Two Clustering Problems Involving the Stochastic Block Model

Ioana Dumitriu

Department of Mathematics
University of Washington (Seattle)

Joint work with Gerandy Brito, Shirshendu Ganguly, Christopher Hoffman, Linh Tran, Maryam Fazel, Roy Han, and Amin Jalali

IPAM
December 7, 2016

1. Intro

2. A General SBM
   - Previous work
   - Our results and improvements
   - Future directions

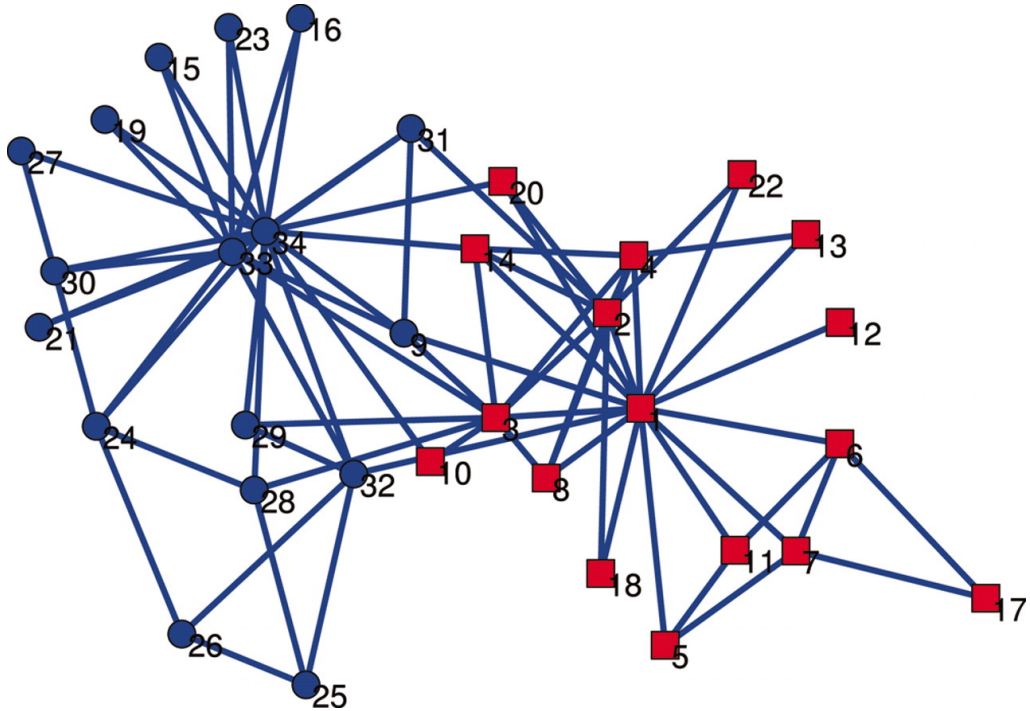
3. A Regular Binary SBM
   - Classical Binary SBM
   - Distinguishability
   - Uniqueness of partition in the model
   - No weak recovery
   - Efficient Strong Recovery
   - Future directions

The Clustering Problem

- Inputs a network with clusters (possibly also overlapping) and asks whether it is possible to detect/recover them accurately and efficiently.
- Applications in machine learning, community detection, synchronization, channel transmission, etc.
- Questions are many and subtle
- Huge body of work: OR, EE, ThCS, Math

The Setup

- One can study actual networks



- Or focus on studying idealized models of networks (e.g., SBM)

The Stochastic Block Model (SBM)

- A.k.a. the "planted partition" model
- Consider K $G(n_i, p_i)$ independent and non-overlapping, joined by a multipartite $G(n_1, \ldots, n_K, q)$.
- Under what sort of conditions on the n_i, p_i, K, q can one recover/approximate/detect the presence of the partition?

