Vadon *et al. Applied Network Science* (2019) 4:42
https://doi.org/10.1007/s41109-019-0149-9

Applied Network Science

RESEARCH Open Access

A new model for overlapping communities with arbitrary internal structure

Viktória Vadon*, Júlia Komjáthy and Remco van der Hofstad

*Correspondence: v.vadon@tue.nl
Technology University of Eindhoven, Department M&CS, Eindhoven, The Netherlands

Abstract

We introduce the *random intersection graph with communities*, a new model for networks with overlapping communities with arbitrary internal structure. We construct the model from a list of arbitrary community graphs that are the building blocks, and a separate list of individuals, each with a prescribed number of community membership tokens. Randomness is introduced by matching these tokens uniformly at random to vertices of the community graphs. We then identify the community members assigned to the same individual, thus overlaps arise due to individuals having several tokens. This gives a highly flexible model for networks with community structure.

We are able to derive a wide range of analytic results on this model. We derive an asymptotic description of the local structure of the graph, which further yields the asymptotic degree distribution, local clustering coefficient, and results on the overlapping structure of the communities. For the global connectivity structure, we identify a phase transition in the size of the largest component. When the largest component constitutes a positive proportion of the graph, we can further characterize its asymptotic local structure. Finally, we study how the connectivity structure changes under a randomized attack, where we remove edges randomly, according to independent coin flips.

Keywords: Random networks, Community structure, Overlapping communities, Local weak convergence, Phase transition, Percolation, Bipartite configuration model

AMS Subject Classification: Primary 05C80; 60C05; 90B15; 05C82

Introduction

Network science is an active and quickly developing field, due to joint efforts from practitioners and theoretical researchers. Empirical studies allow us to understand how real-life networks work in action, explore their defining features and build models based on our findings. Theoretical studies allow us to make predictions or approximations when data analysis is not feasible, and refine our understanding of the causality relations between properties of the network. As such, both sides provide their invaluable contributions, facilitating the progress of one another.

We introduce a new random graph model (for an introduction to random graphs, see (Bollobás 2001; van der Hofstad 2017; 2018+; Janson et al. 2000; Newman 2010)) with the aim to model networks with communities. The aim of this paper is to bring this model to the attention of the community of network practitioners. As such, we present the model and the available results in an intuitive way, focusing on examples, special cases

and possible applications, and comparisons to existing models. For a mathematically precise definition of our model as well as the rigorous proofs of the results discussed here, the reader is referred to (van der Hofstad et al. 2018; 2019) (preprints). Data analysis and comparison to real-world networks are an interesting avenue for future research.

Our model is based on two key features, that we introduce shortly. The choice was inspired by social networks, that serve as our primary motivation, which will also reflect in the terminology used. However, our model is more widely applicable. It has been shown (Guillaume and Latapy 2004; 2006) that many other real-life networks also exhibit a community structure, more specifically, an underlying structure of 'groups' and 'elements' that are part of these groups.

First, we need to specify what we mean by a community. With such a broad definition as 'a more densely interconnected part of the network', several different practical meanings have been attached to this word. In this context, we think of communities as the building blocks of the network, as in the household model (Ball et al. 2009; 2010) or hierarchical configuration model (Stegehuis et al. 2016a; van der Hofstad et al. 2016; Stegehuis et al. 2016b). These communities may be families, workplaces, people with shared hobbies, etc. We represent each one as an arbitrary 'small' graph; small means that the *average* community size is finite, however an individual community, such as people working for a large company, may be large.

The first key feature we want to model is that communities may overlap, i.e., the same individual may be part of several communities. This is natural in the context of a social network: people have their families, colleagues, and possibly several groups of friends from school or from their different hobbies. Such networks are usually modeled by variants of the random intersection graph (RIG) (Bloznelis 2010; 2013; Deijfen and Kets 2009; Godehardt and Jaworski 2003; Karonski et al. 1999; Newman 2003a; Rybarczyk 2011; Singer 1996; Yağan 2016; Yağan and Makowski 2012) (see (Bloznelis et al. 2015) for an overview of the topic), inspired by collaboration networks. However, the random intersection graph has the shortcoming that any two persons within the same community are assumed to be acquainted. Considering a large company, this is unrealistic, as for example the CEO will not know each and every office worker.

The second key feature is that each community has its own *arbitrary, prescribed* internal structure. This removes the assumption that any two members within the same community are acquainted, and allows for the differentiation of roles within a community. Efforts have been made to remove the restriction of complete graphs (Karjalainen et al. 2018; Newman 2003a), however, the only model known to the authors incorporating *arbitrary* communities as building blocks is the hierarchical configuration model (HCM) (Stegehuis et al. 2016a, 2016b; van der Hofstad et al. 2016). The HCM randomizes connections *between* communities, and consequently has the shortcoming that each person can only be part of one community, i.e. communities do not overlap.

To the best knowledge of the authors, the first model to combine the above two key features is the *random intersection graph with communities* (RIGC) that we introduce here. The RIGC fills a gap in the literature to serve as a null model for networks with communities that may overlap and have their own arbitrary internal structure at the same time.

The building blocks of the network are arbitrary small graphs, and they are combined through randomness. Such a model can be used to model networks where local

structure is well-defined, but global structure is much more fluid. In our example of a social network, the inner structure or working of a group is determined by its purpose, but which groups a person is part of is determined by chance encounters. On the large scale, the effect of adding microscopic structures in the form of communities becomes negligible, and macroscopic effects are governed by the added randomness.

In addition to modeling vital aspects of networks with communities, the RIGC is analytically tractable. In the following, we introduce the exact setup of the model, and present some of the available analytic results.

Introduction to the model

In this section, we explain how to generate the random graph from two given lists: a list of individuals and a list of communities.

In the model, each community is represented by a 'small' graph, in the sense described above. Each vertex of these small graphs represents a unique *community role*, and connections between community members are represented by edges. These graphs may be *arbitrary*, as long as they are connected, allowing for different applications. For example, we may assume that each member has a bounded number of connections independently of the community size, which means that the larger the communities grow, the sparser they become. Modeling a network of computer networks may call for specific topologies, such as grids or hierarchical structures. Last but not least, one can take real-life network data as input. We note that the given list may contain several communities with the same structure, for example, each community that is a three-member family would be represented by a triangle.

We also assume the number of community memberships of each individual is prescribed, and give the individual as many *membership tokens*. Intuitively speaking, each community membership represents an identity that the person has, possibly including both offline identities such as a job or family role, and online identities such as an email address or a social media account. We assume that the number of community memberships of each individual is at least one, otherwise we could remove that individual from the model. Overlaps are induced by individuals who are part of more than one community. Necessarily, the total number of tokens given to all individuals has to equal the total number of community roles available, that is, the total size of all given community graphs.

The upcoming construction can be followed step by step in Fig. 1. The graph is randomized by matching the community roles with the membership tokens *uniformly at random*, in a one-to-one fashion. This means that each possible outcome has equal probability to be chosen, and this probability is one over the total number of possibilities. Such a matching can be generated sequentially: in each step, we pick an arbitrary community role, and match it with a membership token chosen uniformly at random. We then remove both objects from the available pool, and repeat the procedure until the available pool becomes empty.

