


## RANDOM INTERSECTION GRAPHS WITH COMMUNITIES

REMCO VAN DER HOFSTAD <sup>\*</sup>, *Eindhoven University of Technology*  
JÚLIA KOMJÁTHY <sup>\*\*</sup>, *Delft University of Technology*  
VIKTÓRIA VADON,<sup>\*\*\*</sup> *University of Miskolc*

### Abstract

Random intersection graphs model networks with communities, assuming an underlying bipartite structure of communities and individuals, where these communities may overlap. We generalize the model, allowing for arbitrary community structures within the communities. In our new model, communities may overlap, and they have their own internal structure described by arbitrary finite community graphs. Our model turns out to be tractable. We analyze the overlapping structure of the communities, show local weak convergence (including convergence of subgraph counts), and derive the asymptotic degree distribution and the local clustering coefficient.

*Keywords:* Random networks; community structure; overlapping communities; random intersection graphs; local weak convergence

2010 Mathematics Subject Classification: Primary 60C05  
Secondary 05C80; 90B15

### 1. Introduction

Communities are local structures that are more densely connected than the network average. They are present in numerous real-life networks [25], such as the internet, collaboration networks, and social networks, and they offer a possible explanation for the frequently observed high clustering (transitivity) [40, Chapters 7.9, 11].

There are several possible reasons why communities arise, e.g. an underlying geometry or properties shared by the vertices. We focus on networks with an underlying structure of individuals and communities that they are part of. While our terminology and examples are mainly taken from social networks, the model is applicable to any network that builds on some kind of community structure. Such structures exist in many real-life networks [27, 28], the most evident example being collaboration networks, like the Internet Movie Database (IMDb) or the ArXiv. In these examples, the ‘individuals’ are the actors and actresses or the authors, and the ‘communities’ are the movies or articles they collaborate on. We can also consider a social network based on communities, where ‘communities’ can represent families, common interests, workplaces, or cities.

---

Received 12 November 2019; revision received 15 February 2021.

<sup>\*</sup> Postal address: Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, The Netherlands. Email address: [r.w.v.d.hofstad@tue.nl](mailto:r.w.v.d.hofstad@tue.nl)

<sup>\*\*</sup> Postal address: Department of Applied Mathematics, Delft University of Technology, Postbus 5, 2600AA, Delft, The Netherlands. Email address: [J.Komjathy@tudelft.nl](mailto:J.Komjathy@tudelft.nl)

<sup>\*\*\*</sup> Postal address: Institute of Mathematics, University of Miskolc, Egyetem ut 1, 3515, Miskolc, Hungary. Email address: [viktoria.vadon@uni-miskolc.hu](mailto:viktoria.vadon@uni-miskolc.hu)

© The Author(s) 2021. Published by Cambridge University Press on behalf of Applied Probability Trust.

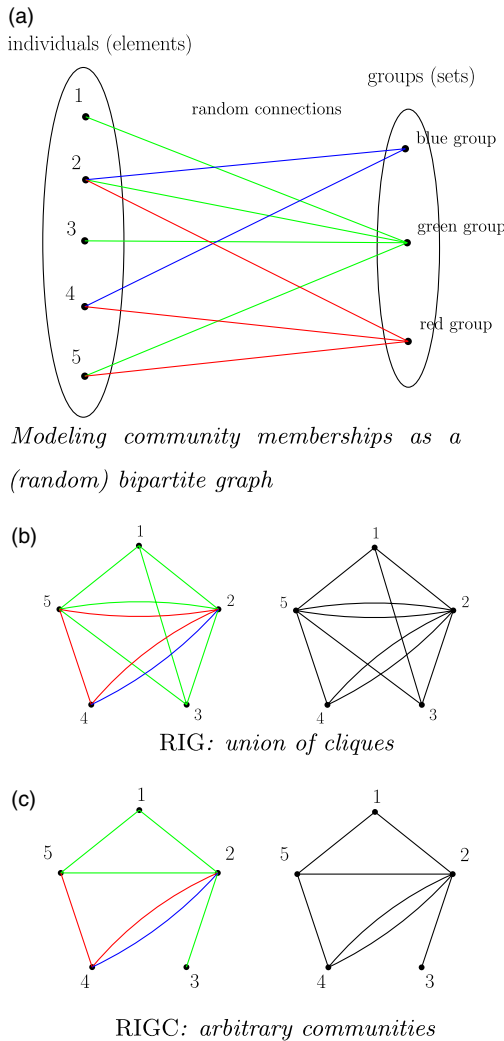


FIGURE 1. Two models for overlapping communities: RIG and RIGC.

Because of the complexity of real-world networks, they are often modeled using *random graphs* [14, 20, 35]. Properties and processes of interest, e.g. distances, clustering, network evolution, and information or epidemic spreading processes, are studied on the random graph models to predict their behavior on real-life networks. An underlying community structure such as the ones mentioned above is modeled using bipartite graphs, where the two partitions correspond to the individuals (people) and the communities (or attributes), and an edge represents a community membership; see Fig. 1a. The historical random graph model for networks with community structure is the random intersection graph (RIG) first introduced in [44]. Over the years, several ways have been introduced to generate the (random) bipartite graph of community memberships [12], ranging from independent percolation on the complete bipartite graph (binomial RIG [21, 37, 44] or inhomogeneous RIG [11, 18]), through pre-assigning the number of community memberships to each individual and connecting them to uniformly chosen

communities (uniform RIG [7, 43] or generalized RIG [8, 9, 10, 13, 26]), to pre-assigning the number of community memberships to each individual as well as the number of community members to each community, then matching these ‘tokens’ uniformly (i.e., the community memberships are generated via the bipartite configuration model) [17, 39]. What all of these models have in common is that once the community memberships are generated, every two individuals that share a community are connected. As a result, communities do overlap, while each community is a complete graph (see Fig. 1b), which may not be a realistic assumption for large communities.

One easy and natural way to go about this is thinning communities [36, 39]; however, this may not give the full generality we desire. The recently introduced hierarchical configuration model [33, 34], which extends the household model [2, 3], offers an alternative approach, using arbitrary communities as building blocks with random connections *between* the communities, resulting in non-overlapping communities. In this paper, we aim to bridge the gap: we introduce a new random graph model, the random intersection graph with community structure (RIGC), which accommodates arbitrary, yet at the same time overlapping, communities, as long as these communities are connected; see Fig. 1c.

The RIGC model is flexible in terms of the choice of parameters, ranging from independent and identically distributed (i.i.d.) random variables to data taken from real-life networks; see Section 2.4 for a brief discussion. The model also turns out to be analytically tractable. In this paper, we keep our assumptions as general as possible, and present results on the overlapping structure and local properties of the model (including local weak convergence, degree structure, and non-trivial clustering). Its global properties, including the existence and quantification of the so-called giant component (a unique linear-sized connected component), and percolation on the RIGC model are studied in the companion paper [32]. In [45], we introduce the model to a more applied audience and state all results informally and without proof. The proofs in this paper rely on the connection to the bipartite configuration model that generates the community memberships. The matching results that we present on the bipartite configuration model are hence both instrumental to the RIGC and of independent interest.

**Outline of the paper.** The rest of this paper is organized as follows. In s:RIGC Section 2, we introduce the random intersection graph with community structure (RIGC), state our results, and provide a brief discussion. We provide the proofs of our results in Section 3.

**Notational conventions.** We will consider a sequence of graphs, and consequently a sequence of input parameters, both indexed by  $n \in \mathbb{N}$ . We note that  $n$  only serves as the index; it does not necessarily mean the size or any other parameter of the graph, which allows for the study of more general (growing) graph sequences. We often omit the dependence on  $n$  to keep the notation light, as long as it does not cause confusion. Throughout this paper, we distinguish the set of positive integers as  $\mathbb{Z}^+ := \{1, 2, 3, \dots\}$  and the set of non-negative integers as  $\mathbb{N} = \{0, 1, 2, \dots\}$ . The notions  $\xrightarrow{\mathbb{P}}$  and  $\xrightarrow{d}$  stand for convergence in probability and convergence in distribution (weak convergence), respectively. We write  $X \stackrel{d}{=} Y$  to mean that the random variables  $X$  and  $Y$  have the same distribution. For an  $\mathbb{N}$ -valued random variable  $X$  such that  $\mathbb{E}[X] < \infty$ , we define its *size-biased* distribution  $X^*$  and the transform  $\tilde{X}$  with the following probability mass functions (PMFs): for all  $k \in \mathbb{N}$ ,

$$\mathbb{P}(X^* = k) = k \mathbb{P}(X = k) / \mathbb{E}[X], \quad \mathbb{P}(\tilde{X} = k) = \mathbb{P}(X^* - 1 = k). \tag{1.1}$$

We say that a sequence of events  $(A_n)_{n \in \mathbb{N}}$  occurs with high probability (w.h.p.) if  $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 1$ . For two (possibly) random sequences  $(X_n)_{n \in \mathbb{N}}$  and  $(Y_n)_{n \in \mathbb{N}}$ , we say that  $X_n = o_{\mathbb{P}}(Y_n)$  if  $X_n/Y_n \xrightarrow{\mathbb{P}} 0$  as  $n \rightarrow \infty$ . We write  $[n] := \{1, 2, \dots, n\}$  and denote the indicator of an event  $A$  by  $\mathbb{1}_A$ . For a graph  $G$ , we denote its vertex set by  $\mathcal{V}(G)$ , its size by  $|G| = |\mathcal{V}(G)|$ , and its edge set by  $\mathcal{E}(G)$ .

## 2. Model and results

In this section, we give a formal definition of the RIGC model and present our results on its local properties, as well as providing a discussion on its applicability.

### 2.1. Definition of the random intersection graph with communities

First, we give a short, intuitive description of the *random intersection graph with communities* (RIGC), followed by a detailed, formal construction. After the parameters are introduced, the construction happens in two steps. First, we construct the *community structure*: an underlying bipartite graph that represents the community memberships, from which all the randomness arises. Then we explain how to derive the RIGC based on the given community structure.

**Intuitive model description.** The aim of the model is to create a network that uses given (arbitrary but connected) community graphs as its building blocks, but at the same time allows them to overlap. We achieve this by thinking of vertices in the community graphs as *community roles* that may be taken by the individuals. We represent the community roles as the right-hand side of an underlying bipartite graph: the set of vertices on the right-hand side of the underlying bipartite graph is the disjoint union of the vertices of a set of communities, labeled by their community number and role within the community. The individuals then are represented as a distinct set of vertices, forming the left-hand side of the bipartite graph, and we allow them to take on (possibly several) community roles by assigning them membership tokens. Each membership token corresponds to one community role taken, and we match membership tokens with community roles one-to-one, uniformly at random (u.a.r.). That is, each role in each community is assumed by a unique individual. The total number of membership tokens given out to individuals has to equal the total number of community roles for this matching to be possible. Finally, to obtain the RIGCs, we identify each individual with all the community roles it takes, ‘gluing’ together the community graphs, which introduces overlaps and creates the (much more interconnected) network.

**Parameters.** Intuitively, we think of the individuals being placed on the left-hand side and the communities on the right-hand side, and consequently we sometimes refer to them as  $\ell$ -vertices and  $r$ -vertices, respectively. We denote the set of individuals by  $\mathcal{V}^\ell = [N_n]$ , where the number of individuals  $N_n$  satisfies  $N_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Similarly, we denote the set of communities by  $\mathcal{V}^r = [M_n]$ , where  $M_n \rightarrow \infty$  is to be defined later.

In this paper, we will encounter three relevant types of degrees, as we work with three types of graphs: the RIGC model itself, the bipartite graph used to generate its community memberships, and the community graphs we use as building blocks. The term ‘degree’ is reserved for the most natural concept, namely, the number of connections of the individual in the resulting RIGC; we sometimes refer to this notion of degree as ‘projected degree’ ( $p$ -degree) for clarity. On the level of the underlying bipartite graph, the role of ‘degrees’ is taken by the number of community memberships (for individuals) and the number of community members (for communities). Hence we introduce the concept of  $\ell$ -degrees and  $r$ -degrees (of

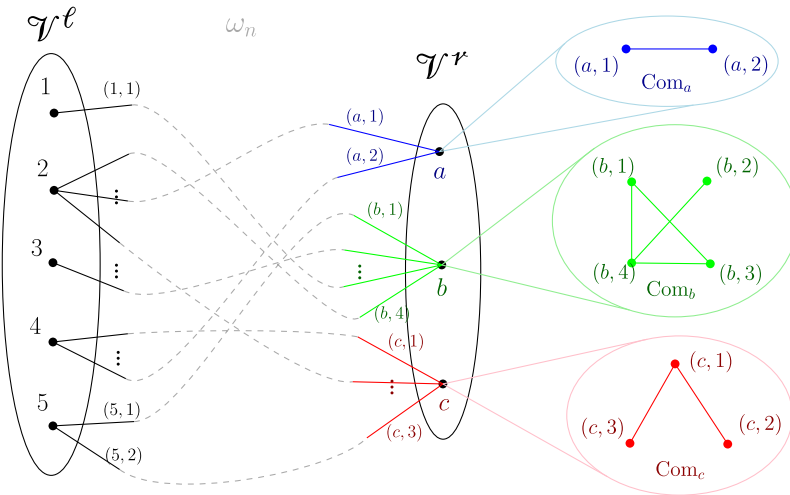


FIGURE 2. An example of the parameters. Individuals form the left-hand side partition  $\mathcal{V}^\ell$ , and their  $\ell$ -degree, i.e., the number of community memberships, is represented by outgoing half-edges. Communities form the right-hand side partition  $\mathcal{V}^r$ , and each is assigned an arbitrary connected community graph. As before, we represent the  $r$ -degree, i.e., the number of community members, by outgoing half-edges. In fact, each half-edge represents a specific vertex (role) in the community graph; thus they are labeled the same way. In the next step, we assign community memberships (community roles) through a (bipartite) matching of the half-edges.

$\ell$ - and  $r$ -vertices, respectively), to which we may collectively refer as bipartite degrees ( $\ell$ -degrees). Within the community graphs, we will refer to the degree of a community vertex as its community degree ( $c$ -degree). We will soon introduce notation for all three types of degree.

As mentioned above, the number of community memberships of an individual  $v \in \mathcal{V}^\ell$  is called its  $\ell$ -degree, and we denote it by  $d_v^\ell$ . For a community  $a \in \mathcal{V}^r$ , we denote its community graph by  $\text{Com}_a$ ; this graph can be arbitrary as long as it is connected. For convenience, we introduce the set of possible community graphs  $\mathcal{H}$ , as the set of (non-empty) simple, finite, connected graphs. Further, we separately equip each graph with its own fixed labeling, i.e., we arbitrarily number the vertices of each graph  $H \in \mathcal{H}$  by the set  $[|H|]$ . We note that we allow several communities to have the same community graph. We call the size  $|\text{Com}_a|$  of the community graph the  $r$ -degree of  $a$ , denoted by  $d_a^r$ . We collect the  $\ell$ - and  $r$ -degrees and the community graphs in the vectors  $\mathbf{d}^\ell := (d_v^\ell)_{v \in \mathcal{V}^\ell}$ ,  $\mathbf{d}^r := (d_a^r)_{a \in \mathcal{V}^r}$ , and  $\mathbf{Com} := (\text{Com}_a)_{a \in \mathcal{V}^r}$ , respectively. Without loss of generality we assume that  $\mathbf{d}^\ell \geq 1$  and  $\mathbf{d}^r \geq 1$  (element-wise) for each  $n$ , as isolated vertices can simply be excluded by adjusting  $N_n$  and  $M_n$ . Also note that  $\mathbf{d}^r$  is derived from  $\mathbf{Com}$ ; thus the RIGC is parametrized by the pair  $(\mathbf{d}^\ell, \mathbf{Com})$ . For a visual representation of the parameters, see Fig. 2.

