Benchmarker recovery theorems for the DC-SBM

Yali Wan
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Marina Meila
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Abstract

There have been many recent theoretical advances in the recovery of communities from random graphs, under the assumptions of the so-called “block models.” For the DC-SBM, we have witnessed a series of recent results consisting of sufficient conditions for recovery, often by spectral algorithms. Since the conditions are not necessary, one expects that cases exist where recovery is possible, but which are not covered by the theory. This paper explores experimentally the limits of the current theory. We generate benchmark cases, for which recovery is possible, and with well-defined parameters controlling their difficulty. Then we verify which of the existing results in the literature predict recovery. If it is hard to find examples not predicted by theory, then we may conclude that the theory is strong. This is not what our experiments show. On the contrary, they suggest that there is much more ground to cover towards deriving sharp thresholds for community recovery in the DC-SBM model. The software tool we created is made publicly available as a tool for researchers. It allows them to create test cases of controlled difficulty, and can easily be extended to test and compare new recovery theorems as they are published.

Motivation

Network modeling for the purpose of community recovery has attracted intense interest in the last decade (Holland et al., 1983; Karypis and Kumar, 1998; Jackson, 2008; Hoff et al., 2002; Goldenberg et al., 2010; Yang and Leskovec, 2015). More recently we have been witnessing rapid progress, and a surge of novel theoretical results on recovery algorithms with recovery guarantees, under some modeling assumptions.

At the center of these results are two familiar models for graphs with communities, the Stochastic Block-Model (SBM) (Holland et al., 1983) and its extension Degree-Corrected SBM (DC-SBM) of (Karrer and Newman, 2011). For the SBM, one is close to establishing thresholds for recovery in various regimes, due to the pioneering work of (Mossel et al., 2014a; Mossel et al., 2014b; Abbe and Sandon, 2015).

For the more general DC-SBM, the recovery thresholds are not known yet. The recent progress has been in obtaining recovery guarantees under weaker and weaker conditions on the model parameters. One drawback of the present results lies in how complicated these conditions are. They are not easy to parse and have implicit dependencies on each other. This makes it hard to understand the space in which the conditions are satisfied, or to compare conditions between different papers and methods. Our present work offers an empirical tool for the theoretician: a software package that generates benchmark graphs with user controlled parameters and performs the numerical verification of the various recovery conditions on these graphs for the existing results in the literature.

We proceed as follows: we generate random graphs from the DC-SBM model, for which we verify empirically (by spectral clustering) that the original clustering is recoverable with low or zero error. Then we check if the theoretical recovery conditions from the papers under consideration hold and discuss the findings. The code we use, which will be made public, is organized with modularity and extensibility in mind, so that it can be reused as new results are published.

While no empirical verification can be complete, we expect to get partial information about a few questions that are tedious or unresolved theoretically, such as which theorems cover more recoverable cases, and which conditions are the most restrictive on the test matrices? Seen this way, our work is one of benchmarking. While most benchmarks are constructed for algorithms, ours is for theorems. Benchmarking and competitions are recognized as drivers of progress in other areas of research. We expect that this benchmarking exercise will also stimulate and guide useful research in community detection.

Background: DC-SBM model, community recovery problem and sparsity regimes

The Degree-Corrected Stochastic Block Model (DC-SBM) To generate a random graph with $n$ nodes and $K$ clusters $C_1, \ldots, C_K$ from a DC-SBM, one constructs a matrix of edge probabilities $S = [S_{ij}]_{i,j=1}^n$, with $S_{ii} = 0$, $S_{ij} = S_{ji}$ in the following way. Each node $i$ is assigned a cluster label, from 1 to $K$, and a weight parameter $w_i > 0$ which indicates how likely the node is to “connect” with other nodes. The affinity between two clusters $C_k, C_l$, $k, l \in [K]$ is characterized by $B_{kl} > 0$, with $B_{kl} = B_{lk}$. Together
$w = (w_i)_{i=1}^n$, $B = [B_{ik}]_{k=1}^K$ represent the parameters of the SBM, and $C = (C_1, \ldots, C_K)$ is the clustering or the community structure. Then, $S_{ij}$ is set to

$$S_{ij} = w_i w_j B_{kl} \quad \text{whenever} \ i \in C_k \text{ and } j \in C_l.$$

To obtain a random graph from the DC-SBM model, one samples each edge $ij$, for $i, j \in [n], i \neq j$ independently, with probability $S_{ij}$.

