Modularity of the ABCD Random Graph Model with Community Structure

Bogumił Kamiński[1], Bartosz Pankratz[2], Paweł Prałat[2], and François Théberge[3]

[1] Warsaw School of Economics, Warsaw, Poland,
e-mail: `bkamins@sgh.waw.pl`,
[2] Ryerson University, Toronto, ON, Canada,
e-mail: `pralat,bartosz.pankratz@ryerson.ca`,
[3] The Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada,
e-mail: `theberge@ieee.org`

Abstract. The **A**rtificial **B**enchmark for **C**ommunity **D**etection graph (**ABCD**) is a random graph model with community structure and power-law distribution for both degrees and community sizes. The model generates graphs with similar properties as the well-known **LFR** one, and its main parameter ξ can be tuned to mimic its counterpart in the **LFR** model, the mixing parameter μ.
In this paper, we investigate various theoretical asymptotic properties of the **ABCD** model. In particular, we analyze the modularity function, arguably, the most important graph property of networks in the context of community detection. Indeed, the modularity function is often used to measure the presence of community structure in networks. It is also used as a quality function in many community detection algorithms, including the widely used *Louvain* algorithm.

Keywords: ABCD model, modularity function, community detection

1 Introduction

One of the most important features of real-world networks is their community structure, as it reveals the internal organization of nodes [9]. In social networks communities may represent groups by interest, in citation networks they correspond to related papers, in the Web communities are formed by pages on related topics, etc. Being able to identify communities in a network could help us to exploit this network more effectively.

Unfortunately, there are very few datasets with ground-truth identified and labelled. As a result, there is need for synthetic random graph models with community structure that resemble real-world networks in order to benchmark and tune clustering algorithms that are unsupervised by nature. The **LFR** (Lancichinetti, Fortunato, Radicchi) model [20, 18] generates networks with communities and at the same time it allows for the heterogeneity in the distributions of both node degrees and of community sizes. It became a standard and extensively used method for generating artificial networks.

In this paper, we analyze the Artificial Benchmark for Community Detection (**ABCD** graph) [14] that was recently introduced and implemented[4], including a fast implementation that uses multiple threads (**ABCDe**)[5]. Undirected variant of **LFR** and **ABCD** produce graphs with comparable properties but **ABCD**/**ABCDe** is faster than **LFR** and can be easily tuned to allow the user to make a smooth transition between the two extremes: pure (disjoint) communities and random graph with no community structure. More importantly from the perspective of this paper, it is easier to analyze theoretically.

The key ingredient for many clustering algorithms is *modularity*, which is at the same time a global criterion to define communities, a quality function of community detection algorithms, and a way to measure the presence of community structure in a network. The definition of modularity for graphs was first introduced by Newman and Girvan in [25].

Despite some known issues with this function such as the "resolution limit" reported in [10], many popular algorithms for partitioning nodes of large graphs use it [8, 24, 19] and perform very well. The list includes one of the mostly used unsupervised algorithms for detecting communities in graphs, the *Louvain* (hierarchical) algorithm [4]. For more details we direct the reader to any book on complex networks, including the following recent additions [15, 17].

1.1 Summary of Results

In this paper, we investigate the modularity function for the **ABCD** model $\mathcal{A}$. The paper is structured as follows. The **ABCD** model is introduced in Subsection 2.2 and the modularity function is defined in Subsection 2.3. Results for other random graph model in the context of the modularity function are summarized in Section 3.

We start analyzing the **ABCD** model by investigating some basic properties—see Section 4. These properties will be needed to establish results for the modularity function but they are important on their own. In particular, we show that the degree distribution is well concentrated around the corresponding expectations. Moreover, we show a concentration for the number of communities and well as the distribution of their sizes. The same generating process is applied in **LFR** so the two results hold for that model as well. The **ABCD** model assigns nodes to communities randomly. Clearly, there is no hope to predict the volumes of small communities of constant size but sufficiently large communities have their volumes as well as the number of internal edges well concentrated around the corresponding expectations.

Then we move to the results for the modularity function. By design of the **ABCD** model, $1 - \xi$ fraction of edges should become community edges and so should end up in some part of the ground truth partition $\mathbf{C}$. (ξ is the main parameter of the model responsible for the level of noise.) It is indeed the case but it turns out that a negligible fraction of the background graph join them

[4] https://github.com/bkamins/ABCDGraphGenerator.jl/

[5] https://github.com/tolcz/ABCDeGraphGenerator.jl/

there. As a result, the modularity function of the ground-truth partition $\mathbf{C}$ is asymptotic to $1 - \xi$, as proved in Theorem 1.

