Self-similarity of Communities of the ABCD Model

Jordan Barrett[*], Bogumił Kamiński[†], Paweł Prałat[‡], François Théberge[§]

September 12, 2024

Abstract

The **A**rtificial **B**enchmark for **C**ommunity **D**etection (**ABCD**) graph is a random graph model with community structure and power-law distribution for both degrees and community sizes. The model generates graphs similar to the well-known **LFR** model but it is faster and can be investigated analytically. In this paper, we show that the **ABCD** model exhibits some interesting self-similar behaviour, namely, the degree distribution of ground-truth communities is asymptotically the same as the degree distribution of the whole graph (appropriately normalized based on their sizes). As a result, we can not only estimate the number of edges induced by each community but also the number of self-loops and multi-edges generated during the process. Understanding these quantities is important as (a) rewiring self-loops and multi-edges to keep the graph simple is an expensive part of the algorithm, and (b) every rewiring causes the underlying configuration models to deviate slightly from uniform simple graphs on their corresponding degree sequences.

***Keywords*—** Random graphs, Complex networks, Configuration model, ABCD, Community structure, Self-similarity, Power-law

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, 17]. In social networks communities may represent groups by interest, in citation networks they correspond to related papers, in the Web graph communities are formed by pages on related topics, etc. Identifying communities in a network is therefore valuable as this information helps us to better understand the network structure.

[*]Department of Mathematics, Toronto Metropolitan University, Toronto, ON, Canada; e-mail: jordan.barrett@torontomu.ca

[†]Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland; email: bkamins@sgh.waw.pl

[‡]Department of Mathematics, Toronto Metropolitan University, Toronto, ON, Canada; e-mail: pralat@torontomu.ca

[§]Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada; email: theberge@ieee.org

Unfortunately, there are very few datasets with ground-truth communities identified and labelled. As a result, there is need for synthetic random graph models with community structure that resemble real-world networks to benchmark and tune clustering algorithms that are unsupervised by nature. The **LFR** (**L**ancichinetti, **F**ortunato, **R**adicchi) model [23, 22] is a highly popular model that generates networks with communities and, at the same time, allows for heterogeneity in the distributions of both node degrees and of community sizes. It became a standard and extensively used method for generating artificial networks.

A similar synthetic network to **LFR**, the **A**rtificial **B**enchmark for **C**ommunity **D**etection (**ABCD**) [16] was recently introduced and implemented[1], including a fast implementation[2] that uses multiple threads (**ABCDe**) [20]. Undirected variants 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 graphs with no community structure. Moreover, it is easier to analyze theoretically—for example, in [15] various theoretical asymptotic properties of the **ABCD** model are investigated including the modularity function that, despite some known issues such as the "resolution limit" reported in [10], is an important graph property of networks in the context of community detection. Finally, the building blocks in the model are flexible and may be adjusted to satisfy different needs. Indeed, the original **ABCD** model was recently adjusted to include potential outliers (**ABCD+o**) [18] and extended to hypergraphs (**h–ABCD**) [19][3]. In the context of this paper, the most important of the above properties is that the **ABCD** model allows for theoretical investigation of its properties.

The **ABCD** model is used by practitioners but, for the reasons mentioned above, it also gains recognition among scientists. For example, [1] suggests to use Adjusted Mutual Information (AMI) between the partitions returned by various algorithms with the ground-truth partitions of synthetically generated random graphs, **ABCD** and **LFR**. In particular, they use both models to compare 30 community detection algorithms, mentioning that *being directly comparable to **LFR**, **ABCD** offers additional benefits including higher scalability and better control for adjusting an analogous mixing parameter*.

Another important aspect of complex networks is self-similarity and scale invariance which are well-known properties of certain geometric objects such as fractals [24]. Scale invariance in the context of complex networks is traditionally restricted to the scale-free property of the distribution of node degrees [2] but also applies to the distributions of community sizes [12, 8], degree-degree distances [32], and network density [5]. Unfortunately, the definition of "scale free" has never reached a single agreement [7, 13] but many experiments provide a statistical significance of these claims such as the experiment on 32 real-world networks that have a wide coverage of economic, biological, informational, social, and technological domains, with their sizes ranging from hundreds to tens of millions of nodes [32].

In search for more complete self-similar descriptions, methods related to the fractal dimension are considered that use box counting methods and renormalization [27, 11, 21]. However, the main issue is that complex networks are still not well defined in a proper geometric sense but one may, for example, introduce the concept of hidden metric spaces to overcome this problem [26].

[1] `https://github.com/bkamins/ABCDGraphGenerator.jl/`

[2] `https://github.com/tolcz/ABCDeGraphGenerator.jl/`

[3] `https://github.com/bkamins/ABCDHypergraphGenerator.jl`

For the context of community structure of complex networks, let us highlight one interesting study of the network of e-mails within a real organization that revealed the emergence of self-similar properties of communities [12]. Such experiments suggest that there is some universal mechanism that controls the formation and dynamics of complex networks.

In this paper, we show that the **ABCD** model exhibits self-similar behaviour: each ground-truth community inherits power-law degree distribution from the distribution of the entire graph (see Theorem 3.1), that is, the power-law exponent as well as the minimum degree of this distribution are preserved. On the other hand, as in all self-similarities mentioned above, some renormalization needs to be applied. In our case, the distribution is truncated so that the maximum degree, corrected by the noise parameter ξ (see Section 2 for its formal definition), does not exceed the community size.

The above observation, interesting and desired on its own, has some immediate implications that are of interest too. Firstly, we can easily compute the expected volume of each community (see Corollary 3.2). Secondly, and more importantly, we can investigate how many self-loops and multi-edges are constructed during the generation process of **ABCD** (see Theorem 3.3). Understanding this quantity is crucial for two reasons. Firstly, removing these self-loops and multi-edges to obtain a simple graph is a time consuming part of the construction algorithm. Secondly, as the **ABCD** construction involves several implementations of the well-known configuration model, the number of self-loops and multi-edges is directly correlated to how "skewed" the final graph is, i.e., more self-loops and multi-edges lead to distributions that are further away from being uniform. We speak about this second reason in more detail in Section 2.4.

The paper is structured as follows. In Section 2, we formally define the **ABCD** model and state one known result about the said model. The main results are presented in Section 3. Then, in Section 4, we present results of simulations that highlight properties that are proved in this paper and show their practical implications. Next, the main result (Theorem 3.1) and its applications (Corollary 3.2 and Theorem 3.3) are proved in Section 5. Finally, some open problems are presented in Section 6.

A preliminary version of this paper will be published in the proceedings of WAW 2024 [3].

2 The ABCD Model

In this section we introduce the **ABCD** model. Its full definition, along with more detailed explanations of its parameters and features, can be found in [16]. We restate the main components of the **ABCD** model here to ensure completeness of the exposition in this article. More accurately, we outline a version of the **ABCD** model that was studied extensively in [15]. In the coming description, all choices made (the truncated power-law, the parameters, etc.) match those in [15]. In fact, there is much flexibility in the **ABCD** model, and we suspect that our results carry over to this more flexible setting. However, we choose to study the version of the **ABCD** model presented in [15] so that (a) we can use previously established results, and (b) we can simplify the statements of our main results.

2.1 Notation

For a given $n \in \mathbb{N} := \{1, 2, \ldots\}$, we use $[n]$ to denote the set consisting of the first n natural numbers, that is, $[n] := \{1, 2, \ldots, n\}$.

Our results are asymptotic by nature, that is, we will assume that $n \to \infty$. For a sequence of events $(E_n, n \in \mathbb{N})$, we say E_n holds *with high probability* (*w.h.p.*) if $\mathbb{P}(E_n) \to 1$ as $n \to \infty$. We say that E_n holds *with extreme probability* (*w.e.p.*) if $\mathbb{P}(E_n) = 1 - \exp(-\Omega(\log^2 n))$. In particular, if there are polynomially many events and each holds w.e.p., then w.e.p. all of them hold simultaneously. To combine this notion with other asymptotic standard notation such as $O(\cdot)$ and $o(\cdot)$, we follow the conventions in [31].

Power-law distributions will be used to generate both the degree sequence and community sizes so let us formally define it. For given parameters $\gamma \in (0, \infty)$, $\delta, \Delta \in \mathbb{N}$ with $\delta \leq \Delta$, we define a truncated power-law distribution $\mathcal{P}(\gamma, \delta, \Delta)$ as follows. For $X \sim \mathcal{P}(\gamma, \delta, \Delta)$ and for $k \in \mathbb{N}$ with $\delta \leq k \leq \Delta$,

$$\mathbb{P}(X = k) = \frac{\int_k^{k+1} x^{-\gamma}\, dx}{\int_\delta^{\Delta+1} x^{-\gamma}\, dx}.$$

2.2 The Configuration Model

The well-known configuration model is an important ingredient of the **ABCD** generation process so let us formally define it here. Suppose then that our goal is to create a graph on n nodes with a given degree distribution $\mathbf{d} := (d_i, i \in [n])$, where $\mathbf{d}$ is a sequence of non-negative integers such that $m := \sum_{i \in [n]} d_i$ is even. We define a random multi-graph $\mathrm{CM}(\mathbf{d})$ with a given degree sequence known as the **configuration model** (sometimes called the **pairing model**), which was first introduced by Bollobás [6]. (See [4, 29, 30] for related models and results.)

We start by labelling nodes as $[n]$ and, for each $i \in [n]$, endowing node i with d_i half-edges. We then iteratively choose two unpaired half-edges uniformly at random (from the set of pairs of remaining half-edges) and pair them together to form an edge. We iterate until all half-edges have been paired. This process yields $G_n \sim \mathrm{CM}(\mathbf{d})$, where G_n is allowed self-loops and multi-edges and thus G_n is a multi-graph.

2.3 Parameters of the ABCD Model

The **ABCD** model is governed by the following eight parameters.

Parameter	Range	Description
n	$\mathbb{N}$	Number of nodes
γ	$(2, 3)$	Power-law degree distribution with exponent γ
δ	$\mathbb{N}$	Min degree as least δ
ζ	$\left(0, \frac{1}{\gamma-1}\right]$	Max degree at most n^ζ
β	$(1, 2)$	Power-law community size distribution with exponent β
s	$\mathbb{N} \setminus [\delta]$	Min community size at least s
τ	$(\zeta, 1)$	Max community size at most n^τ
ξ	$(0, 1)$	Level of noise

2.4 The ABCD Construction

We will use $\mathcal{A} = \mathcal{A}(n, \gamma, \delta, \zeta, \beta, s, \tau, \xi)$ for the distribution of graphs generated by the following 5-phase construction process.

Phase 1: creating the degree distribution

In theory, the degree distribution for an **ABCD** graph can be any distribution that satisfies (a) a power-law with parameter γ, (b) a minimum value of at least δ, and (c) a maximum value of at most n^ζ. In practice, however, degrees are i.i.d. samples from the distribution $\mathcal{P}\left(\gamma, \delta, n^\zeta\right)$.

For $G_n \sim \mathcal{A}$, write $\mathbf{d}_n = (d_i, i \in [n])$ for the chosen degree sequence of G_n with $d_1 \geq \cdots \geq d_n$. Finally, to ensure that $\sum_{i\in[n]} d_i$ is even, we decrease d_1 by 1 if necessary; we relabel as needed to ensure that $d_1 \geq d_2 \geq \cdots \geq d_n$. This potential change has a negligible effect on the properties we investigate in this paper and we thus only present computations for the case when d_1 is unaltered.

Phase 2: creating the communities

We next assign communities to the **ABCD** model. When we construct a community, we assign a number of vertices to said community equal to its size. Initially, the communities will form an empty graph. Then, in Phases 3, 4 and 5, we handle the construction of edges using the degree sequence established in Phase 1.

Similar to the degree distribution, the distribution of community sizes must satisfy (a) a power-law with parameter β, (b) a minimum value of s, and (c) a maximum value of n^τ. In addition, we also require that the sum of community sizes is exactly n. Again, we use a more rigid distribution in practice: communities are generated with sizes determined independently by the distribution $\mathcal{P}(\beta, s, n^\tau)$. We generate communities until their collective size is at least n. If the sum of community sizes at this moment is $n + k$ with $k > 0$ then we perform one of two actions: if the last added community has size at least $k + s$, then we reduce its size by k. Otherwise (that is, if its size is $c < k + s$), then we delete this community, select c old communities and increase their sizes by 1. This again has a negligible effect on the analysis and we thus only present computations for the case when community sizes are unaltered.

For $G_n \sim \mathcal{A}$, write L for the (random) number of communities in G_n and write $\mathbf{C}_n = (C_j, j \in [L])$ for the chosen collection of communities in G_n with $|C_1| \geq \cdots \geq |C_L|$ (again, let us stress the fact that $\mathbf{C}_n$ is a random vector of random length L).