The model defined by (1) becomes the standard Stochastic Block Model when $w_i \equiv 1$. Compared to the SBM model, DC-SBM allows for more degrees of freedom, (order $n$ instead of order $K^2$) and can represent a wider class of network models. Therefore, recovery theorems for the DC-SBM model are well worth attention.

**Definition of “community recovery”** The main research question in community recovery, and in particular regarding the DC-SBM is this: Given a simple undirected graph $G$ on $n$ nodes, with adjacency matrix $A$, sampled from an unknown DC-SBM, can we estimate the clustering $C$ and model parameters $(w, B)$? It is evident that the crux of the problem is finding $C$. Once this is known, the parameters $(w, B)$ can be estimated from $A$ and $C$ by the Maximum Likelihood method. Thus, our paper as well as most results in the literature focus on the community recovery problem. (Chen and Xu, 2014) establishes a scale of definitions of “recovery”; in particular, *strong (or exact) recovery* denotes identifying $C$ exactly; *weak recovery* denotes estimating $C$ with an error $\text{err}^2$ of order $o(n)$; *partial recovery* (or detection) denotes finding $C$ with $\text{err} < 1/2$. The most promising in real applications is the weak recovery; therefore in the rest of the paper the term “recovery” will be understood to mean “weak recovery”. This is also the scenario under which most results about the DC-SBM are obtained.

Weak recovery was shown to be possible (Coja-Oghlan and Lanka, 2009; Rohe et al., 2011; Balcan et al., 2012; Qin and Rohe, 2013; Wan and Meila, 2015) in the dense and sparse regimes, according again to the classification of (Chen and Xu, 2014). These regimes are defined based on the minimum expected degree $d_{min} = \min d_{1:n}$, where $d_i = \sum_j S_{ij}$ is the expected degree of node $i$, and correspond respectively to $d_{min} = \Omega(n)$ and $d_{min} = \Omega(\ln n)$.

The papers cited above vary in the conditions they require to guarantee recovery, due to using different combination of parameters, and sometimes different algorithms. But their requirements lie in four main categories: (1) good separation between the communities, which can be interpreted as the near block-diagonality of $S$; (2) the density of the graphs cannot be too low; (3) the degree distribution within clusters needs to be balanced; and (4), the cluster sizes are also required to be balanced.

Yet, the variations in the conditions make it hard to compare the stringency of the assumptions among different papers. Even the domains of applicability of these assumptions are inexplicit. For instance, it is not always explicit at what values of $n$ some of the asymptotic conditions start holding. And more generally, it is not known how large is the gap between recoverability and the existing conditions. Therefore, this paper sets out to probe the state of the art experimentally.

**Experiment design** Our experiment is designed as follows. First, we define a range of parameters controlling the difficulty of the benchmark problems. For each combination of parameters in this range, we

1. Generate a benchmark DC-SBM and its $S$ matrix, according to Algorithm 1.

2. Verify if the condition holds for the current model and $A$.

   a) Sample an adjacency matrix $A$ from $S$ (multiple times).

   b) For each paper and for each condition in it

   c) By construction all $S$ matrices are perfectly clusterable by standard spectral clustering (Wan and Meila, 2015; Rohe et al., 2011; Ng et al., 2002). In addition, we verify for each $A$ that the clusters can be recovered with small error.