Analyzing the maximum modularity is much more complex. We have two types of results. The first result (Theorem 2) shows that when the level of noise is sufficiently large (ξ close to one), then the maximum modularity $q^*(\mathcal{A})$ is asymptotically larger than $q(\mathbf{C})$, the modularity of the ground-truth. In this regime, the number of edges within community graphs G_i is relatively small so a partition of the background graph into small connected pieces yields a better modularity function. To show this result, we need to investigate the degree distribution of the background graph which might be of independent interest.

The second set of results is concerned with graphs with low level of noise (ξ close to zero). For these graphs, the situation is quite opposite. It turns out that the ground truth partition is asymptotically the best possible, that is, the maximum modularity $q^*(\mathcal{A})$ is only $o(1)$ away from $q(\mathbf{C})$, the modularity of the ground truth partition $\mathbf{C}$; both of them are asymptotic to $1-\xi$ (see Theorem 3). For some technical reason, it is assumed that δ, the minimum degree of $\mathcal{A}$, is sufficiently large: the lower bound of 100 easily works but it may be improved with more detailed treatment. Having said that, it seems that one needs a different approach to uncover the real bottleneck. On the other hand, the above property is not true if $\delta = 1$ (see Theorem 4): if $\delta = 1$, then $q^*(\mathcal{A})$ is substantially larger than $q(\mathbf{C})$, regardless of how close to zero ξ is.

Finally, let us mention that all proofs, statements of various technical lemmas, and results of simulations are omitted in this proceeding version of the paper. For much more details, we direct the reader to the journal counterpart of this short paper that is available on ArXiv [11].

1.2 Simulations

This paper focuses on asymptotic theoretical results of the **ABCD** model. Having said that, we performed a number of simulations and compared asymptotic predictions with graphs generated by computer. These simulations show that the behaviour of small random instances is similar to what is predicted by the theory. This is a good news for practitioners as it shows that, despite the fact that the generative algorithm is randomized, the model has good stability. We discuss the results of simulations in the full journal version of the paper. The code with experiments is accessible on GitHub repository[6].

1.3 Open Problems

Theoretical results and simulations suggest that if δ, the minimum degree of $\mathcal{A}$, satisfies $\delta \geq \delta_0$ for some $\delta_0 \geq 2$, then there exists a constant $\xi_0 = \xi_0(\delta)$ (that possibly depends also on other parameters of the **ABCD** model $\mathcal{A}$) such that the following holds *w.h.p.* (that is, with probability tending to one as $n \to \infty$):

[6] https://github.com/bkamins/ABCDGraphGenerator.jl/

- if $0 < \xi < \xi_0$, then $q^*(\mathcal{A}) \sim q(\mathbf{C})$, where $\mathbf{C}$ is the ground truth partition of the set of nodes of $\mathcal{A}$,
- if $\xi > \xi_0$, then $q^*(\mathcal{A})$ is separated by a constant from $q(\mathbf{C})$.

Our results make the first step towards this conjecture by showing upper and lower bounds for such threshold constant ξ_0, when $\delta_0 = 100$. The bounds for ξ_0 are not close to each other. The next step would be to narrow the gap down or perhaps to determine the threshold value exactly, provided that δ_0 is sufficiently large. Another natural direction would be to decrease the lower bound for δ, that is, to decrease the value of δ_0. We showed that $\delta = 1$ does not have the desired property but maybe $\delta_0 = 2$? Or maybe one can always construct a better partition than $\mathbf{C}$ when $\delta = 2$, regardless how small parameter ξ is? These questions are left as open questions for future investigation.

2 Definitions (of ABCD Model and Modularity)

2.1 Asymptotic Notation

Our results are asymptotic in nature, that is, we will assume that the number of nodes $n \to \infty$. Formally, we consider a sequence of graphs $G_n = (V_n, E_n)$ and we are interested in events that hold *with high probability* (*w.h.p.*), that is, events that hold with probability tending to 1 as $n \to \infty$. It would be also convenient to consider events that hold *with extreme probability* (*w.e.p.*), that is, events that hold with probability at least $1 - \exp(-\Omega((\log n)^2))$. An easy but convenient property is that if a polynomial number of events hold *w.e.p.*, then *w.e.p.* all of them hold simultaneously.