**Community memberships.** Recall that the  $\ell$ -degree of  $v \in \mathcal{V}^\ell$  denotes the number of community memberships of  $v$ ; we intuitively think of this as giving  $d_v^\ell$  membership tokens to  $v$ . We represent these as  $d_v^\ell$   $\ell$ -half-edges incident to  $v$  and label them by

$$(v, i)_{i \leq d_v^\ell}.$$

Let us denote the union of all vertices in community graphs by  $\mathcal{V}(\mathbf{Com})$ ; we call this the set of community roles or community vertices. For a community vertex  $j \in \mathcal{V}(\text{Com}_a)$ , we can

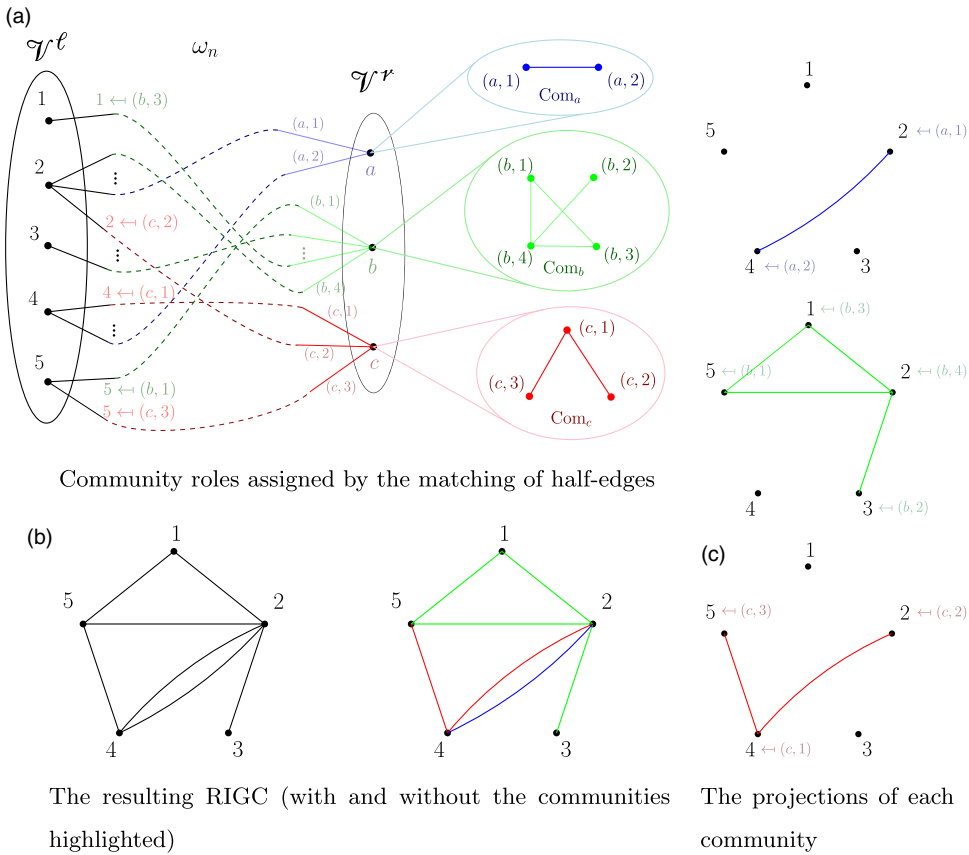


FIGURE 3. The community projection.

uniquely identify  $j$  by the tuple  $(a, l)$ , where  $l$  is the vertex label of  $j$  in  $\text{Com}_a$ . Now, similarly as with individuals, we give each community  $a \in \mathcal{V}^r$   $d_a^r$  community role tokens, represented by  $d_a^r$   $r$ -half-edges incident to  $a$  and labeled by  $(a, l)_{l \leq d_a^r}$ , so that we can represent  $j \in \mathcal{V}(\mathbf{Com})$  by the  $r$ -half-edge  $(a, l)$ .

Next, we introduce the random matching of membership tokens and community role tokens. To ensure that the half-edges can indeed be matched, we assume and define

$$\hbar_n := \sum_{v \in \mathcal{V}^\ell} d_v^\ell = \sum_{a \in \mathcal{V}^r} d_a^r. \tag{2.1}$$

Let  $\Omega_n$  denote the set of all possible bijections between the  $\ell$ -half-edges

$$(v, i)_{i < d_v^\ell, v \in \mathcal{V}^\ell}$$

and the  $r$ -half-edges

$$(a, l)_{l \leq d_a^r, a \in \mathcal{V}^r}.$$

Equivalently, we can think of  $\Omega_n$  as bijections between the  $\ell$ -half-edges and  $\mathcal{V}(\mathbf{Com})$ , since each  $r$ -half-edge  $(a, l)$ ,  $l \leq d_a^r$ ,  $a \in \mathcal{V}^r$  corresponds to a unique community vertex  $j \in \mathcal{V}(\mathbf{Com})$ .

Let the community memberships be determined by a *uniform random bipartite matching* (bipartite configuration)  $\omega_n \sim \text{Unif}[\Omega_n]$ . In fact, we can produce the uniform bipartite matching  $\omega_n$  sequentially, as follows. In each step, we pick an arbitrary unpaired half-edge, and match it to a uniform unpaired half-edge of the opposite type (so that we always match one  $\ell$ -half-edge and one  $r$ -half-edge). The arbitrary choices may even depend on the past of the pairing process, as long as we pair them u.a.r. with one of the remaining half-edges.

**Definition 2.1** (*The ‘underlying BCM’*) Considering the half-edges as tokens to form edges, the bipartite matching  $\omega_n$  also determines a bipartite (multi)graph, defined as follows. For each matched pair of an  $\ell$ -half-edge  $(v,i)$  and  $r$ -half-edge  $(a,l)$ , add an edge with label  $(i,l)$  between  $v$  and  $a$ . We call this edge-labeled graph the *underlying bipartite configuration model (BCM)*. As the edge labels allow us to reconstruct the paired half-edges, the underlying BCM is an equivalent representation of the bipartite matching  $\omega_n$ , and thus encodes the community memberships.

Deleting the edge-labels, we obtain a bipartite version of the configuration model, i.e., the *bipartite configuration model* with degree sequences  $(\mathbf{d}^\ell, \mathbf{d}^r)$ .

**The ‘community projection’.** We now introduce the *community projection*, i.e., the method of projecting the community graphs to the individuals and generating the RIGC model, given the realization of the uniform(ly random) bipartite matching  $\omega_n$ . This procedure is deterministic, and the only randomness of the model comes from the choice of  $\omega_n$ ; thus we can think of the community projection as an operator  $\mathcal{P}$  from  $\Omega_n$  to the space of multigraphs. Alternatively, since the underlying BCM (see Definition 2.1) provides an equivalent representation of the bipartite matching  $\omega_n$ , we can think of the projection as an operator that maps the underlying BCM into the RIGC. This operator can be further generalized as an operator mapping any bipartite graph, which we may interpret as the graph of community memberships, into a network. We will describe the multigraph RIGC by its edge multiplicities.

Recall that the  $r$ -half-edge labeled  $(a,l)$  represents the community role (community vertex)  $j \in \mathcal{V}(\text{Com}_a)$  with vertex label  $l$ , and the  $\ell$ -half-edge  $(v,i)$  is one of the membership tokens of  $v \in \mathcal{V}^\ell$ . Then, if  $(v,i)$  and  $(a,l)$  are matched by  $\omega_n$ , this intuitively means one of the community roles taken by  $v$  is  $j$ . We denote this by  $v \leftarrow j$ . Note that each community role  $j$  is assigned to a *unique* individual  $v$ ; however, each individual  $v$  has  $d_v^\ell$  community roles  $j$  assigned to it. We want to ‘identify’ each individual with all community roles taken, and we carry this out by copying each edge between community roles  $j_1, j_2 \in \mathcal{V}(\text{Com}_a)$  (for each community  $a$ ) to the individuals  $v \leftarrow j_1$  and  $w \leftarrow j_2$ . We emphasize that each community edge is copied individually, even when  $v = w$  or when there is already an edge (or more) between  $v$  and  $w$ ; that is, we allow self-loops and multi-edges (see Section 2.4 for a discussion on multigraphs).

Let us denote the disjoint union of the edges in all community graphs by  $\mathcal{E}(\mathbf{Com})$ ; we refer to this as the *set of community edges*. Now, we shift perspective to obtain the multiplicity  $X(v, w; \omega_n)$  of an edge  $(v,w)$  (for  $v, w \in \mathcal{V}^\ell$ ) for a given bipartite matching  $\omega_n$ . We can do so by counting the number of community edges  $(j_1, j_2)$  such that the community roles  $j_1$  and  $j_2$  are taken by  $v$  and  $w$  (in some order); formally,

$$X(v, w) = X(v, w; \omega_n) := \sum_{(j_1, j_2) \in \mathcal{E}(\mathbf{Com})} \mathbb{1}_{\{v \leftarrow j_1, w \leftarrow j_2\} \cup \{v \leftarrow j_2, w \leftarrow j_1\}}. \tag{2.2}$$

The *random intersection graph with communities*  $\text{RIGC}(\mathbf{d}^\ell, \mathbf{Com})$  is the random multigraph given by the edge multiplicities  $(X(v, w))_{v, w \in \mathcal{V}^\ell}$  determined by the uniform(ly random) bipartite matching  $\omega_n$ .

### 2.2. Notation and assumptions

In this section, we introduce the quantities and assumptions that are crucial throughout the paper.

**Bipartite degrees.** Throughout this paper, we make use of the following description of the  $\ell$ -degree sequences. Let  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$  and  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$  denote uniformly chosen  $\ell$ - and  $r$ -vertices respectively, and denote their random degrees by

$$D_n^\ell := d_{V_n^\ell}^\ell, \quad D_n^r := d_{V_n^r}^r. \tag{2.3}$$

Then the PMF

$$p_k^{(n)} := |\{v \in \mathcal{V}^\ell : d_v^\ell = k\}|/N_n, \tag{2.4a}$$

for  $k \in \mathbb{Z}^+$ , describes the distribution of  $D_n^\ell$  as well as the empirical distribution of  $\mathbf{d}^\ell$ . Similarly, we can describe  $D_n^r$  and  $\mathbf{d}^r$  by the PMF

$$q_k^{(n)} := |\{a \in \mathcal{V}^r : d_a^r = k\}|/M_n. \tag{2.4b}$$

We collect the PMFs in the (infinite-dimensional) probability vectors  $\mathbf{p}^{(n)} = (p_k^{(n)})_{k \in \mathbb{Z}^+}$ ,  $\mathbf{q}^{(n)} = (q_k^{(n)})_{k \in \mathbb{Z}^+}$ .

**The empirical community distribution.** Recall that  $\mathcal{H}$  denotes the set of possible community graphs: simple, connected, finite graphs, each  $H \in \mathcal{H}$  equipped with an arbitrary, fixed labeling using  $[|H|]$  as labels, so that any two community graphs that are isomorphic are labeled in exactly the same way. For a fixed  $H \in \mathcal{H}$ , define

$$\mathcal{V}_H^r := \{a \in \mathcal{V}^r : \text{Com}_a = H\}. \tag{2.5}$$

We introduce the PMF

$$\mu_H^{(n)} := M_n^{-1} |\mathcal{V}_H^r|, \quad \boldsymbol{\mu}^{(n)} = (\mu_H^{(n)})_{H \in \mathcal{H}}, \tag{2.6}$$

so that  $\boldsymbol{\mu}^{(n)}$  describes the empirical PMF of  $\mathbf{Com}$  as well as the PMF of  $\text{Com}_{V_n^r}$ , with  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$ . For  $k \in \mathbb{Z}^+$ , define the (finite) set

$$\mathcal{H}_k := \{H \in \mathcal{H} : |H| = k\}. \tag{2.7}$$

Note that since  $d_a^r = |\text{Com}_a|$ ,  $\mathbf{q}^{(n)}$  from (2.4b) can be obtained by

$$q_k^{(n)} = \sum_{H \in \mathcal{H}_k} \mu_H^{(n)}.$$

**Community degrees and triangles.** Let us denote the disjoint union of the vertices in all community graphs by  $\mathcal{V}(\mathbf{Com})$ ; we refer to this as the *set of community roles*. To a community role  $j \in \mathcal{V}(\mathbf{Com})$  we assign the vector  $(d_j^c, \Delta_j^c)$ , where  $d_j^c$  denotes the degree of  $j$  in its community graph and  $\Delta_j^c$  denotes the number of triangles that  $j$  is part of within its community graph. Let  $J_n \sim \text{Unif}[\mathcal{V}(\mathbf{Com})]$  denote a community role chosen u.a.r. Note that the community that  $J_n$  is part of is chosen in a *size-biased fashion*, and then a vertex in that community is chosen uniformly at random. Define the random vector  $(D_n^c, \Lambda_n^c) := (d_{J_n}^c, \Delta_{J_n}^c)$ , keeping in mind that its coordinates are *dependent*. Define the PMF

$$\varrho_{(k,t)}^{(n)} := \frac{1}{\hat{N}_n} \sum_{j \in \mathcal{V}(\mathbf{Com})} \mathbb{1}_{\{(d_j^c, \Delta_j^c) = (k,t)\}}, \quad \boldsymbol{\varrho}^{(n)} := (\varrho_{(k,t)}^{(n)})_{k \in \mathbb{Z}^+, 0 \leq t \leq \binom{k}{2}}, \tag{2.8}$$



so that  $\mathbf{g}^{(n)}$  describes the joint distribution of  $(D_n^c, \Lambda_n^c)$  as well as the empirical distribution of  $(d_j^c, \Delta_j^c)_{j \in \mathcal{V}(\text{Com})}$ .

**Projected degrees.** For  $v \in \mathcal{V}^\ell$ , its (random) projected degree, i.e. its degree in the RIGC, is by definition given in terms of the edge multiplicities (see (2.2)) as

$$d_v^p = p\text{-deg}(v) := X(v, v) + \sum_{w \in \mathcal{V}^\ell} X(v, w) = 2X(v, v) + \sum_{w \in \mathcal{V}^\ell, w \neq v} X(v, w). \tag{2.9}$$

However, it is more intuitive to look at  $p\text{-deg}(v)$  in terms of the community roles taken by  $v$ . Recall that each community edge incident to some  $j$  such that  $v \leftarrow j$  is added between  $v$  and some other vertex; thus  $j$  contributes  $c\text{-deg}(j)$  to the degree of  $v$ . Then

$$p\text{-deg}(v) = \sum_{j: v \leftarrow j} d_j^c. \tag{2.10}$$

Analogously to  $D_n^c$ , with  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$  as before, we define

$$D_n^p := p\text{-deg}(V_n^\ell). \tag{2.11}$$

Recall that  $p\text{-deg}(v)$  is random for each  $v \in \mathcal{V}^\ell$ , since  $\omega_n$  is random. Thus,  $D_n^p$  has two sources of randomness:  $V_n^\ell$  and  $\omega_n$ . We denote the *random* empirical cumulative distribution function (CDF) of  $D_n^p$  by

$$F_n^p(x) = F_n^p(x; \omega_n) := \frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \mathbb{1}_{\{p\text{-deg}(v) \leq x\}} =: \mathbb{P}(D_n^p \leq x \mid \omega_n), \tag{2.12}$$

where  $\mathbb{P}(\cdot \mid \omega_n)$  denotes the conditional probability with respect to  $\omega_n$ .

