Cluster Extraction using Sparse Recovery

Ming-Jun Lai and **Daniel Mckenzie**
Department of Mathematics, University of Georgia

Why Consider Cluster Extraction?

- Given a graph $G = (V, E)$, finding clusters $C_1, C_2, \ldots, C_k \subset V$ is of interest in data science. Each C_i should have many internal edges, few edges to rest of graph. It is natural to assume that vertices in the same cluster share important properties.
- Typical algorithms (spectral clustering, GenLouvain, hierarchical clustering) assume that $C_1, \ldots, C_k$ do not overlap and $V = C_1 \cup \ldots \cup C_k$, but real-world graphs are more complicated.
- Would like to allow for overlapping clusters, as well as for background vertices that do not belong to any cluster.
- Real-world graphs are also large. If one is only interested in a certain cluster (*e.g.* the community containing a specified user in a social network) it can be computationally wasteful to find all clusters.

Definition 1 (Cluster Extraction Problem) *Given a graph $G = (V, E)$ and a small set of seed vertices $\Gamma \subset V$, find a good cluster C_1 containing Γ.*

Cluster extraction is agnostic about structure of $V \setminus C_1$. Could be background, other clusters etc.

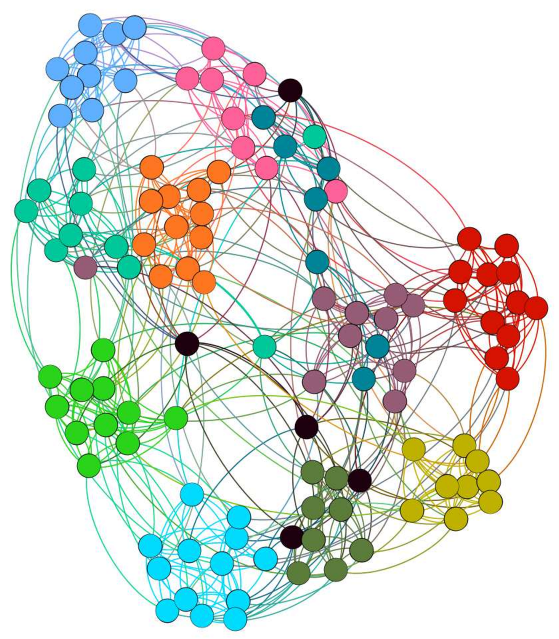

Figure 1: The college football network of [1]. Clusters (indicated by color) correspond to the different conferences. There are five teams, indicated in black, that are independents and should not be assigned to any cluster.

Totally Perturbed Sparse Recovery

For $\mathbf{x}^* \in \mathbb{R}^n$, let $\|\mathbf{x}^*\|_0 := |\{i : x_i^* \neq 0\}|$. If $\|\mathbf{x}^*\|_0$ is small relative to n, we say that $\mathbf{x}^*$ is *sparse*. Given $\mathbf{y} = \Phi\mathbf{x}^*$ and $\Phi \in \mathbb{R}^{m\times n}$ seek to recover $\mathbf{x}^*$ as sparsest solution to linear system $\mathbf{y} = \Phi\mathbf{x}$. Formally:

$$\operatorname{argmin} \|\mathbf{x}\|_0 \text{ such that } \Phi\mathbf{x} = \mathbf{y} \tag{1}$$

In *compressed sensing* $m < n$ so the linear system is underdetermined. Problem (1) is highly non-convex, so either study the convex relaxation (ℓ_1 minimization) or use greedy approach to solve:

$$\operatorname{argmin} \|\Phi\mathbf{x} - \mathbf{y}\|_2 \text{ such that } \|\mathbf{x}\|_0 \leq s := \|\mathbf{x}^*\|_0 \tag{2}$$

Herman & Strohmer [2], Li [4] and others, study problem (2) in presence of *additive* and *multiplicative* noise. That is, suppose $\mathbf{y} = (\Phi + M)\mathbf{x}^* + \mathbf{e}$ and $\mathbf{x}^\#$ is the solution to (2) found using, *e.g.* subspace pursuit or OMP. Is $\mathbf{x}^\# \approx \mathbf{x}^*$?