2.2 ABCD Model

Table 1: Parameters of the **ABCD** model

parameter	range	description
n	$\mathbf{N}$	number of nodes
γ	$(2, 3)$	power-law exponent of degree distribution
δ	$\mathbf{N}$	minimum degree at least δ
ζ	$(0, \frac{1}{\gamma-1}]$	maximum degree at most n^ζ
β	$(1, 2)$	power-law exponent of distribution of community sizes
s	$\mathbf{N} \setminus [\delta]$	community sizes at least s
τ	$(\zeta, 1)$	community sizes at most n^τ
ξ	$(0, 1)$	level of noise

The **ABCD** model is governed by 8 parameters summarized in Table 1. For a fixed set of parameters, we generate the **ABCD** graph $\mathcal{A}$ following the steps outlined below. Each time we refer to graph $\mathcal{A}$ in this paper, we implicitly (or explicitly, but it happens rather rarely) fix all of these parameters.

Degree Distribution Let $\gamma \in (2,3)$, $\delta \in \mathbf{N}$, and $\zeta \in (0,1)$. Degrees of nodes of **ABCD** graph $\mathcal{A}$ are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\gamma, \delta, \zeta)$ with exponent γ, minimum value δ, and maximum value $D = n^{\zeta}$. In order to make sure the sum of degrees is even, if needed, we decrease by one the degree of one node of the largest degree.

It is easy to show that for any $\omega = \omega(n)$ tending to infinity as $n \to \infty$ *w.h.p.* the maximum degree of $\mathcal{A}$ is at most $n^{1/(\gamma-1)}\omega$ (of course, by definition, it is deterministically at most n^{ζ}). As a result, for any two values of $\zeta_1, \zeta_2 \in (\frac{1}{\gamma-1}, 1)$ one may couple the two corresponding ABCD graphs $\mathcal{A}$ so that *w.h.p.* they produce exactly the same graph. Hence, for convenience but without loss of generality, we will later on assume that $\zeta \in (0, \frac{1}{\gamma-1}]$.

Distribution of Community Sizes Let $\beta \in (1,2)$, $s \in \mathbf{N} \setminus [\delta]$, and $\tau \in (\zeta, 1)$. Community sizes of **ABCD** graph $\mathcal{A}$ are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\beta, s, \tau)$ with exponent β, minimum value s, and maximum value $S = n^{\tau}$. Communities are generated with this distribution as long as the sum of their sizes is less than n, the desired number of nodes. Suppose that the last community has size z and after adding it to the remaining ones, the sum of their sizes will exceed n by $k \in \mathbf{N} \cup \{0\}$. If $k = 0$, then there is nothing else to do. If $z - k \geq s$, then the size of the last community is reduced to $z - k$ so that the total number of nodes is exactly n. Otherwise, we select $z - k < s$ old communities at random, increase their sizes by one, and remove the last community so that the desired property holds.

The assumption that $\tau > \zeta$ is introduced to make sure large degree nodes have large enough communities to be assigned to. Similarly, the assumption that $s \geq \delta + 1$ is required to guarantee that small communities are not too small and so that they can accommodate small degree nodes.

Assigning Nodes into Communities At this point, the degree distribution $(w_1 \geq w_2 \geq \ldots \geq w_n)$ and the distribution of community sizes $(c_1 \geq c_2 \geq \ldots \geq c_\ell)$ are already fixed. The final **ABCD** graph $\mathcal{A}$ will be formed as the union of $\ell + 1$ independent graphs: ℓ community graphs $G_i = (C_i, E_i)$, $i \in [\ell]$, and a single background graph $G_0 = (V, E_0)$, where $V = \bigcup_{i \in [\ell]} C_i$. Roughly ξw_i edges incident to node i will, by definition, belong to its own community but a few additional edges from the background graph might end up in that community. In order to create enough room for these edges, node of degree w_i will be allowed to be assigned to a community of size c_j if the following inequality is satisfied:

$$\lceil (1 - \xi\phi) w_i \rceil \leq c_j - 1, \qquad \text{where } \phi = 1 - \sum_{k \in [\ell]} (c_k/n)^2.$$

Note that this condition is equivalent to the following one:

$$w_i \le \frac{c_j - 1}{1 - \xi\phi}. \tag{1}$$

An assignment of nodes into communities will be called *admissible* if the above inequality is satisfied for all nodes. We show that there are many admissible assignments. In particular, there are linearly many nodes of degree δ but, fortunately, *w.h.p.* communities of size more than n^{ζ} (more than the maximum degree) have space for almost all nodes. We select one admissible assignment uniformly at random. Sampling uniformly one of such assignments turns out to be relatively easy from both theoretical and practical points of view.