Dictionary of terms (always, whp)

- *strong recovery regime*: it is possible to reconstruct the partition completely
-
-
-
-

Dictionary of terms (always, whp)

- *strong recovery regime*: it is possible to reconstruct the partition completely
- *weak recovery regime*: it is possible to reconstruct the partition up to $o(n)$ vertices which may be mislabeled
-
-
-

Dictionary of terms (always, whp)

- *strong recovery regime*: it is possible to reconstruct the partition completely
- *weak recovery regime*: it is possible to reconstruct the partition up to $o(n)$ vertices which may be mislabeled
- *partial recovery regime*: it is possible to reconstruct a constant fraction of the vertices; the rest may be mislabeled
-
-

Dictionary of terms (always, whp)

- *strong recovery regime*: it is possible to reconstruct the partition completely
- *weak recovery regime*: it is possible to reconstruct the partition up to $o(n)$ vertices which may be mislabeled
- *partial recovery regime*: it is possible to reconstruct a constant fraction of the vertices; the rest may be mislabeled
- *detection regime*: it is possible to find a partition correlated to the original one (but non-quantifiably so)
-

Dictionary of terms (always, whp)

- *strong recovery regime*: it is possible to reconstruct the partition completely
- *weak recovery regime*: it is possible to reconstruct the partition up to $o(n)$ vertices which may be mislabeled
- *partial recovery regime*: it is possible to reconstruct a constant fraction of the vertices; the rest may be mislabeled
- *detection regime*: it is possible to find a partition correlated to the original one (but non-quantifiably so)
- *impossibility regime*: it is impossible to find a partition correlated to the original (generally, because of indistinguishability reasons)

SBM Analysis

- Recovery:
  - Huge body of literature in OR/EE/ThCS; possibility of recovery studied via the Maximum Likelihood Estimator (MLE) and convex relaxations using semidefinite programming (SDPs); multiple-structure SDPs (sparse+low-rank, e.g. Vinayak, Oymak, Hassibi (2014)).
  - Most general analysis for recovery via information-theoretic impossibility bounds and a convex relaxation for the MLE in Chen and Xu (2014); various order-sharp bounds for K *equivalent* clusters (K may grow with n).

SBM Analysis

- Recovery:
  - *Threshold criterion* for K fixed and clusters of size $O(n)$ by Abbe, Sandon (2015).
  - Various work on clusters of different sizes and connectivities, but no general threshold criteria.

SBM Analysis

- Weak recovery:
  - *Threshold criterion* for finite K, with all clusters $O(n)$, and all $p_i = p$ by Yun, Proutiere (2014).
- Partial recovery/approximation:
  - Bounds by spectral methods (Coja-Oghlan (2010), Le, Levina, Vershynin (2015)), SDPs (Guedon & Vershynin (2015)).
- Detectability:
  - Empirically-based conjectures by Decelle, Krzakala, Moore, Zdeborova (2011); proved by Abbe, Sandon (2015) together with information-theoretic gap for more than 4 communities

SBM Analysis

The only case that so far has been completely solved, in terms of all various thresholds, is the two "equal" cluster (binary) case. Mossel, Neeman, Sly (2012-2014), Massoulié (2013), Abbe, Bandeira, Hall (2014), Coja-Oghlan (2010).

A General SBM

Model

We assume a partition $V_1, \ldots, V_K$ of the n vertices, $V_i = n_i$. Connect u to v with probability

$$P(u \sim v) = \begin{cases} p_i, & \text{if } \exists i \text{ s.t. } u, v \in V_i \\ q, & \text{otherwise.} \end{cases}$$

No restrictions on the growth of V_is. Find the recovery regimes (when is recovery possible? efficiently possible? impossible?)

Chen, Xu (2014)

- Considered equivalent clusters ($|V_i| = |V_j|, p_i = p_j$ for all i, j)
- Obtained lower bound for impossibility of recovery threshold by information-theoretical means
- Obtained upper bounds for recovery (using MLE) and efficient recovery (using a convex relaxation for the MLE)
- Bounds work up/down to the connectivity threshold order $p = \ln n/n$.
- Hinted at the fact that *by choosing the minimal partition set size and minimal p_i, results can be generalized*.