Phase 3: assigning degrees to nodes

At this point in the construction of $G_n \sim \mathcal{A}$ we have a degree sequence $\mathbf{d}_n$ and a collection of communities $\mathbf{C}_n$ with community C_j containing $|C_j|$ *unassigned* nodes, i.e., nodes that have not been assigned a label or a degree. We then iteratively assign labels and degrees to nodes as follows. Starting with $i = 1$, let U_i be the collection of unassigned nodes at step i. At step i choose a node uniformly at random from the set of nodes u in U_i that satisfy

$$d_i \leq \frac{|C(u)| - 1}{1 - \xi\phi},$$

where $C(u)$ is the community containing u and

$$\phi = 1 - \frac{1}{n^2} \sum_{j \in [L]} |C_j|^2 ,$$

and assign this node label i and degree d_i; we have that $U_{i+1} = U_i \setminus \{u\}$. We bound the degrees assignable to node u in community C to ensure that there are enough nodes in $C \setminus \{u\}$ for u to pair with, preventing guaranteed self-loops or guaranteed multi-edges during phase 4 of the construction. The details of this bound are quite involved and are not overly important for our results. Thus, we point the reader to either [15] or [16] for a full explanation of the bound.

Phase 4: creating edges

At this point G_n contains n nodes labelled as $[n]$, partitioned by the communities $\mathbf{C}_n$, with node $i \in [n]$ containing d_i unpaired half-edges. The last step is to form the edges in G_n. Firstly, for each $i \in [n]$ we split the d_i half-edges of i into two distinct groups which we call *community* half-edges and *background* half-edges. For $a \in \mathbb{Z}$ and $b \in [0, 1)$ define the random variable $\lfloor a + b \rceil$ as

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

Now define $Y_i := \lfloor (1 - \xi) d_i \rceil$ and $Z_i := d_i - Y_i$ (note that Y_i and Z_i are random variables with $\mathbb{E}[Y_i] = (1 - \xi) d_i$ and $\mathbb{E}[Z_i] = \xi d_i$) and, for all $i \in [n]$, split the d_i half-edges of i into Y_i community half-edges and Z_i background half-edges. Next, for all $j \in [L]$, construct the *community graph* $G_{n,j}$ as per the configuration model on node set C_j and degree sequence $(Y_i, i \in C_j)$. Finally, construct the *background graph* $G_{n,0}$ as per the configuration model on node set $[n]$ and degree sequence $(Z_i, i \in [n])$. In the event that the sum of degrees in a community is odd, we pick a maximum degree node i in said community and replace Y_i with $Y_i + 1$ and Z_i with $Z_i - 1$. As we show in the proof of Theorem 3, this minor adjustment also has a negligible effect on the analysis and we thus assume that none of these sums are odd. Note that $G_{n,j}$ is a graph and C_j is the set of nodes in this graph; we refer to C_j as a *community* and $G_{n,j}$ as a *community graph*. Note also that $G_n = \bigcup_{0 \le j \le n} G_{n,j}$.

Phase 5: rewiring self-loops and multi-edges

Note that, although we are calling $G_{n,0}, G_{n,1}, \ldots, G_{n,L}$ *graphs*, they are in fact *multi-graphs* at the end of phase 4. To ensure that G_n is simple, we perform a series of *rewirings* in G_n. A rewiring takes two edges as input, splits them into four half-edges, and creates two new edges distinct from the input. We first rewire each community graph $G_{n,j}$ independently as follows.

1. For each edge $e \in E(G_{n,j})$ that is either a loop or contributes to a multi-edge, we add e to a *recycle* list that is assigned to $G_{n,j}$.
2. We shuffle the *recycle* list and, for each edge e in the list, we choose another edge e' uniformly from $E(G_{n,j}) \setminus \{e\}$ (not necessarily in the *recycle* list) and attempt to rewire these two edges. We save the result only if the rewiring does not lead to any further self-loops or multi-edges, otherwise we give up. In either case, we then move to the next edge in the *recycle* list.

3. After we attempt to rewire every edge in the *recycle* list, we check to see if the new *recycle* list is smaller. If yes, we repeat step 2 with the new list. If no, we give up and move all of the "bad" edges from the community graph to the background graph.

We then rewire the background graph $G_{n,0}$ in the same way as the community graphs, with the slight variation that we also add edge e to *recycle* if e forms a multi-edge with an edge in a community graph or, as mentioned previously, if e was moved to the background graph as a result of giving up during the rewiring phase of its community graph. At the end of phase 5, we have a simple graph $G_n \sim \mathcal{A}$.

Note that phase 5 of the **ABCD** construction process exists only to ensure that G_n is simple. Thus, if one were satisfied with a multi-graph G_n that had all of the properties $\mathcal{A}$ offers, one could simply terminate the process after phase 4. However, for most practical uses such as community detection, we require a simple graph and thus require phase 5. As mentioned in Section 1, phase 5 is a time consuming part of the algorithm. Theorem 3.3 gives us some insight as to why that is the case, namely, because with high probability the number of self-loops and multi-edges generated during phase 4 is at least $\Omega(L)$. Theorem 3.3 is therefore quite valuable as it lets us know when our choice of γ, β, ζ and τ will yield a best-case-scenario number of self-loops and multi-edges (in expectation).

Theorem 3.3 is also valuable for helping us understand how "skewed" the community graphs, along with the background graph, are with respect to graphs generated uniformly at random from the set of simple graphs on the respective degree sequences. In [14], Janson shows that if a graph is constructed as the configuration model on degree sequence $\mathbf{d}$, followed by a series of rewirings, then a relatively small number of rewirings yields a distribution that is asymptotically equal (with respect to the total variation distance) to the uniform distribution on simple graphs with degree sequence $\mathbf{d}$. By extrapolating this result, we can infer that the number of rewirings required in phase 5 of the **ABCD** construction process is directly correlated with how "skewed" the resulting graph is.

2.5 A Known Result for ABCD

A result from [15] that we use often in this paper is a tight bound on the number of communities generated by the **ABCD** model.

Theorem 2.1 ([15] Corollary 5.5 (a)). *Let $G_n \sim \mathcal{A}$ and let L be the number of communities in G_n. Then w.e.p. the number of communities, L, is equal to*

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

where

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

Note that the concentration in Theorem 2.1 is a consequence of the bound $|C_j| \leq n^\tau$ for all communities C_j and fails if this bound is omitted.

3 Main Result

Our main result is a stochastic bound on the degree sequence of a given community in $\mathcal{A}$. For $G_n \sim \mathcal{A}$ with degree sequence $\mathbf{d}_n$, and for community graph $G_{n,j}$ with nodes from C_j, we make the following distinction: the *degree sequence of* $G_{n,j}$ is the degree sequence of the community graph $G_{n,j}$, whereas the *degree sequence of* C_j is the subset of $\mathbf{d}_n$ containing the degrees of nodes in C_j. Hence, the degree sequence of C_j is $(d_v, v \in C_j)$ and the degree sequence of $G_{n,j}$ is $(Y_v, v \in C_j)$ where we recall that $Y_v = \lfloor (1-\xi)d_v \rceil$. The following two results, Theorem 3.1 and Corollary 3.2, are stated in terms of the degree sequences $(d_v, v \in C_j)$. However, both results can be easily restated in terms of the degree sequences $(Y_v, v \in C_j)$.

Theorem 3.1. *Let $G_n \sim \mathcal{A}$. Let C_j be a community in G_n with $|C_j| = z$ and let $\mathbf{c}_j$ be the degree sequence of community C_j. Next, let $\epsilon = \epsilon(n) = n^{-(\tau-\zeta)(2-\beta)/2} = o(1)$, let*

$$\Delta_z = \min\left\{\frac{z-1}{1-\xi\phi}, n^{\zeta}\right\}, \qquad \text{where } \phi = 1 - \frac{1}{n^2}\sum_{j\in[L]} |C_j|^2,$$

and let X^- and X^+ be random variables with the following probability distribution functions on $\{\delta, \dots, \Delta_z\}$:

$$\mathbb{P}\left(X^- = k\right) = \frac{\int_k^{k+1} x^{-\gamma}\, dx}{\int_\delta^{\Delta_z+1} x^{-\gamma}\, dx}, \text{ and}$$

$$\mathbb{P}\left(X^+ = k\right) = \frac{\left(1-\epsilon\mathbf{1}_{[k=\delta]}\right)\int_k^{k+1} x^{-\gamma}\, dx}{(1-\epsilon)\int_\delta^{\delta+1} x^{-\gamma}\, dx + \int_{\delta+1}^{\Delta_z+1} x^{-\gamma}\, dx} = (1+o(1))\mathbb{P}\left(X^- = k\right),$$

where $\mathbf{1}_{[k=\delta]}$ is the Kronecker delta (function of two variables, k and δ, that is equal to 1 if $k = \delta$ and equal to 0 otherwise). Finally, let X be a uniformly random element of $\mathbf{c}_j$. Then w.h.p. X is stochastically bounded below by X^- and above by X^+.

The power of Theorem 3.1 is that it allows us to compare the structure of community graphs in $G_n \sim \mathcal{A}$ with the structure of graphs constructed via the configuration model on an i.i.d. degree sequence that is well understood. In this paper we provide two uses of this new and powerful tool. Aside from these two uses, this theorem, combined with the fact that communities in the **ABCD** model are generated independently by the simple and analyzable configuration model, provides a vehicle to future analysis of other important properties such as clustering coefficient, spreading of information, expansion properties, robustness, etc. The first illustration of its power is a sharpening of Lemma 5.6 in [15], describing the volumes of communities in $G_n \sim \mathcal{A}$. For $X \sim \mathcal{P}(\gamma, \delta, \Delta)$, write

$$\mu_\ell(\gamma, \delta, \Delta) = \mathbb{E}\left[X^\ell\right], \tag{1}$$

and note in particular that $\mu_1(\gamma, \delta, n^\zeta)$ is the expected degree of a node in $G_n \sim \mathcal{A}$. Next, for community C_j, define

$$\mathrm{vol}(C_j) := \sum_{v \in C_j} d_v .$$

Corollary 3.2. *Let $G_n \sim \mathcal{A}$, let C_j be a community in G_n with $|C_j| = z$, and let*

$$\Delta_z = \min\left\{\frac{z-1}{1-\xi\phi}, n^{\zeta}\right\}.$$

Then, conditioned on the stochastic domination in Theorem 3.1,

$$\frac{\mathbb{E}\left[\mathrm{vol}(C_j)\right]}{z} = (1+o(1))\,\mu_1(\gamma,\delta,\Delta_z) = \begin{cases}(1+o(1))\,\mu_1(\gamma,\delta,n^{\zeta}) & \text{if } z(n)\to\infty\,,\ \text{and}\\ \Theta\left(\mu_1(\gamma,\delta,n^{\zeta})\right) & \text{otherwise.}\end{cases}$$

The second use of Theorem 3.1 that we present here is an analysis of the number of self-loops and multi-edges that are created during phase 4 of the construction process of $G_n \sim \mathcal{A}$. In practice, phase 5 of the **ABCD** construction can be computationally expensive. It is therefore valuable to study the number of collisions (self-loops and multi-edges) generated during phase 4 of the construction. The following theorem tells us that, although w.h.p. we can never do better than generating $\Omega(L)$ collisions, where L is the number of communities, we expect to see *at most* $O(L)$ collisions under certain restrictions on γ, β, ζ, and τ.

Theorem 3.3. *Let $G_n \sim \mathcal{A}$ and define the following five variables depending on G_n.*

$S_c :=$ *The number of self-loops in community graphs after phase 4.*
$M_c :=$ *The number of multi-edge pairs in community graphs after phase 4.*
$S_b :=$ *The number of self-loops in the background graph after phase 4.*
$M_b :=$ *The number of multi-edge pairs in the background graph after phase 4.*
$M_{bc} :=$ *The number of background edges that are also community edges after phase 4.*

Then, conditioned on the stochastic domination in Theorem 3.1,

1. $\mathbb{E}\left[S_c\right] = O\Big((n^{1-\tau(2-\beta)})(1+n^{\zeta(4-\gamma-\beta)})\Big)$,
2. $\mathbb{E}\left[M_c\right] = O\Big((n^{1-\tau(2-\beta)})(1+n^{\zeta(7-2\gamma-\beta)})\Big)$,
3. $\mathbb{E}\left[S_b\right] = O(n^{\zeta(3-\gamma)})$,
4. $\mathbb{E}\left[M_b\right] = O(n^{\zeta(6-2\gamma)})$, *and*
5. $\mathbb{E}\left[M_{bc}\right] = o(\mathbb{E}\left[M_c\right])$.