Distribution of Weights Parameter $\xi \in (0,1)$ reflects the amount of noise in the network. It controls the fraction of edges that are between communities. Indeed, asymptotically (but not exactly) $1-\xi$ fraction of edges are going to end up within one of the communities. Each node will have its degree w_i split into two parts: *community degree* y_i and *background degree* z_i ($w_i = y_i + z_i$). Our goal is to get $y_i \approx (1-\xi)w_i$ and $z_i \approx \xi w_i$. However, both y_i and z_i have to be non-negative integers and for each community $C \subseteq V$, $\sum_{i \in C} y_i$ has to be even. Note that since $\sum_{i\in V} w_i$ is even, so is

$$\sum_{i\in V} z_i = \sum_{i\in V}(w_i - y_i) = \sum_{i \in V} w_i - \sum_{C}\sum_{i\in C} y_i.$$

For each community $C \subseteq V$ we identify the *leader*, a node of the largest degree w_i associated with community C. (If many nodes in C have the largest degree, then we arbitrarily select one of them to be the leader.) For non-leaders we split the weights as follows:

$$y_i = \Big\lfloor (1-\xi)w_i \Big\rceil \qquad \text{and} \qquad z_i = w_i - y_i,$$

where for a given integer $a \in \mathbf{Z}$ and real number $b \in [0,1)$ the random variable $\lfloor a+b \rceil$ is defined as

$$\lfloor a+b \rceil = \begin{cases} a & \text{with probability } 1-b \\ a+1 & \text{with probability } b. \end{cases} \tag{2}$$

(Note that $\mathbf{E}[\lfloor a+b \rceil] = a(1-b)+(a+1)b = a+b$.) For the leader of community C we round $(1-\xi)w_i$ up or down so that the sum of weights in each cluster is even. If $(1-\xi)w_i \in \mathbf{N}$ and the sum of weights y_i in C is odd, then we randomly make a decision whether subtract or add one to make the sum to be even.

Creating Graphs As already mentioned, the final **ABCD** graph $\mathcal{A} = (V,E)$ will be formed as the union of $\ell+1$ independent graphs: ℓ community graphs $G_i = (C_i, E_i)$, $i \in [\ell]$, and a single background graph $G_0 = (V, E_0)$, where $V =$

$\bigcup_{i\in[\ell]} C_i$, that is, $E = \bigcup_{i\in[\ell]\cup\{0\}} E_i$. Each of these $\ell+1$ graphs will be created independently. The partition $\mathbf{C} = \{C_1, C_2, \ldots, C_\ell\}$ will be called a *ground-truth* partition.

Suppose then that our goal is to create a graph on n nodes with a given degree distribution $\mathbf{w} := (w_1, w_2, \ldots, w_n)$, where $\mathbf{w}$ is any vector of non-negative integers such that $w := \sum_{i\in[n]} w_i$ is even. We define a random multi-graph $\mathcal{P}(\mathbf{w})$ with a given degree sequence known as the **configuration model** (sometimes called the **pairing model**), which was first introduced by Bollobás [5]. (See [3, 27, 28] for related models and results.) We start with w *points* that are partitioned into n *buckets* labelled with labels $v_1, v_2, \ldots, v_n$; bucket v_i consists of w_i points. It is easy to see that there are $\frac{w!}{(w/2)!2^w}$ pairings of points. We select one of such pairings uniformly at random, and construct a multi-graph $\mathcal{P}(\mathbf{w})$, with loops and parallel edges allowed, as follows: nodes are the buckets $v_1, v_2, \ldots, v_n$, and a pair of points xy corresponds to an edge v_iv_j in $\mathcal{P}(\mathbf{w})$ if x and y are contained in the buckets v_i and v_j, respectively.

2.3 Modularity Function

The modularity function favours partitions of the set of nodes of a graph G in which a large proportion of the edges fall entirely within the parts but benchmarks it against the expected number of edges one would see in those parts in the corresponding **Chung-Lu** random graph model [7] which generates graphs with the expected degree sequence following exactly the degree sequence in G.