Alternatively, we can pick an arbitrary membership token, and match it with a community role chosen uniformly at random; or may even arbitrarily pick one of these two methods in each step. This in particular allows us to build the network from the point of view of an individual, first finding the communities that the individual is part of, then finding further individuals that are part of these communities, and so on.

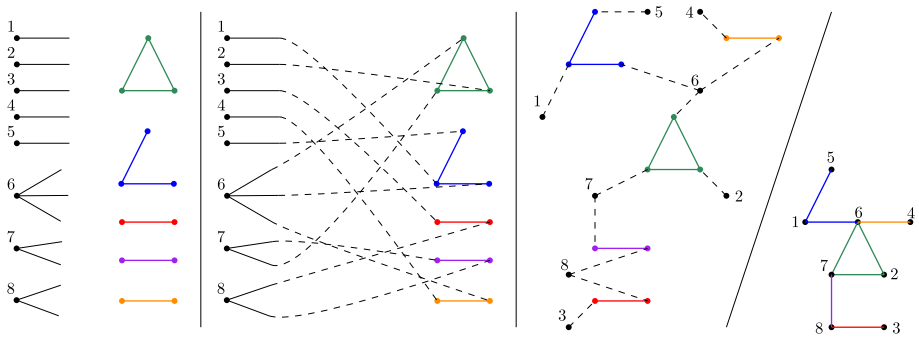

Fig. 1 A small example for constructing the RIGC model. From left to right, we see each step of the construction. Leftmost, we represent the list of community graphs in different colors, and the individuals by the numbered vertices; the number of stubs drawn for each individual corresponds to the number of membership tokens. The second subfigure represents the matching, with the imaginary edges drawn dashed. An imaginary edge between an individual and a vertex in a community means that one of the tokens of that individual was matched with that community role. The third subfigure is simply a redrawing of the second for the sake of clarity. Rightmost, we see the resulting RIGC, after the imaginary edges have been contracted

Visually, we can represent the matching by imaginary edges (see the second picture in Fig. 1): each imaginary edge represents a membership token and a community role that are matched, and connects a vertex representing an individual to a vertex in a community graph. That is, the community roles taken by an individual will be connected to the individual by imaginary edges. An individual has as many imaginary edges as their prescribed number of tokens, and since each community role is matched with exactly one token, each is incident to exactly one imaginary edge. We note that this is not yet the random network that we are looking for, but an important intermediate step. It reflects that the only source of randomness is the matching, represented by the imaginary edges, and is a great tool throughout the analysis.

Finally, we obtain the random network by identifying each individual with the set of community roles that this individual takes. As these community roles, i.e., vertices in communities, are connected to the vertex representing the individual by imaginary edges, we can achieve this by contracting all imaginary edges (see the last picture in Fig. 1).

Note that in a random matching, it might occur that we match two or more tokens of the same individual to community roles within the same community. However, this is quite unlikely. As we increase the network size, the probability for a typical individual to take several roles within any community vanishes. Also note that this may also happen in the real world, as some people may have separate work and private accounts for the same service. It may also happen that two individuals are present in several communities together, and are connected by two or more edges in the RIGC, which potentially makes the graph a *multigraph*. Again, we can argue that this is not unrealistic, and can be interpreted as a stronger interaction between those two individuals. Additionally, this phenomenon also becomes rare as the network size grows, allowing us to *study the network 'as it is,' without deleting duplicate edges and self-loops*. Since the difference before and after deleting duplicate edges becomes negligible in the large-network limit, our asymptotic results apply to both variants of the model.

As hinted before, we study the model in the *large-network limit*, that is, when the number of individuals tends to infinity. In this limit, we derive asymptotic results,

that in turn serve as crude approximations for finite, but large, network sizes. (The convergence takes place in abstract graph and distribution spaces, thus the validation of these approximations is up to future empirical research.) To allow us to compare networks of different sizes, we assume some 'consistency' of the parameters, (the list of communities and the list of individuals with their number of community memberships):

Assumption 1 (The parameters) *We assume that in the large-network limit, i.e., as the number of individuals* $N \to \infty$*,*

a) *the number of communities equals* $M = \gamma N + \varepsilon(N)$ *for some constant* γ*, and* $\varepsilon(N)$ *is an error term such that* $\varepsilon(N)/N \to 0$*;*
b) *for every positive integer k, the proportion of individuals with k membership tokens converges to some* $p_k = \mathbb{P}(T = k) \geq 0$*, where T is a random variable with finite mean* $\mathbb{E}[T] < \infty$*. Intuitively, T represents the asymptotic distribution of the number of membership tokens, arising as the limit of the empirical distributions;*
c) *for every finite connected graph H, the relative frequency of H in the list of communities converges to some* $\mu_H \geq 0$*, such that the collection* (μ_H) *over finite connected graphs H is a probability mass function. Thus, the proportion of communities with size k, i.e., k members, converges to some* $q_k = \mathbb{P}(S = k) \geq 0$*, where S is a random variable with finite mean* $\mathbb{E}[S] < \infty$*. Intuitively, S describes the asymptotic distribution of community sizes.*

Note that there necessarily is a relation between γ, $\mathbb{E}[T]$ and $\mathbb{E}[S]$. The reason is that the total number of membership tokens must be equal to the total size of the communities: $N(\mathbb{E}[T] + \varepsilon_1) = M \cdot (\mathbb{E}[S] + \varepsilon_2)$, where the error terms $\varepsilon_1 \to 0$ and $\varepsilon_2 \to 0$ appear because the empirical averages may be slightly different from their respective limits. Consequently, $\gamma = \mathbb{E}[T] / \mathbb{E}[S]$.

A good incentive for working with pre-assigned numbers of community membership tokens for each individual is that this method allows us complete control over the distribution of T. An alternative approach inspired by the random intersection graph literature (in particular, the "active and passive" model) (Bloznelis 2010; 2013; Godehardt and Jaworski 2003; Rybarczyk 2011) assigns each community role to a uniformly chosen individual. While this approach makes generating the model easier, it has the disadvantage that T always follows a Poisson distribution, which is light-tailed. This means that the number of community memberships per individual has very small variability, which may be an undesirable property. Certain applications may call for heavy-tailed or power-law distributions (more on power laws later), where *hubs* appear: individuals that are part of a large number of communities (up to N^α, for some $1/2 < \alpha < 1$). The existence of such hubs is often crucial to the connectivity structure of the network or in making it a small world, or even an ultra-small world (see (Newman (2003b), Section 3.1) for an introduction to small and ultra-small worlds and further references). It is thus advantageous to use membership tokens, as this method allows for greater generality, including both light-tailed and heavy-tailed distributions for T.

Results

Next, we introduce our analytic results on the RIGC model. First, we focus on local properties, that is, behavior in the neighborhood of a 'typical' vertex. This includes

basic properties such as the degree distribution and local clustering, and model-specific properties about the overlapping structure of communities. We then move to global properties, that is, structural properties of the network. In particular, we study its connected components, and how resilient the largest component is under randomized attack to the network.