Theorem 2 (Cor. 1 in [4], simplified) *Let $\hat{\Phi} = \Phi + M$ and $\hat{\mathbf{y}} = \hat{\Phi}\mathbf{x}^*$ where $\|\mathbf{x}^*\|_0 = s$. Suppose that signal $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ is received. Define*

$$\epsilon_{\mathbf{y}} := \|\mathbf{e}\|_2/\|\hat{\mathbf{y}}\|_2 \text{ and } \epsilon_\Phi := \|M\|_{2\to2}/\|\Phi\|_{2\to2}$$

Let $\mathbf{x}^\#$ denote the solution to the following problem, found using subspace pursuit:

$$\operatorname{argmin} \|\Phi\mathbf{x} - \mathbf{y}\|_2 \text{ such that } \|\mathbf{x}\|_0 \leq s \tag{3}$$

Assuming $\delta_{3s} := \delta_{3s}(\Phi) \leq 0.4859$ then:

$$\|\mathbf{x}^* - \mathbf{x}^\#\|_2 \leq C(\delta_s, \epsilon_\Phi, \epsilon_{\mathbf{y}})\|\mathbf{x}^*\|_2 \tag{4}$$

Turning Cluster Extraction into Sparse Recovery

$L = I - D^{-1}A$ denotes the (normalized) Laplacian of G. Let L^{in} denote the Laplacian of $G^{\text{in}} \subset G$ where G^{in} is obtained by deleting all edges between clusters. Note that clusters $C_1, \ldots, C_k$ of G are now connected components of G^{in}.

If $\mathbf{1}_{C_a}$ denotes the *indicator vector* of C_a, then a theorem in spectral graph theory states that $L^{\text{in}}\mathbf{1}_{C_a} = 0$. Importantly, note that $\|\mathbf{1}_{C_a}\|_0 = |C_a| =: n_a$ hence if $|C_a|$ is small relative to $|V|$, $\mathbf{1}_{C_a}$ is *sparse*. Assume, wlog, that $v_1 \in C_1$. We can find $\mathbf{1}_{C_1}$ as solution to:

$$\operatorname{argmin} \|L^{\text{in}}\mathbf{x}\|_2 \text{ subject to } \|\mathbf{x}\|_0 \leq n_1 \text{ and } x_1 = 1 \tag{5}$$

Of course L^{in} is unknown. In [3] we show that $L = L^{\text{in}} + M$ with $\|M\|_{2\to2}$ small. One would hope that if $\mathbf{x}^\#$ is the solution to:

$$\operatorname{argmin} \|L\mathbf{x}\|_2 \text{ subject to } \|\mathbf{x}\|_0 \leq n_1 \text{ and } x_1 = 1 \tag{6}$$

Then by Theorem 2 $\mathbf{x}^\# \approx \mathbf{1}_{C_1}$. Unfortunately problem (6) turns out to be poorly conditioned. Thus we first use the seed vertices $\Gamma \subset C_1$ to find a rough approximation $\Omega \supset C_1$ and then solve a related sparse recovery problem to extract C_1 from Ω.

Semi-Supervised Cluster Pursuit (SSCP) [3]

1. **Input:** Adjacency matrix A, $\Gamma \subset C_1$ and $\hat{n}_1 \approx |C_1|$
2. Compute $L^+ = I + D^{-1}A$ and $\mathbf{b} = \sum_{i\in\Gamma} \ell_i^+$.
3. Let $\mathbf{v} = (L^+_{\Gamma^c})^\top \mathbf{b}$
4. Define $\Omega = \{i : v_i \text{ among } 1.1\hat{n}_1 \text{ largest entries in } \mathbf{v}\} \cup \Gamma$
5. Compute $L = I - D^{-1}A$ and $\mathbf{y} = \sum_{i\in\Omega} \ell_i$
6. Find $\mathbf{x}^\#$ as the solution to $\operatorname{argmin}\{\|L_\Omega\mathbf{x} - \mathbf{y}\|_2 : \|\mathbf{x}\|_0 \leq 0.1\hat{n}_1\}$
7. Let $W^\# = \{i : x_i^\# > 0.5\}$
8. **Output:** $C_1^\# = \Omega \setminus W^\#$, an approximation to C_1