Formally, for a graph $G = (V, E)$ and a given partition $\mathbf{A} = \{A_1, A_2, \ldots, A_\ell\}$ of V, the *modularity function* is defined as follows:

$$q(\mathbf{A}) = \sum_{A_i\in\mathbf{A}} \frac{e(A_i)}{|E|} - \sum_{A_i\in\mathbf{A}} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)}\right)^2, \tag{3}$$

where for any $A \subseteq V$, $e(A) = |\{uv \in E : u, v \in A\}|$ is the number of edges in the subgraph of G *induced by* set A, and $\mathrm{vol}(A) = \sum_{v\in A} \deg(v)$ is the *volume* of set A. In particular, $\mathrm{vol}(V) = 2|E|$. The first term in (3), $\sum_{A_i\in\mathbf{A}} e(A_i)/|E|$, is called the *edge contribution* and it computes the fraction of edges that fall within one of the parts. The second one, $\sum_{A_i\in\mathbf{A}}(\mathrm{vol}(A_i)/\mathrm{vol}(V))^2$, is called the *degree tax* and it computes the expected fraction of edges that do the same in the corresponding random graph (the null model). The modularity measures the deviation between the two.

The maximum *modularity* $q^*(G)$ is defined as the maximum of $q(\mathbf{A})$ over all possible partitions $\mathbf{A}$ of V; that is, $q^*(G) = \max_{\mathbf{A}} q(\mathbf{A})$. In order to maximize $q(\mathbf{A})$ one wants to find a partition with large edge contribution subject to small degree tax. If $q^*(G)$ approaches 1 (which is the trivial upper bound), we observe a strong community structure; conversely, if $q^*(G)$ is close to zero (which is the trivial lower bound), there is no community structure. The definition in (3) can be generalized to weighted edges by replacing edge counts with sums of edge weights. It can also be generalized to hypergraphs [12, 13].

3 Related Results for Random Graphs

Analyzing the maximum modularity $q^*(G)$ for sparse random graphs is a challenging task. The most attention was paid to **random d-regular graphs** $\mathcal{G}_{n,d}$ but even for this family of graphs we only know upper and lower bounds for $q^*(\mathcal{G}_{n,d})$ that are quite apart from each other. For example, for random 3-regular graph $\mathcal{G}_{n,3}$ we only know that *w.h.p.*

$$0.667026 \le q^*(\mathcal{G}_{n,3}) \le 0.789998.$$

These bounds were recently proved in [21] but the main goal of that paper was to confirm the conjecture from [22] that *w.h.p.* $q^*(\mathcal{G}_{n,3}) \ge 2/3+\varepsilon$ for some $\varepsilon > 0$. We refer the reader to [22, 26] for numerical bounds on $q^*(\mathcal{G}_{n,d})$ for other values of $d \ge 3$ and for some explicit but weaker bounds. It is also known that *w.h.p.* $q^*(\mathcal{G}_{n,2}) \sim 1$ [22].

The **binomial random graphs** $\mathcal{G}(n,p)$ were studied in [23] where it was shown that *w.h.p.* $q^*(\mathcal{G}(n,p)) \sim 1$, provided that $pn \le 1$, On the other hand, *w.h.p.* $q^*(\mathcal{G}(n,p)) = \Theta(1/\sqrt{pn})$, provided that $pn \ge 1$ and $p < 1-\varepsilon$ for some $\varepsilon > 0$. The modularity of the well-known **Preferential Attachment (PA) model** [2] and the **Spatial Preferential Attachment (SPA) model** [1] was studied in [26]. Finally, the modularity of a model of random geometric graphs on the hyperbolic plane [16], known as the **KPKBV model** after its inventors, was recently studied in [6].