Degrees and clustering

The upcoming sections are based on (van der Hofstad et al. 2018). Degrees, i.e., the number of connections an individual has, contain crucial information about the local structure of a network, and are easy to measure in real-world networks. For several random graph models, it has been proven that the degree distribution can make a difference between a small-world graph, where distances scale as $\log N$ in a graph of size N, and an ultra-small-world graph, where distances grow even slower than $\log N$, often scaling as $\log\log N$ (see (van der Hofstad 2017), Section 1.4.3) for a rigorous definition); scale-free networks (defined later) are often ultra small (Bollobás and Riordan 2003b; Cohen and Havlin 2003). While there is no direct relation between degrees and distances, when these predictions fail, that indicates the presence of further structure in the network. Thus the degree distribution often serves as the first benchmark when tuning a model to fit a network.

Thus, the first property of the RIGC model that we study is the empirical degree distribution, which is random itself. There are two ways to represent this empirical distribution. The conventional representation is the sequence of proportions (n_k/N), where N denotes the total number of individuals, and n_k denotes the random number of individuals with degree k, for all positive integers k. However, in this case working with a random sequence is not convenient. Alternatively, we can represent this empirical distribution with a single random variable. Choose a vertex V uniformly at random, and denote its degree by D_V. Then, given a realization of the graph, the probability for D_V to equal k is the probability of picking a vertex with degree k, which is exactly n_k/N. Thus understanding the distribution of D_V is the same as understanding the random sequence (n_k/N).

Note that in a random graph, there are two sources of randomness in D_V: the choice of the uniform individual, and the choice of the graph itself, and these are independent. In fact, these two sources of randomness help us describe the asymptotic distribution of D_V. Choosing an individual randomly means that it has a random number of membership tokens. For each token, a random community role is assigned, which adds a certain number of connections to V from that community. To formalize this, we introduce the random variable Γ, such that $\mathbb{P}(\Gamma = c)$ is the limit of the proportion of community roles that have c neighbors within their community graph. Recall the random variable T that describes the limiting distribution of the number of membership tokens.

Theorem 1 (Limiting degree distribution) *Let $\Gamma_i, i \geq 1$ be random variables independent of each other and of T, with the same distribution as Γ defined above. Then, for every $k \geq 1$, in the large-network limit $N \to \infty$,*

$$\mathbb{P}(D_V = k) \to \mathbb{P}\left(\sum_{i=1}^{T} \Gamma_i = k\right).$$

We define the random variable D_0 with the limiting distribution $\mathbb{P}(D_0 = k) := \mathbb{P}\left(\sum_{i=1}^{T} \Gamma_i = k\right)$. The proof of Theorem 1 relies on a more in-depth description of the neighborhood of a uniformly chosen individual that we intuitively discuss in the next section.

We discuss when the RIGC showcases a special type of degree distribution that has been observed (Newman 2005) in real-world networks: *power laws*. We say that D_0 follows a power law with exponent $\delta > 1$, if $\mathbb{P}(D_0 > x)$ scales as $x^{1-\delta}$, that is, $cx^{1-\delta} < \mathbb{P}(D_0 > x) < Cx^{1-\delta}$ for some constants $0 < c < C$ and x large enough. Power laws are heavy-tailed distributions, meaning that higher moments are infinite. The precise threshold is $\delta - 1$ (which may be a fractional moment): for $s \geq \delta - 1, \mathbb{E}\left[D_0^s\right] = \infty$, but for any $s < \delta - 1, \mathbb{E}\left[D_0^s\right] < \infty$. That is, the smaller δ is, the "heavier" the tail is, and the fewer moments exist. If for a random variable, the tail $\mathbb{P}(D_0 > x)$ decays faster than any power of x, we call the distribution light-tailed. In this case, all moments are finite, and we define the "power-law exponent" of such a distribution as infinity.

If the empirical degree distribution of a network of size N follows a power law with exponent δ, the maximal degree scales as $N^{1/(\delta-1)}$. For the infinite-variance case $\delta < 3$, including the *scale-free* case $2 < \delta < 3$ when the mean is finite but the variance is infinite, this means that the largest degree is much larger than $N^{1/2}$. Such highly-connected vertices, called hubs, heavily influence the structure of the network, making the study of power-law and particularly scale-free degrees important. We identify a sufficient condition for the RIGC to have power-law degrees. Recall T, the asymptotic distribution of the number of membership tokens, and Γ, the asymptotic distribution of within-community neighbors.

Corollary 1 (Power laws) *If T or Γ (or both) follow a power-law distribution, then the asymptotic degree distribution D_0 also follows a power law, and its exponent is the smaller of the two exponents; intuitively speaking, the "heavier tail wins".*

When the asymptotic average degree $\mathbb{E}[D_0]$ is finite, the number of edges in the graph scales as $\mathbb{E}[D_0]N/2$, that is, the RIGC is sparse. This happens exactly when $\mathbb{E}[\Gamma]$ is finite.

Next we study clustering, also known as transitivity, in the RIGC model. It has been observed (see e.g. Watts and Strogatz (1998)) that real-world networks often contain significantly more triangles than traditional random graphs with the same degree distribution. In social networks, this can be explained by the fact that people who share a common friend are more likely to meet each other, through this common friend, than other people in the network. The *local clustering coefficient* aims to capture this. For a vertex v with degree $D_v \geq 2$, its local clustering coefficient $\mathrm{Cl}(v)$ is the proportion of pairs of neighbors of v that are also directly connected. This is the same as the number of triangles that v is part of, divided by $\binom{D_v}{2}$, the total number of pairs of neighbors (if the vertex has less than two neighbors, we set $\mathrm{Cl}(v) = 0$). We describe the network by the average local clustering coefficient, which is the average of local clustering coefficients of each vertex.

We study the random empirical distribution of the local clustering coefficient in the RIGC. Again, we represent the random empirical distribution by a single random variable, with two sources of randomness. Recall that V denotes an individual chosen uniformly at random, and consider its local clustering coefficient $\mathrm{Cl}(V)$. With this representation, the average local clustering is $\mathbb{E}[\mathrm{Cl}(V)]$.

We continue to describe the distribution of $\mathrm{Cl}(V)$. Again, we recognize that due to the uniform choice of V, it has a random number of membership tokens. Due to the random matching, we match these tokens to a set of community roles chosen uniformly at random, and each community role is part of a certain number of triangles within its community. We find and show that the number of otherwise arising triangles is negligible, due to the structure of neighborhoods explained in the next section.

To formalize the result, we introduce the random variable Δ, such that $\mathbb{P}(\Delta = d)$ is the limit of the proportion of community roles that are part of d triangles within their own community. Note that a community role with c within-community neighbors is part of at most $\binom{c}{2}$ triangles. Thus, with the random variable Γ that describes the limiting distribution of within-community connections, (Γ, Δ) form a random vector with *dependent* coordinates. Recall that the random variable T describes the asymptotic distribution of the number of membership tokens.