Our results

In [JHDF16], we use a combination of the ideas in Chen, Xu (2014) and Vinayak, Oymak, Hassibi (2014) to generalize all their bounds.
Main contributions:

- Lower bounds on impossibility threshold, upper bounds on recovery and efficient recovery thresholds, in terms of all involved parameters.
- Identified crucial relative density quantity $\rho_i = n_i(p_i - q)$; all bounds can be expressed in terms of it. All ρ_i must be at least logarithmic in n for recovery.

Improvements over previous work and other results

1. Showed that general bounds are actually useful. Using the "smallest n_i, p_i" intuition is wrong!

   Example 1: $n_1 = n - \sqrt{n}, n_2 = \sqrt{n}, p_1 = n^{-2/3}, p_2 = 1/\log n$, $q = n^{-2/3-0.01}$. Tough case b/c small second cluster, large background noise.

   As per Chen, Xu (2014), using the "smallest n_i, p_i" intuition would indicate this should be an impossible case. Our theorems place this in the recoverable regime.

Improvements over previous work and other results

2. We demonstrated that communities are efficiently recoverable by convex methods up to size $\sqrt{\log n}$. Maybe first example in literature with less than $\log n$, which was taken as threshold. *This may suggest no hard boundary.*

   Example 2: k clusters of size $n_1 = \sqrt{\log n}$ and $p_1 = O(1)$, respectively, $O(\sqrt{n})$ clusters of size $n_2 = O(\sqrt{n})$ and $p_2 = \log n/\sqrt{n}$; $q = O(\log n/n)$.

Improvements over previous work and other results

3. Showed that convex recovery is possible in cases not covered by "peeling" strategies (Ailon, Chen, Xu '13).

   Example 3: many $O(n^{1-\epsilon})$ small communities of size $O(n^{\epsilon})$ and density $O(1)$; one $n/2$-sized community with density $O(n^{-\alpha}\log n)$; background $q = O(n^{-\beta}\log n)$.

   E.g., can pick ϵ about $1/4$ to get α about $1/2$ and β almost 1.

Improvements over previous work and other results

4. Weak communities are recoverable. All probabilities are $O(1)$; convex programming by default cannot recover communities under $O(\sqrt{n})$, but we can make $p_{\min}$ very close to q, provided n_{min} is between $O(\sqrt{n})$ and $O(n)$. Matches known results for both upper and lower bound.

Methodology

- Getting the most out of convex optimization.
- Used state-of-the-art spectral bounds for random matrices (Chatterjee 2012, Tomozei, Massoulié 2010).
- Circumvented use of matrix Bernstein inequalities on block-diagonal matrices (not tight).

Future work

- Degree-corrected SBM.
- Outlier nodes (random or adversarial, Cai, Li (2014)).
- Use of adaptive algorithms (bounded number of edge queries allowed, Yun, Proutiere (2014))
- Overlapping clusters.

A Regular Binary SBM

Classical Binary SBM: Definition and Thresholds

Start with two independent $G(n,p)$ and connect them via a bipartite $G(n,n,q)$. Assume $p = p_n = \frac{a}{n}$, $q = q_n = \frac{b}{n}$. Then, whp:

- Strong recovery. (Roughly $\sqrt{a} - \sqrt{b} - 1 > 0$.) Mossel, Neeman, Sly (2014); Abbe, Bandeira, Hall (2014).
- Weak recovery. (Roughly $n(a-b)^2/(a+b) \to \infty$.) Mossel, Neeman, Sly (2014); Yun, Proutiere (2014).

Classical Binary SBM: Definition and Thresholds

Start with two independent $G(n,p)$ and connect them via a bipartite $G(n,n,q)$. Assume $p = p_n$, $q = q_n$. Then, whp:

- Partial recovery. $\frac{n(p-q)^2}{p+q} \geq C$ for C large allows for $f(C) \cdot n$ vertices to be recovered; (Coja-Oghlan (2010), Mossel, Neeman, Sly (2012)). As $C \to 2, f(C) \to 1/2$.
- Detectability/impossibility. $p = a/n,\ q = b/n$, threshold at $(a-b)^2 \geq 2(a+b)$, Mossel, Neeman, Sly (2012, 2013), Massoulié (2013). Under threshold, indistinguishable from $G(n,(p+q)/2)$.