**Generating the benchmark matrices** All DC-SBM we generate have $K = 5$ clusters, and sizes $n = 300, \ldots, 30,000$. Node weights in each cluster are sampled from the same Generalized Pareto (GPareto) distribution. The remaining input parameters are:

- **cluster relative sizes**: balanced (all equal) or unbalanced (in ratio 4 : 7 : 10 : 13 : 16)

- **the cluster level relations are parametrized by the spectrum $\Lambda = (\lambda_1 = 1 > \lambda_2 \geq \ldots \geq \lambda_K > 0)$ and the distribution $\rho = (\rho_1, \ldots, \rho_K)$**

- **weight distribution**: balanced ($\sigma_1 = 0$), perturbed ($\sigma_1 = 0.01$), unbalanced ($\sigma_1 = 0.1$) with respect to $\rho$ (see Algorithm 1)

The parameters $\Lambda$ control how separate the clusters are via the value $\lambda_K$, known as the *eigengap*. It was shown in (Meilă and Shi, 2001; Rohe et al., 2011) that for $S$ constructed as above, the Laplacian matrix $L = diag(d_{1:n})^{-1/2} S diag(d_{1:n})^{-1/2}$, which plays a central role in spectral clustering of graphs, has exactly $K$ non-zero eigenvalues given by $\Lambda$. We do the experiments with three different sets of eigenvalues $\Lambda$, having $\lambda_K = 0.01, 0.4, 0.99$ respectively; the last value corresponding to an almost exactly block diagonal matrix $S$.

The variation from balanced to unbalanced degree distribution is further controlled by the magnitude of noise added to the cluster weight volume. In Figure 1 we exemplify the $B$ and $S$ matrices we generate. The figure also shows that in this simulation, $d_i$ increases linearly with respect to $n$. 

---

1 It is standard to assume that $K$ the number of clusters is known; however, several notable recovery algorithms do not require knowing $K$.

2 The recovery error between the true $C$ and the estimated $\hat{C}$ is defined as $\text{err}^2 = 1 - \frac{1}{2} \max_{k \in [K]} |C_k - \hat{C}_k|$. We note however that several more recent results break the sparsity barrier, by proving recovery is possible when only a limited number of node degrees are below the $\ln n$ threshold (Coja-Oghlan and Lanka, 2009; Qin and Rohe, 2013).

3 In (Wan and Meila, 2015) the role of $\rho$ is explained in detail. Essentially, $\rho$ is a “default” cluster size distribution.
Checking the conditions

In this study, the theorems we compare come from (Coja-Oghlan and Lanka, 2009; Rohe et al., 2011; Balcan et al., 2012; Chaudhuri et al., 2012; Qin and Rohe, 2013; Wan and Meila, 2015). We also include two spectral clustering papers (Ng et al., 2002; Balakrishnan et al., 2011) which provide recovery guarantees and are compatible with our experimental setup.

For the main recovery theorem in each paper we check for each test case if the conditions of the theorem are satisfied. Table 1 describes the specific conditions we tested for each paper, for a total of about 20 conditions. For the spectral clustering papers, $S$ is treated like the adjacency matrix of a weighted graph, in other words as a similarity matrix, and consequently we check the recovery conditions directly on $S$, without sampling.

Some results, such as (Coja-Oghlan and Lanka, 2009; Ng et al., 2002), depend on unspecified constants; in such cases, we calculate upper or lower bounds on these constants from the data and check if the interval obtained is non-empty. The next section describes the conditions in more detail and summarizes our findings.

Results

Throughout our simulation, we find that most of the papers fail to cover even a single test case; the exceptions are (Wan and Meila, 2015) and (Balcan et al., 2012). The other papers approach the satisfaction of the conditions as the parameters are tuned towards their favor.

(Wan and Meila, 2015) Since $S$ is generated from the DC-SBM model, assumptions 1, 2 and 5 hold immediately. From Figure 2, we observe that as $n$ gets larger, assumptions 3 and 4 hold more often, which is because the node degree which grows with $n$ is faster than $\log(n)$ in the assumptions. Another interesting fact is that these two assumptions prefer harder case where $S$ is less block-diagonal, because as the off-diagonal entries of $S$ increase, the magnitudes of $S$ spread out, as a result the node degree increases above the assumption thresholds. Assumption 6 is highly associated with the balance of the degree distribution across various clusters. It can tolerate slight perturbation to the degree distribution but fails with significant unbalance. Comparing the balanced and unbalanced cluster size setting from figure 2 and 3, unbalanced case violates assumption 3 and 4 more often, while assumption 6 stays the same.