Theorem 2 (Asymptotic local clustering) *Let* (Γ_i, Δ_i), $i \geq 1$ *be independent copies of the dependent random vector* (Γ, Δ), *also independent of T. For any* $x \in [0, 1]$, *in the large-network limit* $N \to \infty$,

$$\mathbb{P}(\mathrm{Cl}(V) \leq x) \to \mathbb{P}\left(\frac{\sum_{i=1}^{T} \Delta_i}{\binom{\sum_{i=1}^{T} \Gamma_i}{2}} \leq x\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{T} \Delta_i}{\binom{D_0}{2}} \leq x\right) =: \mathbb{P}(\mathrm{Cl}(0) \leq x),$$

where we introduce the random variable $\mathrm{Cl}(0)$ *to describe the limiting distribution. The average local clustering also converges: in the large-network limit* $N \to \infty$,

$$\mathbb{E}[\,\mathrm{Cl}(V)] \to \mathbb{E}[\,\mathrm{Cl}(0)]\,.$$

For many real-life networks, the average local clustering coefficient is positive (Watts and Strogatz 1998). It is thus a desirable property for network models to have positive asymptotic clustering, i.e., a positive limit of average clustering as the network size $N \to \infty$. The simplest random graph models, the Erdős-Rényi random graph (Erdős and Rényi 1959; Gilbert 1959) and the configuration model (Bollobás 1980; Molloy and Reed 1995) have vanishing clustering, i.e., the average clustering goes to 0 as the network size grows. On the other hand, classical random intersections graphs and the hierarchical configuration model (with adequate parameters) produce positive asymptotic clustering. This raises interest in where the RIGC falls on this scale.

Corollary 2 (Condition for positive asymptotic clustering) *The RIGC produces positive asymptotic clustering exactly when there is a positive asymptotic proportion of community graphs that contain one or more triangles.*

It has also been observed (see e.g. Vázquez et al. (2002)) in real-world networks that local clustering of a vertex scales inversely with the degree of the vertex. If all communities are complete graphs, i.e., the special case of the traditional random intersection graph, and sufficient regularity (the asymptotic community size S has finite variance), the RIGC reproduces this property. This is also true for other variants of the classical random intersection graph (generalized RIG) (Bloznelis 2013). However, this is a non-trivial open question in the general case, when the communities are arbitrary.

Neighborhoods and overlapping structure

In this section, we provide a more in-depth description of neighborhoods in the graph. As demonstrated by our results on degrees and (local) clustering, this is useful in deriving various properties of the graph. However, such neighborhoods are also of independent interest, to gain insight into the structure of the graph, or to measure similarity of graphs. The asymptotics of the local neighborhood structure can be conveniently described using *local weak convergence* (Benjamini and Schramm 2001). We provide a brief, intuitive description here; for details and rigor of applying this notion to the RIGC, the reader is referred to (van der Hofstad et al. (2018), Sections 2.2, 4.1).

Our aim is to describe the asymptotic behavior of the neighborhood of a typical vertex in the large-network limit, and we do so by considering finite neighborhoods (see Fig. 2 for a small example) of an individual chosen uniformly at random. To understand such neighborhoods, we now recall how we construct the random graph, and in particular, a neighborhood. The list of individuals, each with a given number of membership tokens, and the list of community graphs are given. Randomness solely comes from matching the membership tokens with community roles uniformly at random. We can match these two types of objects sequentially, and in each step, we can arbitrarily pick an unmatched object of either type, as long as its match is chosen uniformly at random from the unmatched objects of the other type.

As we are in a random graph, we can *explore* the neighborhood of an individual by *building* it, according to the construction rules of the random graph. Given an individual whose neighborhood we want to explore, we distinguish it as the *root*, and we start by matching each of its membership tokens (in some arbitrary order) to community roles chosen uniformly at random. To focus on the larger-scale structure, we enclose each community by a *supervertex*, and preserve the community graph itself as a *decoration* of the supervertex. Further, we give the supervertex as many distinguished *community role*

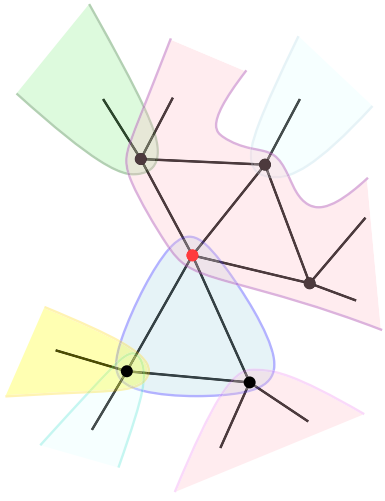

Fig. 2 A neighborhood in the RIGC model. We demonstrate a small neighborhood of the red vertex in the middle, cut off before the second neighbors; the colorful backgrounds illustrate different communities. The figure sheds light on the local structure of the model: note that the red vertex is part of two communities, which consequently overlap with each other. Through these communities, we reach the neighbors of the red vertex. Through the neighbors, we may again enter new communities, and so on

tokens as the size of the enclosed community. In our *structural graph*, the supervertices representing communities that the root is part of become direct neighbors of the root and form *layer 1*.

Next, we explore all further individuals in the communities of layer 1; we match each community role (in some arbitrary order) to a uniformly chosen membership token. These membership tokens belong to some individuals, and these individuals form *layer 2*. Matching the remaining membership tokens of the layer 2 individuals, we reach another set of communities, that become the supervertices of layer 3, and so on. Every odd layer contains supervertices that represent communities first entered through individuals in the previous layer; every even layer contains individuals first found in communities in the previous layer. In the structural graph that we build this way, edges represent a match between a membership token and a community role token, thus they correspond to imaginary edges from the construction.

This structural graph is *locally tree-like*: up to a finite number of layers, it is less and less likely to see a cycle as the network size grows. (Note that this holds on the structural level only; cycles contained in the community graphs are preserved in the network.) The reason behind this phenomenon is that we match membership tokens and community roles *uniformly* at random, thus the probability of choosing from a set is proportional to its size. The community sizes and the number of tokens are small relative to the network size, thus the set of tokens within a finite number of layers becomes negligible compared to the total pool that increases with the network size. Consequently, re-connecting to an already seen individual or supervertex and thus creating a cycle becomes unlikely.

This observation suggests that in the limit, the neighborhood of a typical vertex becomes a decorated random *tree*, where the tree describes the membership relations of individuals and communities, and the decorations describe the communities. In the following, we explain the law of this decorated tree, which is in fact a branching process (see (Athreya and Ney 2004) for an introduction to branching processes). For our purposes here, it is sufficient to think of it as a random tree that is constructed layer by layer from a root, and each vertex has an *independent* random number of *offspring* in the next layer.

The root corresponds to a individual chosen uniformly at random, thus the number of its offspring is drawn from the asymptotic distribution of tokens T. However, the offspring distribution of further layers is affected by the fact that we match the *membership tokens* and *community roles* uniformly at random. Thus, a community with size k is k times more likely to be chosen than a community of size 1. This effect is called *size-biasing*. Also note that a community of size k produces $k-1$ individuals as offspring, since one of its members is in the previous layer.