Moreover, for all valid $\gamma, \beta, \zeta, \tau$,

$$\mathbb{E}\left[S_c\right] = \Omega(L)\,,$$

if $\gamma + \beta > 4$ then

$$\mathbb{E}\left[S_c + M_c + M_{bc}\right] = \Theta(L)\,,$$

if $2\zeta(3-\gamma) + \tau(2-\beta) \le 1$ then

$$\mathbb{E}\left[S_b + M_b\right] = O(L)\,,$$

and if both inequalities are satisfied then

$$\mathbb{E}\left[S_c + M_c + S_b + M_b + M_{bc}\right] = \Theta(L)\,.$$

The proofs of Theorem 3.1, Corollary 3.2 and Theorem 3.3, are presented in Section 5.

4 Simulation Corner

In this section, we present a few experiments highlighting the properties that are proved to hold with high probability. The experiments show that the asymptotic predictions are useful even for graphs on a moderately small number of nodes.

4.1 The Coupling

Our main result (Theorem 3.1) shows that the degree distribution of a community of size z in $G_n \sim \mathcal{A}$ is stochastically sandwiched between $(X_i^-, i \in [z])$ and $(X_i^+, i \in [z])$ where $X_i^- \sim \mathcal{P}(\gamma, \delta, \Delta_z)$ and $X_i^+ \xrightarrow{d} X_i^-$ as $n \to \infty$. For two random variables X and Y, $X \xrightarrow{d} Y$ is used to indicate the convergence in distribution which means that the cumulative distribution function (CDF) of X converges to the CDF of Y.) To compare the degree distribution of communities in **ABCD** graphs to the stochastic lower-bound $(X_i^-, i \in [z])$, we perform the following experiment. We generate three **ABCD** graphs G_n, G_n^* and G_n^{**}. Consistent in all three graphs are the parameters $n = 2^{20}, \delta = 5, \zeta = 0.4, s = 50, \tau = 0.6$, and $\xi = 0.5$. The graph G_n has parameters $\gamma = 2.1$ and $\beta = 1.1$, the graph G_n^* has $\gamma = 2.5$ and $\beta = 1.5$, and G_n^{**} has $\gamma = 2.9$ and $\beta = 1.9$. For each graph, we plot the complementary cumulative distribution function (ccdf) of degrees of (a) the whole graph, (b) the union of all smallest communities (G_n had 8 communities of size $s = 50$, G_n^* had 29, and G_n^{**} had 82), and (c) the unique largest community (sizes 4074, 4073, and 3903 in respective graphs G_n, G_n^*, and G_n^{**}). We then plot, in parallel, the expected ccdfs for the three graphs; for the whole graph the ccdf is that of $\mathcal{P}\left(\gamma, \delta, n^{\zeta}\right)$, and for the community graphs we use the expected ccdf of the stochastic lower-bound $(X_i^-, i \in [z])$, i.e., the function $\bar{F} : \{\delta, \ldots, \Delta_z\} \to [0, 1]$ where

$$\bar{F}(k) = \frac{\int_k^{\Delta_z+1} x^{-\gamma}\,dx}{\int_\delta^{\Delta_z+1} x^{-\gamma}\,dx} = \frac{k^{1-\gamma} - (\Delta_z + 1)^{1-\gamma}}{\delta^{1-\gamma} - (\Delta_z + 1)^{1-\gamma}}\,.$$

The results are presented in Figure 1. From these results, we see that the distribution of $(X_i^-, i \in [z])$ is a very good approximation of the distribution of degrees in a community of smallest size as well as a community of largest size. We note that, since $(X_i^-, i \in [z])$ is a lower-bound, we expect the theoretical ccdf to sit slightly above the empirical ccdf, and this is confirmed by the experiment.

4.2 Volumes of Communities

Next, to investigate how well Corollary 3.2 predicts the volume of a particular community, we perform the following experiment. We generate three **ABCD** graphs G_n, G_n^* and G_n^{**}. Consistent in all three graphs are the parameters $n = 2^{20}$, $\delta = 5$, $\zeta = 0.6$, $s = 50$, $\tau = 0.9$, and $\xi = 0.5$. The graph G_n has parameters $\gamma = 2.1$ and $\beta = 1.1$, the graph G_n^* has $\gamma = 2.5$ and $\beta = 1.5$, and G_n^{**} has $\gamma = 2.9$ and $\beta = 1.9$. In each graph, we sorted communities with respect to their size (from the smallest to the largest) and then grouped them into 10 buckets as equal as possible (that is, the number of communities in any pair of buckets differs by at most one). For each bucket we compute the average degree and the standard deviation over all communities in that bucket. We compare it with the asymptotic prediction based on Corollary 3.2, that is, for each community of size z we compute $\mu_1(\gamma, \delta, \Delta_z)$, and take the average over all communities in the bucket. The results are presented in Figure 2. We see that $n = 2^{20}$ is large enough and simulations match the theoretical predictions almost exactly.

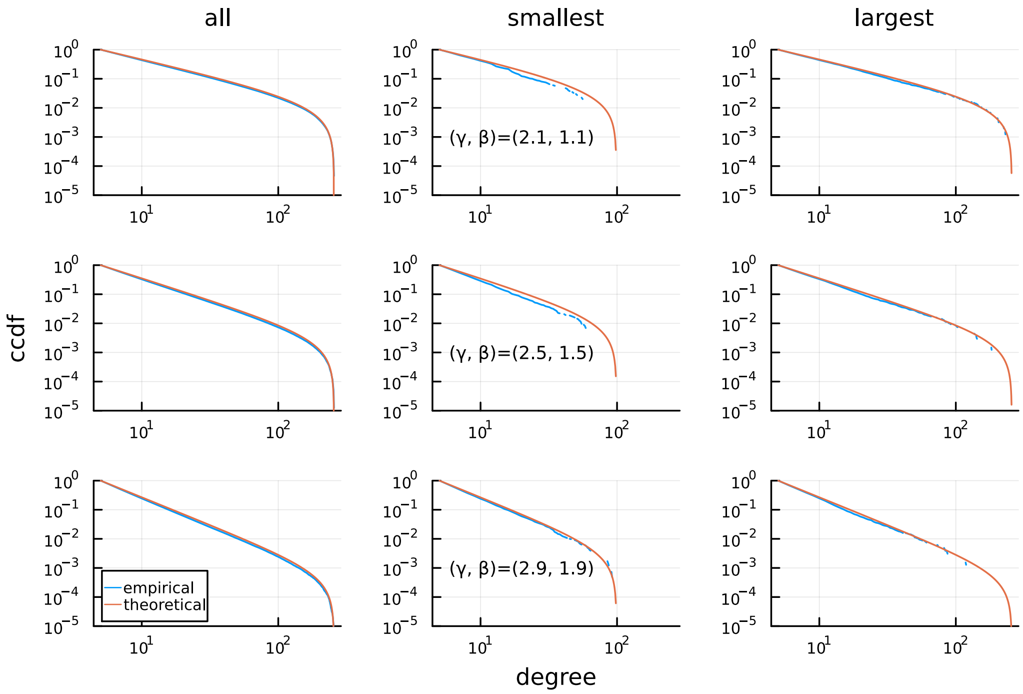

Figure 1: The ccdf for the three different **ABCD** graphs G_n (top), G_n^* (middle), and G_n^{**} (bottom), and for three different subsets of nodes in each graph, namely, the whole graph (left), the union of smallest community graphs (middle), and the unique largest community graph (right). Each function is drawn on a log–log scale. The blue curves are the empirical data and the orange curves are the theoretical predictions.

4.3 Self-loops and Multi-edges

Finally, to investigate the number of collisions (of various types) generated during phase 4 of the **ABCD** construction as functions of n, we perform the following experiment. For each $n \in \{2^{15}, 2^{16}, 2^{17}, 2^{18}, 2^{19}, 2^{20}\}$, we generate three sequences of 20 **ABCD** graphs $(G_n(i), i \in [20])$, $(G_n^*(i), i \in [20])$, and $(G_n^{**}(i), i \in [20])$. Consistent in all three sequences are the parameters $\delta = 5$, $\zeta = 0.6$, $s = 50$, $\tau = 0.9$, and $\xi = 0.5$. The graphs in sequence $(G_n(i), i \in [20])$ have $\gamma = 2.1$ and $\beta = 1.1$, the graphs in $(G_n^*(i), i \in [20])$ have $\gamma = 2.5$ and $\beta = 1.5$, and the graphs in $(G_n^{**}(i), i \in [20])$ have $\gamma = 2.9$ and $\beta = 1.9$. We compare the growth of S_c/L, M_c/L, S_b/L, and M_b/L (the average values and the corresponding standard deviations over 20 graphs), as functions of n, for all three sequences. Each sequence represents a different scenario in expectation based on Theorem 3.3, and we comment on each result separately.

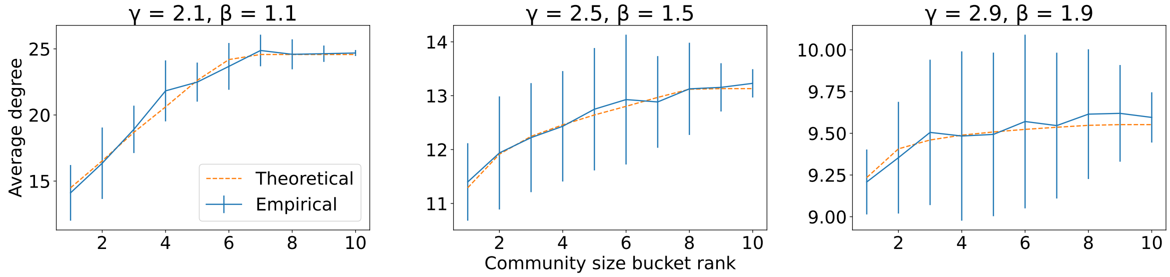

Figure 2: The average degrees in communities for G_n (left), G_n^* (middle), and G_n^{**} (right). The communities are ranked by their size and grouped into 10 buckets as equal as possible. The blue line with error bars is the average degree and standard deviation among all communities in each bucket. Note that the errors, in absolute values, are largest for the leftmost plot and smallest for the rightmost plot. The orange dashed line shows the expected volumes for the stochastic lower-bound $(X_i^-, i \in [z])$, computed for each community size and bucketed in the same way as the empirical data.

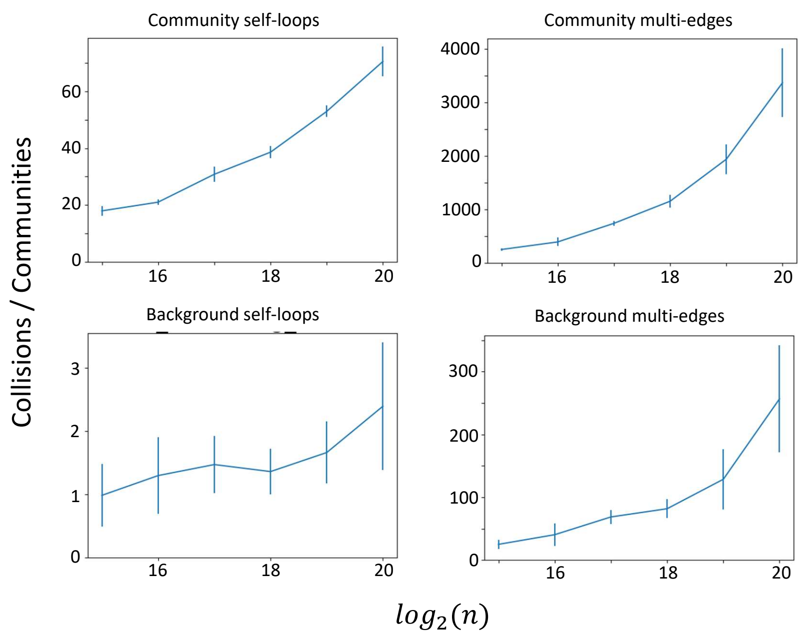

Figure 3: In reading order: $S_c/L, M_c/L, S_b/L$ and M_b/L vs. $\log_2(n)$ for $(G_n(i), i \in [20])$ with $\gamma = 2.1$ and $\beta = 1.1$, averaged over the 20 graphs.