**Assumptions.** Recall (2.3), (2.4), and (2.6). We can now summarize our assumptions on the model parameters, in particular, the conditions under which our results hold.

**Assumption 2.2** *The conditions for the empirical distributions are summarized as follows:*

- (A) *There exists a random variable  $D^\ell$  with PMF  $\mathbf{p}$  such that  $\mathbf{p} \rightarrow \mathbf{p}^{(n)}$  pointwise as  $n \rightarrow \infty$ , i.e.,*

$$D_n^\ell \xrightarrow{d} D^\ell. \tag{2.13}$$

- (B)  *$\mathbb{E}[D^\ell]$  is finite, and as  $n \rightarrow \infty$ ,*

$$\mathbb{E}[D_n^\ell] \rightarrow \mathbb{E}[D^\ell]. \tag{2.14}$$

- (C) *There exists a PMF  $\boldsymbol{\mu}$  on  $\mathcal{H}$  such that  $\boldsymbol{\mu}^{(n)} \rightarrow \boldsymbol{\mu}$  pointwise as  $n \rightarrow \infty$ . (1) Consequently, since  $q_k^{(n)} = \sum_{H \in \mathcal{H}_k} \mu_H^{(n)}$ , with the finite set  $\mathcal{H}_k$  from (2.7), there exists a random variable  $D^r$  with PMF  $\mathbf{q}$  such that  $\mathbf{q}^{(n)} \rightarrow \mathbf{q}$  pointwise as  $n \rightarrow \infty$ , or equivalently,*

$$D_n^r \xrightarrow{d} D^r. \tag{2.15}$$

- (D)  *$\mathbb{E}[D^r]$  is finite, and as  $n \rightarrow \infty$ ,*

$$\mathbb{E}[D_n^r] \rightarrow \mathbb{E}[D^r]. \tag{2.16}$$

**Remark 2.3** (Consequences of Assumption 2.2) We note the following:

- (i) By its definition in (2.1),  $\ell_n = N_n \mathbb{E}[D_n^\ell] = M_n \mathbb{E}[D_n^r]$ . By Assumption 2.2(B,D),

$$M_n/N_n = \mathbb{E}[D_n^\ell]/\mathbb{E}[D_n^r] \rightarrow \mathbb{E}[D^\ell]/\mathbb{E}[D^r] =: \gamma \in \mathbb{R}^+. \tag{2.17}$$

- (ii) Since  $\boldsymbol{q}^{(n)}$  (see (2.8)) can be obtained from  $\boldsymbol{\mu}^{(n)}$ , Assumption 2.2(C) also implies that there exists a random variable  $(D^c, \Lambda^c)$  with PMF  $\boldsymbol{q}$  such that  $\boldsymbol{q}^{(n)} \rightarrow \boldsymbol{q}$  pointwise as  $n \rightarrow \infty$ , or equivalently,  $(D_n^c, \Lambda_n^c) \xrightarrow{d} (D^c, \Lambda^c)$ .
- (iii) Assumption 2.2(A,B) imply that  $d_{\max}^\ell := \max_{v \in \mathcal{V}^\ell} d_v^\ell = o(\ell_n)$ . This implication is proved for a similar setting in [29, Exercise 6.3]. Similarly, the conditions (C1,D) imply that  $d_{\max}^r := \max_{a \in \mathcal{V}^r} d_a^r = o(\ell_n)$ .

**Remark 2.4** (Random parameters) The results in Section 2.3 below remain valid when the sequence of parameters  $(\boldsymbol{d}^\ell, \mathbf{Com})$  (resp.,  $(\boldsymbol{d}^\ell, \boldsymbol{d}^r)$ ) is itself random. In this case, we require that  $N_n \rightarrow \infty$  and  $M_n \rightarrow \infty$  almost surely, and we replace Assumption 2.2(A–D) (resp., Assumption 2.2(A,B,C1,D)) by the conditions  $\boldsymbol{p}^{(n)} \xrightarrow{\mathbb{P}} \boldsymbol{p}$  pointwise,  $\mathbb{E}[D_n^\ell | \boldsymbol{d}^\ell] \xrightarrow{\mathbb{P}} \mathbb{E}[D^\ell]$ ,  $\boldsymbol{\mu}^{(n)} \xrightarrow{\mathbb{P}} \boldsymbol{\mu}$  pointwise (resp.,  $\boldsymbol{q}^{(n)} \xrightarrow{\mathbb{P}} \boldsymbol{q}$ ), and  $\mathbb{E}[D_n^r | \boldsymbol{d}^r] \xrightarrow{\mathbb{P}} \mathbb{E}[D^r]$ , where we assume the limiting PMFs  $\boldsymbol{p}$  and  $\boldsymbol{\mu}$  (resp.,  $\boldsymbol{q}$ ) to be deterministic. For a similar setting in the configuration model, see [29, Remark 7.9 on ‘regularity of random degrees’], where this is spelled out in more detail.

Note that analogously to Remark 2.3(i), under the conditions of Remark 2.4,  $M_n/N_n \xrightarrow{\mathbb{P}} \gamma$ .

### 2.3. Results

In this section, we state our results on local properties of the RIGC. The main result is the local weak convergence (LWC) of the RIGC (defined shortly), which is equivalent to the convergence of subgraph counts (neighborhood counts). LWC also implies the convergence of degrees and local clustering, and provides some insight into the overlapping structure of communities. We use the following notions throughout this section. Recall that  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$  denotes an  $\ell$ -vertex chosen u.a.r., and  $\mathbb{P}(\cdot | \omega_n)$  denotes conditional probability with respect to  $\omega_n$ . Let  $\mathbb{E}_{V_n^\ell}[\cdot | \omega_n]$  denote the corresponding conditional expectation, that is, empirical averages for a given  $\omega_n$ .

**Local weak convergence.** First, we give a brief definition of LWC to state our results. We give a much more detailed introduction to the concept in Section 3.1.

**Definition 2.5** (Rooted (multi)graph, rooted isomorphism, and neighborhood)

- (i) We call a pair  $(G, o)$  a rooted (multi)graph if  $G$  is a locally finite, connected (multi)graph and  $o$  is a distinguished vertex of  $G$ .
- (ii) We say that two multigraphs  $G_1$  and  $G_2$  are isomorphic if there exists a bijection  $\varphi : \mathcal{V}(G_1) \rightarrow \mathcal{V}(G_2)$  such that for any  $v, w \in \mathcal{V}(G_1)$ , the number of edges between  $v, w$  in  $G_1$  equals the number of edges between  $\varphi(v), \varphi(w)$  in  $G_2$ .
- (iii) We say that the rooted (multi)graphs  $(G_1, o_1) \simeq (G_2, o_2)$ , are rooted isomorphic if there exists a graph-isomorphism between  $G_1$  and  $G_2$  that maps  $o_1$  to  $o_2$ .

- (iv) For some  $r \in \mathbb{N}$ , we define  $B_r(G, o)$ , the (closed)  $r$ -ball around  $o$  in  $G$  or  $r$ -neighborhood of  $o$  in  $G$ , as the subgraph of  $G$  spanned by all vertices of graph distance at most  $r$  from  $o$ . We think of  $B_r(G, o)$  as a rooted (multi)graph with root  $o$ .

**Definition 2.6.** (Local weak convergence in probability) Let  $(G_n)_{n \in \mathbb{N}}$  with size  $|G_n| \xrightarrow{\mathbb{P}} \infty$  be a sequence of random (multi)graphs, and let  $U_n | G_n \sim \text{Unif}[\mathcal{V}(G_n)]$ . By  $|G_n| \xrightarrow{\mathbb{P}} \infty$ , we mean that for all  $K \in \mathbb{R}^+$ ,  $\mathbb{P}(|G_n| \geq K) \rightarrow 1$  as  $n \rightarrow \infty$ . Let  $(\mathcal{R}, o)$  denote a random element of the set of rooted (multi)graphs, which we call a random rooted (multi)graph. We say that  $(G_n, U_n)$  converges to  $(\mathcal{R}, o)$  in probability in the LWC sense, and write  $(G_n, U_n) \xrightarrow{\mathbb{P}\text{-loc}} (\mathcal{R}, o)$ , if for any fixed rooted (multi)graph  $(G, o)$  and  $r \in \mathbb{N}$ ,

$$\begin{aligned} \mathbb{P}(B_r(G_n, U_n) \simeq B_r(G, o) \mid G_n) &:= \frac{1}{|G_n|} \sum_{u \in \mathcal{V}(G_n)} \mathbb{1}_{\{B_r(G_n, u) \simeq B_r(G, o)\}} \\ &\xrightarrow{\mathbb{P}} \mathbb{P}(B_r(\mathcal{R}, o) \simeq B_r(G, o)). \end{aligned} \tag{2.18}$$

We also say that  $(\mathcal{R}, o)$  is the local weak limit in probability of  $(G_n, U_n)$ .

We can now state our first main result on the LWC of the RIGC model.

**Theorem 2.7. (Local weak convergence of the RIGC)** Consider  $\text{RIGC}_n = \text{RIGC}(d^\ell, \mathbf{Com})$  under Assumption 2.2. Then, with  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$ , as  $n \rightarrow \infty$ ,

$$(\text{RIGC}_n, V_n^\ell) \xrightarrow{\mathbb{P}\text{-loc}} (\text{CP}, o), \tag{2.19}$$

where  $(\text{CP}, o)$  is a random rooted graph with distribution as specified in Section 3.3.

The proof of Theorem 2.7 is completed in Section 3.3. The local weak limit generalizes random clique trees, which are shown to be the local weak limit of RIGs (see e.g. [38]). Our construction replaces the cliques by arbitrary connected graphs and is slightly more involved, hence Section 3.3. In the following, we present some corollaries of Theorem 2.7.

**Degrees.** Recall (2.11) and (2.12). We define the random variable  $D^p$  and its distribution function

$$D^p \stackrel{d}{=} \sum_{i=1}^{D^\ell} D_{(i)}^c, \quad F^p(x) := \mathbb{P}(D^p \leq x), \tag{2.20}$$

with  $D^\ell$  from Assumption 2.2(A), and  $D_{(i)}^c$  are i.i.d. copies of  $D^c$  from Remark 2.3(ii).

**Corollary 2.8.** (Degrees in the RIGC) Consider  $\text{RIGC}(d^\ell, \mathbf{Com})$  under the conditions of Theorem 2.7. Then, as  $n \rightarrow \infty$ ,

$$\|F_n^p - F^p\|_\infty = \sup_{x \in \mathbb{R}} |F_n^p(x) - F^p(x)| \xrightarrow{\mathbb{P}} 0, \tag{2.21}$$

and consequently,

$$D_n^p \xrightarrow{d} D^p. \tag{2.22}$$

Corollary 2.8 is almost a direct consequence of Theorem 2.7. Pointwise convergence in probability of the empirical CDF follows directly; however, the convergence of the *sup-norm*

requires a proof that we provide in Section 3.4.1. We remark that Corollary 2.8 can alternatively be proved independently through a first and second moment method under weaker conditions. In particular, Assumption 2.2(C) can be replaced by  $D_n^c \xrightarrow{d} D^c$ . Let us also note that while (2.22) is more intuitive, (2.21) is a stronger statement. Indeed, (2.21) implies that the *random empirical degree distribution*, i.e., the observed degree sequence, is close to its theoretical limit w.h.p.

**Clustering.** We proceed by studying the clustering in the RIGC, in particular focusing on local clustering. For an arbitrary individual  $v \in \mathcal{V}^\ell$ , let  $\Delta^p(v)$  denote the (random) number of triangles that  $v$  is part of in the RIGC. Here we also include degenerate triangles, where one or more vertices are the same, and count triangles with multiplicity, i.e., all possible ways we can choose the three edges. We define the local clustering at  $v$  as

$$Cl(v) := \frac{\Delta^p(v)}{\binom{p-\text{deg}(v)}{2}}, \tag{2.23}$$

with the convention that  $Cl(v) := 0$  whenever  $p\text{-deg}(v) < 2$ . Define the empirical local clustering coefficient  $\zeta_n := Cl(V_n^\ell)$  and denote its random empirical CDF by

$$F_n^\zeta(x) = F_n^\zeta(x; \omega_n) := \frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \mathbb{1}_{\{Cl(v) \leq x\}} = \mathbb{P}(\zeta_n \leq x \mid \omega_n). \tag{2.24}$$

We introduce

$$\zeta \stackrel{d}{=} \left( \sum_{i=1}^{D^c} \Lambda_{(i)}^c \right) / \binom{\sum_{i=1}^{D^c} D_{(i)}^c}{2}, \quad F^\zeta(x) := \mathbb{P}(\zeta \leq x), \tag{2.25}$$

where  $(D_{(i)}^c, \Lambda_{(i)}^c)$  are i.i.d. copies of the random vector  $(D^c, \Lambda^c)$  from Remark 2.3(ii) and are independent of  $D^\ell$  (see Assumption 2.2(A)).

**Corollary 2.9.** (Local clustering in the RIGC) *Consider RIGC( $d^\ell, \mathbf{Com}$ ) under the conditions of Theorem 2.7. Then, as  $n \rightarrow \infty$ ,*

$$\|F_n^\zeta - F^\zeta\|_\infty = \sup_{x \in \mathbb{R}} |F_n^\zeta(x) - F^\zeta(x)| \xrightarrow{\mathbb{P}} 0. \tag{2.26}$$

*In particular,  $\zeta_n \xrightarrow{d} \zeta$  and thus the average local clustering converges:*

$$\mathbb{E}[\zeta_n] \rightarrow \mathbb{E}[\zeta]. \tag{2.27}$$

Analogously to Corollary 2.8, Corollary 2.9 is almost a direct consequence of Theorem 2.7. *Pointwise* convergence in probability of the empirical CDF follows directly; however, the convergence of the *sup-norm* requires a proof that we provide in Section 3.4.1. We note that in fact Corollary 2.9 still holds if we replace Assumption 2.2(C) by the conditions in Assumption 2.2(C1) and Remark 2.3(ii). The intuition behind Corollary 2.9 is that triangles typically arise within one community; that is, triangles containing edges from different communities make a negligible contribution as the model size grows. This is due to the ‘locally tree-like’ structure of the underlying BCM (see Proposition 3.2 below). We remark that under our general conditions, we cannot establish that the local clustering scales inversely with the degree (as in

e.g. [9, 39]); however, the inverse degree serves as an upper bound for the clustering. In the following, we establish when the model has positive asymptotic clustering.