Thus, for the asymptotic distribution of community sizes S, we introduce its size-biased version $\widetilde{S}$ with mass function $\mathbb{P}\left(\widetilde{S}=k-1\right)=k\mathbb{P}(S=k)/\mathbb{E}[S]$. Similarly for the asymptotic number of tokens T, we define $\widetilde{T}$ with mass function $\mathbb{P}\left(\widetilde{T}=k-1\right)=k\mathbb{P}(T=k)/\mathbb{E}[T]$. This describes the random tree, and we decorate it as follows. Given that a community has size k (offspring $k-1$), its community graph is chosen with probabilities proportional to the limiting frequencies μ_H (see Assumption 1), for connected graphs H on k vertices. That is, the probability of choosing a particular graph H on k vertices equals $\mu_H/\mathbb{P}(S=k)$.

Above, we have described a decorated branching process, which is the limit of the structural graph. We now explain how to obtain the limit of the RIGC, and we refer to this

step as the *projection*. First, we 'blow up' the supervertices into the communities they represent. This graph corresponds to the intermediate step in the construction, where the matching is represented by imaginary edges; the role of the imaginary edges is now taken by the edges of the random tree. Thus, the second step is to contract the original edges of the random tree, and we obtain the limit of the network, from the perspective of a typical individual.

Next, we discuss how the locally tree-like nature of the structural graph impacts the overlapping structure of communities. We say that two communities overlap, or are neighbors, when they share one or more individuals; the size of the overlap is the number of individuals shared.

Theorem 3 (Single-overlap property) *The number of overlapping pairs of communities scales linearly with the network size N, that is, a community chosen uniformly at random overlaps with constantly many others.*
The "typical" overlap between overlapping communities is a single individual.

More precisely, with probability tending to 1 in the large-network limit $N \to \infty$:

i) an individual chosen uniformly at random is not part of an overlap larger than 1, i.e., it is not part of several communities together with any other individual;
ii) a community chosen uniformly at random does not overlap in more than one individual with any other community.

This means that overlaps of size 2 or higher may still occur, but their number is sublinear in the network size N, and consequently also sublinear in the number of communities M, as well as the total number of overlaps.

The single-overlap property allows for community detection with good accuracy in a special case, inspired by clique-percolation (Derényi et al. 2005). We say that a graph is triangle-connected, if we can connect any two vertices by a chain of triangles such that each consecutive pair shares an edge. (This is a special case of k-clique-connected graphs for $k = 3$.) Assume that all community graphs are triangle-connected; this is a slightly stronger condition than assuming that the graph is connected and each vertex is part of at least one triangle. In fact, such a graph can be constructed by adding triangles one by one so that each new triangle shares an edge with one of the old ones. This construction can serve as a greedy algorithm to find such a community. If an arbitrary community only shares overlaps of size 1 with any neighboring community, it cannot share an edge with them, and thus our greedy algorithm will detect the boundaries of this community exactly. Considering the RIGC with triangle-connected communities, with probability tending to 1, the graph realization is "nice" and only a sublinear proportion of communities share an overlap of 2 or larger. Thus only a sublinear proportion of communities are misdetected.

Connected components

The upcoming sections are based on (van der Hofstad et al. 2019). We move onto studying the global structure of the model, in particular, the *connected components*. A connected component represents a set of individuals that are able to communicate, interact and influence one another; it is thus a natural question to ask how large these interacting sets are.

First, we observe that the internal structure of the communities does not have any influence on the global connectivity structure. The reason behind is that we have assumed each community to be connected in itself, thus it provides a path between any two community members, even if they are not directly connected. Indeed, our results below only depend on the community sizes and the distribution of membership tokens of individuals.

We find that component sizes are closely related to the decorated branching process, introduced in the section above (p. 10). Recall that this branching process captures the local structure of the graph, treating communities as supervertices, and the network is obtained by a projection: 'blowing up' the supervertices into the communities they represent, and then contracting the tree edges. For simplicity, we will utilize and refer to the random tree, keeping in mind this distinction. Intuitively, the component of an individual is its neighborhood, for which a corresponding decorated branching process provides an approximate description. Roughly speaking, if the tree is small, then the component is small; if the tree grows unbounded, i.e., is infinite, then the component is large compared to the network size. (We omit the technicalities why this reasoning stands even when the component size is *not* negligible compared to the network size; however, this is not surprising, as similar phenomena appear in commonly used random graph models such as the Erdős-Rényi random graph (Alon and Spencer (2008), Section 10.5) or the configuration model (Molloy and Reed 1995)).

Thus, we first investigate the behavior of this random tree, which is largely defined by the offspring distributions $\widetilde{T}$ and $\widetilde{S}$, and we discuss the effect of the root having offspring T later. It is well known that a branching process undergoes a *phase transition*, i.e., it shows largely different behavior based on the choice of a parameter. The behavior we are interested in is whether the tree has a positive probability of growing infinitely. The phase transition happens at $\mathbb{E}[\widetilde{T}]\,\mathbb{E}[\widetilde{S}] = 1$.

The reason for this is that $\mathbb{E}[\widetilde{T}]\,\mathbb{E}[\widetilde{S}]$ is the expected growth ratio between two consecutive layers that contain individuals. Thus, if $\mathbb{E}[\widetilde{T}]\,\mathbb{E}[\widetilde{S}] > 1$, the expected size of layers containing individuals grows exponentially; in this case, the tree is infinite with some positive probability ξ, and we say that the branching process is supercritical. On the contrary, if $\mathbb{E}[\widetilde{T}]\,\mathbb{E}[\widetilde{S}] < 1$, the expected size of layers containing individuals decreases exponentially, and the random tree is almost surely finite; we call the branching process subcritical.

The phase transition of the branching process translates to the component sizes in the RIGC model as follows. In the subcritical case, when the random tree is almost surely finite, the graph consists of small components; most are of constant size, and all are sublinear in the network size. In the supercritical case, the random tree has probability ξ to be infinite, and a ξ proportion of individuals has such a corresponding branching process. Clearly, individuals within a finite graph cannot have an infinite component, instead all such vertices join together and form a *giant component*: a unique linear-sized component. The remaining components are all 'small', in the same sense as in the subcritical case.

We summarize our findings in the theorem below:

Theorem 4 (The largest component of the RIGC) *The size of the largest component in the RIGC exhibits a phase transition, depending on the choice of the parameters:*

i) *In the subcritical (and critical) case* $\mathbb{E}[\widetilde{S}]\,\mathbb{E}[\widetilde{T}] \leq 1$, *all components of the RIGC are sublinear in the network size N.*
ii) *In the supercritical case* $\mathbb{E}[\widetilde{S}]\,\mathbb{E}[\widetilde{T}] > 1$, *the largest component contains a proportion* ξ *of all individuals. This component is unique: all other components are sublinear in the network size N.*

The statements above hold with probability tending to 1 as the network size $N \to \infty$.

We note that there is a lower order probabilistic error in the size of the giant component; in particular, $\xi = 1$ does not necessarily mean that the whole graph is connected, but the individuals outside the giant only make up a sublinear proportion. (Identifying the conditions under which the RIGC is connected is outside the scope of this paper.)