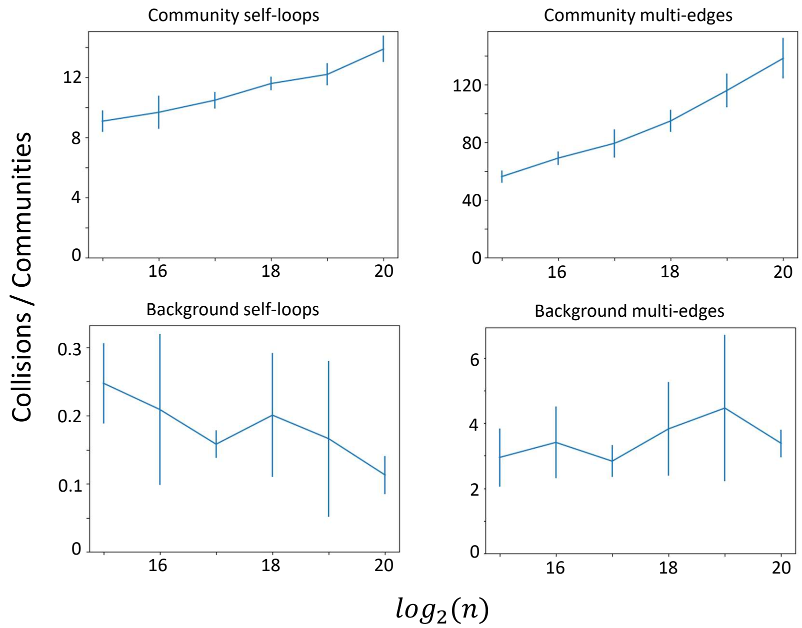

Figure 4: In reading order: $S_c/L, M_c/L, S_b/L$ and M_b/L vs. $\log_2(n)$ for $(G_n^*(i), i \in [20])$ with $\gamma = 2.5$ and $\beta = 1.5$, averaged over the 20 graphs.

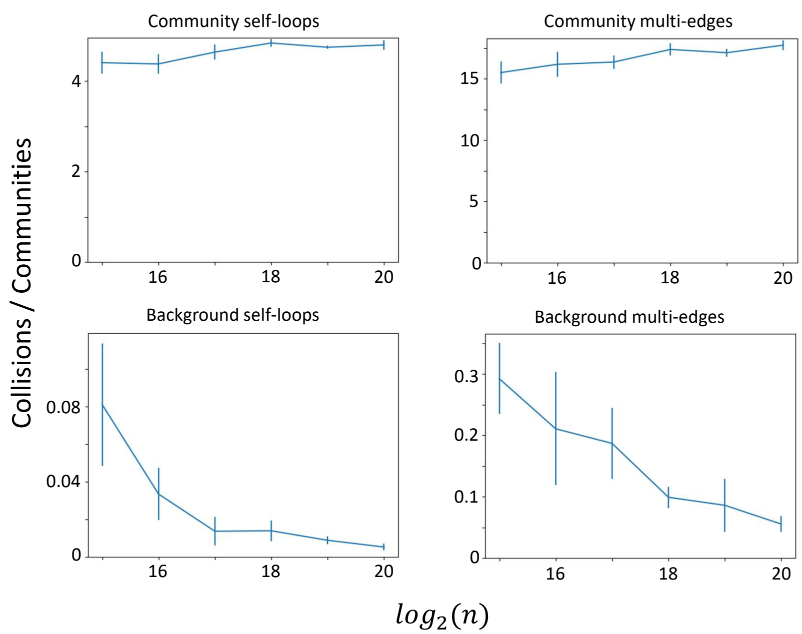

Figure 5: In reading order: $S_c/L, M_c/L, S_b/L$ and M_b/L vs. $\log_2(n)$ for $(G_n^{**}(i), i \in [20])$ with $\gamma = 2.9$ and $\beta = 1.9$, averaged over the 20 graphs.

- For $(G_n(i), i \in [20])$ with $\gamma = 2.1$ and $\beta = 1.1$, we have $\gamma + \beta < 4$ and $2\zeta(3-\gamma) + \tau(2-\beta) >$

$\zeta(3-\gamma)+\tau(2-\beta)>1$ and so we expect each of the variables S_c/L, M_c/L, S_b/L, and M_b/L to be unbounded. In Figure 3 we see that, indeed, each of the four variables seem to grow with n in the simulations.

- For $(G_n^*(i), i \in [20])$ with $\gamma = 2.5$ and $\beta = 1.5$, we have $\gamma+\beta = 4$ and $2\zeta(3-\gamma)+\tau(2-\beta) > 1 > \zeta(3-\gamma)+\tau(2-\beta)$ and so we expect S_b/L to be bounded and $S_c/L, M_c/L, M_b/L$ to be unbounded. As Figure 4 shows, the simulations are consistent with the theory for $S_c/L, M_c/L$ and S_b/L. However, the trend of M_b/L is unclear. Considering that $2\zeta(3-\gamma)+\tau(2-\beta) = 1.05$ in this case, it is reasonable that the growth of M_b/L should not reveal itself at this scale of n.

- For $(G_n^{**}(i), i \in [20])$ with $\gamma = 2.9$ and $\beta = 1.9$, we have $\gamma+\beta > 4$ and $1 > 2\zeta(3-\gamma)+\tau(2-\beta) > \zeta(3-\gamma)+\tau(2-\beta)$ and so we expect all of $S_c/L, M_c/L, S_b/L, M_b/L$ to be bounded. Figure 5 again shows us that theory matches simulations. We note the very slight upward trend of S_c/L and M_c/L, likely due to n being too small to see the asymptotic bound take hold.

We conclude that Theorem 3.3 does a good job at telling us the behaviour of S_c/L, M_c/L, S_b/L, and M_b/L for various γ and β, although the results are not as clear as the other experiments which would likely be resolved by taking larger values of n.

5 Proofs

5.1 The coupling (proof of Theorem 3.1)

Before we set up a coupling that sandwiches the **ABCD** construction process in order to control the degree sequence of any community C_j, we need to show that almost all nodes belong to large communities. Such communities are large enough such that they can be assigned nodes of any degree. Indeed, since the maximum degree in G_n is (deterministically) at most n^ζ, only communities of size less than $n^\zeta(1-\xi\phi)+1 \le n^\zeta$ might *not* be available during the entire phase 3 of the **ACBD** construction process.

Lemma 5.1. *Let $\omega = \omega(n)$ be any function such that $\omega \to \infty$ sufficiently slowly as $n \to \infty$. Next, let $G_n \sim \mathcal{A}$ and let $V' \subseteq V(G_n)$ be the set of nodes in communities of size at most n^ζ. Then, w.h.p. $|V'| < \omega n^{1-(\tau-\zeta)(2-\beta)} = o(n^{1-(\tau-\zeta)(2-\beta)/2}) = o(n)$.*

Proof. Recall that $0 < \zeta < \tau < 1$ and $1 < \beta < 2$. Pick a community $C \in \mathbf{C}_n$ uniformly at random and let $X = |C|$ if $|C| \le n^\zeta$; otherwise, $X = 0$. Then, for $s \le m \le n^\zeta$,

$$
\begin{aligned}
\mathbb{P}\left(X=m\right) &= \frac{\int_m^{m+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
&= (\beta-1)\frac{\int_m^{m+1} y^{-\beta}\,dy}{s^{1-\beta}-(n^\tau+1)^{1-\beta}} \\
&= \left(1+O(n^{\tau(1-\beta)})\right)(\beta-1)s^{\beta-1}\int_m^{m+1} y^{-\beta}\,dy\,,
\end{aligned}
$$

and hence

$$
\begin{aligned}
\mathbb{E}\left[X\right] &= \left(1+O(n^{\tau(1-\beta)})\right)(\beta-1)s^{\beta-1}\sum_{m=s}^{\lfloor n^\zeta \rfloor} m\int_m^{m+1} y^{-\beta}dy \\
&\leq \left(1+O(n^{\tau(1-\beta)})\right)(\beta-1)s^{\beta-1}\int_s^{n^\zeta+1} y^{1-\beta}dy \\
&= \left(1+O(n^{\tau(1-\beta)})\right)\frac{(\beta-1)s^{\beta-1}}{2-\beta}\left((n^\zeta+1)^{2-\beta}-s^{2-\beta}\right) \\
&= \left(1+O(n^{\tau(1-\beta)})+O(n^{\tau(\beta-2)})\right)\frac{(\beta-1)s^{\beta-1}}{2-\beta}n^{\zeta(2-\beta)} \\
&= \Theta\left(n^{\zeta(2-\beta)}\right) .
\end{aligned}
$$

Finally, since w.e.p. $L = \Theta\left(n^{1-\tau(2-\beta)}\right)$ (see Theorem 2.1), we get that

$$
\mathbb{E}\left[V'\right] = O\left(n\exp(-\log^2 n) + n^{1-\tau(2-\beta)}\mathbb{E}\left[X\right]\right) = O\left(n^{1-(\tau-\zeta)(2-\beta)}\right) ,
$$

and the lemma now follows from Markov's inequality:

$$
\mathbb{P}\left(|V'| \geq \omega n^{1-(\tau-\zeta)(2-\beta)}\right) \leq \frac{\mathbb{E}\left[V'\right]}{\omega n^{1-(\tau-\zeta)(2-\beta)}} = O\left(\frac{1}{\omega}\right) \to 0
$$

as $n \to \infty$. □

We will also need the following simple fact about the distribution $\mathcal{P}(\gamma,\delta,\Delta)$.

Fact 5.2. *Fix $\gamma > 0$ and $1 \leq \delta \leq \delta' \leq \Delta' \leq \Delta$. Then $X \sim \mathcal{P}(\gamma,\delta,\Delta)$, conditioned on $\delta' \leq X \leq \Delta'$, has distribution $\mathcal{P}(\gamma,\delta',\Delta')$.*

The remainder of Section 5.1 is dedicated to proving Theorem 3.1. In the coming arguments, we say sequence $(X_i, i \in I)$ is stochastically dominated by sequence $(Y_i, i \in I)$ if, for uniform $X \in (X_i, i \in I)$ and uniform $Y \in (Y_i, i \in I)$, X is stochastically dominated by Y. Furthermore, with respect to phase 3 of the **ABCD** construction process, we refer to a community C as *locked* at step i if $d_i > (|C|-1)/(1-\xi\phi)$ and otherwise we refer to C as *unlocked* at step i. We say that a node is locked/unlocked at step i if its corresponding community is locked/unlocked at step i. Note that, since $d_1 \leq n^\zeta$, all communities of size at least $n^\zeta(1-\xi\phi)+1$ are always unlocked.

We start with the modified version of phase 3 of the **ABCD** construction process that will be used to prove the lower bound in Theorem 3.1. Fix z with $s \leq z \leq n^\tau$ and define the construction process $\mathcal{A}^-(z)$, yielding a collection of degrees assigned to a collection of communities notated as G_n^-, as follows.

1. Copy phases 1 and 2 of the **ABCD** construction process to get a degree distribution $\mathbf{d}_n = (d_i, i \in [n])$ and a collection of communities $\mathbf{C}_n = (C_j, j \in [L])$ each containing unassigned nodes (recall that unassigned nodes are nodes that have not yet been assigned a label or a degree).

2. Copy phase 3 of the **ABCD** construction process until the communities of size z are unlocked. This event occurs at step i where i is the smallest label satisfying $d_i \leq \frac{z-1}{1-\xi\phi}$ (recall that the degree sequence $\mathbf{d}_n = (d_i, i \in [n])$ is non-increasing and that label i and degree d_i are assigned to an unassigned node at time i). At this point, all communities of size at least z are unlocked and $i-1$ nodes that belong to communities of size at least $z+1$ have been assigned a label and a degree.

3. Now unlock all communities and assign labels $i, \ldots, n$ and corresponding degrees $d_i, \ldots, d_n$ to the unlabelled nodes in $[n]$ uniformly at random.

We will first show that a community C_j in G_n^- of size z has the desired degree distribution.

Lemma 5.3. *Fix $z = z(n)$ such that $s \leq z \leq n^\tau$. Let $G_n^- \sim \mathcal{A}^-(z)$, let C_j be a community in G_n^- with $|C_j| = z$ and with degree sequence $\mathbf{c}_z^-$, and let $(X_i^-, 1 \leq i \leq z)$ be the i.i.d. sequence defined in Theorem 3.1. Then, $\mathbf{c}_z^- \stackrel{d}{=} (X_i^-, 1 \leq i \leq z)$.*

Proof. To prove the lemma, we will use the well-known Principle of Deferred Decisions. This simple but very useful technique is often used in analysis of randomized algorithms. The idea behind the principle is that the entire set of random choices are not made in advance, but rather fixed only as they are revealed to the algorithm [25]. In our context, a simple but useful observation is that when constructing G_n^- one can defer exposing some information about the degree sequence $\mathbf{d}_n$ to the very end. Indeed, during phase 1 of the **ABCD** construction, we may only expose information whether $d_i \leq \frac{z-1}{1-\xi\phi}$ or not; if $d_i > \frac{z-1}{1-\xi\phi}$, then we expose d_i but otherwise we only reveal that $d_i \leq \frac{z-1}{1-\xi\phi}$. This partial information is enough to continue with the auxiliary process of constructing G_n^-.