(Balcan et al., 2012) The main assumption proposed by this paper is that the clusters are self-determined communities, aka each node is more connected with the nodes within the same cluster than outside the same cluster. Because node weights are unequal, this is hard to satisfy for low weight nodes, occurring only in the cases when $\lambda_K$ is almost 1, and $S$ is almost block-diagonal.

(Coja-Oghlan and Lanka, 2009) Assumptions C0, C1, C5 automatically hold from the DC-SBM model configura-
Figure 2: Results for Wan&Meila with balanced cluster sizes. For each of the 6 conditions, T (F) indicates whether the condition is true (false).

Table 2: Changes in the range of $D$ over different $n$ ranges.

<table>
<thead>
<tr>
<th>The range of $n$</th>
<th>$\text{mean}(\min(w) / \max(w))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${1000, 3000, 10000, 30000}$</td>
<td>0.08</td>
</tr>
<tr>
<td>${300, 1000, 3000, 10000}$</td>
<td>0.07</td>
</tr>
<tr>
<td>${300, 1000, 3000, 10000, 30000}$</td>
<td>0.03</td>
</tr>
</tbody>
</table>

We use $C2$ and $C3$ to normalize the $w$ and fix $\epsilon$. We then record $\bar{w}$ the maximum value of the unknown $D$. Since this value should be independent of $n$, we examine its range over various sets of $n$. From table 2, we observe that $\bar{w}$ decreases significantly as the range of $n$ increases. This suggests that asymptotically the value of $D$, if it exists, could be very small.

(Chaudhuri et al., 2012) The paper assumes the extended planted partition model, which is more restricted than DC-SBM by reducing $K \times K$ parameters in $B$ to only 2 parameters $p = B_{kk}$ and $q = B_{kl}$, $k \neq l$, $p > q$ (we do not test for this condition). Since it assumes simpler model structure, the other assumptions for recovery should be easier to satisfy.

Assumption 1 requires the balance of cluster sizes; Assumptions 2-4 put restrictions on the degree distribution, among which Assumption 4 is the hardest to satisfy, since it requires the degree distribution having small variance in a squared degree setting normalized by the average degree. We denote as Assumption 5 the extra assumption inside Theorem 3:

$$E[d_i] \geq \frac{128}{9} \ln(6n/\delta), \text{ for } i \in V$$

where $\delta \ll 1$ is the probability of success. From figure 4, we see that Assumptions 1-3 hold and Assumptions 4-5 fail regardless of the value of $n$. We further plot a critical value coming from equation (5)

$$1 - \min(d_i) / \left[ \frac{128}{9} \ln(6n/\delta) \right] \geq \epsilon,$$

which should be negative for Assumption 5 to hold. This value is getting smaller when $n$ gets larger, which indicates that assumption 5 will hold when $n$ is sufficiently large.

(Qin and Rohe, 2013) Assumption (a) lower bounds the smallest eigenvalue. Assumption (b) lower bounds the expected node degree. From Figure 1 and Figure 5, we observe that, as $n$ increases, so does the average degree and assumption (b) starts to hold. Assumption (a) is not met regardless of the value of $n$. Assumption (a) is

$$\frac{1}{8\sqrt{3}} \sqrt{\frac{K \ln(4n/\epsilon)}{\min d_{\min} + \gamma}} \leq \lambda_K, \quad (7)$$

which puts a lower bound on $\lambda_K$ depending on the minimum $d_{\min}$. The mis-clustering rate is bounded with probability $(1 - \epsilon)$ if these assumptions hold; $\gamma$ is a constant. We set $\gamma = d_{\min}$ as suggested by the paper. Figure 5 displays the
lower bound, which stays larger than 1 in all cases. As a result assumption (a) fails since \( \lambda_K < 1 \). As \( n \) increases, lower bound in (7) decreases. We may anticipate the satisfaction of assumption (b) when \( n \) is sufficiently large. Meanwhile, the balanced cluster size setting performs better than the unbalanced setting.