**Corollary 2.10** (Condition for positive asymptotic clustering) *Under the conditions of Corollary 2.9, the asymptotic average clustering  $\mathbb{E}[\zeta]$  is positive if and only if  $\mathbb{P}(\Lambda^c \geq 1) > 0$ , with  $\Lambda^c$  from Remark 2.3(ii).*

*Proof of Corollary 2.10.* Note that  $\mathbb{P}(\Lambda^c \geq 1) > 0$  happens exactly when the assigned communities are not  $\mu$ -almost surely triangle-free with  $\mu$  from Assumption 2.2(C); i.e.,  $\mu_H > 0$  for at least one  $H \in \mathcal{H}$  such that  $H$  contains at least one triangle. Clearly, this is a necessary condition, but also sufficient, as it implies that any vertex has a positive probability to be part of a triangle and have bounded degree at the same time.  $\square$

Another measure of clustering is the so-called global clustering coefficient, defined as three times the total number of triangles in the graph divided by the total number of connected triples (paths of length 2, often called ‘wedges’); formally,

$$Cl_{\text{glob}} := \frac{3\Delta_{\text{total}}^p}{\sum_{v \in \mathcal{V}^\ell} \binom{p-\text{deg}(v)}{2}} = \frac{\sum_{v \in \mathcal{V}^\ell} \Delta^p(v)}{\sum_{v \in \mathcal{V}^\ell} \binom{p-\text{deg}(v)}{2}}. \tag{2.28}$$

Note the relation with the local clustering coefficient defined in (2.23) as the ratio of  $\Delta^p(v)$  and  $\binom{p-\text{deg}(v)}{2}$ ; in (2.28), we instead consider the ratio of the *sum* of these quantities over all individuals. Also note that we can think of the global clustering coefficient as the *ratio of the averages* of  $\Delta^p(v)$  and  $\binom{p-\text{deg}(v)}{2}$ ,

$$Cl_{\text{glob}} = \frac{\frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \Delta^p(v)}{\frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \binom{p-\text{deg}(v)}{2}}, \tag{2.29}$$

while the average local clustering is given by the *average of the ratios* of the same quantities,

$$\mathbb{E}[Cl(V_n^\ell) \mid \omega_n] = \frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} Cl(v) = \frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \frac{\Delta^p(v)}{\binom{p-\text{deg}(v)}{2}}. \tag{2.30}$$

While the global clustering coefficient and average local clustering coefficient embrace similar concepts, their behaviors are different. By [38, Corollary 4.4], in addition to LWC, convergence of the global clustering coefficient requires the stronger condition of

$$\mathbb{E}[(p-\text{deg}(V_n^\ell))^2 \mid \omega_n] = \mathbb{E}[(D_n^p)^2 \mid \omega_n] \xrightarrow{\mathbb{P}} \mathbb{E}[(D^p)^2],$$

which can be reduced to  $\mathbb{E}[(D_n^\ell)^2] \rightarrow \mathbb{E}[(D^\ell)^2]$  and  $\mathbb{E}[(D_n^c)^2] \rightarrow \mathbb{E}[(D^c)^2]$  (by Corollary 2.8). Under these conditions,  $Cl_{\text{glob}}$  converges in probability to the *ratio of expectations* of the numerator and denominator of  $\zeta$  in (2.25), i.e.,

$$Cl_{\text{glob}} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\sum_{i=1}^{D^\ell} \Lambda_{(i)}^c\right] / \mathbb{E}\left[\binom{\sum_{i=1}^{D^\ell} D_{(i)}^c}{2}\right], \tag{2.31}$$

which in general is different from the limiting average local clustering  $\mathbb{E}[\zeta]$ .

**The overlapping structure.** Next, we turn our attention to the overlapping structure of the communities, which is one of the main motivators for the RIGC model. By an overlap, we mean

two or more communities having one or more individuals in common. From this definition, it is clear that the internal structure of the communities does not play a role in the overlapping structure; thus the following discussion applies to the RIG model as well. By the construction of the model, i.e. the inclusion of individuals in several communities, it is clear that overlaps are present. We will study first the number of overlaps, and later the typical size of the overlaps as well. Let us introduce some notation. For  $v \in \mathcal{V}^\ell$  and  $a \in \mathcal{V}^r$ , we say that  $v$  is part of  $\text{Com}_a$  and write  $v \leftarrow \text{Com}_a$  if  $v \leftarrow j$  for some  $j \in \text{Com}_a$ . Let us denote the size of the overlap between  $a, b \in \mathcal{V}^r, a \neq b$ , by

$$\mathbb{O}(a, b) := \sum_{v \in \mathcal{V}^\ell} \mathbb{1}_{\{v \leftarrow \text{Com}_a\} \cap \{v \leftarrow \text{Com}_b\}}. \tag{2.32}$$

We define the set of communities overlapping with the community  $a$  as

$$\mathcal{N}(a) := \{b \in \mathcal{V}^r : b \neq a, \mathbb{O}(a, b) \geq 1\}. \tag{2.33}$$

For  $k \in \mathbb{Z}^+$ , we introduce the set of unordered pairs of (at least)  $k$ -fold overlapping communities:

$$\mathcal{L}_k = \mathcal{L}_k^{(n)} := \{\{a, b\} : a, b \in \mathcal{V}^r, a \neq b, \mathbb{O}(a, b) \geq k\}. \tag{2.34}$$

Note that  $\mathcal{L}_k \supseteq \mathcal{L}_{k+1}$  for all  $k \in \mathbb{Z}^+$  and  $\mathcal{L}_1$  contains all overlapping pairs, regardless of the size of overlap they share. Recall that  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$ , and further recall that  $\mathbb{P}(\cdot \mid \omega_n)$  denotes the conditional probability with respect to  $\omega_n$  and  $\mathbb{E}[\cdot \mid \omega_n]$  denotes the corresponding conditional expectation. We can now state our result on the number of overlaps.

**Proposition 2.11.** (Number of overlaps) *Consider  $\text{RIGC}(d^\ell, \mathbf{Com})$  under Assumption 2.2. In addition, assume that, as  $n \rightarrow \infty$ ,*

$$\mathbb{E}[(D_n^\ell)^2] \rightarrow \mathbb{E}[(D^\ell)^2] < \infty. \tag{2.35}$$

*Then, as  $n \rightarrow \infty$ , the average number of communities overlapping with a ‘typical’ one converges:*

$$\frac{2|\mathcal{L}_1|}{M_n} = \mathbb{E}[|\mathcal{N}(V_n^r)| \mid \omega_n] \xrightarrow{\mathbb{P}} \mathbb{E}[D^r] \mathbb{E}[\tilde{D}^\ell]. \tag{2.36}$$

Note that (2.35) ensures that  $\mathbb{E}[\tilde{D}^\ell] < \infty$ , so the right-hand side of (2.36) is finite. We prove Proposition 2.11 in Section 3.4.2 using LWC. Intuitively, (2.36) asserts that a typical community  $V_n^r$  overlaps with constantly many others, and thus the number of overlapping pairs of communities is linear in the total number of communities.

Next, we assert that the ‘typical’ overlap size is 1; we call this the *single-overlap property*. There are several ways to interpret what the ‘typical overlap’ means, leading to slightly different statements, as follows.

**Theorem 2.12.** (Single-overlap property) *Consider  $\text{RIGC}(d^\ell, \mathbf{Com})$  under Assumption 2.2. Then the single-overlap property holds, in the following ways:*

- (i) *Vertex perspective. For a uniform individual  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$ , the communities that  $V_n^\ell$  is part of w.h.p. overlap only at  $V_n^\ell$ . Formally, as  $n \rightarrow \infty$ ,*

$$\mathbb{P}(\exists \{a, b\} \in \mathcal{L}_2 : V_n^\ell \leftarrow \text{Com}_a, V_n^\ell \leftarrow \text{Com}_b \mid \omega_n) \xrightarrow{\mathbb{P}} 0. \tag{2.37}$$

- (ii) *Community perspective.* For a uniform community  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$ , the communities that  $V_n^r$  overlaps with w.h.p. share only a single individual with  $V_n^r$ . Formally, as  $n \rightarrow \infty$ ,

$$\mathbb{P}(\exists b \in \mathcal{N}(V_n^r) : \mathcal{O}(V_n^r, b) \geq 2 \mid \omega_n) \xrightarrow{\mathbb{P}} 0. \tag{2.38}$$

- (iii) *Global perspective.* Assume additionally the condition (2.35), and let  $\{A_n, B_n\} \sim \text{Unif}[\mathcal{L}_1]$  denote a pair of communities chosen u.a.r. from among all distinct pairs of overlapping communities. Then w.h.p. their overlap is one individual. Formally, as  $n \rightarrow \infty$ ,

$$\mathbb{P}(\mathcal{O}(A_n, B_n) \geq 2 \mid \omega_n) = |\mathcal{L}_2| / |\mathcal{L}_1| \xrightarrow{\mathbb{P}} 0. \tag{2.39}$$

We complete the proof in Section 3.4.2 but discuss the statement now. The extra second moment condition (2.35) in Theorem 2.12(iii) suggests a substantial difference from Parts (2.12)–(2.12). Indeed, Parts (2.12)–(2.12) establish local properties and follow directly from LWC, which is not true for Part (2.12). The difficulty is in relating the choice of the pair  $(A_n, B_n) \sim \text{Unif}[\mathcal{L}_1]$  to the choice of a single uniform vertex (and further choices in its neighborhood). This problem is nontrivial and further regularity is required. Also note that Proposition 2.11 requires the same second moment condition for  $\mathbb{E}[\tilde{D}^\ell]$  to be finite that is used in identifying the asymptotics for  $|\mathcal{L}_1|$ , that is, the denominator in (2.39). In the underlying BCM (see Definition 2.1),  $|\mathcal{L}_1|$  is the number of pairs of communities that are at graph distance 2; however, the fluctuations of this quantity are an open problem in the case when the variance of the degrees diverges.

**Relationship with the ‘passive’ random intersection graph.** The overlapping structure may be represented as a graph on  $\mathcal{V}^r$  by adding an edge between a pair of communities for each individual they are both connected to. This leads to a ‘dual’ RIG, defined on the communities, that is sometimes referred to as the ‘passive model’ in the literature [26]. The sizes of the overlaps  $\mathcal{O}(a, b)$  and the number of overlapping pairs  $|\mathcal{L}_1|$  can be seen as the edge multiplicities and total number of edges in the passive model, respectively; in particular,  $2|\mathcal{L}_1|/M_n$  gives the average degree. Note that in this regard, applying Theorem 2.12 with the roles of left-hand side and right-hand side reversed (and also replacing (2.35) by  $\mathbb{E}[(D_n^r)^2] \rightarrow \mathbb{E}[(D^r)^2] < \infty$  in Theorem 2.12(ii)) provides some insight on the number of multi-edges in the ‘active’ RIG (with complete graph communities) on the  $\ell$ -vertices. In turn, this provides an upper bound for the number of multi-edges in the RIGC model as well, but obtaining a lower bound is nontrivial: since the communities are not complete graphs, the fact that two individuals are together in several communities does not necessarily mean that they are connected by multiple edges, and finer properties of the measure  $\mu$  (see Assumption 2.2(C)) come into play. It further complicates the situation that if we condition on having several communities that both individuals are part of, we also introduce a bias to the  $\beta$ -degrees involved.

## 2.4. Discussion on the random intersection graph with communities

In this section, we discuss the relationship of our model to other network models and shed light on possible applications and their limitations.

**Parameter choices.** Working with prescribed parameters provides a wide range of applicability. As Corollaries 2.8 and 2.9 suggest, the degree distribution and clustering of the RIGC model are tunable to match our observations of real-world networks; however, the choice of  $d^\ell$  and **Com** is hard to infer. One way of obtaining these parameters explicitly is through

community-detection algorithms [22, 23]. For theoretical research, one may be interested in generating the input parameters randomly; we give two examples of this. A simple idea is to use i.i.d. random variables with distribution  $D^\ell$  and  $\text{Com}$  to generate the sequences  $\mathbf{d}^\ell$  and  $\mathbf{Com}$ , respectively. However, the parameters must satisfy (2.1). If both  $\text{Var}(D^\ell) < \infty$  and  $\text{Var}(D^r) < \infty$ , we can use the algorithm proposed by Chen and Olvera-Cravioto in [16] to generate the sequences  $\mathbf{d}^\ell, \mathbf{Com}$  in such a way that the sums of the  $\ell$ - and  $r$ -degrees are equal, while the entries are asymptotically independent. While the algorithm in [16] was designed for the directed configuration model, it is straightforwardly applicable to the BCM.

Our second example is generating a matching pair of  $\mathbf{d}^\ell$  and  $\mathbf{d}^r$  in a *dependent* way through a bipartite version of the generalized random graph [15], or a Norros–Reittu model [41]. Once  $\mathbf{d}^r$  is given, we have to generate  $\mathbf{Com}$  in a compatible way, i.e., such that the community sizes are indeed the  $r$ -degrees. Assumption 2.2(C) that there exists a family of conditional measures

$$\mu_{H|k} = \mathbb{P}(\text{Com}_a \simeq H \mid d_a^r = |\text{Com}_a| = k), \quad \mu_{\cdot|k} = (\mu_{H|k})_{H \in \mathcal{H}_k}, \quad (\mu_{\cdot|k})_{k \in \mathbb{Z}^+}, \quad (2.40)$$

that describe the conditional distribution of community graphs for each given community size. In fact  $\mu_{H|k} = \mu_H/q_k$ , with  $\mu$  and  $q$  from Parts (C) and (C1), respectively, of Assumption 2.2. (We note that because of this relation, under Assumption 2.2(C1), the implication is reversible, i.e., the existence of  $(\mu_{\cdot|k})_{k \in \mathbb{Z}^+}$  implies Assumption 2.2(C).) Thus we can generate each  $\text{Com}_a$  according to the measure  $\mu_{\cdot|d_a^r}$ , all independently of each other.

**Overlaps.** The motivation behind RIGs is to generate *overlapping* communities, which is clearly satisfied by Proposition 2.11. However, Theorem 2.12 asserts the single-overlap property of the RIGC and RIG graphs, which limits the applicability of these models. For example, they may not be a good fit for scientific collaboration networks, where the same authors often collaborate on several papers and with several other collaborators. However, the RIGC may be used for social networks when the different communities of the same person tend to be separate: their family members, their colleagues, their sports club friends, etc., typically do not know each other.

On the other hand, the single-overlap property may be used to optimize community detection; for example, consider the C-finder algorithm based on the clique percolation method [19, 42], which we explain briefly. A  $k$ -clique in a graph is a complete subgraph on  $k$  vertices, and we call two  $k$ -cliques adjacent if they share  $k - 1$  vertices. A component in  $k$ -clique percolation is a maximal set of vertices that are connected through a chain of adjacent  $k$ -cliques. We remark that such components may overlap, as long as the intersection does not contain a  $(k - 1)$ -clique; the simplest case is when the overlap has fewer than  $k - 1$  vertices. The C-finder algorithm outputs such components as possibly overlapping communities in the network. Now suppose each community of the RIGC is 3-clique connected, i.e., built up from edge-adjacent triangles. Due to the single-overlap property of the RIGC, a typical community will be a component of 3-clique percolation by itself, i.e., no other communities will be 3-clique adjacent to it, allowing detection with great accuracy. Thus, such an RIGC works very well in conjunction with the C-finder algorithm, either in first generating the RIGC and then detecting its communities, or in running C-finder on the dataset for which one wishes to use the RIGC as a null model.