Recall that community C_j is locked as long as $d_i > \frac{z-1}{1-\xi\phi}$. Let i be the smallest label such that $d_i \leq \frac{z-1}{1-\xi\phi}$. (Note that, in particular, if $n^\zeta \leq \frac{z-1}{1-\xi\phi}$, then C_j is immediately unlocked, that is $i = 1$.) Once we unlock C_j in G_n^- at step i, we unlock all communities and assign degrees $d_i, \ldots, d_n$ uniformly to the set of unassigned nodes in $[n]$. Thus, $\mathbf{c}_z^-$ is a uniform subsequence of $(d_i, \ldots, d_n)$ of size z. Now, we finally expose the degrees in this subsequence. By Fact 5.2, each d_i follows precisely a truncated power law with upper bound $\Delta_z = \min\left\{\frac{z-1}{1-\xi\phi}, n^\zeta\right\}$ and lower bound δ. Thus, $\mathbf{c}_z^- \stackrel{d}{=} (X_i^-, 1 \leq i \leq z)$, proving the lemma. □

We are now ready to couple the auxiliary process constructing G_n^- with the original process generating G_n, the **ABCD** graph. This will prove the lower bound in Theorem 3.1.

Proof of Theorem 3.1 (lower bound). Construct $G_n^- \sim \mathcal{A}^-(z)$ with nodes labelled as $[n]$, degree sequence $\mathbf{d}_n = (d_i, i \in [n])$, and community sequence $\mathbf{C}_n = (C_j, j \in [L])$. Next, for all $i \in [n]$ define $z_i = \lceil d_i(1-\xi\phi) + 1 \rceil$; note that a community C is unlocked in phase 3 of the **ABCD** construction at the first step i for which $|C| \geq z_i$). Now construct G_n in parallel with G_n^- as follows.

1. Let G_n have degree sequence $\mathbf{d}_n$ and community sequence $\mathbf{C}_n$.

2. Copy the degree assignment process of G_n^- until the communities of size z are unlocked. Let i be the smallest label satisfying $d_i \leq \frac{z-1}{1-\xi\phi}$. Instead of unlocking all communities as we do in $G_n^- \sim \mathcal{A}^-(z)$, we will unlock only those communities C satisfying

$$|C| \geq z_i = \lceil d_i(1-\xi\phi) + 1 \rceil$$

as we do in $\mathcal{A}$. (Note that, if $|C| = z \geq \lceil n^\zeta(1-\xi\phi)+1\rceil$, then $i = 1$ and C is unlocked from the start.)

3. Now, for $j \in \{i, \dots, n\}$ starting with $j = i$, we first unlock all communities C satisfying $|C| \geq z_j$. We then partition the nodes into four sets. We say that node v is *open* in G_n at step j if v is both unlocked and unlabelled before step j, and otherwise we say v is closed at step j (and similarly for G_n^-). The four sets are as follows:

$$\begin{aligned} V_j^{++} &= \{v : v \text{ is open in both } G_n^- \text{ and } G_n \text{ at step } j\}\ , \\ V_j^{+-} &= \{v : v \text{ is open in } G_n^- \text{ and closed in } G_n \text{ at step } j\}\ , \\ V_j^{-+} &= \{v : v \text{ is closed in } G_n^- \text{ and open in } G_n \text{ at step } j\}\ , \\ V_j^{--} &= \{v : v \text{ is closed in both } G_n^- \text{ and } G_n \text{ at step } j\}\ . \end{aligned}$$

Note that V_i^{+-} is the set of nodes in communities of size at most $z_i - 1$ and $V_i^{-+} = \emptyset$. However, all four sets will change with j. We now choose a node v in G_n^- to receive label j and degree d_j as per the $\mathcal{A}^-(z)$ construction (note that v is a uniform element of $V_j^{++} \cup V_j^{+-}$). We then choose a node in G_n to receive label j and degree d_j as follows.

- If $v \in V_j^{++}$, then we give label j and degree d_j to v in G_n.
- If $v \in V_j^{+-}$, then we give label j and degree d_j to a uniform node in V_j^{-+} with probability p_j, and to a uniform node in V_j^{++} with probability $1 - p_j$, where

$$p_j = \frac{|V_j^{++}||V_j^{-+}| + |V_j^{+-}||V_j^{-+}|}{|V_j^{++}||V_j^{+-}| + |V_j^{+-}||V_j^{-+}|}\,;$$

  we will later verify that $p_j \leq 1$.

4. Once all nodes have been assigned a degree, create the community edges and background edges in G_n as per the usual $\mathcal{A}$ construction process.

We claim (a) that $G_n \sim \mathcal{A}$, and (b) that any community $C \in \mathbf{C}_n$ of size z with G_n-degree sequence $\mathbf{c}_z$ and G_n^--degree sequence $\mathbf{c}_z^-$ satisfies $\mathbf{c}_z \geq \mathbf{c}_z^-$ point-wise.

Starting with claim (a), it is clear by the construction process $\mathcal{A}^-(z)$ that $\mathbf{d}_n$ and $\mathbf{C}_n$ are valid sequences for $G_n \sim \mathcal{A}$. We must then verify that, for $j = i, \dots, n$, the node in G_n chosen to receive label j and degree d_j is a uniform node from the set of unlabelled nodes in communities of size at least $d_j(1-\xi\phi)+1$. Note that this set of nodes is precisely $V_j^{++} \cup V_j^{-+}$, and so we need only show that, for $u, v \in V_j^{++} \cup V_j^{-+}$, the probability of labelling u and the probability of labelling v are equal. We will first show that $p_j \leq 1$ by showing that $|V_j^{-+}| \leq |V_j^{+-}|$ for all $j \in \{i, \dots, n\}$. In fact, we will show a stronger result, namely, that $|V_j^{+-}| - |V_j^{-+}|$ is precisely the number of nodes in communities that are locked in G_n at time j.

As mentioned earlier, when $j = i$, V_j^{+-} is the set of nodes in communities that are still locked (that is, of size at most $z_i - 1$) and $V_j^{-+} = \emptyset$, so the desired property holds. Now suppose the property holds up to some time $j \geq i$. At step j, if $v \in V_j^{++}$ receives label j and degree d_j in G_n^-, then v also receives this label and degree in G_n, and thus v is moved from V_j^{++} to V_{j+1}^{--} ($|V_{j+1}^{+-}| - |V_{j+1}^{-+}|$ is unaffected by this event). On the other hand, if $v \in V_j^{+-}$ receives label j and degree d_j at step j, then v is moved from V_j^{+-} to V_{j+1}^{--} and we have two sub-cases to consider. If some node $u \in V_j^{-+}$ receives label j and degree d_j in G_n, then u is moved from V_j^{-+} to V_j^{--} (V_{j+1}^{+-} and V_{j+1}^{-+} each lose one node in this case); if some node $u \in V_j^{++}$ receives label j and degree d_j in G_n, then u is moved from V_j^{++} to V_j^{+-} (V_{j+1}^{+-} loses a node and gains a different node in this case). Thus, in any case, $|V_{j+1}^{+-}| - |V_{j+1}^{-+}|$ is unaffected by the process of assigning labels and degrees. Finally, we need to investigate what happens when communities are unlocked. Any node in a locked community at step j is in V_j^{+-} or V_j^{--}. Once a community is unlocked, all of the corresponding nodes in V_j^{+-} move to V_{j+1}^{++} and all of the corresponding nodes in V_j^{--} move to V_{j+1}^{-+}. Thus, every node in a newly unlocked community decreases V_{j+1}^{+-} by one or increases V_{j+1}^{-+} by one, but not both. Therefore, $\left(|V_j^{+-}| - |V_j^{-+}|\right) - \left(|V_{j+1}^{+-}| - |V_{j+1}^{-+}|\right)$ is precisely the number of nodes in communities unlocked at step $j+1$. The claim now follows by induction.

We have established that

$$p_j = \frac{|V_j^{++}||V_j^{-+}| + |V_j^{+-}||V_j^{-+}|}{|V_j^{++}||V_j^{+-}| + |V_j^{+-}||V_j^{-+}|} \leq 1.$$

Next, consider a node $v \in V_j^{-+}$. Then v is given label j and degree d_j in G_n if and only if some node V_j^{+-} is chosen in G_n^-, the label is redirected to V_j^{-+} in G_n, and v is then chosen uniformly from the set V_j^{-+} to receive the label in G_n. Thus, the probability that $v \in V_j^{-+}$ is assigned label j and degree d_j is

$$\left(\frac{|V_j^{+-}|}{|V_j^{++}| + |V_j^{+-}|}\right)\left(\frac{|V_j^{++}||V_j^{-+}| + |V_j^{+-}||V_j^{-+}|}{|V_j^{++}||V_j^{+-}| + |V_j^{+-}||V_j^{-+}|}\right)\left(\frac{1}{|V_j^{-+}|}\right) = \frac{1}{|V_j^{++}| + |V_j^{-+}|}.$$

Consequently, a node v in V_j^{++} is labelled in G_n at step j with probability

$$\left(1 - \frac{|V_j^{-+}|}{|V_j^{++}| + |V_j^{-+}|}\right)\left(\frac{1}{|V_j^{++}|}\right) = \left(\frac{|V_j^{++}|}{|V_j^{++}| + |V_j^{-+}|}\right)\left(\frac{1}{|V_j^{++}|}\right) = \frac{1}{|V_j^{++}| + |V_j^{-+}|}.$$

Therefore, at every step $i \leq j \leq n$, the node chosen to receive label j and degree d_j is a uniform element of $V_j^{++} \cup V_j^{-+}$, the set of unlocked and unlabelled (that is, open) nodes in G_n at step j. Lastly, the remaining part of the construction process of G_n is equivalent to that of $\mathcal{A}$, and hence $G_n \sim \mathcal{A}$.

We continue with the proof of claim (b). Let $C \in \mathbf{C}_n$ satisfy $|C| = z$. Then the coupling ensures that C is unlocked in both G_n and G_n^- before there is any deviation in the assignment process. Hence, if a node $v \in C$ receives label j and degree d_j in G_n^-, then v will receive the same label in G_n unless v has already been labelled. If v was already labelled in G_n then this label is some $j' < j$. Since $d_1 \geq \cdots \geq d_n$, $d_{j'} \geq d_j$. Therefore, the degree sequence $\mathbf{c}_z^-$ of C in G_n^- is bounded above point-wise by the degree sequence $\mathbf{c}_z$ in G_n. The proof now follows from Lemma 5.3. □

We continue with another modified version of phase 3 of the **ABCD** construction process. This new version will be used to prove the upper bound in Theorem 3.1. Fix z with $s \leq z \leq n^\tau$ and define the construction process $\mathcal{A}^+(z)$, yielding a collection of degrees assigned to a collection of communities notated as G_n^+, as follows.

1. Copy phases 1 and 2 of the **ABCD** construction process to get a degree distribution $\mathbf{d}_n = (d_i, i \in [n])$ and a collection of communities $\mathbf{C}_n = (C_j, j \in [L])$ each containing unassigned nodes.

2. Copy phase 3 of the **ABCD** construction process until the communities of size z are unlocked. This event occurs at step i where i is the smallest label satisfying $d_i \le \frac{z-1}{1-\xi\phi}$. Let n' be the number of locked nodes, i.e., the number of nodes in communities of size at most $z_i - 1$ (recall that $z_i = \lceil d_i(1-\xi\phi)+1 \rceil$). At this point, of the $n - n'$ unlocked nodes, we have assigned $i-1$ of them labels $1, \ldots, i-1$ and corresponding degrees $d_1, \ldots, d_{i-1}$ in some order.

3. Now keep the communities of size at most $z_i - 1$ locked and assign labels $i, \ldots, n-n'$ and corresponding degrees $d_i, \ldots, d_{n-n'}$ uniformly at random to the collection of unlocked and unassigned nodes.

4. Finally, unlock the communities of size at most $z_i - 1$ and assign the n' unassigned nodes labels $n-n'+1, \ldots, n$ and corresponding degrees $d_{n-n'+1}, \ldots, d_n$ in any order (we will later show that w.h.p. $d_{n-n'+1} = \cdots = d_n = \delta$).

Note that, by the end of step 3, all nodes in communities of size z have been assigned a label and a degree. This labelling is all we need to complete the proof, and we include step 4 only for the sake of completeness.

We first show that a community C_j in $G_n^+ \sim \mathcal{A}^+(z)$ with z nodes has the desired degree distribution. Our statement this time is not as strong as Lemma 5.3, though thanks to Lemma 5.1 we can still stochastically bound the degree sequence of a community of size z in G_n^+.

Lemma 5.4. *Let $G_n^+ \sim \mathcal{A}^+(z)$, let C_j be a community in G_n^+ with $|C_j| = z$ and with degree sequence $\mathbf{c}_z^+$, and let $(X_i^+, 1 \le i \le z)$ be the i.i.d. sequence defined in Theorem 3.1. Then w.h.p. $\mathbf{c}_z^+$ is stochastically bounded above by $(X_i^+, 1 \le i \le z)$.*

Proof. As in the proof of Lemma 5.3, we will use the Principle of Deferred Decisions, that is, at the beginning we only uncover some partial information about the degree sequence $\mathbf{d}_n$. As before, we first expose whether or not $d_i > \frac{z-1}{1-\xi\phi}$ and, if the inequality holds, then we expose the value of d_i. However, if $d_i \le \frac{z-1}{1-\xi\phi}$, then we reveal d_i only if $d_i = \delta$, and otherwise we do not expose additional information about d_i.