**Figure 5:** Typical results for (Qin and Rohe, 2013). Left: the satisfaction of the 2 conditions versus \( n \); middle: critical value of (7) vs. \( n \) in balanced cluster size setting; right: the same in unbalanced cluster setting.

**Figure 6:** Typical results for (Rohe et al., 2011), with largest eigengap \( \lambda_K = 0.99 \) and balanced degree distribution. Left: the satisfaction of the 2 conditions versus \( n \); middle: critical value of (8) vs. \( n \) in balanced cluster size setting; right: the same in unbalanced cluster setting.

**Discussion, conclusion and further work**

In summary, because each of the eight papers we studied gives sufficient (but not necessary) guarantees for recovery, our experiments consist of generating networks for which recovery is possible and check which theory predicts it more often. We were able to generate such examples because it has been known for a long time, empirically, how to generate cases that are (likely to be) clusterable. Thus, this experiment tested the limits of the current theory. As we already mentioned, if necessary and sufficient conditions for community recovery were known, i.e. *sharp recovery thresholds*, such experiments would have been uninformative. If the current results were close to the unknown thresholds, then it would have been difficult to find clusterable examples not covered by the theory. Our results show that this is not the case.

We started with the aim of providing empirical comparisons between theoretical results in order to help readers understand their strengths and weaknesses. We were also expecting to see a gradual degradation of the agreement between theory and reality as the test cases became harder. To our surprise, it turned out that we had to create the trivially easy \( \lambda_K = 0.99 \) set of test cases in order to observe some theorems predict recoverability.

We were also expecting that the theoretical result by and large improve with time, as newer results build on previous
ones. This was partially confirmed by the most recent paper, (Wan and Meila, 2015), whose conditions are the only ones to cover a number of instances with non-trivially separated clusters.

We now turn to examining if any particular type of condition can be held responsible for the negative results. Before we start, we need to caution the reader on drawing hasty conclusions from examining Figures 2–7. As we have already mentioned, several theorems have interdependent conditions, or conditions that depend on the same unknown value. Between these, one can trade-off violating one condition for satisfying another.

For every one of the eight papers studied, the requirements for cluster separation failed to be met, always or occasionally. This suggest that in many cases they are too severe. Interestingly enough, (Wan and Meila, 2015), which fares by far the best in terms of tolerance to intercluster edges, the separation conditions involve neither the off-diagonal blocks of $S$, nor $\lambda_K$. Rather, they are based on the separation in the spectral mapping obtained by spectral clustering, and depend on the imbalance between the sums of the node degrees in each cluster, with respect to the distribution $\rho$.

Furthermore, it appears that all four types of conditions previously mentioned were violated for some of the theorems. The biggest surprise is that the requirements on graph density were also occasionally too restrictive. For instance, even though Figure 1 shows that in our examples the average degree grows linearly with $n$, and the graphs are relatively dense (average $d \approx 0.1n$), equation (7) does not hold for any $n$ up to 30,000. Extrapolating from our graphs, we see that it may start holding if $n$ increases by another 2–4 orders of magnitude.

What we have observed with these benchmarking experiments suggests that the current results are not close to the yet unknown thresholds for recovery. They suggest that our understanding of the problem is not complete, and that the existing conditions do not yet align with the actual combinations of parameters that make recovery challenging.

As any benchmarking work, this one, too, can be improved on. Our second contribution is the Matlab code we wrote. This is being made available as the package ThmBench on github/mmp2. The code is written with ease of use and modularity in mind. We invite researchers in the field to construct their own test cases, which may offer new perspectives on the limits of our current understanding in this area. We also have made it easy to add new testing modules, so that as new results are published, their conditions can be benchmarked as well.

References


