ \ge \; & \frac{1}{s}\|h_T\|_1^2 - \alpha \|h_{T^c}\|_1 \|h_T\|_\infty \\ \ge \; & \frac{1}{s}\|h_T\|_1^2 - \alpha \|h_{T^c}\|_1 \|h_T\|_1 \\ \ge \; & \left(\frac{1}{s} - \alpha\right) \|h_T\|_1^2 \end{aligned}$$

where the last step is due to $\|h_T\|_1 \ge \|h_{T^c}\|_1$ in the inequality (4). On the other hand,

$$\begin{aligned} & |\langle Ah, A_T (A_T^* A_T)^{-1} h_T \rangle| \\ \le \; & \|Ah\|_\infty \|A_T (A_T^* A_T)^{-1} h_T\|_1 \\ \le \; & (\delta + \epsilon) \|A_T (A_T^* A_T)^{-1}\|_1 \|h_T\|_1 \end{aligned}$$

using (3). Combining these two inequalities yields

$$\|h_T\|_1 \le \frac{s(\delta+\epsilon)}{1-\alpha s} \|A_T (A_T^* A_T)^{-1}\|_1,$$

as desired. ⋄

Proof of Corollary 2 This corollary follows follows from the Lemma above. Note that when the conditions in Theorem 2 hold, $A_T^* A_T = I$ and $\|A_T\|_1 \le \sqrt{\binom{k}{j}} = \sqrt{j+1}$.

Now it suffice to eatablish the fact that in this special case, we have

$$\|A_{T^c}^* A_T (A_T^* A_T)^{-1}\|_\infty \le \frac{1}{j+1} < 1$$

Note that since any $\sigma_1, \sigma_2 \in T$ satisfy $|\sigma_1 \cap \sigma_2| \le j-2$, we have $A_T^* A_T$ is an identity matrix. So $\|A_{T^c}^* A_T (A_T^* A_T)^{-1}\|_\infty = \|A_{T^c}^* A_T\|_\infty$. Now assume $\tau \in T^c$, let $S_\tau = \{\sigma : |\sigma \cap \tau| \ge j, \sigma \in T\}$, then $|S_\tau| \le 1$. This is because otherwise, suppose $\{\sigma_1, \sigma_2\} \subset S_\tau$ such that $|S_\tau| \ge 2$, then we have

$$\begin{aligned} |\tau| &\ge |\tau \cap (\sigma_1 \cup \sigma_2)| = |\tau \cap \sigma_1| + |t \cap \sigma_2| - |t \cap \sigma_1 \cap \sigma_2| \\ &\ge j + j - (j-2) = j+2 \end{aligned}$$

which contradicts with the fact that τ is a $j+1$-subset. So there exist at most one $\sigma_0 \in T$ such that $|\tau \cap \sigma| \ge j$. Let v_τ be the row vector of $A_{T^c}^* A_T$ with row index correspond to τ. Then $\|v_\tau\|_\infty \le \frac{\binom{j}{j}}{\binom{j+1}{j}} = \frac{1}{j+1} < 1$. ⋄

A.5 Proof of Theorem 4

We prove under the condition that any $\sigma_1, \sigma_2 \in T$ satisfy $|\sigma_1 \cap \sigma_2| = 0$, then solve $\mathcal{P}_1$ will exactly identify x_0.

For simplicity, given any $\tau \in T^c$, we define

$$\mu_\tau = \sum_{\sigma \in T} \frac{1}{\binom{|\tau|}{j}\binom{|\sigma|}{j}} \binom{|\tau \cap \sigma|}{j}$$

Note that the intersection of σ_1 and σ_2 is zero implies that $A_T^* A_T = I$, moreover, given $\tau \in T^c$, the colleciton of sets $\{\tau \cap \sigma | \sigma \in T\}$ are disjoint. Note that if there is only one σ_0 satisfies $|\tau \cap \sigma_0| \ge j$, then

$$\mu_\tau = \frac{1}{\sqrt{\binom{|\tau|}{j}\binom{|\sigma_0|}{j}}} \binom{|\tau \cap \sigma_0|}{j} < 1$$

because it is the inner product of two column vectors corresponds to τ and σ_0 of A, where there are no two columns in A are identical.

Now suppose there are at least two σ's satisfy, $|\tau \cap \sigma| \ge j$, then we have

$$\mu_\tau = \sum_{\sigma \in T} \frac{1}{\sqrt{\binom{|\tau|}{j}\binom{|\sigma|}{j}}} \binom{|\tau \cap \sigma|}{j}$$

$$\leq \sum_{\sigma\in T,|\tau\cap\sigma|\geq j} \frac{1}{\sqrt{\binom{|\tau|}{j}\binom{|\tau\cap\sigma|}{j}}}\binom{|\tau\cap\sigma|}{j}$$
$$= \sum_{s_k\in T,|\tau\cap\sigma|\geq j} \frac{\sqrt{\binom{|\tau\cap\sigma|}{j}}}{\sqrt{\binom{|\tau|}{j}}}$$

Since the collection of sets $\{\tau\cap\sigma|\sigma\in T\}$ are disjoint, so if we can prove $\sqrt{\binom{|\tau\cap\sigma_1|}{j}}+\sqrt{\binom{|\tau\cap\sigma_2|}{j|}} < \sqrt{\binom{|\tau\cap(\sigma_1\cup\sigma_2)|}{j}}$, then we know that

$$\begin{aligned}\mu_\tau &\leq \textstyle\sum_{\sigma\in T,|\tau\cap\sigma|\geq j} \frac{\sqrt{\binom{|\tau\cap\sigma|}{j}}}{\sqrt{\binom{|\tau|}{j}}} \\ &< \sqrt{\binom{|\tau\cap(\cup_{\sigma\in T,|\tau\cap\sigma|\geq j}\sigma)|}{j}}/\sqrt{\binom{|\tau|}{j}} \leq 1\end{aligned}$$

So now we only need to prove the following inequality: suppose $j\geq 2$, given $n_1\geq j, n_2\geq j$, we need to prove $\sqrt{\binom{n_1}{j}}+\sqrt{\binom{n_2}{j}} < \sqrt{\binom{n_1+n_2}{j}}$

The case of $j=2$ can be verified directly, while for $j\geq 3$, we square both sides and we now that we only need to prove $\binom{n_1}{j}+\binom{n_2}{j}+2\sqrt{\binom{n_1}{j}\binom{n_2}{j}} < \binom{n_1+n_2}{j}$. Since $\binom{n_1+n_2}{j}=\sum_{s=0^j}\binom{n_1}{j-s}\binom{n_2}{s}$. So we know we only need to prove $2\sqrt{\binom{n_1}{j}\binom{n_2}{j}} < n_2\binom{n_1}{j-1}+n_1\binom{n_2}{j-1}$. Since $n_2\binom{n_1}{j-1}+n_1\binom{n_2}{j-1} \geq 2\sqrt{n_1n_2\binom{n_1}{j-1}\binom{n_2}{j-1}}$, so we only need to verify $n_1\binom{n_1}{j-1} > \binom{n_1}{j}$, this can be easily verified by writing out explicitly both sides.