In fact, we can identify ξ, the probability that the branching process produces an infinite tree, in terms of *probability generating functions* (PGF). We define the PGF of a non-negative integer valued random variable X as $\mathrm{PGF}_X(z) := \sum_{k=0}^{\infty} z^k \mathbb{P}(X = k)$. Recall that T denotes the asymptotic distribution of tokens, and the root of the branching process has offspring T. Also recall that, due to matching membership tokens and community roles uniformly, the rest of the offspring are size-biased, and we introduced the distribution $\mathbb{P}\left(\widetilde{T} = k - 1\right) = k\mathbb{P}(T = k)/\mathbb{E}[T]$. Similarly, for the asymptotic community size S, we introduced $\widetilde{S}$. Denote by η the smallest non-negative solution to the fixed-point equation $z = \mathrm{PGF}_{\widetilde{S}}\left(\mathrm{PGF}_{\widetilde{T}}(z)\right)$, then ξ is given by $\xi = 1 - \mathrm{PGF}_T(\eta)$.

In the following, we obtain further properties of the giant component by applying the previous reasoning: that an individual is part of the giant component exactly when the corresponding branching process produces an infinite tree. (This is true, except for a negligible sublinear proportion of individuals.) Keep in mind that the limit of the neighborhood is obtained from the decorated branching process by a projection: 'blowing up' the supervertices into the communities they represent and then contracting the edges of the branching process.

Corollary 3 (Degrees and edges in the giant) *Consider the empirical degree distribution of the giant component, or equivalently, the degree of an individual chosen uniformly at random within the giant component. This converges to the degree of the root in the projection of the decorated branching process that is conditioned on being infinite.*

If the mean of this limiting degree distribution is $m < \infty$, then the number of edges in the giant component scales as $m\xi N/2$. If this mean is infinite, the number of edges in the giant component is superlinear in the network size.

We intuitively explain why conditioning the random tree on being infinite generally *changes* the distribution of the degree of the root in the projection. Typically, a supercritical branching process either grows slowly initially and stops after a small number of layers, or grows quickly initially and produces an infinite tree; the ultimate behavior highly depends on the early stages of development. Thus, the random tree is more likely to be infinite when the offspring in lower layers are large. Thus, conversely, conditioning on an infinite tree induces a bias towards larger offspring. In particular, this affects both the number of communities in layer 1 as well as their size, and these are the communities the root takes its community roles from. (In many cases, we expect the average degree in

the giant component to be larger than the average degree in the entire network. However, as the communities are arbitrary, the opposite may occur.)

Information spread and attack vulnerability

In the previous section, we have focused on connected components, as each component is a set of individuals that may interact. On the other hand, certain interactions, such as the spread of a virus, be it a biological or computer virus, may very well be non-deterministic. As a simplistic model for a random spread, we consider *percolation*: for each edge, we flip a p-coin, independently of each other, to randomize whether that edge is able to transmit. With probability $q = 1 - p$, we consider an edge unable to transmit and remove it from the graph; we call the remaining subgraph the *percolated* graph. We emphasize that when we talk about p-percolation, $p \in [0, 1]$ is the probability that an edge is considered to be able to transmit and is kept (retained); in particular, $p = 1$ yields the original graph and $p = 0$ yields the empty graph.

Consider a simple epidemic spread, called SI-epidemic, named after the two possible states of individuals: susceptible or infected. We start the spread by setting a single individual as infected. Time progresses in discrete steps and the dynamics afterwards are as follows: all the individuals that became infected in the previous step attempt to transmit the infection through all incident edges, all of which succeed independently with probability p. If a yet susceptible neighbor is reached, then we set it as infected. Each individual only attempts to spread the infection once. When no new individual becomes infected, the process stops.

For the SI-infection model, even such a simple static model as percolation is able to capture the final infected cluster: all vertices that are connected to the source of infection in the percolated graph will eventually be infected. This, again, raises interest in the component sizes of the percolated graph: if there is a large component after percolation, there is a possibility for a large viral outbreak, in case the source is chosen from this large component.

There is always a trade-off in modeling: realistic models tend to be more complex, while simple models are easy to analyze. In the case of percolation, its simplicity has yet another advantage: it leaves room for different interpretations. We can also consider the independent removal of edges as a randomized attack on the network, in which case we are interested in how well the network can withstand such attacks. Once again, we arrive at the same question from a different perspective: what remains of the giant component, if we randomly remove a proportion $1 - p$ of the edges?

By definition, percolation on the RIGC model means that we first generate the random graph, and then, conditionally on the graph realization, randomly retain or remove each edge. However, recall that we construct the graph by adding imaginary edges randomly between individuals and community roles, and then contracting these imaginary edges. Thus, the edges of the resulting graph correspond one-to-one to the collection of all edges from all community graphs. This means that the percolation in fact does *not* depend on the graph realization. Moreover, we may even change the order: percolate the edges within the community graphs first, and then generate the RIGC, see Fig. 3.

We thus recognize that percolating the RIGC is the same as creating an RIGC with the list of percolated communities. There is one technicality to take care of. When we randomly remove edges from the communities, they may become *disconnected*. To satisfy

the condition that each community must be connected, we separate each connected component as its own community. This creates a new, random list of communities, that typically contains more, but smaller communities than the original unpercolated list, and we refer to it as the *percolated community list*. We can now formalize our observations as follows:

Proposition 1 (Percolated RIGC) *Percolation on the RIGC model is another RIGC, with the percolated community list and the original list of individuals and original number of membership tokens. In particular, percolation on the classical random intersection graph, i.e., when each community is a complete graph, is also an RIGC.*

With Proposition 1 and Theorem 4 in hand, we have the recipe for identifying the supercritical and subcritical cases of *percolation* on the RIGC. We assume that the original RIGC is supercritical, so that a giant component exists, and study for which choices of p the percolated graph is supercritical.

We note that the percolated community list sensitively depends on the original structure of the communities. This is in contrast with the observation that the global connectivity structure did not depend on the internal structure of communities, *as long as they are connected*. At the same time, this comparison also serves as an explanation why community structure now matters: it determines how the community falls apart under percolation, in particular, how many and how large components are produced.

Let us introduce the limiting community size distribution of the *percolated* community list and denote it by $S(p)$. Recall the limiting distribution T of the number of membership tokens, and its size-biased version $\widetilde{T}$, given by $\mathbb{P}\left(\widetilde{T} = k-1\right) = k\mathbb{P}(T=k)/\mathbb{E}[T]$. Similarly, define $\widetilde{S(p)}$ by $\mathbb{P}\left(\widetilde{S(p)} = k-1\right) = k\mathbb{P}(S(p)=k)/\mathbb{E}[\,S(p)]$. It follows from Theorem 4 that the condition for supercriticality is $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S(p)}\right] > 1$. Keep in mind that we consider this as an implicit condition on p.