4 Some Properties of ABCD

4.1 Degree Distribution

Let $\gamma \in (2,3)$, $\delta \in \mathbf{N}$, and $\zeta \in (0,1)$. Recall that the degrees of nodes of the **ABCD** model are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\gamma,\delta,\zeta)$ with exponent γ, minimum value δ, and maximum value $D = n^\zeta$. More precisely, if $X \in \mathcal{P}(\gamma,\delta,\zeta)$, then for any $k \in \{\delta, \delta+1, \ldots, D\}$,

$$\begin{aligned} q_k = \Pr(X=k) &= \frac{\int_k^{k+1} x^{-\gamma}dx}{\int_\delta^{D+1} x^{-\gamma}dx} = \frac{k^{1-\gamma}-(k+1)^{1-\gamma}}{\delta^{1-\gamma}-(D+1)^{1-\gamma}} \\ &= \left(1+\mathcal{O}(n^{-\zeta(\gamma-1)})+\mathcal{O}(k^{-1})\right)\, k^{-\gamma}(\gamma-1)\delta^{\gamma-1}. \end{aligned} \tag{4}$$

The first lemma provides an upper bound for the maximum degree, which justifies our assumption that $\zeta \in (0, 1/(\gamma-1)]$. The second lemma shows that the degree distribution is well concentrated around the expectation. Since the statements are quite technical, we omit them. However, let us mention the following corollary. The volume of all nodes in $\mathcal{A}$ is *w.e.p.* equal to

$$\mathrm{vol}(V) = \sum_{k=\delta}^{D} kY_k = (1+\mathcal{O}((\log n)^{-1}))\; dn, \qquad \text{where } d := \sum_{k=\delta}^{D} kq_k.$$

4.2 Distribution of Community Sizes

Let $\beta \in (1,2)$, $s \in \mathbf{N}$, and $\tau \in (\zeta, 1)$. Recall that community sizes of the **ABCD** model are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\beta, s, \tau)$ with exponent β, minimum value s, and maximum value $S = n^{\tau}$. More precisely, if $X \in \mathcal{P}(\beta, s, \tau)$, then after following exactly the same computation as in (4) we get that for any $k \in \{s, s+1, \ldots, S\}$,

$$\begin{aligned} p_k = \Pr(X = k) &= \frac{\int_k^{k+1} x^{-\beta} dx}{\int_s^{S+1} x^{-\beta} dx} = \frac{k^{1-\beta} - (k+1)^{1-\beta}}{s^{1-\beta} - (S+1)^{1-\beta}} \\ &= \left(1 + \mathcal{O}(n^{-\tau(\beta-1)}) + \mathcal{O}(k^{-1})\right) k^{-\beta} (\beta - 1) s^{\beta - 1}. \end{aligned} \tag{5}$$

Our next lemma shows that community sizes of **ABCD** are well concentrated around their expectation. Again, we omit technical statements only reporting that *w.e.p.* the number of communities is equal to

$$\ell = \ell(n) = (1 + \mathcal{O}((\log n)^{-1}))\, \hat{c}\, n^{1-\tau(2-\beta)},$$

where

$$\hat{c} = \frac{2 - \beta}{(\beta - 1) s^{\beta - 1}}.$$

4.3 Assigning Nodes into Communities and Distribution of Weights

Recall that at this point of the process, the degree distribution $(w_1 \geq w_2 \geq \ldots \geq w_n)$ and the distribution of community sizes $(c_1 \geq c_2 \geq \ldots \geq c_\ell)$ are already fixed. In order to assign nodes to communities we will use the following easy and natural algorithm. We consider nodes, one by one, starting from w_1 (high degree node) and finishing with w_n (low degree node). Recall that node i of degree w_i has to be assigned to a community of size c_j so that inequality (1) holds. We assign node w_i randomly to one of the communities that have size larger than $\lceil (1 - \xi\phi) w_i \rceil$ and still have some "available spots". We do it with probability proportional to the number of available spots left. One can show that the above simple algorithm generates one of the admissible assignments uniformly at random.

The volumes of small communities are not well concentrated around their means. On the other hand, the volumes of very large communities are well concentrated around their means, as our next lemma shows. As usually, we skip the statement directing the reader to the journal version of this paper.

5 Modularity

5.1 Modularity of the Ground-truth Partition: $q(\mathbf{C})$

Let us start by investigating the modularity of the ground-truth partition of $\mathcal{A}$.

Theorem 1. *Let* $\mathbf{C} = \{C_1, C_2, \ldots, C_\ell\}$ *be the ground-truth partition of the set of nodes of* $\mathcal{A}$. *Then,* w.e.p.

$$q^*(\mathcal{A}) \geq q(\mathbf{C}) = (1 + \mathcal{O}((\log n)^{-(\gamma-2)}))\,(1 - \xi).$$

5.2 Maximum Modularity: $q^*(G)$

As mentioned in Section 3, analyzing the maximum modularity $q^*(G)$ for sparse random graphs is a challenging task and typically only bounds for $q^*(G)$ are known that are far apart from each other. Since the **ABCD** model $\mathcal{A}$ is more complex than other sparse random graphs, especially random d-regular graphs, there is no hope for tight bounds for the maximum modularity function but we will make some interesting observations below.