By the construction of $G_n^+ \sim \mathcal{A}^+(z)$, we know that the sequence of degrees in C_j is a uniform subsequence of $(d_i, \ldots, d_{n-n'})$, where i is the smallest labelled node satisfying $d_i \le \frac{z-1}{1-\xi\phi}$ and n' is the number of nodes in communities of size at most $z_i - 1$. Then, letting V' be as in Lemma 5.1, we have that $n' \le |V'|$ and that w.h.p. by Lemma 5.1, $|V'| < \omega n^{1-(\tau-\zeta)(2-\beta)}$ for any function $\omega = \omega(n) \to \infty$. Thus, w.h.p. $n' = o\left(n^{1-(\tau-\zeta)(2-\beta)/2}\right) = o(\epsilon n)$. (Recall that $\epsilon = n^{-(\tau-\zeta)(2-\beta)/2}$.) Since we aim for a statement that holds w.h.p., we may condition on this event.

Let n'' be the number of nodes of degree δ. Note that n'' is simply a $\mathrm{Binomial(n-i, p_\delta)}$ random variable with

$$p_\delta = \frac{\int_\delta^{\delta+1} x^{-\gamma}\, dx}{\int_\delta^{\Delta_z+1} x^{-\gamma}\, dx},$$

where $\Delta_z = \min\left\{\frac{z-1}{1-\xi\phi}, n^\zeta\right\}$. It follows immediately from Chernoff's bound that w.h.p. we have

$$n'' = (n-i)p_\delta + \omega\sqrt{n} = (n-i)p_\delta + o(\epsilon n) = (n-i)p_\delta(1+o(\epsilon)),$$

the second equality holding since $1-(\tau-\zeta)(2-\beta)/2 > 1/2$. We may condition on this event too.

Let us now summarize our situation. The degree distribution of C_j is a uniform subsequence of length z of the sequence

$$(d_i, \ldots, d_{n-n'}) = (d_i, \ldots, d_{n-n''})^\frown(d_{n-n''+1}, \ldots, d_{n-n'})$$

of $n - n' - i = (n-i)(1-o(\epsilon))$ degrees. ($\mathbf{x}^\frown\mathbf{y}$ is the concatenation of sequences $\mathbf{x}$ and $\mathbf{y}$.) The subsequence $(d_i, \ldots, d_{n-n''})$ consists of degrees that are at least $\delta+1$ and at most Δ_z; recall that, since we have not yet exposed these degrees, by Fact 5.2 they are i.i.d. random variables with distribution $\mathcal{P}(\gamma, \delta+1, \Delta_z)$. On the other hand, $(d_{n-n''+1}, \ldots, d_{n-n'})$ is simply a sequence of $n'' - n' = (n-i)p_\delta(1-o(\epsilon))$ copies of δ.

Now, let us provide a more careful argument to show that a uniform subsequence of $(d_i, \ldots, d_{n-n''})$ of length z satisfies the stochastic domination in the statement of the theorem. We sample z times uniformly at random from this sequence (that may be viewed as a multi-set) without replacement and observe that each time we select δ with probability at least

$$\begin{aligned}\frac{n''-n'-z}{n-n'-i-z} = p_\delta(1-o(\epsilon)) &= \frac{(1-\epsilon-o(\epsilon^2))\int_\delta^{\delta+1} x^{-\gamma}\,dx}{(1-\epsilon)\int_\delta^{\delta+1} x^{-\gamma}\,dx + (1-\epsilon)\int_{\delta+1}^{\Delta_z+1} x^{-\gamma}\,dx} \\ &> \frac{(1-\epsilon)\int_\delta^{\delta+1} x^{-\gamma}\,dx}{(1-\epsilon)\int_\delta^{\delta+1} x^{-\gamma}\,dx + \int_{\delta+1}^{\Delta_z+1} x^{-\gamma}\,dx}\,.\end{aligned}$$

If we select a value other than δ, then our selected degree has distribution $\mathcal{P}(\gamma, \delta+1, \Delta_z)$. Therefore, w.h.p. the random subsequence $\mathbf{c}_z^+$ is stochastically bounded from above by the i.i.d. sequence $(X_i^+, 1 \le i \le z)$ defined in Theorem 3.1, and the proof of the lemma is finished. □

We will now couple the constructions of $G_n \sim \mathcal{A}$ and $G_n^+ \sim \mathcal{A}^+(z)$ and prove the upper bound in Theorem 3.1. Contrast to the proof of the lower bound, we will first construct $G_n \sim \mathcal{A}$ and couple this construction with another construction G_n^+ which we will later show satisfies $G_n^+ \sim \mathcal{A}^+(z)$.

Proof of Theorem 3.1 (upper bound). Construct $G_n \sim \mathcal{A}$ with nodes labelled as $[n]$, degree sequence $\mathbf{d}_n = (d_i, i \in [n])$, and community sequence $\mathbf{C}_n = (C_j, j \in [L])$, and construct G_n^+ in parallel as follows.

1. Let G_n^+ have degree sequence $\mathbf{d}_n$ and community sequence $\mathbf{C}_n$.

2. Copy the degree assignment process of G_n until the communities of size z are unlocked. Let i be the smallest labelled node satisfying $d_i \le \frac{z-1}{1-\xi\phi}$ and let n' be the number of nodes in communities of size at most $z_i - 1$ (recall that $z_i = \lceil d_i(1-\xi\phi)+1\rceil$). Instead of unlocking communities progressively as we do in $G_n \sim \mathcal{A}$, we will keep the n' nodes locked until we have assigned label $n-n'$ and degree $d_{n-n'}$ as we do in $\mathcal{A}^+(z)$.

3. Now, for $j \in \{i, \ldots, n-n'\}$ starting with $j = i$, we first partition the nodes into three sets as follows.

$$\begin{aligned} V_j^{++} &= \left\{v : v \text{ is open in both } G_n \text{ and } G_n^+ \text{ at step } j\right\}\,, \\ V_j^{+-} &= \left\{v : v \text{ is open in } G_n \text{ and closed in } G_n^+ \text{ at step } j\right\}\,, \\ V_j^{--} &= \left\{v : v \text{ is closed in both } G_n \text{ and } G_n^+ \text{ at step } j\right\}\,. \end{aligned}$$

Note, distinct from the lower-bound, that $V_i^{+-} = \emptyset$, and that that there is no set V_j^{-+}. We need not define V_j^{+-}, as we will never encounter a scenario where a node is assigned in G_n but unassigned in G_n^+. We now choose a node v in G_n to receive label j and degree d_j as per the $\mathcal{A}$ construction process (note that v is chosen uniformly at random from $V_j^{++} \cup V_j^{+-}$). We then choose a node in G_n^+ to receive label j and degree d_j as follows.

- If $v \in V_j^{++}$, we give label j and degree d_j to v in G_n^+.
- If $v \in V_j^{+-}$, we give label j and degree d_j to a uniform node in V_j^{++} in G_n^+.

4. Finally, unlock the n' locked nodes in G_n^+ and assign labels $n - n' + 1, \ldots, n$ and degrees $d_{n-n'+1}, \ldots, d_n$ uniformly among these newly unlocked nodes, independent of how these labels and degrees are assigned in G_n.

Similar to the previous coupling, the last step of the coupling is given only for the sake of completeness and has no bearing on the proof. We claim (a) that $G_n^+ \sim \mathcal{A}^+(z)$, and (b) that any community $C \in \mathbf{C}_n$ of size z with degree sequence $\mathbf{c}_z^+$ in G_n^+ and degree sequence $\mathbf{c}_z$ in G_n satisfies $\mathbf{c}_z^+ \geq \mathbf{c}_z$ point-wise.

Starting with claim (a), it is clear by the construction process $\mathcal{A}$ that $\mathbf{d}_n$ and $\mathbf{C}_n$ are valid sequences for $G_n^+ \sim \mathcal{A}^+(z)$. It is also clear that the degree assignment process in G_n^+ for nodes in communities of size at most $z_i - 1$ is valid, since this assignment process is identical to that of $\mathcal{A}$ (which is identical to that of $\mathcal{A}^+(z)$ as well). We must then verify that, for $j \in \{i, \ldots, n - n'\}$, the node in G_n^+ chosen to receive label j and degree d_j is a uniform node from the set of unassigned nodes in communities of size at least z_i. Note that this set of nodes is precisely V_j^{++}. For $u \in V_j^{++}$, u is assigned label j and degree d_j in G_n^+ if u is assigned this label and degree in G_n or if a node $v \in V_j^{+-}$ is assigned this label and degree in G_n and this label and degree is redirected to u in G_n^+. Thus, the probability that $u \in V_j^{++}$ is labelled at step j is

$$\frac{1}{|V_j^{++}| + |V_j^{+-}|} + \left(\frac{|V_j^{+-}|}{|V_j^{++}| + |V_j^{+-}|} \right) \left(\frac{1}{|V_j^{++}|} \right) = \frac{1}{|V_j^{++}|} ,$$

and, in particular, the probability is equal for all $u \in V_j^{++}$. Therefore, at every step $i \leq j \leq n$, the node chosen to receive label j and degree d_j is a uniform element from the set of unlocked and unlabelled nodes in G_n^+ at step j, and this proves claim (a).

We continue with the proof of claim (b). Let $C \in \mathbf{C}_n$ satisfy $|C| = z$. Then the coupling ensures that C is unlocked in both G_n^+ and G_n before there is any deviation in the assignment process. Hence, if a node $v \in C$ receives label j and degree d_j in G_n, then v will receive the same label and degree in G_n^+ unless v has already been given some label $j' < j$ and degree $d_{j'} \geq d_j$ in G_n^+. Therefore, the degree sequence $\mathbf{c}_z$ of C in G_n is bounded above point-wise by the degree sequence $\mathbf{c}_z^+$ in G_n^+. The proof now follows from Lemma 5.4. □

5.2 Volumes of Communities (Proof of Corollary 3.2)

Let $X \sim \mathcal{P}(\gamma, \delta, \Delta)$ and recall that $\mu_\ell(\gamma, \delta, \Delta) = \mathbb{E}\left[X^\ell\right]$. Unfortunately, there is no closed formula for $\mu_\ell(\gamma, \delta, \Delta)$. However, in the coming proofs, we use the following standard technique to bound

$\mu_\ell(\gamma,\delta,\Delta)$ (and other related values) from above and below:

$$\mu_\ell(\gamma,\delta,\Delta) = \sum_{k=\delta}^{\Delta} k^\ell \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx} \le \sum_{k=\delta}^{\Delta} \frac{\int_k^{k+1} x^{\ell-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx} = \frac{\int_\delta^{\Delta+1} x^{\ell-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx}\,, \text{ and}$$

$$\mu_\ell(\gamma,\delta,\Delta) = \sum_{k=\delta}^{\Delta} k^\ell \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx} \ge \sum_{k=\delta}^{\Delta} \left(\frac{k}{k+1}\right)^\ell \frac{\int_k^{k+1} x^{\ell-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx} \ge \left(\frac{\delta}{\delta+1}\right)^\ell \frac{\int_\delta^{\Delta+1} x^{\ell-\gamma}\,dx}{\int_\delta^{\Delta+1} x^{-\gamma}\,dx}\,.$$

Proof of Corollary 3.2. Let $G_n \sim \mathcal{A}$ with degree sequence $\mathbf{d}_n$, let C_j be a community in G_n with $|C_j| = z$, let $\mathbf{c}_j$ be the degree sequence of C_j, and let

$$\Delta_z = \min\left\{\frac{z-1}{1-\xi\phi}, n^\zeta\right\}, \text{ where } \phi = 1 - \frac{1}{n^2}\sum_{j\in[L]} |C_j|^2\,.$$