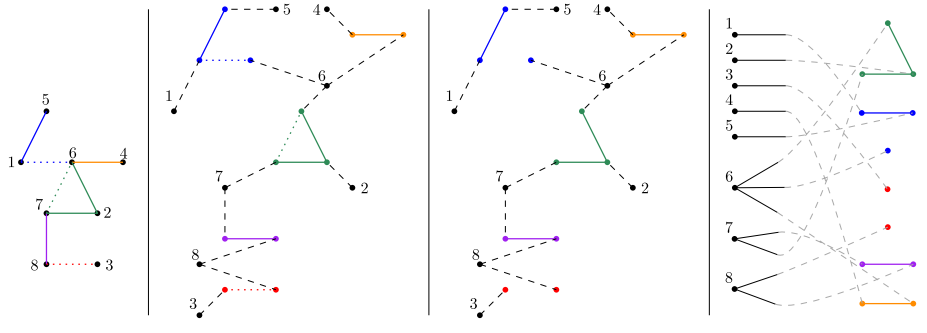

Fig. 3 Percolation on the RIGC. We demonstrate percolation on the example of Figure 1. From left to right, we trace back why percolation on the RIGC is another RIGC with a different community list. Leftmost, we randomly choose edges to delete from the resulting RIGC; we represent these edges as dotted. In the second subfigure, we trace back each edge of the RIGC to the corresponding edge within a community, as shown in the step before contracting the imaginary edges. In the third subfigure, we remove the percolated edges. Note that the individuals, their number of membership tokens, and where the tokens are matched, are unaffected. Communities that became disconnected due to removing edges are now replaced by their connected components, which we treat as separate communities. Rightmost, we recognize the matching for another RIGC, with the same list of individuals with the same number of membership tokens, and with the new (random) percolated community list

The closer p is to 1, the more edges we keep, and the larger the percolated community sizes are, thus $\mathbb{E}\left[\widetilde{S(p)}\right]$ is increasing in p. Consequently, we can find a threshold value (critical percolation parameter) p_c and write the supercriticality condition explicitly as $p > p_c$. This threshold p_c is the smallest p (more precisely, the infimum of p values) such that $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S(p)}\right] > 1$. Under some mild regularity conditions, e.g. $\mathbb{E}\left[T^2\right] < \infty$ and $\mathbb{E}\left[S^2\right] < \infty$ is sufficient, p_c is simply the solution to the implicit equation $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}[\widetilde{S(p_c)}] = 1$. (We give a counterexample later.)

Corollary 4 (Percolation phase transition) *Percolation on the supercritical RIGC exhibits a phase transition as we vary the percolation parameter p. With* $0 \leq p_c < 1$ *defined above,*

i) *in the subcritical case* $p < p_c$*, the largest component of the percolated graph is sublinear in the network size N;*
ii) *in the supercritical case* $p > p_c$*, there exists a unique, linear-sized component that contains a proportion* $0 < \xi(p) \leq \xi$ *of the individuals.*

The statements above hold with probability tending to 1 as $N \to \infty$*.*

In the following, we provide a discussion on Corollary 4 and some special cases, with an emphasis on attack vulnerability.

The case $p = p_c$ was excluded, as it is open in some particular instances. Under some regularity conditions, i.e., $\mathbb{E}\left[S^2\right] < \infty$ and $\mathbb{E}\left[T^2\right] < \infty$, we have $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}[\widetilde{S(p_c)}] = 1$, thus the graph shows so-called critical behavior. It is alike with the subcritical case in the sense that the largest component is sublinear, but shows more subtle differences. However, an in-depth discussion of critical behavior is beyond the scope of this paper.

The fact that $p_c < 1$ always holds means that the network does not exhibit *instant failure*. This means that the network may lose a linear proportion of edges, and as long as this fraction is small enough, a smaller, but still linear-sized giant component is retained. The intuitive reason for this is that when p is close enough to 1, small communities do not suffer too much damage, and that is sufficient to hold a good proportion of the network together.

On the other end of the spectrum, it is possible that $p_c = 0$, depending on the choice of parameters, which we discuss later in more detail. If $p_c = 0$, then for arbitrarily small ε, if at least an ε proportion of edges is kept, a small, but linear-sized component persists. This phenomenon is called *robustness*, and a robust network can withstand randomized attacks very well. It has been observed that many *scale-free* networks and models are robust (Bollobás and Riordan 2003a; Cohen et al. 2000; Solé and Montoya 2001). Scale-free, as before, refers to a network with a power-law degree distribution with exponent $\tau \in (2, 3)$; that is, degrees with finite mean but infinite variance. In such a network, the hubs, i.e., the highest-degree vertices with degree larger than $N^{1/2}$, hold the network together against random attacks. In the case of the RIGC, the hubs may be both large communities and individuals with a large number of membership tokens.

In the following, we discuss how the distribution of the number of tokens T and the community sizes S affect the robustness of the network. In particular, corresponding to

the traditional finite variance condition, we consider whether $\mathbb{E}\left[T^2\right]$ and $\mathbb{E}\left[S^2\right]$ are finite. However, there is another reason for considering these quantities: their relation to $\mathbb{E}\left[\widetilde{T}\right]$ and $\mathbb{E}\left[\widetilde{S}\right]$ respectively, explained below.

As discussed after Proposition 1, percolation with edge retention probability p is supercritical when $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S(p)}\right] > 1$. As discussed before, $\mathbb{E}\left[\widetilde{S(p)}\right]$ is a non-decreasing function of p; since $p = 0$ yields the empty graph and $p = 1$ yields the original graph, the values of $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S(p)}\right]$ range from 0 to $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S}\right]$. It is thus crucial whether $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S}\right]$ is finite. Recall that we have defined the size-biased version $\widetilde{T}$ as $\mathbb{P}(\widetilde{T} = k-1) = k\mathbb{P}(T = k)/\mathbb{E}[T]$. Consequently, $E\left[\widetilde{T}\right] = \mathbb{E}[T(T-1)]/\mathbb{E}[T]$ is finite exactly when $\mathbb{E}\left[T^2\right]$ is finite, and similarly, $\mathbb{E}\left[\widetilde{S}\right]$ is finite exactly when $\mathbb{E}\left[S^2\right]$ is finite.

The simplest, 'regular' case is when $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S}\right]$ is finite, i.e., both $\mathbb{E}\left[\widetilde{T}\right]$ and $\mathbb{E}\left[\widetilde{S}\right]$ are finite. This corresponds to the classical non-scale-free case and the graph is *non-robust*.

If $\mathbb{E}\left[\widetilde{T}\right]$ is infinite, irrespective of whether $\mathbb{E}\left[\widetilde{S}\right]$ is finite, the graph is *robust*. The reason is that some individuals are hubs and are part of a polynomially large number of communities, and have at least one incident edge in each. Thus, as long as an arbitrarily small, but positive proportion of edges is retained, these individuals remain hubs and hold a considerable proportion of the network connected. Indeed, for any small but positive p, $\mathbb{E}\left[\widetilde{S(p)}\right]$ is positive and thus $\mathbb{E}\left[\widetilde{T}\right]\mathbb{E}\left[\widetilde{S(p)}\right]$ is infinite. Also note that in this case the equation $\mathbb{E}\left[\widetilde{S(p)}\right]\mathbb{E}\left[\widetilde{T}\right] = 1$ does not have a solution, as at $p = 0$, we have an empty graph and this quantity becomes 0.