Complexity

All thresholds are achievable via efficient (polynomial-time) algorithms.

Regular SBM

Definition

Start with two independent $G(n, d_1)$ (uniformly d_1-regular graphs) and connect them by an independently chosen bipartite $G(n, n, d_2)$ (uniformly d_2 regular bipartite graph). The resulting model is $G(n, d_1, d_2)$.

Why?

- Diametrically opposite of general heterogeneous case.
- All bounded degrees. Classical SBM allows for high-degree vertices which do not generally appear in real-world networks.
- Edge dependence (albeit weak).
- Structure is a lot more rigid. Can one obtain recovery in "lower" regimes?

$G(n, d_1, d_2)$ and $G(2n, d_1 + d_2)$

Recall the impossibility regime for classical binary SBM due to indistinguishability from $G(2n, (p+q)/2)$.
By contrast,

Lemma

(BDGHT16) Let $P_n = \mathcal{P}(n, d_1, d_2)$ and $P'_n = \mathcal{P}(2n, d_1 + d_2)$ be the uniform measures corresponding to the above two graph models. Then P_n and P'_n are orthogonal for all values of $d_1, d_2 \geq 3$.

Proof.

Count the number of graphs. Get exponentially small ratio as $n \to \infty$. □

Discussion

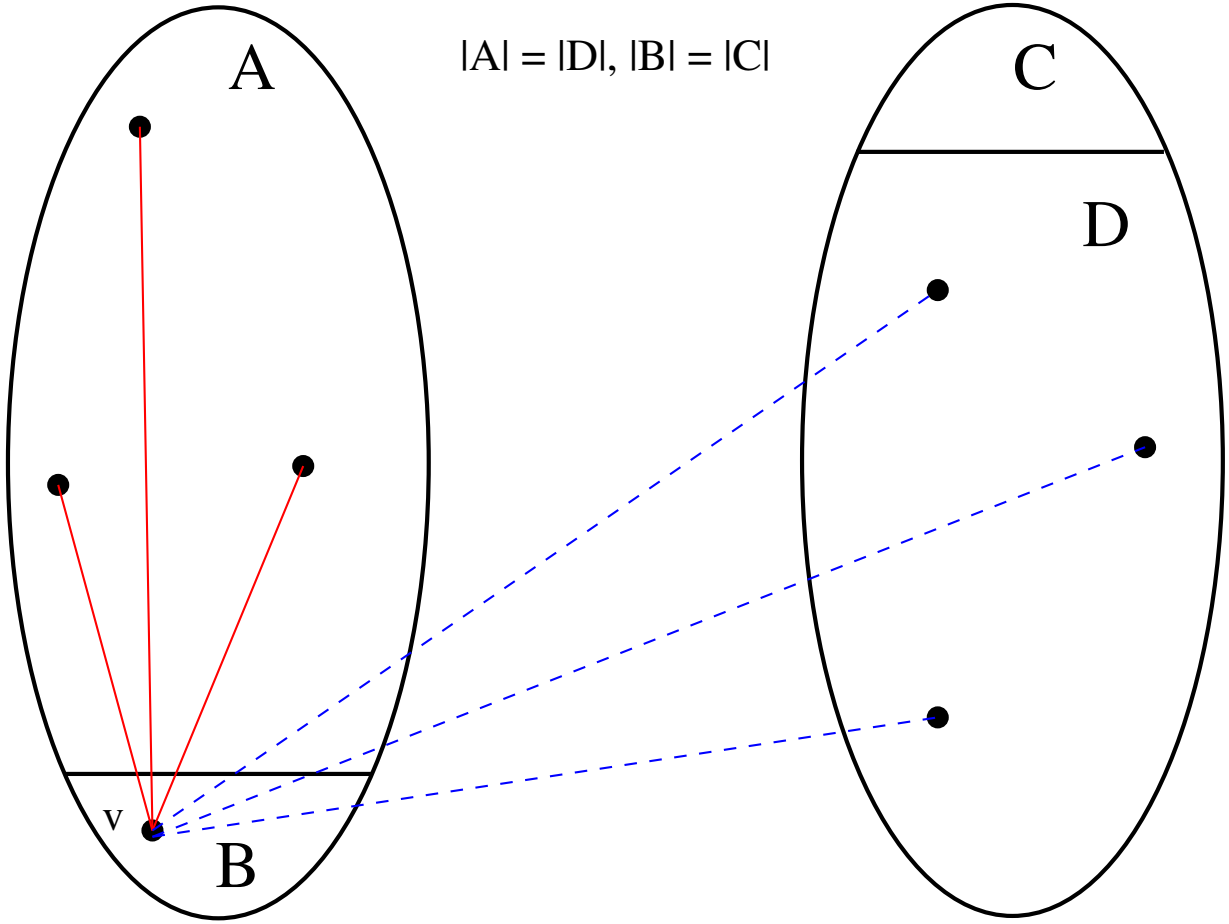
- Distinguishability has no computational value.
- The graph is *checkable*!
- If we could prove *(asymptotical) uniqueness* of partition whp, we would have (*not necessarily efficient*) strong recoverability.