Now let $(X_i^-, 1 \le i \le z)$ and $(X_i^+, 1 \le i \le z)$ be as in Theorem 3.1. Then, conditional on the stochastic domination in Theorem 3.1,

$$\frac{\mathbb{E}\left[\mathrm{vol}(C_j)\right]}{z} \ge \frac{1}{z}\,\mathbb{E}\left[\sum_{i=1}^{z} X_i^-\right] = \mu_1(\gamma,\delta,\Delta_z)\,,$$

and

$$\frac{\mathbb{E}\left[\mathrm{vol}(C_j)\right]}{z} \le \frac{1}{z}\,\mathbb{E}\left[\sum_{i=1}^{z} X_i^+\right] = (1+o(1))\frac{1}{z}\,\mathbb{E}\left[\sum_{i=1}^{z} X_i^-\right] = (1+o(1))\mu_1(\gamma,\delta,\Delta_z)\,,$$

which establishes the first claim in Corollary 3.2. Next, we have

$$\begin{aligned}
\mu_1(\gamma,\delta,n^\zeta) - \mu_1(\gamma,\delta,\Delta_z) &= \left(\sum_{k=\delta}^{n^\zeta} k \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{n^\zeta+1} x^{-\gamma}\,dx} - \sum_{k=\delta}^{\Delta_z} k \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{\Delta_z+1} x^{-\gamma}\,dx}\right) \\
&= \left(1 + O(\Delta_z^{1-\gamma})\right)\left(\sum_{k=\delta}^{n^\zeta} k \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{n^\zeta+1} x^{-\gamma}\,dx} - \sum_{k=\delta}^{\Delta_z} k \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{n^\zeta+1} x^{-\gamma}\,dx}\right) \\
&= \left(1 + O(\Delta_z^{1-\gamma})\right) \sum_{k=\Delta_z+1}^{n^\zeta} k \frac{\int_k^{k+1} x^{-\gamma}\,dx}{\int_\delta^{n^\zeta+1} x^{-\gamma}\,dx} \\
&\le \left(1 + O(\Delta_z^{1-\gamma})\right) \frac{\int_{\Delta_z+1}^{n^\zeta+1} x^{1-\gamma}\,dx}{\int_\delta^{n^\zeta+1} x^{-\gamma}\,dx} \\
&= O(\Delta_z^{2-\gamma})\,.
\end{aligned}$$

The second claim in Corollary 3.2 now follows since

$$\begin{aligned}
\frac{\mathbb{E}\left[\mathrm{vol}(C_j)\right]}{z} &= (1+o(1))\Big(\mu_1(\gamma,\delta,n^\zeta) - \Big(\mu_1(\gamma,\delta,n^\zeta) - \mu_1(\gamma,\delta,\Delta_z)\Big)\Big) \\
&= (1+o(1))\Big(\mu_1(\gamma,\delta,n^\zeta) - O(\Delta_z^{2-\gamma})\Big)
\end{aligned}$$

and, since $\Delta_z = \Theta(\min\{z, n^\zeta\})$, we have that $\Delta_z \to \infty$ as $z \to \infty$. □

5.3 Loops and Multi-edges (Proof of Theorem 3.3)

Throughout this section, it will be useful to refer to the multi-graph generated by the first four phases of the **ABCD** construction. Write $G_n \sim \mathcal{A}^{(4)}$ to mean G_n is the hypergraph generated by the first four phases.

Before tackling the upper-bounds in Theorem 3.3, we first prove that the number of self-loops and multi-edges in $G_n \sim \mathcal{A}^{(4)}$ is asymptotically bounded from below by the number of communities. In fact, we show that the number of self-loops in community graphs alone is asymptotically bounded in this way.

Lemma 5.5. *Let $G_n \sim \mathcal{A}^{(4)}$ with L communities and let S_c be the number of self-loops in community graphs in G_n. Then w.h.p.*

$$S_c = \Omega(L)\,.$$

Proof. Fix a constant z large enough so that $z \geq s$ and $\lfloor (1-\xi)\Delta_z \rfloor \geq 2$ and let $G_{n,j}$ be a community graph in G_n with $|C_j| = z$ and with degree sequence $(Y_i, i \in C_j)$ (recall that $Y_i = \lfloor (1-\xi)d_i \rceil$ where $\lfloor \cdot \rceil$ is a random rounding function). Then, by the lower bound in Theorem 3.1, a uniformly random degree Y_i is stochastically bounded from below by $\lfloor (1-\xi)X \rfloor$ where $X \sim \mathcal{P}(\gamma, \delta, \Delta_z)$. Thus, by the stochastic bound, we have

$$\mathbb{P}\left(Y_i = \lfloor (1-\xi)\Delta_z \rfloor\right) \geq \mathbb{P}\left(X = \Delta_z\right) > 0\,.$$

Thus, w.h.p. a linear proportion of community graphs with z nodes contain at least one node v with $\deg(v) = \lfloor (1-\xi)\Delta_z \rfloor \geq 2$. Furthermore, a node with this degree generates a loop in $G_n \sim \mathcal{A}^{(4)}$ with positive probability, and so w.h.p. a linear proportion of community graphs with z nodes contain at least one loop. Finally, as the number of communities of size z is w.h.p. $\Theta(L)$, the lemma follows. □

We continue now with the upper-bounds. The heart of Theorem 3.3 is the following lemma.

Lemma 5.6. *Fix $z > \Delta > \delta > 0$ and $\gamma \in (2,3)$. Let $\mathbf{q}_z = (q_i, i \in [z])$ be a sequence of i.i.d. random variables with $q_i \sim \mathcal{P}(\gamma, \delta, \Delta)$ and let H_z be sampled as the configuration model with degree sequence $\mathbf{q}_z$. Let S and M be the number of self-loops and, respectively, multi-edges in H_z. Then*

$$\begin{aligned}\mathbb{E}[S] &\leq (1 + O(\Delta^{\gamma-3}))\, c(\gamma,\delta)\Delta^{3-\gamma}\,, \text{ and}\\ \mathbb{E}[M] &\leq (1 + O(\Delta^{\gamma-3}))\, c(\gamma,\delta)^2\Delta^{6-2\gamma}\,,\end{aligned}$$

where

$$c(\gamma,\delta) = \frac{(\gamma-1)\delta^{\gamma-2}}{2(3-\gamma)}\,.$$

Proof. We begin with known bounds for S and M. We have

$$\mathbb{E}[S \mid \mathbf{q}_z] = \frac{\sum_{i\in[z]} q_i(q_i-1)}{2\left(\sum_{i=1}^{z} q_i - 1\right)} \leq \frac{1}{2}\frac{\sum_{i\in[z]} q_i^2}{\sum_{i=1}^{z} q_i - 1}\,, \tag{2}$$

and

$$\mathbb{E}[M \mid \mathbf{q}_z] \leq \frac{\sum_{1\leq i<j\leq z} q_i(q_i-1)q_j(q_j-1)}{2\left(\sum_{i=1}^{z} q_i - 1\right)\left(\sum_{i=1}^{z} q_i - 3\right)} \leq \frac{1}{2}\frac{\sum_{1\leq i<j\leq z} q_i^2 q_j^2}{\left(\sum_{i=1}^{z} q_i - 3\right)^2}\,. \tag{3}$$

See Chapter 7 in [28] for a detailed study on the number of self-loops and multi-edges in the configuration model. In particular, the equality in (2) and the first inequality in (3) come from respective equations (7.3.21) and (7.3.26) in [28].

For independent $X, Y \sim \mathcal{P}(\gamma, \delta, \Delta)$ we have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{k=\delta}^{\Delta} \frac{k^2 \int_k^{k+1} x^{-\gamma}\, dx}{\int_\delta^{\Delta+1} x^{-\gamma}\, dx} \\
&\leq \frac{\int_\delta^{\Delta+1} x^{2-\gamma}\, dx}{\int_\delta^{\Delta+1} x^{-\gamma}\, dx} \\
&= \left(\frac{\gamma-1}{3-\gamma}\right)\left(\frac{(\Delta+1)^{3-\gamma} - \delta^{3-\gamma}}{\delta^{1-\gamma} - (\Delta+1)^{1-\gamma}}\right) \\
&= (1 + O(\Delta^{\gamma-3} + \Delta^{1-\gamma}))\left(\frac{\gamma-1}{3-\gamma}\right)\delta^{\gamma-1}\Delta^{3-\gamma} \\
&= (1 + O(\Delta^{\gamma-3}))\left(\frac{\gamma-1}{3-\gamma}\right)\delta^{\gamma-1}\Delta^{3-\gamma}\,,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[X^2Y^2\right] &= \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] \\
&= (1 + O(\Delta^{\gamma-3}))\left(\frac{\gamma-1}{3-\gamma}\right)^2 \delta^{2\gamma-2}\Delta^{6-2\gamma}\,.
\end{aligned}$$

Now, since $\mathbf{q}_z$ contains i.i.d. random variables, and since $\sum_{i=1}^{z} q_i \geq \delta z$, it follows from (2) that

$$\begin{aligned}
\mathbb{E}\left[S\right] &= \mathbb{E}\left[\mathbb{E}\left[S \mid \mathbf{q}_z\right]\right] \\
&\leq \frac{1}{2}\,\mathbb{E}\left[\frac{\sum_{i\in[z]} q_i^2}{\sum_{i\in[z]} q_i - 1}\right] \\
&\leq \frac{1}{2(\delta z - 1)} \sum_{i\in[z]} \mathbb{E}\left[q_i^2\right] \\
&\leq (1 + O(\Delta^{\gamma-3}))\left(\frac{1}{2\delta z}\right)\left(z\left(\frac{\gamma-1}{3-\gamma}\right)\delta^{\gamma-1}\Delta^{3-\gamma}\right) \\
&= (1 + O(\Delta^{\gamma-3}))\left(\frac{(\gamma-1)\delta^{\gamma-2}}{2(3-\gamma)}\right)\Delta^{3-\gamma}\,,
\end{aligned}$$

and from (3) that

$$
\begin{aligned}
\mathbb{E}[M] &= \mathbb{E}[\mathbb{E}[M \mid \mathbf{q}_z]] \\
&\le \frac{1}{2}\mathbb{E}\left[\frac{\sum_{1\le i<j\le z} q_i^2 q_j^2}{\left(\sum_{i=1}^z q_i - 3\right)^2}\right] \\
&\le \frac{1}{2(\delta z-3)^2}\sum_{1\le i<j\le z}\mathbb{E}\left[q_i^2 q_j^2\right] \\
&\le (1+O(\Delta^{\gamma-3}))\left(\frac{1}{2\delta^2 z^2}\right)\binom{z}{2}\left(\frac{\gamma-1}{3-\gamma}\right)^2 \delta^{2\gamma-2}\Delta^{6-2\gamma} \\
&\le (1+O(\Delta^{\gamma-3}))\left(\frac{(\gamma-1)\delta^{\gamma-2}}{2(3-\gamma)}\right)^2 \Delta^{6-2\gamma}\,.
\end{aligned}
$$

Note that, in the first computation, we use the fact that

$$
\frac{1}{2(\delta z-1)} = (1+O(z^{-1}))\frac{1}{2\delta z} = (1+O(\Delta_z^{\gamma-3}))\frac{1}{2\delta z}\,,
$$

and in the second computation, we use the fact that

$$
\frac{1}{2(\delta z-3)^2} = (1+O(z^{-1}))\frac{1}{2\delta^2 z^2} = (1+O(\Delta_z^{\gamma-3}))\frac{1}{2\delta^2 z^2}\,.
$$

This finishes the proof of the lemma. □

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. Let $G_n \sim \mathcal{A}^{(4)}$ with degree sequence $\mathbf{d}_n = (d_i, i\in[n])$, and let S_c, M_c, S_b, M_b and M_{bc} be as in the statement of the theorem. Starting with S_b and M_b, note that the degree sequence in $G_{n,0}$ is $(Z_i, i\in[n])$ where $Z_i = \lfloor \xi d_i \rceil$. Thus, $Z_i \le d_i$, meaning by Lemma 5.6 that

$$
\begin{aligned}
\mathbb{E}[S_b] &\le \left(1+O(n^{\zeta(\gamma-3)})\right)c(\gamma,\delta)\left(n^{\zeta}\right)^{3-\gamma} = O(n^{\zeta(3-\gamma)})\,,\ \text{and} \\
\mathbb{E}[M_b] &\le \left(1+O(n^{\zeta(\gamma-3)})\right)c(\gamma,\delta)^2\left(n^{\zeta}\right)^{6-2\gamma} = O(n^{\zeta(6-2\gamma)})\,,
\end{aligned}
$$

proving claims 3. and 4.