We believe that we can also use the clique percolation approach to make the RIGC a better fit than the traditional RIG for collaboration networks, in particular for scientific collaboration networks of authors and the papers they collaborate on. Rather than considering each paper as



its own community, which leads to cliques with a typical overlap size larger than one, we can instead merge cliques with more than a single overlap into one community, which, in fact, uses the components of clique percolation as communities. Then we can think of each community as the collaboration network of a subcommunity of authors who often collaborate with one another, and the collaboration network as a network with hierarchical structure.

**Multigraphs.** The usual criticism that the configuration model receives is that it may produce a multigraph, and this happens w.h.p. in the case when the degrees have infinite (asymptotic) variance [29, Chapter 7]. As the RIGC uses a BCM in its construction, we are bound to deal with multigraphs on the level of community memberships, and possibly on the level of the projection as well. One classical remedy is to condition the graph on simplicity; it is however outside the scope of this paper to study this conditional measure (which we conjecture is non-uniform) or to study whether the simplicity probability remains bounded away from 0 as the graph size grows. Another classical approach applied to the configuration model is erasure, and analogously, we can define the erased RIGC by removing self-loops and collapsing multi-edges into a simple edge, i.e., redefining the edge multiplicities from (2.2) as  $X'_{v,v} = 0$  and  $X'_{v,w} = \mathbb{1}_{\{X_{v,w} \geq 1\}}$ . A different way of erasing would be to erase multiple edges in the underlying BCM, but note that such erasure does *not* ensure that the resulting RIGC is a simple graph; multi-edges may still arise from two individuals being part of two (or more) communities together. Hence erasing directly in RIGC is the natural/better choice. In this paper, however, we choose to study the RIGC as a multigraph, and argue that we do not see the effect of this in the local behavior; indeed, subject to Theorem 2.7, the local weak limit of the RIGC is simple (a distribution on rooted simple graphs). This means that a typical individual will w.h.p. not see a self-loop or multi-edge in its finite neighborhood. Based on this observation, our results extend to the erased RIGC without any modification.

### 3. Proofs

In this section, we prove the above results. For this purpose, we first provide a more detailed overview of the concept of LWC in Section 3.1. We then prove LWC for the RIGC as a consequence of the LWC of the underlying BCM, as has been done for other RIG models. Finally we include the proofs of our further results. For some of these results, we include a sketch of the proof here, and the more rigorous albeit tedious details can be found in an extended version of this paper [31].

#### 3.1. Preliminaries: marked graphs and local weak convergence

In order to prove our results, we first introduce the concepts that we rely on in our proof, the most central one being *local weak convergence* (LWC), a notion of convergence for sparse graph sequences. The usefulness of LWC comes from the fact that numerous properties of the finite graph(s) can be determined or approximated based on the limiting object alone [6, 24]. As its name suggests, LWC describes the graph from a *local* point of view; indeed, in Definition 2.6, we have defined LWC in probability in terms of convergence of frequencies of graph neighborhoods. We cover some of the theory behind the notion of LWC, in fact in a more general setting of *marked graphs*, which we also define shortly. The theory of LWC presented is partially based on [1, 4, 5] and [30, Chapter 2], but generalized and tailored to our needs. We start by defining marked graphs.

**Marked graphs.** *Marks* provide a general framework for indicating additional information on the edges and/or vertices of a (multi)graph, such as edge weights, edge directions, graph coloring, etc. In our case, we use marks to include edge labels of the underlying BCM, as well as indicate the community graphs assigned to each  $r$ -vertex. We formally define marked (multi)graphs below. For simplicity, we will simply write graphs. Keep in mind that edges between the same pair of vertices receive marks separately and may have different marks.

Let  $\mathcal{G}$  denote the set of all locally finite (multi)graphs on a countable (finite or countably infinite) vertex set. Let the set of marks  $\mathcal{M}$  be an arbitrary countable set that contains the special symbol  $\emptyset$ , which is to be interpreted as ‘no mark’. A marked graph is a pair  $(G, \Xi)$ , where  $\Xi$  is the mark function that maps elements of  $G$  into  $\mathcal{M}$ ; in particular, for  $v \in \mathcal{V}(G)$ ,  $\Xi(v) \in \mathcal{M}$ , and for  $e \in \mathcal{E}(G)$ ,  $\Xi(e) \in \mathcal{M}^2$ . It is common to associate two marks to each edge, with one mark associated to each endpoint, which is often interpreted as associating separate marks to the two directions of a bi-directed edge. Since we work with the BCM, it is more useful to think of the marks as being associated to the half-edges that form the edge. We denote the set of graphs with marks from the mark set  $\mathcal{M}$  by  $\mathcal{G}(\mathcal{M})$ .

We remark that any graph in  $\mathcal{G}$  (which we may refer to as the set of *unmarked graphs*, for clarity) can be turned into a marked graph by assigning the ‘no mark’ symbol  $\emptyset$  to each vertex and half-edge; thus, results and definitions formulated for marked graphs apply straightforwardly to (unmarked) graphs.

**Rooted marked graph, isomorphism, and  $r$ -neighborhood.** We now generalize Definition 2.5 to marked graphs.

- (i) Choose a vertex  $o$  in a marked graph  $(G, \Xi)$  to be distinguished as the root; if  $G$  is not connected, we restrict ourselves to the connected component of  $o$ , and denote the rooted marked graph by  $(G, \Xi, o)$ .

Denote the set of rooted marked graphs by  $\mathcal{G}_o(\mathcal{M})$ . We call a random element of  $\mathcal{G}_o(\mathcal{M})$  a random rooted marked graph.

- (ii) We say that the rooted marked graphs  $(G_1, \Xi_1, o_1)$  and  $(G_2, \Xi_2, o_2)$  are isomorphic, and denote this by  $(G_1, \Xi_1, o_1) \simeq (G_2, \Xi_2, o_2)$ , if there is a graph-isomorphism between them that also maps root to root and preserves marks. When there are multiple edges between the same pair of vertices, we require that there be the same number of edges with any given mark between the corresponding pairs of vertices in the two graphs.
- (iii) The (closed) ball  $B_r(G, \Xi, o)$  can be defined analogously to the unmarked graph ball (Definition 2.5(iv)), by restricting the mark function to the subgraph as well.

**Distance and topology.** We are now ready to define a metric on  $\mathcal{G}_o(\mathcal{M})$ . For two elements  $(G_1, \Xi_1, o_1), (G_2, \Xi_2, o_2) \in \mathcal{G}_o(\mathcal{M})$ , we define the largest radius  $r$  such that the  $r$ -neighborhoods of the roots are isomorphic:

$$r_{\max} := \begin{cases} -1 & \text{if } \Xi_1(o_1) \neq \Xi_2(o_2), \\ +\infty & \text{if } (G_1, \Xi_1, o_1) \simeq (G_2, \Xi_2, o_2), \\ \sup\{r \in \mathbb{N} : B_r(G_1, \Xi_1, o_1) \simeq B_r(G_2, \Xi_2, o_2)\} & \text{otherwise.} \end{cases} \tag{3.1}$$

We then define the distance between the rooted marked graphs as

$$d_{\text{loc}}((G_1, \Xi_1, o_1), (G_2, \Xi_2, o_2)) := 2^{-r_{\text{max}}} \in [0, 2]. \tag{3.2}$$

The distance  $d_{\text{loc}}$  is a metric on the *isomorphism classes* of  $\mathcal{G}_o(\mathcal{M})$ , which turns this space into a Polish space, i.e., a complete, separable metric space (see [1] or [30, Theorem A.6]).

**Local weak convergence of deterministic graphs.** Let  $(G_n, \Xi_n)_{n \in \mathbb{N}}, (G_n, \Xi_n) \in \mathcal{G}(\mathcal{M})$ , be a sequence of (deterministic) finite marked graphs such that  $|G_n| \rightarrow \infty$ . For each  $n$ , let  $U_n$  be a vertex of  $G_n$  chosen u.a.r., and consider the *measures* defined by  $(G_n, \Xi_n, U_n)$  on  $(\mathcal{G}_o(\mathcal{M}), d_{\text{loc}})$ . We will define the LWC of  $(G_n, \Xi_n)_{n \in \mathbb{N}}$  as the weak convergence of the above measures, which can be defined in the standard way. Let  $(\mathbb{R}, d_{\text{euc}})$  denote the Polish space of the real numbers equipped with the Euclidean distance, and introduce the set of test functionals

$$\Phi = \{\varphi : \mathcal{G}_o(\mathcal{M}) \rightarrow \mathbb{R} : \varphi \text{ is bounded and continuous}\}. \tag{3.3}$$

We remark that a special case of continuous functionals are those that only depend on a finite neighborhood of the root. We say that  $(G_n, \Xi_n, U_n)_{n \in \mathbb{N}}$  converges in the LWC sense to a (possibly random) element  $(G, \Xi, o) \in \mathcal{G}_o(\mathcal{M})$ , denoted by  $(G_n, \Xi_n, U_n) \xrightarrow{\text{loc}} (G, \Xi, o)$ , if for all  $\varphi \in \Phi$ , as  $n \rightarrow \infty$ ,

$$\mathbb{E}[\varphi(G_n, \Xi_n, U_n)] \rightarrow \mathbb{E}[\varphi(G, \Xi, o)]. \tag{3.4}$$

This statement is equivalent (see e.g. [30, Theorem 2.6]) to the convergence of neighborhood counts; that is, the following statement is an equivalent definition of LWC: for any  $r \in \mathbb{N}$  and any fixed  $(G', \Xi', o') \in \mathcal{G}_o(\mathcal{M})$ , as  $n \rightarrow \infty$ ,

$$\mathbb{P}(B_r(G_n, \Xi_n, U_n) \simeq B_r(G', \Xi', o')) \rightarrow \mathbb{P}(B_r(G, \Xi, o) \simeq B_r(G', \Xi', o')). \tag{3.5}$$

**Local weak convergence of random graphs.** We now generalize Definition 2.6 for marked graphs (simultaneously generalizing (3.5) for random graphs). Let  $(G_n, \Xi_n)_{n \in \mathbb{N}}, (G_n, \Xi_n) \in \mathcal{G}_o(\mathcal{M})$ , be a sequence of (finite) random marked graphs such that  $|G_n| \xrightarrow{\mathbb{P}} \infty$ , and let  $U_n | (G_n, \Xi_n) \sim \text{Unif}[\mathcal{V}(G_n)]$  be a uniformly chosen vertex. Let  $\mathbb{P}(\cdot | (G_n, \Xi_n))$  denote conditional probability with respect to the marked graph (i.e., the free variable is  $U_n$ ). We say that  $(G_n, \Xi_n, U_n)_{n \in \mathbb{N}}$  converges in probability in the local weak sense to a (possibly) random element  $(G, \Xi, o) \in \mathcal{G}_o(\mathcal{M})$ , and write  $(G_n, \Xi_n, U_n) \xrightarrow{\mathbb{P}\text{-loc}} (G, \Xi, o)$ , if the empirical neighborhood counts converge in probability, i.e., if for any fixed  $r \in \mathbb{N}$  and fixed  $(G', \Xi', o') \in \mathcal{G}_o(\mathcal{M})$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned} \mathbb{P}(B_r(G_n, \Xi_n, U_n) \simeq B_r(G', \Xi', o') | (G_n, \Xi_n)) &:= \frac{1}{|G_n|} \sum_{u \in \mathcal{V}(G_n)} \mathbb{1}_{\{B_r(G_n, \Xi_n, U_n) \simeq B_r(G', \Xi', o')\}} \\ &\xrightarrow{\mathbb{P}} \mathbb{P}(B_r(G, \Xi, o) \simeq B_r(G', \Xi', o')). \end{aligned} \tag{3.6}$$

We can also generalize (3.4) for an equivalent definition (again, see e.g. [30, Theorem 2.12] for a proof of the equivalence) of LWC in probability. Let  $\mathbb{E}[\cdot | (G_n, \Xi_n)]$  denote conditional expectation corresponding to the conditional probability measure  $\mathbb{P}(\cdot | (G_n, \Xi_n))$ . Then  $(G_n, \Xi_n, U_n) \xrightarrow{\mathbb{P}\text{-loc}} (G, \Xi, o)$  exactly when for all test functionals  $\varphi \in \Phi$  (see (3.3)),

$$\mathbb{E}[\varphi(B_r(G_n, \Xi_n, U_n)) | (G_n, \Xi_n)] \xrightarrow{\mathbb{P}} \mathbb{E}[\varphi(B_r(G, \Xi, o))]. \tag{3.7}$$

**Extensions.** We remark that there exist other notions of LWC for random graphs. *Almost sure* LWC can be defined by replacing the convergence in probability by almost sure convergence in (3.6). LWC *in distribution* is defined as

$$\mathbb{P}(B_r(G_n, \Xi_n, U_n) \simeq B_r(G', \Xi', o')) \rightarrow \mathbb{P}(B_r(G, \Xi, o) \simeq B_r(G', \Xi', o')), \tag{3.8}$$

where we note the lack of conditioning on the left-hand side. In this paper, we use LWC in probability, as it is not too restrictive while being strong enough to imply *asymptotic independence* of the neighborhoods of two uniformly chosen vertices. Such asymptotic independence is *not* guaranteed by LWC in distribution; see e.g. [30, Section 2.3.1] for a discussion of the differences between these notions.

**Remark 3.1** (Different root distributions) In certain cases, it is meaningful and interesting to study the convergence of subgraph counts around a vertex  $W_n$  chosen according to a non-uniform distribution, for example size-biased by degree or chosen within a (large enough) subset of vertices. Our motivation is to restrict the choice of the root to one partition of the BCM. In the following, for a random vertex  $W_n$  with an arbitrary distribution on  $\mathcal{V}(G_n)$ , we shall write  $(G_n, \Xi_n, W_n) \xrightarrow{\mathbb{P}\text{-loc}} (G, \Xi, o)$  to mean that the neighborhood counts around  $W_n$  converge, i.e., for all  $r \in \mathbb{N}$  and all  $(G', \Xi', o') \in \mathcal{E}_o(\mathcal{M})$ , as  $n \rightarrow \infty$ ,

$$\mathbb{P}(B_r(G_n, \Xi_n, W_n) \simeq B_r(G', \Xi', o') \mid (G_n, \Xi_n)) \xrightarrow{\mathbb{P}} \mathbb{P}(B_r(G, \Xi, o) \simeq B_r(G', \Xi', o')). \tag{3.9}$$

**3.2. Local weak convergence of the underlying bipartite configuration model**

Analogously to classical RIGs, we approach LWC of the RIGC via the LWC of the underlying BCM; see Definition 2.1 in Section 2.1. Without the community graphs assigned to each community and the edge labels encoding the assigned community roles, the LWC of the BCM follows from known results that we recall shortly. To formally state the LWC of the BCM, we first define the limiting object.

Recall  $D^\ell$  and  $D^r$  from Parts (A) and (C1), respectively, of Assumption 2.2, and also recall (1.1). We define a discrete-time branching process  $\text{BP}_\ell$  with a single root  $\underline{0}$  in generation 0. The offspring of any two individuals are independent but not identically distributed. The root has offspring  $D^\ell$ , every other individual in even generations has offspring distributed as  $\tilde{D}^\ell$ , and every individual in odd generations has offspring distributed as  $\tilde{D}^r$ . We denote the resulting family tree by  $(\text{BP}_\ell, \underline{0})$ , which we think of as a rooted (unmarked) tree: the family tree is obtained by adding edges between each individual and all its offspring. Then we have the following result.