Uniqueness for large d_1, d_2

Theorem

(BDGHT15) There exists a constant $C < d_1 < d_2$ such that the partition in the model is unique whp as $n \to \infty$.

A heuristic; "proof by picture"



$\deg_A(v) = \deg_D(v)$

(A, B) and (C, D)

also

(A, C) and (B, D)

Majority Rule Algorithm

- Given a partition of the vertex set into A and B, label the vertices in A by 1 and in B by -1.
- For each vertex v, replace the label σ_v of v by $sgn(\sum_{w \sim v} \sigma_w)$.
- Return the new partition into 1 and -1.

Majority Rule Lemma and Corollary

Lemma

(BDGHT16) If $d_1 > d_2 + 4$, there exists an $\epsilon = \epsilon(d_1)$ such that for any partition (A, B) well-correlated to the true partition $(\mathcal{A}, \mathcal{B})$, i.e., $|\mathcal{A} \cap B| \leq \epsilon n$, one iteration of the majority rule algorithm will construct a new partition (A_1, B_1) such that $|\mathcal{A} \cup B_1| \leq .9\epsilon n$.

Corollary

If an algorithm exists for $(1 - \epsilon)$-partial recovery (or weak recovery), then it can be boosted via majority rule to a strong recovery algorithm. ***No weak recovery threshold.***

Recovery theorem

Theorem

If $d_1 - d_2 > 2\sqrt{d_1 + d_2 - 1}$, *the partition is strongly recoverable in polynomial time.*

We adapted the methods of Massoulié (except his algorithm for $G(n, p, q)$ works to *detect* the partition, and in our case the same works to *recover* it).
The method is spectral; the condition is a limitation of the method.

Comments

- This shows that, subject to the bound, a spectral algorithm can be employed to obtain weak recovery; after which Majority Rule will obtain strong recovery.
- Note that uniqueness of the partition is a consequence! two different partitions would have to overlap almost everywhere, and as we saw this cannot happen.

Recap

We showed that in the $d_1 - d_2 > 2\sqrt{d_1 + d_2 - 1}$, strong recovery is possible in polynomial time.

We believe that recovery is always possible (not necessarily efficiently), but can only show it so far for $d_1 > d_2 + 4$ with d_1, d_2 large.

Next steps

- Can we keep pushing d_2 down? Since if $d_1 - d_2 > 2\sqrt{d_1 + d_2 - 1}$ we DO have uniqueness, there is reason to strongly suspect it's true for (almost) all values $d_1, d_2 \geq 3$ (for $d_2 = 2$ one can easily find counterexamples; roughly since the bipartite part of the graph is not connected).
- Is $d_1 - d_2 > 2\sqrt{d_1 + d_2 - 1}$ the complexity/efficiency barrier? Is there such a barrier?
- Generalize to different degree distributions.
- (with Gerandy Brito and Kameron Decker) use similar methodology to show bipartite biregular graph spectral gap; apply to more general types of "regular" graphs.
- (with Yizhe Zhu) look at the empirical spectral distributions of SBM; rate of convergence.