Continuing with S_c and M_c, for community graph $G_{n,j}$ with $|C_j| = z$ let $S_{c,j}$ and $M_{c,j}$ be the number of self-loops and multi-edges in $G_{n,j}$. Note that, for any node $i\in C_j$, the degree of i in $G_{n,j}$ is $Y_i \le d_i$. Thus, by Theorem 3.1, Y_i is stochastically bounded from above by the random variable $Y \sim \mathcal{P}(\gamma, \delta+1, \Delta_z)$. Then, again by Lemma 5.6, we have that

$$
\begin{aligned}
\mathbb{E}[S_{c,j} \mid |C_j| = z] &\le \left(1+O\left(\Delta_z^{\gamma-3}\right)\right)c(\gamma,\delta+1)\Delta_z^{3-\gamma}\,,\ \text{and} \\
\mathbb{E}[M_{c,j} \mid |C_j| = z] &\le \left(1+O(\Delta_z^{\gamma-3})\right)c(\gamma,\delta+1)^2\Delta_z^{6-2\gamma}\,.
\end{aligned}
$$

For the remainder of the proof, we write $c = c(\gamma,\delta+1)$ to simplify notation. Recall from phase 2 of the construction process of G_n that $|C_j| \sim \mathcal{P}(\beta, s, n^{\tau})$. Therefore,

$$
\begin{aligned}
\mathbb{E}[S_{c,j}] &= \sum_{z=s}^{n^{\tau}} \mathbb{E}[S_{c,j} \mid |C_j| = z]\,\mathbb{P}(|C_j| = z) \\
&\le \sum_{z=s}^{n^{\tau}} \left(1+O\left(\Delta_z^{\gamma-3}\right)\right) c\,\Delta_z^{3-\gamma}\frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^{\tau}+1} y^{-\beta}\,dy}\,.
\end{aligned}
$$

We split the sum at the community size z^*, where z^* is minimal with the property that

$$\left\lfloor \frac{z^*-1}{1-\xi\phi} \right\rfloor \geq n^{\zeta}\,.$$

Note that, for $z \leq z^*$, $\Delta_z = \Theta(z)$, and for $z \geq z^*$, $\Delta_z = n^{\zeta}$. Let c' be a constant satisfying $\Delta_z^{3-\gamma} \leq c' z^{3-\gamma}$ for all $s \leq z \leq z^*$. For the first part of the sum, we have

$$\begin{aligned}
&\sum_{z=s}^{z^*} \left(1+O\left(\Delta_z^{\gamma-3}\right)\right) c\,\Delta_z^{3-\gamma} \frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
\geq\ &\sum_{z=s}^{z^*} \left(1+O\left(z^{\gamma-3}\right)\right) cc' z^{3-\gamma} \frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
\leq\ &\left(1+O\left(s^{\gamma-3}\right)\right) cc' \sum_{z=s}^{z^*} \frac{\int_z^{z+1} y^{3-\gamma-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\left(1+O\left(s^{\gamma-3}\right)\right) cc' \frac{\int_s^{z^*+1} y^{3-\gamma-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\left(1+O\left(s^{\gamma-3}\right)\right) cc' \left(\beta-1\right) s^{1-\beta} \left(\frac{(z^*+1)^{4-\gamma-\beta} - s^{4-\gamma-\beta}}{4-\gamma-\beta} \right) \\
&= O\left(1+(z^*)^{4-\gamma-\beta}\right) \\
&= O\left(1+n^{\zeta(4-\gamma-\beta)}\right)\,.
\end{aligned}$$

For the second part of the sum, we have

$$\begin{aligned}
&\sum_{z=z^*+1}^{n^\tau} \left(1+O\left(\Delta_z^{\gamma-3}\right)\right) c\,\Delta_z^{3-\gamma} \frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\sum_{z=z^*+1}^{n^\tau} \left(1+O\left(n^{\zeta(\gamma-3)}\right)\right) c\,n^{\zeta(3-\gamma)} \frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\left(1+O\left(n^{\zeta(\gamma-3)}\right)\right) cn^{\zeta(3-\gamma)} \sum_{z=z^*+1}^{n^\tau} \frac{\int_z^{z+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\left(1+O\left(n^{\zeta(\gamma-3)}\right)\right) cn^{\zeta(3-\gamma)} \frac{\int_{z^*+1}^{n^\tau+1} y^{-\beta}\,dy}{\int_s^{n^\tau+1} y^{-\beta}\,dy} \\
=\ &\left(1+O\left(n^{\zeta(\gamma-3)}\right)\right) cn^{\zeta(3-\gamma)} \frac{(z^*+1)^{1-\beta} - (n^\tau+1)^{1-\beta}}{s^{1-\beta} - (n^\tau+1)^{1-\beta}} \\
&= O\left(n^{\zeta(3-\gamma)} (z^*)^{1-\beta}\right) \\
&= O\left(n^{\zeta(3-\gamma)} n^{\zeta(1-\beta)}\right) \\
&= O\left(n^{\zeta(4-\gamma-\beta)}\right)\,,
\end{aligned}$$

and thus, $\mathbb{E}[S_{c,j}] = O(1 + n^{\zeta(4-\gamma-\beta)})$. An analogous calculation shows that $\mathbb{E}[M_{c,j}] = O(1 + n^{\zeta(7-2\gamma-\beta)})$. Lastly, we handle the fact that an extra half-edge can be added to a community during Phase 4 of the construction process. This increase can add at most one self-loop or multi-edge, meaning the statement $\mathbb{E}[M_{c,j}] = O(1 + n^{\zeta(7-2\gamma-\beta)})$ remains true after accounting for the potential extra half-edges. Claims 1. and 2. now follow from linearity of expectation, along with the fact that w.e.p. the number of communities in G_n is $\Theta(n^{1-\tau(2-\beta)})$.

Claim 5. states that $\mathbb{E}[M_{bc}] = o(M_c)$. To see this, let C_j be a community in G_n and let $u, v \in C_j$. Now let $M_c(u,v)$ be the number of $\{u,v\}$ multi-edge pairs in $G_{n,j}$, and let M_{bc} be the number of $\{u,v\}$ multi-edge pairs with one edge in $G_{n,j}$ and the other in $G_{n,0}$. Then

$$\mathbb{E}[M_c(u,v) \mid \mathbf{d}_n] = \Theta\left(\frac{d_u^2 d_v^2}{\left(\sum_{i\in C_j} d_i\right)^2}\right),$$

whereas

$$\mathbb{E}[M_{bc}(u,v) \mid \mathbf{d}_n] = \Theta\left(\frac{d_u^2 d_v^2}{\left(\sum_{i\in C_j} d_i\right)\left(\sum_{i\in [n]} d_i\right)}\right).$$

Since $\sum_{i\in C_j} d_i = o\left(\sum_{i\in[n]} d_i\right)$ for all communities C_j, we get that $\mathbb{E}[M_{bc}(u,v)] = o(\mathbb{E}[M_c(u,v)])$, and Claim 5. follows from linearity of expectation.

Finally, we know that w.e.p. $L = \Theta(n^{1-\tau(2-\beta)})$ and that $\mathbb{E}[S_c] = \Omega(L)$. Now suppose that $\gamma + \beta > 4$ and that $2\zeta(3-\gamma) + \tau(2-\beta) \le 1$, and note that these two inequalities imply that $2\gamma + \beta > 3 + \gamma + \beta > 7$ and that $3 - \gamma + \tau(2-\beta) \le 1$. Therefore, under this assumption, and conditioned on the stochastic domination, we have

$$\begin{aligned}
\mathbb{E}[S_c] &= O\left((n^{1-\tau(2-\beta)})(1 + n^{\zeta(4-\gamma-\beta)})\right) = O\left(n^{1-\tau(2-\beta)}\right), \\
\mathbb{E}[M_c + M_{bc}] &= (1+o(1))\mathbb{E}[M_c] = O\left((n^{1-\tau(2-\beta)})(1 + n^{\zeta(7-2\gamma-\beta)})\right) = O\left(n^{1-\tau(2-\beta)}\right), \\
\mathbb{E}[S_b] &= O\left(n^{\zeta(3-\gamma)}\right) = O\left(n^{1-\tau(2-\beta)}\right), \text{ and} \\
\mathbb{E}[M_b] &= O\left(n^{2(\zeta(3-\gamma))}\right) = O\left(n^{1-\tau(2-\beta)}\right),
\end{aligned}$$

which proves the final claim. □

6 Conclusion

Let us finish the paper with some open problems. We have shown two examples of how Theorem 3.1 can help us understand the nature of **ABCD** graphs. There are more applications of Theorem 3.1 that we do not explore here. Essentially, any result that holds for a configuration model on an i.i.d. degree sequence, sampled as $\mathcal{P}(\gamma, \delta, \Delta)$ for some $\gamma \in (2,3)$, should hold for a community graph in $G_n \sim \mathcal{A}$ modulo some discrepancy involving the rewiring phase of the **ABCD** construction. With additional work, it may also be true that such results hold for a community graph in $G_n \sim \mathcal{A}$. Possible avenues for $G_n \sim \mathcal{A}$ include studying its diameter, its diffusion rate, its clustering coefficient, etc.

Our results in Corollary 3.2 and Theorem 3.3 are results only in expectation, though our experiments indicate that the behaviour of at least community volumes is quite tight. Given that the truncated power-law $\mathcal{P}\left(\gamma, \delta, n^{\zeta}\right)$ has unbounded second moment, and that $\mathcal{P}\left(\beta, s, n^{\tau}\right)$ has unbounded first moment, any study involving concentration will prove to be challenging. However, considering that the collection of community degree sequences partition the degree sequence of the whole graph, it is possible that these sequences exhibit self-correcting behaviour, and this is a potential road-map to a tighter version of our results.

In Theorem 3.3 we only show that collisions are bounded below asymptotically by $\Omega(L)$. On the other hand, our experimental results suggest that the number of collisions is, in fact, $\omega(L)$ when $\gamma + \beta \leq 4$ or when $2\zeta(3-\gamma) + \tau(2-\beta) > 1$. Thus, there is potential room to improve Theorem 3.3 by tightening the lower-bound.

Acknowledgements

This research is supported in part by the Tutte Institute for Mathematics and Computing (TIMC), a division of the Communications Security Establishment (CSE). TIMC does not specifically endorse the contents of this work, or any other work by the authors. Any opinions or positions represented herein do not represent the official position of CSE or the Government of Canada.

References

[1] Samin Aref, Hriday Chheda, and Mahdi Mostajabdaveh. The bayan algorithm: Detecting communities in networks through exact and approximate optimization of modularity. *arXiv preprint arXiv:2209.04562*, 2022.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] Jordan Barrett, Bogumił Kamiński, Paweł Prałat, and François Théberge. Self-similarity of communities of the abcd model. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 17–31. Springer, 2024.

[4] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[5] Neli Blagus, Lovro Šubelj, and Marko Bajec. Self-similar scaling of density in complex real-world networks. *Physica A: Statistical Mechanics and its Applications*, 391(8):2794–2802, 2012.

[6] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[7] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

[8] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[9] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[10] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41, 2007.

[11] Lazaros K Gallos, Chaoming Song, and Hernán A Makse. A review of fractality and self-similarity in complex networks. *Physica A: Statistical Mechanics and its Applications*, 386(2):686–691, 2007.

[12] Roger Guimera, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. Self-similar community structure in a network of human interactions. *Physical review E*, 68(6):065103, 2003.

[13] Petter Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature communications*, 10(1):1016, 2019.

[14] Svante Janson. Random graphs with given vertex degrees and switchings. *Random Structures & Algorithms*, 57(1):3–31, 2020.

[15] Bogumił Kamiński, Bartosz Pankratz, Paweł Prałat, and François Théberge. Modularity of the abcd random graph model with community structure. *Journal of Complex Networks*, 10(6):cnac050, 2022.

[16] Bogumił Kamiński, Paweł Prałat, and François Théberge. Artificial benchmark for community detection (abcd)—fast random graph model with community structure. *Network Science*, pages 1–26, 2021.

[17] Bogumił Kamiński, Paweł Prałat, and François Théberge. Mining complex networks. 2021.

[18] Bogumił Kamiński, Paweł Prałat, and François Théberge. Artificial benchmark for community detection with outliers (abcd+ o). *Applied Network Science*, 8(1):25, 2023.

[19] Bogumił Kamiński, Paweł Prałat, and François Théberge. Hypergraph artificial benchmark for community detection (h–abcd). *Journal of Complex Networks*, 11(4):cnad028, 2023.

[20] Bogumił Kamiński, Tomasz Olczak, Bartosz Pankratz, Paweł Prałat, and François Théberge. Properties and performance of the abcde random graph model with community structure. *Big Data Research*, 30:100348, 2022.

[21] JS Kim, K-I Goh, G Salvi, E Oh, B Kahng, and D Kim. Fractality in complex networks: Critical and supercritical skeletons. *Physical Review E*, 75(1):016110, 2007.

[22] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, 2009.

[23] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[24] Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

[25] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.* Cambridge university press, 2017.

[26] M Ángeles Serrano, Dmitri Krioukov, and Marián Boguná. Self-similarity of complex networks and hidden metric spaces. *Physical review letters*, 100(7):078701, 2008.

[27] Chaoming Song, Shlomo Havlin, and Hernan A Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.

[28] Remco Van Der Hofstad. *Random graphs and complex networks*, volume 43. Cambridge university press, 2016.

[29] Nicholas C Wormald. Generating random regular graphs. *Journal of algorithms*, 5(2):247–280, 1984.

[30] Nicholas C Wormald et al. Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298, 1999.

[31] Nicholas Charles Wormald. Random graphs and asymptotics. In J.L. Gross and J. Yellen, editors, *Handbook of Graph Theory (Section 8.2)*, pages 817–836. CRC Press, 2004.

[32] Bin Zhou, Xiangyi Meng, and H Eugene Stanley. Power-law distribution of degree–degree distance: A better representation of the scale-free property of complex networks. *Proceedings of the national academy of sciences*, 117(26):14812–14818, 2020.