**Proposition 3.2** Consider  $\text{BCM}_n = (\text{BCM}(\mathbf{d}^\ell, \mathbf{d}^r))$  under Assumption 2.2 (A,B,C1,D) and recall  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$ . Then, in the generalized meaning of the notion  $\xrightarrow{\mathbb{P}\text{-loc}}$  from Remark 3.1,

$$(\text{BCM}_n, V_n^\ell) \xrightarrow{\mathbb{P}\text{-loc}} (\text{BP}_\ell, \underline{0}). \tag{3.10}$$

*Proof.* The statement follows from the proof of [38, Theorem 3.1(iv)]. □

The referenced proof uses a very precise coupling argument of a breadth-first search exploration of the graph and an appropriate branching process. In an extended version of this paper [31], we provide our alternative proof, in which we apply a first and second moment method to the neighborhood counts.

By symmetry, a statement analogous to Proposition 3.2 holds for  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$ . Define the branching process  $\text{BP}_r$  analogously to  $\text{BP}_\ell$ , with the roles of  $\ell$  and  $r$  reversed. That is, the root  $\underline{0}$  has offspring  $D^r$ , every other vertex in even generations has offspring  $\tilde{D}^r$ , and every vertex in odd generations has offspring  $\tilde{D}^\ell$ . Then

$$(\text{BCM}_n, V_n^r) \xrightarrow{\mathbb{P}\text{-loc}} (\text{BP}_r, \underline{0}), \tag{3.11}$$

where  $\xrightarrow{\mathbb{P}\text{-loc}}$  is again meant in the generalized sense of Remark 3.1. From (3.10) and (3.11), we can also obtain the local weak limit of the BCM with a uniform root  $V_n^\delta \sim \text{Unif}[\mathcal{V}^\ell \cup \mathcal{V}^r]$ . We define a mixing variable  $\delta$  and a mixture of BP family trees  $\text{BP}_\delta$  as follows:

$$\mathbb{P}(\delta = \ell) = 1/(1 + \gamma), \quad \mathbb{P}(\delta = r) = \gamma/(1 + \gamma), \tag{3.12a}$$

$$(\text{BP}_\delta, \underline{0}) \stackrel{d}{=} \mathbb{1}_{\{\delta=\ell\}}(\text{BP}_\ell, \underline{0}) + \mathbb{1}_{\{\delta=r\}}(\text{BP}_r, \underline{0}). \tag{3.12b}$$

Then, by (3.10), (3.11), Definition 2.6, and Remark 3.1,

$$(\text{BCM}_n, V_n^\delta) \xrightarrow{\mathbb{P}\text{-loc}} (\text{BP}_\delta, \underline{0}), \tag{3.13}$$

where  $\xrightarrow{\mathbb{P}\text{-loc}}$  now applies in the original sense of Definition 2.6.

### 3.3. Local weak convergence of the random intersection graph with communities

In this section, we prove the LWC of the RIGC. We start by constructing the limiting object  $(\text{CP}, o)$ . The notation is inspired by the fact that  $(\text{CP}, o)$  is the ‘community projection’ (see Section 2.1) of a random rooted marked tree  $(\text{BP}_\ell, \Xi^P, \underline{0})$  defined below, in the same way that the RIGC is the ‘community projection’ of the underlying BCM. It is then not surprising that  $(\text{BP}_\ell, \Xi^P, \underline{0})$  is the local weak limit of the *underlying* BCM, including the community graphs, which we represent as a marked graph (formally introduced below). The limiting object  $(\text{BP}_\ell, \Xi^P, \underline{0})$  is obtained from the BP family tree  $(\text{BP}_\ell, \underline{0})$ , introduced in Section 3.2, by equipping it with a mark function  $\Xi^P$  defined shortly. In the following, we give a formal definition of these objects, starting with the marked graph representation of the underlying BCM.

**The underlying BCM as a marked graph.** We introduce the mark function  $\Xi^c$  on BCM to represent the community graphs and the assignment of community roles. Recall the set of possible community graphs  $\mathcal{H}$  and the ‘no mark’ symbol  $\emptyset$ . Let the set of marks be  $\mathcal{M}^P := \mathcal{H} \cup \mathbb{Z}^+ \cup \{\emptyset, \ell\}$ . We mark each  $v \in \mathcal{V}^\ell$  by  $\ell :=: \Xi^c(v)$  and each  $a \in \mathcal{V}^r$  by its community graph  $\text{Com}_a :=: \Xi^c(a)$ . Recall that an edge of the underlying BCM formed by  $\ell$ -half-edge  $(v, i)$  and  $r$ -half-edge  $(a, l)$  is labeled by  $(i, l)$ ; we also mark this edge by the tuple  $(i, l)$ . We refer to  $(\text{BCM}, \Xi^c)$  as the *community-marked BCM*, and we note that it encodes all information necessary for constructing the RIGC. Indeed, the community graphs are given as the marks of  $r$ -vertices, and edge-marks encode the assigned community roles: if  $\ell$ -vertex  $v$  is connected to  $r$ -vertex  $a$  by an edge marked  $(i, l)$ , we know that  $v$  takes on the community role of the vertex with label  $l$  in  $\text{Com}_a$ . Thus the community projection operator of Section 2.1 can be naturally redefined as  $\widehat{\mathcal{P}} : (\text{BCM}, \Xi^c) \mapsto \text{RIGC}$ . For some  $v \in \mathcal{V}^\ell$ , we write  $\widehat{\mathcal{P}} : (\text{BCM}, \Xi^c, v) \mapsto (\text{RIGC}, v)$  for the rooted version of the projection.

**The local weak limit of the underlying BCM.** We now introduce the marked BP family tree  $(\text{BP}_\ell, \Xi^P, \underline{0})$ . Recall  $(\text{BP}_\ell, \underline{0})$  from Section 3.2; conditionally on this (possibly infinite) BP family tree, we now define the random mark function  $\Xi^P$ , using the set of marks  $\mathcal{M}^P$  from

above. We mark vertices in even generations by  $\ell$  and vertices in odd generations by a random  $H \in \mathcal{H}$ , determined as follows. Recall the family of conditional measures  $(\mu_{\cdot|k})_{k \in \mathbb{Z}^+}$  from (2.40), and let us denote the degree of  $a \in \mathcal{V}(\text{BP}_\ell)$  by  $d_a$ . Independently of everything else, we mark  $a$  according to the measure  $\mu_{\cdot|d_a}$ . This determines the marks on the vertices of  $\text{BP}_\ell$ . Now we describe the marks on the edges of  $\text{BP}_\ell$ . We mark each edge  $e$  by a random tuple  $(i, l) \in (\mathbb{Z}^+)^2$ , and we determine  $i$  and  $l$  separately. (In particular, the two numbers in each tuple marking an edge are independent of *each other*, but are *not* independent of everything else.) Denote the endpoint of  $e$  in an even generation by  $v$  and the endpoint in an odd generation by  $a$ , where  $v$  and  $a$  intuitively correspond to an  $\ell$ - and  $r$ -vertex, respectively. We think of  $i$  and  $l$  as the marks of the  $\ell$ - and  $r$ -half-edges incident to  $v$  and  $a$ , respectively. For any vertex  $u$ , we mark families of half-edges incident to  $u$  in a dependent way, so that each mark in  $[d_u]$  is used once, but independently of all other families. For  $u \neq \underline{0}$ , we first mark the half-edge that is part of the edge connecting  $u$  to its parent, by a uniformly chosen mark  $K \sim \text{Unif}[d_u]$ . Since  $(\text{BP}_\ell, \underline{0})$  is a BP family tree, we may assume the children of each individual are ordered, which provides an ordering for those half-edges incident to  $u$  that are part of edges connecting  $u$  to its children. We mark these half-edges by the remaining marks  $[d_u] \setminus \{K\}$  in increasing order. For the root  $\underline{0}$ , we mark all its half-edges by  $[d_0]$  in increasing order, analogously.

This defines the law of  $\Xi^p$  conditionally on  $(\text{BP}_\ell, \underline{0})$ , and consequently the joint law  $(\text{BP}_\ell, \Xi^p, \underline{0})$ .

**Proposition 3.3** *Consider the underlying BCM represented as a marked graph  $(\text{BCM}_n, \Xi^c)$  under Assumption 2.2, and recall that  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$ . Then, in the generalized sense of  $\xrightarrow{\mathbb{P}\text{-loc}}$  from Remark 3.1,*

$$(\text{BCM}_n, \Xi^c, V_n^\ell) \xrightarrow{\mathbb{P}\text{-loc}} (\text{BP}_\ell, \Xi^p, \underline{0}). \tag{3.14}$$

*Proof.* The statement follows analogously to Proposition 3.2, by generalizing the proof of [38, Theorem 3.1(iv)]. The breadth-first search exploration of the marked graph can be coupled to the marked branching process, by extending the coupling of the degrees to include the marks. More specifically, we have to couple the following three quantities in the BP to the corresponding ones in the graph exploration:  $D^\ell$  for the root; the joint distribution of  $\tilde{D}^\ell$  and  $K$  for other vertices in even generations; the joint distribution of  $\tilde{D}^r$ ,  $H$ , and  $K$  for vertices in odd generations. By the empirical convergence assumed in Assumption 2.2, the error of this coupling vanishes as  $n \rightarrow \infty$ . We omit further details.  $\square$

**The local weak limit of the RIGC.** Recall that in representing the underlying BCM as a marked graph, we also reinterpreted the community projection  $\hat{\mathcal{P}} : (\text{BCM}, \Xi^c, v) \mapsto (\text{RIGC}, v)$  acting on a rooted, suitably marked bipartite graph. (Note that this operation is well-defined even when this bipartite graph is infinite, as each community graph is inserted by a local operation.) Hence we can define  $(\text{CP}, o)$  as the community-projection  $\hat{\mathcal{P}}$  of  $(\text{BP}_\ell, \Xi^p, \underline{0})$  defined above, analogously to random clique trees in [38]. It follows from the construction that  $(\text{CP}, o)$  is a simple, locally finite rooted graph with countable (possibly infinite) vertex set  $\mathcal{V}(\text{CP}) = \{v \in \text{BP}_\ell, v \text{ is in an even generation}\}$ . We obtain the following insight on the overlapping structure of the communities: each vertex  $v \in \mathcal{V}(\text{CP})$  is part of exactly  $d_v$  communities; however, by the tree structure of  $\text{BP}_\ell$ , any two of these communities share only  $v$  as a common vertex, i.e., the proposed local weak limit CP has the *single-overlap* property.

We now prove the LWC.

*Proof of Theorem 2.7.* We show that Theorem 2.7 follows from Proposition 3.3. For some  $r \in \mathbb{N}$  and  $v \in \mathcal{V}^\ell$ , consider the neighborhood  $B_r(\text{RIGC}, v)$  (see Definition 2.5). We first argue

that this neighborhood is fully determined by  $B_{2r+1}(\text{BCM}, \Xi^c, v)$ . We show that any vertex in  $B_r(\text{RIGC}, v)$  and any community containing an edge in  $B_r(\text{RIGC}, v)$  is at distance at most  $2r + 1$  from  $v$  in the underlying BCM. Consider a vertex  $u$  or an edge  $e$  in  $B_r(\text{RIGC}, v)$ , and construct the shortest possible chain of incident edges and vertices connecting it to  $v$ : either  $v = v_0, e_0, v_1, e_1, \dots, e_{k-1}, v_k = u$ , or  $v = v_0, e_0, v_1, e_1, \dots, e_{k-1}, v_k, e_k = e$ . (The length of a chain is the total number of elements in it.) The chain is the longest possible when we consider an edge  $e$  between vertices that are both at distance  $r$  from  $v$ . In this case, the shortest chain is  $v = v_0, e_0, v_1, e_1, \dots, e_{k-1}, v_r, e_r = e$ , with length  $2r + 2$ . Let  $a_i$  denote the community that  $e_i$  is part of; then the endpoints  $v_i$  and  $v_{i+1}$  must be members of community  $a_i$ . Hence, in the underlying BCM,  $v_i$  and  $v_{i+1}$  are both neighbors of  $a_i$ . That is, vertices that are neighbors in the RIGC are second neighbors in the BCM. Hence we have a (not necessarily self-avoiding) path  $v = v_0, a_0, v_1, a_1, \dots, a_{k-1}, v_k = u$  or  $v = v_0, a_0, v_1, a_1, \dots, a_{k-1}, v_k, a_k \ni e_k$  in the underlying BCM. Because of the correspondence between  $e_i$  and  $a_i$ , the number of  $\beta$ -vertices in the path in the BCM is the same as the length of the chain in the RIGC. We have already established that this length is at most  $2r + 2$ ; hence the number of edges along the path in the BCM, i.e., the graph distance between its endpoints, is at most  $2r + 1$ . This implies that, indeed,  $B_{2r+1}(\text{BCM}, \Xi^c, v)$  fully determines  $B_r(\text{RIGC}, v)$ .

From now on, we consider the graph sequences  $\text{BCM}_n$  and  $\text{RIGC}_n$ , and we use a tightness argument to prove that the convergence of neighborhood frequencies of  $B_{2r+1}(\text{BCM}_n, \Xi^c, V_n^\ell)$  implies the convergence of neighborhood frequencies of  $B_r(\text{RIGC}, V_n^\ell)$ . We now establish this tightness and an appropriate truncation, so that we can rely on a finite subset of possible neighborhoods in our proof. Theorem 3.3 implies that degrees are tight in the random graph sequence  $B_r(\text{BCM}_n, \Xi^c, V_n^\ell)$ , as follows. Recall the limit  $(\text{BP}_\ell, \Xi^p, \underline{Q})$  from Section 3.2. By construction, all degrees in  $\text{BP}_\ell$  are distributed as either  $D^\ell, \tilde{D}^\ell$ , or  $\tilde{D}^r$  (see (1.1) and Assumption 2.2(A,C1)). By Assumption 2.2, these random variables are almost surely finite; hence the maximal degree in  $B_r(\text{BP}_\ell, \Xi^p, \underline{Q})$  is almost surely finite. Note that for any  $K \in \mathbb{Z}^+$ , the functional  $\mathbb{1}_{\{\text{maximal degree in } r\text{-ball} > K\}}$  is bounded and continuous in the metric space  $(\mathcal{G}_o, d_{\text{loc}})$  (see Section 3.1). Thus by (3.4) and Proposition 3.3,

$$\begin{aligned} &\mathbb{P}(\max\{\beta\text{-deg}(v) : v \in B_r(\text{BCM}_n, \Xi^c, V_n^\ell)\} > K \mid \omega_n) \\ &\xrightarrow{\mathbb{P}} \mathbb{P}(\max\{\beta\text{-deg}(v) : v \in B_r(\text{BP}_\ell, \Xi^p, \underline{Q})\} > K), \end{aligned} \tag{3.15}$$

where the right-hand side vanishes as  $K \rightarrow \infty$ , as we have shown that the maximal degree in the  $r$ -ball is almost surely finite. Hence for any  $\varepsilon > 0$ , there exists  $K = K(\varepsilon, r) \in \mathbb{Z}^+$  such that

$$\mathbb{P}(\max\{\beta\text{-deg}(v) : v \in B_{2r+1}(\text{BP}_\ell, \Xi^p, \underline{Q})\} > K) < \varepsilon/6, \tag{3.16a}$$

$$\mathbb{P}(\max\{\beta\text{-deg}(v) : v \in B_{2r+1}(\text{BCM}_n, \Xi^c, V_n^\ell)\} > K \mid \omega_n) < \varepsilon/3 \quad \text{w.h.p.} \tag{3.16b}$$