Large Level of Noise Let us start with investigating graphs with a large level of noise, that is, with ξ close to one. For such graphs, one should focus on the background graph G_0 which involves all but a small fraction of edges. It turns out that G_0 is connected *w.h.p.*, provided that its minimum degree is at least 3, or otherwise *w.h.p.* it has a giant component. By restricting ourselves to a spanning tree of the giant component of G_0, we may partition the set of nodes into small parts such that each part induces a connected graph. This is not much, but for noisy graphs it yields the modularity that is larger than the modularity of the ground-truth partition.

Theorem 2. *Let $\gamma \in (2,3)$, $\delta \in \mathbf{N}$, $\zeta \in (0, \frac{1}{\gamma-1}]$, and $\xi \in (0,1)$.*

(a) If $\xi\delta \geq 3$, then set $\alpha = 1$.
(b) If $\xi\delta < 3$, then there exists a universal constant $\alpha > 0$ which depends on the parameters of the model but it is always separated from 0 (that is, α is not *a function of n).*

There exists a partition $\mathbf{C}$ of the set of nodes V of $\mathcal{A}$ such that the following properties hold w.h.p.

$$q^*(\mathcal{A}) \geq q(\mathbf{C}) \geq (1 + \mathcal{O}(n^{-(1-\zeta)/2}))\frac{2\alpha n}{vol(V)}$$
$$= (1 + \mathcal{O}((\log n)^{-1}))\ \frac{2\alpha}{d}, \qquad \text{where } d = \sum_{k=\delta}^{D} kq_k.$$

(Note that q_i is defined in (4).)

Recall that the modularity function of the ground-truth partition is *w.e.p.* asymptotic to $1-\xi$. The above theorem implies that if $\delta \geq 4$ and the graph has a large level of noise, namely, $\xi \geq 3/\delta$ and $\xi > 1 - 2/d$, then *w.h.p.* the modularity function obtained from dissecting the spanning tree of G_0 is larger! The same conclusion can be derived when $\delta \leq 3$ by considering ξ sufficiently close to one.

Low Level of Noise This time we will investigate graphs with a low level of noise, that is, with ξ close to zero. Let us fix a value of $\delta \in \mathbf{N}$ such that $\delta \geq 100$. For any $a \in \mathbf{N}$ and $b \in \mathbf{N} \setminus \{1,2\}$ such that $ab < \delta$, let

$$c(a,b) := \frac{b - 2\sqrt{b-1}}{2b}\ \frac{ab}{ab+b-1} - \frac{b-1}{ab+b-1} - 0.011. \qquad (6)$$

Let

$$\xi_0(\delta) := \max_{a\in\mathbf{N}, b\in\mathbf{N}\setminus\{1,2\}, ab<\delta} \min\left(1-\frac{ab}{\delta}, \frac{c(a,b)}{4}, \frac{1}{20}\right). \tag{7}$$

It is clear that $\xi_0(\delta)$ is a non-decreasing function of δ. Moreover, $\xi_0(100) \approx 0.0217$ (the maximum is achieved for $a = 8$ and $b = 12$), and $\xi_0(\delta) = 1/20$ for $\delta \geq 340$.

Our first result says that **ABCD** graph $\mathcal{A}$ with minimum degree $\delta \geq 100$ and $\xi \in (0, \xi_0(\delta))$ has *w.h.p.* the maximum modularity $q^*(\mathcal{A})$ asymptotically equal to the modularity function on the ground-truth.

Theorem 3. *Let $\delta \in \mathbf{N}$ such that $\delta \geq 100$ and $0 < \xi < \xi_0(\delta)$, where $\xi_0(\delta)$ is defined in (7). Let $\mathbf{C} = \{C_1, C_2, \ldots, C_\ell\}$ be the ground-truth partition of the set of nodes of $\mathcal{A}$. Then,* w.h.p. $q^*(\mathcal{A}) \sim q(\mathbf{C}) \sim 1 - \xi$.