Remark 3 *In step 6 we use subspace pursuit to take advantage of Theorem 2. Using other sparse recovery algorithms is certainly possible.*

Theoretical Guarantees

We consider graphs drawn from the *Symmetric Stochastic Block Model*, $G \sim \text{SSBM}(n, k, p, q)$, where G has k disjoint, equally sized clusters: $V = C_1 \cup \ldots \cup C_k$ and edge $\{v_i, v_j\}$ inserted with probability p if $v_i, v_j \in C_a$ and q if $v_i \in C_a$ and $v_j \in C_b$ for $a \neq b$. Here $|V| = n$ so $|C_a| = n/k$. Using Theorem 2 we prove:

Theorem 4 ([3]) *Suppose $G \sim SSBM(n, k, p, q)$ with k constant, $q = \log(n)/n$ and $p = \omega\log(n)/n$ for any $\omega \to \infty$. Let $C_1^\#$ be the output of SSCP with inputs A, $\Gamma \subset C_1$ where $|\Gamma| = g|C_1|$ for any fixed $g \in (0, 1)$ and $\hat{n}_1 = |C_1|$*

1. $\dfrac{|C_1^\# \setminus C_1| + |C_1 \setminus C_1^\#|}{|C_1|} = o(1)$ *almost surely*
2. `SSCP` *find $C_1^\#$ in $O(n\log^3(n))$ operations.*

Numerical Results

We compared the performance of `SSCP` against several state-of-the-art cluster extraction methods (Tables 1–3). Full experimental details are contained in [3]. Jaccard $:= |C_1^\# \cap C_1|/|C_1^\# \cup C_1|$.

	`SSCP`		`HKGrow`		`LOSP++`		`ESSC`	
	Jaccard	Time	Jaccard	Time	Jaccard	Time	Jaccard	Time
$n = 1000$	0.73	0.01	0.34	0.02	0.66	0.03	0.79	0.32
$n = 2000$	0.85	0.04	0.84	0.01	0.78	0.01	0.70	1.21
$n = 3000$	0.88	0.08	1	0.02	0.81	0.05	0.80	2.34
$n = 4000$	0.92	0.22	1	0.03	0.84	0.1	0.99	2.49
$n = 5000$	0.94	0.34	1	0.03	0.87	0.13	0.94	6.6

Table 1: Results for $G \sim \text{SSBM}(n, 10, p, q)$ with p and q as in Theorem 4.

	`SSCP`		`HKGrow`		`LOSP++`		`ESSC`	
	Jaccard	Time	Jaccard	Time	Jaccard	Time	Jaccard	Time
Caltech	0.43	0.01	0.27	0.004	0.38	0.01	0.43	3.72
Smith	0.33	0.02	0.06	0.02	0.31	0.04	-	-
Rice	0.39	0.14	0.43	0.03	0.42	0.10	-	-
UCSC	0.28	0.35	0.16	0.04	0.28	0.31	-	-

Table 2: Results for four social networks from the `facebook100` data set. Quantities displayed are averaged over ten independent trials per cluster and over all clusters.

	`SSCP`		`HKGrow`		`LOSP++`	
	Jaccard	Time	Jaccard	Time	Jaccard	Time
1%	0.80	3.11	0.63	0.05	0.67	0.93
2%	0.84	3.65	0.65	0.05	0.66	1.61
5%	0.90	3.65	0.75	0.06	0.75	3.48

Table 3: Results for 20 000 MNIST images, averaged over ten independent trials per digit and over all ten digits. Amount of labeled data varied from 1% to 5%.

Concluding Remarks

- I am currently extending this approach to *Dynamic Graphs*: $\mathbb{G} = \{G^{(1)}, \ldots, G^{(T)}\}$.
- All code available at: `danielmckenzie.github.io`
- Questions or comments? `danmac29@uga.edu`

References

[1] Girvan & Newman, Community structure in social and biological networks, (2002).
[2] Herman & Strohmer, General deviants: An analysis of perturbations in compressed sensing, (2010).
[3] Lai & Mckenzie, Semi-Supervised Cluster Extraction via a Compressive Sensing Approach arXiv preprint arXiv:1808.05780 (2018).
[4] H. Li, Improved analysis of SP and CoSaMP under total perturbations, (2016).