Consider some  $(H, o) \in \mathcal{G}_o$  that is a possible outcome of  $B_r(\text{RIGC}, V_n^\ell)$ . Denote by  $\mathcal{X}$  the set of all possible outcomes  $(G, \Xi_G, o)$  of the random graph  $B_{2r+1}(\text{BCM}_n, \Xi^c, V_n^\ell)$ . Further, denote by  $\mathcal{Y}$  the subset of all  $(G, \Xi_G, o) \in \mathcal{X}$  such that the  $r$ -ball in their  $\widehat{\mathcal{P}}$ -projection is  $(H, o)$ . That is,  $B_r(\text{RIGC}, V_n^\ell) \simeq (H, o)$  exactly when  $B_{2r+1}(\text{BCM}, \Xi^c, V_n^\ell) \simeq (G, \Xi_G, o)$  for some  $(G, \Xi_G, o) \in \mathcal{Y}$ . Similarly,  $B_r(\text{CP}, o) \simeq (H, o)$  exactly when  $B_{2r+1}(\text{BP}_\ell, \Xi^p, \underline{Q}) \simeq (G, \Xi_G, o)$  for some  $(G, \Xi_G, o) \in \mathcal{Y}$ . Thus, to prove Theorem 2.7, it is sufficient to show that

$$\begin{aligned} &\sum_{(G, \Xi_G, o) \in \mathcal{Y}} \mathbb{P}(B_{2r+1}(\text{BCM}, \Xi^c, V_n^\ell) \simeq (G, \Xi_G, o) \mid \omega_n) \\ &\xrightarrow{\mathbb{P}} \sum_{(G, \Xi_G, o) \in \mathcal{Y}} \mathbb{P}(B_{2r+1}(\text{BP}_\ell, \Xi^p, \underline{Q}) \simeq (G, \Xi_G, o)). \end{aligned} \tag{3.17}$$

We prove this using the above truncation. With  $K = K(\varepsilon, r)$  above, denote the subset of elements in  $\mathcal{Y}$  where the maximal degree is at most  $K$  by

$$\mathcal{Y}(\leq K) := \left\{ (G, \Xi_G, o) : \max\{\delta\text{-deg}(v) : v \in (G, \Xi_G, o)\} \leq K \right\}, \tag{3.18}$$

and let  $\mathcal{Y}(>K) := \mathcal{Y} \setminus \mathcal{Y}(\leq K)$ . Define  $\mathcal{X}(\leq K)$  and  $\mathcal{X}(>K)$  analogously: all elements of  $\mathcal{X}$  where the maximal degree is at most  $K$  and larger than  $K$ , respectively. Note that  $\mathcal{Y}(\leq K) \subseteq \mathcal{X}(\leq K)$  is *finite*; hence by Proposition 3.3, the sum over this finite set converges:

$$\begin{aligned} & \sum_{(G, \Xi_G, o) \in \mathcal{Y}(\leq K)} \mathbb{P}\left(B_{2r+1}(\text{BCM}, \Xi^c, V_n^\ell) \simeq (G, \Xi_G, o) \mid \omega_n\right) \\ & \xrightarrow{\mathbb{P}} \sum_{(G, \Xi_G, o) \in \mathcal{Y}(\leq K)} \mathbb{P}\left(B_{2r+1}(\text{BP}_\ell, \Xi^P, \underline{Q}) \simeq (G, \Xi_G, o)\right). \end{aligned} \tag{3.19}$$

We bound the tail of the left-hand side of (3.17) as

$$\begin{aligned} & \sum_{(G, \Xi_G, o) \in \mathcal{Y}(>K)} \mathbb{P}\left(B_{2r+1}(\text{BCM}, \Xi^c, V_n^\ell) \simeq (G, \Xi_G, o) \mid \omega_n\right) \\ & \leq \sum_{(G, \Xi_G, o) \in \mathcal{X}(>K)} \mathbb{P}\left(B_{2r+1}(\text{BCM}, \Xi^c, V_n^\ell) \simeq (G, \Xi_G, o) \mid \omega_n\right) \\ & = \mathbb{P}(\max\{\delta\text{-deg}(v) : v \in B_{2r+1}(\text{BCM}_n, \Xi^c, V_n^\ell)\} > K \mid \omega_n) < \varepsilon/3 \quad \text{w.h.p.,} \end{aligned} \tag{3.20}$$

by (3.16b). Analogously, the tail of the right-hand side of (3.17) can be bounded as

$$\begin{aligned} & \sum_{(G, \Xi_G, o) \in \mathcal{Y}(>K)} \mathbb{P}\left(B_{2r+1}(\text{BP}_\ell, \Xi^P, \underline{Q}) \simeq (G, \Xi_G, o)\right) \\ & \leq \sum_{(G, \Xi_G, o) \in \mathcal{X}(>K)} \mathbb{P}\left(B_{2r+1}(\text{BP}_\ell, \Xi^P, \underline{Q}) \simeq (G, \Xi_G, o)\right) \\ & = \mathbb{P}(\max\{\delta\text{-deg}(v) : v \in B_{2r+1}(\text{BP}_\ell, \Xi^P, \underline{Q})\} > K) < \varepsilon/6, \end{aligned} \tag{3.21}$$

by (3.16a). Note that (3.19) implies that the difference between the finite sums is at most  $\varepsilon/2$  w.h.p. Combining this observation with (3.20) and (3.21) via the triangle inequality implies that (3.17) holds. Since we have previously reduced Theorem 2.7 to this statement, this concludes the proof of Theorem 2.7. □

### 3.4. Proofs of further results on the random intersection graph with communities

We now provide the proofs of our results on the local properties of the RIGC model as consequences of LWC. Namely, we prove the convergence of the degree and local clustering coefficient and study the overlapping structure in Sections 3.4.1 and 3.4.2, respectively.

3.4.1 *Degrees and clustering.* Recall the definition of  $(\text{CP}, o)$ , the local weak limit of the RIGC, as the  $\widehat{\mathcal{P}}$ -projection of  $(\text{BP}_\ell, \Xi^P, \underline{Q})$  from Section 3.3. By this construction, it is clear that  $D^P$  (see (2.20)) and  $\zeta$  (see (2.25)) respectively describe the degree and local clustering coefficient of  $o \in \text{CP}$ . Further recall the empirical degree  $D_n^P$  (see (2.11)–(2.12)) and empirical local clustering  $\zeta_n$  (see (2.24)–(2.25)). Since

$$(\text{RIGC}_n, V_n^\ell) \xrightarrow{\mathbb{P}\text{-loc}} (\text{CP}, o),$$



it is intuitive that  $D_n^p \xrightarrow{d} D^p$  and  $\zeta_n \xrightarrow{d} \zeta$ . We complete the formal proof of the stronger statements (2.21) and (2.26) below.

*Proof of Corollaries 2.8 and 2.9.* We show that *pointwise* convergence of the empirical CDFs follows directly from Theorem 2.7, by the following reasoning. Recall that  $\mathbb{P}(\cdot | \omega_n)$  denotes conditional probability with respect to  $\omega_n$  and  $\mathbb{E}[\cdot | \omega_n]$  denotes the corresponding conditional expectation. Further, denote by  $\mathbb{P}_o$  and  $\mathbb{E}_o$  the probability measure of  $(\text{CP}, o)$  and the corresponding expectation. For arbitrary *fixed*  $x \in \mathbb{R}$ , we define the functionals

$$\varphi_x, \psi_x : \mathcal{G}_o \rightarrow \{0, 1\}, \quad \varphi_x(G, o) := \mathbb{1}_{\{\text{deg}(o) \leq x\}}, \quad \psi_x(G, o) := \mathbb{1}_{\{\text{Cl}(o) \leq x\}}. \tag{3.22}$$

Clearly, both functionals are bounded and only depend on a finite neighborhood of  $o$ ; thus they are continuous in the metric space  $(\mathcal{G}_o, d_{\text{loc}})$  (see Section 3.1). Note that we can express the (empirical) CDFs from (2.12), (2.20), (2.24), and (2.25), respectively, as

$$\begin{aligned} F_n^p(x) &= \mathbb{E}[\varphi_x(\text{RIGC}_n, V_n^\ell) | \omega_n], & F^p(x) &= \mathbb{E}_o[\varphi_x(\text{CP}, o)], \\ F_n^\zeta(x) &= \mathbb{E}[\psi_x(\text{RIGC}_n, V_n^\ell) | \omega_n], & F^\zeta(x) &= \mathbb{E}_o[\psi_x(\text{CP}, o)]. \end{aligned} \tag{3.23}$$

Theorem 2.7 asserts that  $(\text{RIGC}_n, V_n^\ell) \xrightarrow{\mathbb{P}\text{-loc}} (\text{CP}, o)$ ; thus, using the equivalent definition of LWC in probability (3.7), for any *fixed*  $x \in \mathbb{R}$ , as  $n \rightarrow \infty$ ,

$$F_n^p(x) \xrightarrow{\mathbb{P}} F^p(x), \quad F_n^\zeta(x) \xrightarrow{\mathbb{P}} F^\zeta(x). \tag{3.24}$$

That is, indeed, the empirical CDF of degrees and local clustering coefficient converge in probability pointwise. This can be strengthened to the convergence in sup-norm in (2.21) and (2.26) by a truncation argument. The details can be found in the extended version of this paper [31]. □

**3.4.2. The overlapping structure.** In this section, we prove Proposition 2.11 and Theorem 2.12 on the typical number and size of overlaps in the RIGC model. Theorem 2.12(i)–(ii) follows easily from Theorem 2.7 by counting 4-cycles through the root. For a detailed argument, we refer the interested reader to the extended version of this paper [31]. We do provide the proof of Proposition 2.11 and Theorem 2.12(iii), which rely on the additional second moment condition (2.35) and require a slightly different approach. We make use of the following notation. Recall that  $V_n^\ell \sim \text{Unif}[\mathcal{V}^\ell]$ ,  $V_n^r \sim \text{Unif}[\mathcal{V}^r]$ , and  $V_n^b \sim \text{Unif}[\mathcal{V}^\ell \cup \mathcal{V}^r]$ . Further recall that  $\mathbb{P}(\cdot | \omega_n)$  denotes conditional probability with respect to  $\omega_n$  (i.e., conditionally on the graph realization), and  $\mathbb{E}[\cdot | \omega_n]$  denotes the corresponding conditional expectation (i.e., partial average over the choice of the uniform vertex).

*Sketch of proof of Proposition 2.11.* As before, we reduce Proposition 2.11 to LWC. Thus, we define the functional  $\varphi$  on  $\mathcal{G}_o$  (see Section 3.1 for the notation) that counts the number of vertices at graph distance 2 from the root; i.e., for  $(G, o) \in \mathcal{G}_o$ ,

$$\varphi(G, o) := |\partial B_2(G, o)|. \tag{3.25}$$

Recall (2.32)–(2.33). We can rewrite the left-hand side of (2.36) as

$$\frac{2|\mathcal{L}_1|}{M_n} = \mathbb{E}[|\mathcal{N}(V_n^r)| | \omega_n] = \mathbb{E}[\varphi(\text{BCM}_n, V_n^r) | \omega_n]. \tag{3.26}$$

Recall  $(BP_r, \underline{0})$  from Section 3.2 and note that

$$\mathbb{E}[\varphi(BP_r, \underline{0})] = \mathbb{E}[D^r] \mathbb{E}[\tilde{D}^\ell], \tag{3.27}$$

which is exactly the proposed limit of (3.26). The proof can now be completed by applying Proposition 3.3 and (3.4) as before; however, some technical difficulties arise. Namely, the functional  $\varphi$  is *not* bounded, so a truncation argument is necessary. The technicalities of this truncation argument are included in the extended version of this paper [31]. This concludes the proof of Proposition 2.11.  $\square$

*Proof of Theorem 2.12(iii).* Recall  $\mathbb{O}(a, b)$  and  $\mathcal{L}_k$  from (2.32)–(2.34). By Proposition 2.11,  $|\mathcal{L}_1|$  is of order  $M_n$ ; thus, to show that  $|\mathcal{L}_2|/|\mathcal{L}_1| = o_{\mathbb{P}}(1)$ , it is sufficient to prove that  $|\mathcal{L}_2| = o_{\mathbb{P}}(M_n)$ , which we carry out via a first moment method. We compute

$$2 \mathbb{E}[|\mathcal{L}_2|] = \mathbb{E}\left[ \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b}} \mathbb{1}_{\{\mathbb{O}(a, b) \geq 2\}} \right] = \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b}} \mathbb{P}(\mathbb{O}(a, b) \geq 2). \tag{3.28}$$

With some  $K$  to be chosen later, we split the sum

$$\sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b}} \mathbb{P}(\mathbb{O}(a, b) \geq 2) = \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b \\ d_a^r \leq K}} \mathbb{P}(\mathbb{O}(a, b) \geq 2) + \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b \\ d_a^r > K}} \mathbb{P}(\mathbb{O}(a, b) \geq 2). \tag{3.29}$$

We start by bounding the first term. Recall that  $v \leftarrow \text{Com}_a$  denotes the event that  $v$  takes a community role in  $\text{Com}_a$ . For individuals  $v_1, \dots, v_k$  and communities  $a_1, \dots, a_l$ , denote the event that all  $k$  individuals are in all  $l$  communities by

$$\{v_1, \dots, v_k\} \xleftarrow{\otimes} \{\text{Com}_{a_1}, \dots, \text{Com}_{a_l}\} := \bigcap_{i \leq k} \bigcap_{j \leq l} \{v_i \leftarrow \text{Com}_{a_j}\}. \tag{3.30}$$

Further recall  $d_v^\ell$  and  $d_a^r$  from Section 2.1 and  $\mathfrak{h}_n$  from (2.1). By the union bound, we obtain

$$\begin{aligned} \mathbb{P}(\mathbb{O}(a, b) \geq 2) &= \mathbb{P}(\exists v, w \in \mathcal{V}^\ell, v < w : \{v, w\} \xleftarrow{\otimes} \{\text{Com}_a, \text{Com}_b\}) \\ &\leq \frac{1}{2} \sum_{\substack{v, w \in \mathcal{V}^\ell \\ v \neq w}} \mathbb{P}(\{v, w\} \xleftarrow{\otimes} \{\text{Com}_a, \text{Com}_b\}) \\ &\leq \sum_{\substack{v, w \in \mathcal{V}^\ell \\ v \neq w}} \frac{d_a^r (d_a^r - 1) d_b^r (d_b^r - 1) d_v^\ell (d_v^\ell - 1) d_w^\ell (d_w^\ell - 1)}{2 \cdot \mathfrak{h}_n (\mathfrak{h}_n - 1) (\mathfrak{h}_n - 2) (\mathfrak{h}_n - 3)} \end{aligned} \tag{3.31}$$

by counting the suitable versus the available choices of half-edges. Using (1.1), (2.3), and the fact that  $\mathfrak{h}_n = \mathbb{E}[D_n^\ell] N_n$  by Remark 2.3(i), we have

$$\sum_{v \in \mathcal{V}^\ell} \frac{d_v^\ell (d_v^\ell - 1)}{\mathfrak{h}_n} = \frac{1}{N_n} \sum_{v \in \mathcal{V}^\ell} \frac{d_v^\ell (d_v^\ell - 1)}{\mathbb{E}[D_n^\ell]} = \frac{\mathbb{E}[D_n^\ell (D_n^\ell - 1)]}{\mathbb{E}[D_n^\ell]} = \mathbb{E}[\tilde{D}_n^\ell]. \tag{3.32}$$