The lower bound of 100 for δ as well as the constants $\xi_0(\delta)$ are not tuned for the strongest result. Since the proof technique we use will not allow us to close the gap anyway, we aimed for a simple argument that works for large enough δ and relatively simple constants. Having said that, the above property is not true for $\delta = 1$; that is, if $\mathcal{A}$ has minimum degree $\delta = 1$, then one may find a partition of the nodes of $\mathcal{A}$ that yields larger modularity than the one associated with the ground-truth.

Theorem 4. *Fix $\delta = 1$ and let $0 < \xi < 1$. Let $\mathbf{C} = \{C_1, C_2, \ldots, C_\ell\}$ be the ground-truth partition of the set of nodes of $\mathcal{A}$. Then,* w.e.p.

$$\begin{aligned} q^*(\mathcal{A}) &\geq (1 + \mathcal{O}((\log n)^{-(\gamma-2)})) \left((1-\xi) + \frac{\xi q_1}{d}\left(2 - \frac{q_1}{d}\right)\right) \\ &> (1 + \mathcal{O}((\log n)^{-(\gamma-2)}))\,(1-\xi) = q(\mathbf{C}), \end{aligned}$$

where q_k is defined in (4) and $d = \sum_{k=\delta}^{D} kq_k$.

References

1. Aiello, W., Bonato, A., Cooper, C., Janssen, J., Prałat, P.: A spatial web graph model with local influence regions. Internet Mathematics 5(1-2), 175–196 (2008)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
3. Bender, E.A., Canfield, E.R.: The asymptotic number of labeled graphs with given degree sequences. Journal of Combinatorial Theory, Series A 24(3), 296–307 (1978)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)
5. Bollobás, B.: A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. European Journal of Combinatorics 1(4), 311–316 (1980)
6. Chellig, J., Fountoulakis, N., Skerman, F.: The modularity of random graphs on the hyperbolic plane. Journal of Complex Networks 10(1), cnab051 (2022)
7. Chung Graham, F., Lu, L.: Complex graphs and networks. No. 107, American Mathematical Soc. (2006)

8. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E 70(6), 066111 (2004)
9. Fortunato, S.: Community detection in graphs. Physics reports 486(3-5), 75–174 (2010)
10. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proceedings of the national academy of sciences 104(1), 36–41 (2007)
11. Kamiński, B., Pankratz, B., Prałat, P., Théberge, F.: Modularity of the abcd random graph model with community structure. arXiv:2203.01480 (2022)
12. Kamiński, B., Poulin, V., Prałat, P., Szufel, P., Théberge, F.: Clustering via hypergraph modularity. PloS one 14(11), e0224307 (2019)
13. Kamiński, B., Prałat, P., Théberge, F.: Community detection algorithm using hypergraph modularity. In: International Conference on Complex Networks and Their Applications. pp. 152–163. Springer (2020)
14. Kamiński, B., Prałat, P., Théberge, F.: Artificial benchmark for community detection (abcd)—fast random graph model with community structure. Network Science pp. 1–26 (2021)
15. Kamiński, B., Prałat, P., Théberge, F.: Mining complex networks (2021)
16. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguná, M.: Hyperbolic geometry of complex networks. Physical Review E 82(3), 036106 (2010)
17. Lambiotte, R., Schaub, M.: Modularity and dynamics on complex networks (2021)
18. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E 80(1), 016118 (2009)
19. Lancichinetti, A., Fortunato, S.: Limits of modularity maximization in community detection. Physical review E 84(6), 066122 (2011)
20. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical review E 78(4), 046110 (2008)
21. Lichev, L., Mitsche, D.: On the modularity of 3-regular random graphs and random graphs with given degree sequences. arXiv preprint arXiv:2007.15574 (2020)
22. McDiarmid, C., Skerman, F.: Modularity of regular and treelike graphs. Journal of Complex Networks 6(4), 596–619 (2018)
23. McDiarmid, C., Skerman, F.: Modularity of erdős-rényi random graphs. Random Structures & Algorithms 57(1), 211–243 (2020)
24. Newman, M.E.: Fast algorithm for detecting community structure in networks. Physical review E 69(6), 066133 (2004)
25. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E 69(2), 026113 (2004)
26. Prokhorenkova, L.O., Prałat, P., Raigorodskii, A.: Modularity of complex networks models. Internet Mathematics (2017)
27. Wormald, N.C.: Generating random regular graphs. Journal of algorithms 5(2), 247–280 (1984)
28. Wormald, N.C., et al.: Models of random regular graphs. London Mathematical Society Lecture Note Series pp. 239–298 (1999)