Since  $\hat{\ell}_n \rightarrow \infty$  as  $n \rightarrow \infty$ , we have that  $2\hat{\ell}_n(\hat{\ell}_n - 1)(\hat{\ell}_n - 2)(\hat{\ell}_n - 3) \geq \hat{\ell}_n^4$  for  $n$  large enough; thus, combining (3.31)–(3.32), we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{O}(a, b) \geq 2) &\leq \frac{d_a^r(d_a^r - 1)d_b^r(d_b^r - 1)}{\hat{\ell}_n^2} \sum_{v, w \in \mathcal{V}^\ell} \frac{d_v^\ell(d_v^\ell - 1)}{\hat{\ell}_n} \frac{d_w^\ell(d_w^\ell - 1)}{\hat{\ell}_n} \\ &\leq \frac{d_a^r(d_a^r - 1)d_b^r(d_b^r - 1)}{\hat{\ell}_n^2} (\mathbb{E}[\tilde{D}_n^\ell])^2. \end{aligned} \tag{3.33}$$

Then, using the condition  $d_a^r \leq K$ , the definition of  $d_{\max}^r$  from Remark 2.3(iii), and the fact that  $\hat{\ell}_n = \sum_{b \in \mathcal{V}^r} d_b^r$  by definition, we have

$$\begin{aligned} \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b \\ d_a^r \leq K}} \mathbb{P}(\mathcal{O}(a, b) \geq 2) &\leq (\mathbb{E}[\tilde{D}_n^\ell])^2 \sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b \\ d_a^r \leq K}} \frac{d_a^r(d_a^r - 1)d_b^r(d_b^r - 1)}{\hat{\ell}_n^2} \\ &< (\mathbb{E}[\tilde{D}_n^\ell])^2 K^2 M_n \sum_{b \in \mathcal{V}^r} \frac{d_b^r(d_{\max}^r - 1)}{\hat{\ell}_n^2} \leq (\mathbb{E}[\tilde{D}_n^\ell])^2 K^2 M_n \frac{d_{\max}^r}{\hat{\ell}_n}. \end{aligned} \tag{3.34}$$

We continue by bounding the second term in (3.29), where  $d_a^r > K$ . Using Markov’s inequality, we obtain an alternative bound for the probability

$$\mathbb{P}(\mathcal{O}(a, b) \geq 2) \leq \mathbb{E}[\mathcal{O}(a, b)]/2. \tag{3.35}$$

Taking expectation in (2.32) and again using (3.32), we have

$$\mathbb{E}[\mathcal{O}(a, b)] = \sum_{v \in \mathcal{V}^\ell} \mathbb{P}(v \overset{\otimes}{\leftarrow} \{\text{Com}_a, \text{Com}_b\}) \leq \sum_{v \in \mathcal{V}^\ell} \frac{d_a^r d_v^\ell (d_v^\ell - 1) d_b^r}{\hat{\ell}_n (\hat{\ell}_n - 1)} = \frac{d_a^r d_b^r}{\hat{\ell}_n - 1} \mathbb{E}[\tilde{D}_n^\ell]. \tag{3.36}$$

Combining (3.35)–(3.36), and using that  $\sum_{b \in \mathcal{V}^r} d_b^r = \hat{\ell}_n \leq 2(\hat{\ell}_n - 1)$  for  $n$  large enough, we get

$$\sum_{\substack{a, b \in \mathcal{V}^r \\ a \neq b \\ d_a^r > K}} \mathbb{P}(\mathcal{O}(a, b) \geq 2) \leq \frac{\mathbb{E}[\tilde{D}_n^\ell]}{2} \sum_{b \in \mathcal{V}^r} \frac{d_b^r}{\hat{\ell}_n - 1} \sum_{\substack{a \in \mathcal{V}^r \\ d_a^r > K}} d_a^r \leq \mathbb{E}[\tilde{D}_n^\ell] \sum_{a \in \mathcal{V}^r} d_a^r \mathbb{1}_{\{d_a^r > K\}}. \tag{3.37}$$

Combining (3.29), (3.34), and (3.37), we have

$$\begin{aligned} \frac{2 \mathbb{E}[\mathcal{L}_2]}{M_n} &\leq (\mathbb{E}[\tilde{D}_n^\ell])^2 K^2 \frac{d_{\max}^r}{\hat{\ell}_n} + \mathbb{E}[\tilde{D}_n^\ell] \frac{1}{M_n} \sum_{a \in \mathcal{V}^r} d_a^r \mathbb{1}_{\{d_a^r > K\}} \\ &= (\mathbb{E}[\tilde{D}_n^\ell])^2 K^2 \frac{d_{\max}^r}{\hat{\ell}_n} + \mathbb{E}[\tilde{D}_n^\ell] \mathbb{E}[D_n^r \mathbb{1}_{\{D_n^r > K\}}]. \end{aligned} \tag{3.38}$$

We show that  $\mathbb{E}[\mathcal{L}_2]/M_n \rightarrow 0$  by showing that it can be made arbitrarily small for  $n$  large enough. Fix an arbitrary  $\varepsilon > 0$ ; we will choose first  $K$ , then  $n$ , so that the obtained upper bound is smaller than  $\varepsilon$ . Under the second moment condition (2.35),  $\mathbb{E}[\tilde{D}_n^\ell] \rightarrow \mathbb{E}[\tilde{D}^\ell] < \infty$ ;

thus  $(\mathbb{E}[\tilde{D}_n^\ell])_{n \in \mathbb{N}}$  is bounded. By Assumption 2.2(D),  $(D_n^r)_{n \in \mathbb{N}}$  is uniformly integrable; thus we can choose  $K = K(\varepsilon)$  large enough so that for all  $n$  large enough,

$$\mathbb{E}[\tilde{D}_n^\ell] \mathbb{E}[D_n^r \mathbb{1}_{\{D_n^r > K\}}] \leq \varepsilon. \quad (3.39)$$

Again using that  $(\mathbb{E}[\tilde{D}_n^\ell])_{n \in \mathbb{N}}$  is bounded, and further that  $K$  is now fixed and  $d_{\max}^r/\ell_n \rightarrow 0$  by Remark 2.3(iii), we conclude that for  $n$  large enough,

$$(\mathbb{E}[\tilde{D}_n^\ell])^2 K^2 \frac{d_{\max}^r}{\ell_n} \leq \varepsilon. \quad (3.40)$$

Therefore, for  $n$  large enough,  $\mathbb{E}[|\mathcal{L}_2|]/M_n \leq \varepsilon$ , which is equivalent to  $\mathbb{E}[|\mathcal{L}_2|] = o(M_n)$ . By Markov's inequality,  $|\mathcal{L}_2| = o_{\mathbb{P}}(M_n)$ , which combined with Proposition 2.11 implies  $|\mathcal{L}_2|/|\mathcal{L}_1| = o_{\mathbb{P}}(1)$ . This concludes the proof of Theorem 2.12(iii).  $\square$

### Acknowledgements

This work is supported by the Netherlands Organisation for Scientific Research (NWO) through the VICI grant 639.033.806 (R. v. d. H.), the VENI grant 639.031.447 (J. K.), the Gravitation Networks grant 024.002.003 (R. v. d. H.), and the TOP grant 613.001.451 (V. V.). V. V. thanks Lorenzo Federico and Clara Stegehuis for helpful discussions throughout the project.

### References

- [1] ALDOUS, D. AND STEELE, J. M. (2004). The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on Discrete Structures* (Encyclopaedia Math. Sci., Vol. 110), Springer, Berlin, pp. 1–72.
- [2] BALL, F., SIRL, D. AND TRAPMAN, P. (2009). Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Prob.* **41**, 765–796.
- [3] BALL, F., SIRL, D. AND TRAPMAN, P. (2010). Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Math. Biosci.* **224**, 53–73.
- [4] BENJAMINI, I., LYONS, R. AND SCHRAMM, O. (2015). Unimodular random trees. *Ergod. Theory Dynam. Systems* **35**, 359–373.
- [5] BENJAMINI, I. AND SCHRAMM, O. (2001). Recurrence of distributional limits of finite planar graphs. *Electron. J. Prob.* **6**, paper no. 23, 13 pp.
- [6] BERGER, N., BORGS, C., CHAYES, J. T. AND SABERI, A. (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *Ann. Prob.* **42**, 1–40.
- [7] BLACKBURN, S. R. AND GERKE, S. (2009). Connectivity of the uniform random intersection graph. *Discrete Math.* **309**, 5130–5140.
- [8] BLOZNELIS, M. (2010). Component evolution in general random intersection graphs. *SIAM J. Discrete Math.* **24**, 639–654.
- [9] BLOZNELIS, M. (2013). Degree and clustering coefficient in sparse random intersection graphs. *Ann. Appl. Prob.* **23**, 1254–1289.
- [10] BLOZNELIS, M. (2017). Degree–degree distribution in a power law random intersection graph with clustering. *Internet Math.* Available at <https://doi.org/10.24166/im.03.2017>.
- [11] BLOZNELIS, M. AND DAMARACKAS, J. (2013). Degree distribution of an inhomogeneous random intersection graph. *Electron. J. Combinatorics* **20**, paper no. 3, 13 pp.
- [12] BLOZNELIS, M. *et al.* (2015). Recent progress in complex network analysis: properties of random intersection graphs. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Springer, Heidelberg, pp. 79–88.
- [13] BLOZNELIS, M., JAWORSKI, J. AND KURAUSKAS, V. (2013). Assortativity and clustering of sparse random intersection graphs. *Electron. J. Prob.* **18**, paper no. 38, 24 pp.
- [14] BOLLOBÁS, B. (2001). *Random Graphs*, 2nd edn. Cambridge University Press.
- [15] BRITTON, T., DEIJFEN, M. AND MARTIN-LÖF, A. (2006). Generating simple random graphs with prescribed degree distribution. *J. Statist. Phys.* **124**, 1377–1397.

- [16] CHEN, N. AND OLVERA-CRAVIOTO, M. (2013). Directed random graphs with given degree distributions. *Stoch. Systems* **3**, 147–186.
- [17] COUPECHOUX, E. AND LELARGE, M. (2015). Contagions in random networks with overlapping communities. *Adv. Appl. Prob.* **47**, 973–988.
- [18] DEIJFEN, M. AND KETS, W. (2009). Random intersection graphs with tunable degree distribution and clustering. *Prob. Eng. Inf. Sci.* **23**, 661–674.
- [19] DERÉNYI, I., PALLA, G. AND VICSEK, T. (2005). Clique percolation in random networks. *Phys. Rev. Lett.* **94**, 160202.
- [20] DURRETT, R. (2007). *Random Graph Dynamics*. Cambridge University Press.
- [21] FILL, J. A., SCHEINERMAN, E. R. AND SINGER-COHEN, K. B. (2000). Random intersection graphs when  $m = \omega(n)$ : an equivalence theorem relating the evolution of the  $G(n, m, p)$  and  $G(n, p)$  models. *Random Structures Algorithms* **16**, 156–176.
- [22] FORTUNATO, S. (2010). Community detection in graphs. *Phys. Rep.* **486**, 75–174.
- [23] FORTUNATO, S. AND HRIC, D. (2016). Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44.
- [24] GARAVAGLIA, A., VAN DER HOFSTAD, R. AND LITVAK, N. (2020). Local weak convergence for PageRank. *Ann. Appl. Prob.* **30**, 40–79.
- [25] GIRVAN, M. AND NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Nat. Acad. Sci. USA* **99**, 7821–7826.
- [26] GODEHARDT, E. AND JAWORSKI, J. (2003). Two models of random intersection graphs for classification. In *Exploratory Data Analysis in Empirical Research*, Springer, Berlin, Heidelberg, pp. 67–81.
- [27] GUILLAUME, J.-L. AND LATAPY, M. (2004). Bipartite structure of all complex networks. *Inf. Process. Lett.* **90**, 215–221.
- [28] GUILLAUME, J.-L. AND LATAPY, M. (2006). Bipartite graphs as models of complex networks. *Physica A* **371**, 795–813.
- [29] VAN DER HOFSTAD, R. (2017). *Random Graphs and Complex Networks*, Vol. 1. Cambridge University Press.
- [30] VAN DER HOFSTAD, R. (2020+). *Random Graphs and Complex Networks*, Vol. 2. In preparation. Available at <http://www.win.tue.nl/rhofstad/NotesRGCNII.pdf>. 16 November 2020 version referenced.
- [31] VAN DER HOFSTAD, R., KOMJÁTHY, J. AND VADON, V. (2018). *Random intersection graphs with communities, extended version*. Preprint. Available at <https://arxiv.org/abs/1809.02514>.
- [32] VAN DER HOFSTAD, R., KOMJÁTHY, J. AND VADON, V. (2019). *Phase transition in random intersection graphs with communities*. Preprint. Available at <https://arxiv.org/abs/1905.06253>.
- [33] VAN DER HOFSTAD, R., VAN LEEUWAARDEN, J. S. H. AND STEGEHUIS, C. (2016). Hierarchical configuration model. *Internet Math.* Available at <https://doi.org/10.24166/im.01.2017>.
- [34] VAN DER HOFSTAD, R., VAN LEEUWAARDEN, J. S. H. AND STEGEHUIS, C. (2016). Power-law relations in random networks with communities. *Phys. Rev. E* **94**, 012302.
- [35] JANSON, S., ŁUCZAK, T. AND RUCIŃSKI, A. (2000). *Random Graphs*. John Wiley, New York.
- [36] KARJALAINEN, J., VAN LEEUWAARDEN, J. S. H. AND LESKELÄ, L. (2018). Parameter estimators of sparse random intersection graphs with thinned communities. In *International Workshop on Algorithms and Models for the Web-Graph (WAW 2018)*, Springer, Cham, pp. 44–58.
- [37] KARONSKI, M., SCHEINERMAN, E. R. AND SINGER-COHEN, K. B. (1999). On random intersection graphs: the subgraph problem. *Combinatorics Prob. Comput.* **8**, 131–159.
- [38] KURAUSKAS, V. (2015). *On local weak limit and subgraph counts for sparse random graphs*. Preprint. Available at <https://arxiv.org/abs/1504.08103>.
- [39] NEWMAN, M. E. J. (2003). Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121.
- [40] NEWMAN, M. E. J. (2010). *Networks*. Oxford University Press.
- [41] NORROS, I. AND REITTU, H. (2006). On a conditionally Poissonian graph process. *Adv. Appl. Prob.* **38**, 59–75.
- [42] PALLA, G., DERÉNYI, I., FARKAS, I. AND VICSEK, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818.
- [43] RYBARCZYK, K. (2011). Diameter, connectivity, and phase transition of the uniform random intersection graph. *Discrete Math.* **311**, 1998–2019.
- [44] SINGER, K. B. (1996). *Random intersection graphs*. Doctoral Thesis, Johns Hopkins University.
- [45] VADON, V., KOMJÁTHY, J. AND VAN DER HOFSTAD, R. (2019). A new model for overlapping communities with arbitrary internal structure. *Appl. Network Sci.* **4**, article no. 42.