The most interesting and most complex case is when $\mathbb{E}\left[\widetilde{S}\right]$ is infinite, but $\mathbb{E}\left[\widetilde{T}\right]$ is finite. This means that the hubs are the communities, and the communities only, thus depending on the exact limiting distribution of community graphs (μ_H), the RIGC *may or may not be robust*. Intuitively, when the large community graphs are dense enough, e.g. when all communities are complete graphs, the total edge count of the RIGC becomes superlinear, and consequently the graph is robust. On the other hand, when large community graphs are sparse, e.g. have bounded within-community degree independent of the community size, these communities fracture into many small pieces under percolation, and the graph is not robust. It remains an intriguing open question what happens in between, as the dependence on (μ_H) is hard to quantify.

Conclusion

The model we introduce, the *random intersection graph with communities*, fills a gap in the literature in modeling networks with community structure, where communities are allowed to overlap and at the same time have their own internal structure. Built in a random fashion with arbitrary small graphs as building blocks, it is well-fitted to model networks with well-defined local structure but a more fluid global structure. Consider a social network where the internal structure of a community is determined by its purpose, e.g. whether it is a family or a workplace, while the variability in the combination of roles people take on is so vast that it can be considered random over the population. This large-scale randomness allows us to carry out exact asymptotic calculations.

The local structure of the model, such as degree distribution and local clustering, is defined by an interplay between the internal structure of communities, as well as the randomness arising from the combination of roles taken by each individual in the network. Under mild conditions, we have a sparse graph with positive clustering, well-suited

for modeling real-world networks. In the global structure, such as connectivity, generic macroscopic effects emerge despite the particular microscopic structures. However, the existence these microscopic structures, in particular, the fragility of communities, once again plays a crucial role in spreading processes or randomized attacks on the network.

Abbreviations
HCM: hierarchical configuration model; RIG: random intersection graph; RIGC: random intersection graph with communities

Acknowledgements
The authors thank the anonymous reviewers for their helpful insights and suggestions that helped to improve the presentation of this paper.

Authors' contributions
VV performed a major part of the mathematical analysis of the proposed model, as well as of the writing of the paper. JK contributed to the mathematical analysis of the proposed model as well as revised the manuscript critically for important intellectual content. RvdH has contributed to the design of the model and its mathematical analysis. All authors have read and approved the final manuscript.

Funding
This work is supported by the Netherlands Organisation for Scientific Research (NWO) through VICI grant 639.033.806 (RvdH), VENI grant 639.031.447 (JK), the Gravitation NETWORKS grant 024.002.003 (RvdH), and TOP grant 613.001.451 (VV). The Gravitation NETWORKS grant and TOP grant played a major role in conceptualizing the study of networks with community structures, in particular spreading processes on such networks. None of the funding influenced the outcome of the mathematical analysis.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 12 November 2018 Accepted: 28 May 2019
Published online: 27 June 2019

References
Alon N, Spencer JH (2008) The Probabilistic Method. 3rd ed. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, Hoboken. With an appendix on the life and work of Paul Erdős. https://doi.org/10.1002/9780470277331
Athreya KB, Ney PE (2004) Branching Processes. Dover Publications, Inc., Mineola. Reprint of the 1972 original [Springer, New York; MR0373040]
Ball F, Sirl D, Trapman P (2009) Threshold behaviour and final outcome of an epidemic on a random network with household structure. Adv Appl Probab 41(3):765–796. https://doi.org/10.1239/aap/1253281063
Ball F, Sirl D, Trapman P (2010) Analysis of a stochastic SIR epidemic on a random network incorporating household structure. Math Biosci 224(2):53–73. https://doi.org/10.1016/j.mbs.2009.12.003
Benjamini I, Schramm O (2001) Recurrence of distributional limits of finite planar graphs. Electron J Probab 6:23–13. https://doi.org/10.1214/EJP.v6-96
Bloznelis M (2010) Component evolution in general random intersection graphs. SIAM J Discrete Math 24(2):639–654. https://doi.org/10.1137/080713756
Bloznelis M (2013) Degree and clustering coefficient in sparse random intersection graphs. Ann Appl Probab 23(3):1254–1289. https://doi.org/10.1214/12-AAP874
Bloznelis M, Godehardt E, Jaworski J, Kurauskas V, Rybarczyk K (2015) Recent progress in complex network analysis: properties of random intersection graphs. In: Data Science, Learning by Latent Structures, and Knowledge Discovery. Stud. Classification Data Anal. Knowledge Organ. Springer, Heidelberg. pp 79–88
Bollobás B (1980) A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. Eur J Comb 1(4):311–316
Bollobás B (2001) Random Graphs, 2nd edn. Cambridge Studies in Advanced Mathematics, vol. 73. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511814068
Bollobás B, Riordan O (2003a) Robustness and vulnerability of scale-free random graphs. Internet Math 1(1):1–35
Bollobás B, Riordan OM (2003b) Mathematical results on scale-free random graphs. In: Handbook of Graphs and Networks. Wiley-VCH, Weinheim. pp 1–34
Cohen R, Havlin S (2003) Scale-free networks are ultrasmall. Phys Rev Lett 90:058701. https://doi.org/10.1103/PhysRevLett.90.058701
Cohen R, Erez K, Ben-Avraham D, Havlin S (2000) Resilience of the internet to random breakdowns. Phys Rev Lett 85:4626–4628. https://doi.org/10.1103/PhysRevLett.85.4626
Deijfen M, Kets W (2009) Random intersection graphs with tunable degree distribution and clustering. Probab Engrg Inform Sci 23(4):661–674. https://doi.org/10.1017/S0269964809990064
Derényi I, Palla G, Vicsek T (2005) Clique percolation in random networks. Phys Rev Lett 94:160202. https://doi.org/10.1103/PhysRevLett.94.160202

Erdős P, Rényi A (1959) On random graphs. I. Publ Math Debrecen 6:290–297
Gilbert EN (1959) Random graphs. Ann Math Stat 30(4):1141–1144
Godehardt E, Jaworski J (2003) Two models of random intersection graphs for classification(Schwaiger M, Opitz O, eds.). Springer, Berlin. https://doi.org/10.1007/978-3-642-55721-7_8
Guillaume J-L, Latapy M (2004) Bipartite structure of all complex networks. Inform Process Lett 90(5):215–221. https://doi.org/10.1016/j.ipl.2004.03.007
Guillaume J-L, Latapy M (2006) Bipartite graphs as models of complex networks. Phys A Stat Mech Appl 371(2):795–813
van der Hofstad R, Komjáthy J, Vadon V (2018) Random intersection graphs with communities. arXiv:1809.02514
van der Hofstad R (2017) Random Graphs and Complex Networks. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 43. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316779422
van der Hofstad R (2018+) Random Graphs and Complex Networks. Vol. 2. Available at http://www.win.tue.nl/~rhofstad/NotesRGCNII.pdf
van der Hofstad R, Komjáthy J, Vadon V (2019) Phase transition in random intersection graphs with communities. arXiv:1905.06253
van der Hofstad R, van Leeuwaarden JSH, Stegehuis C (2016) Hierarchical configuration model. Internet Math. https://doi.org/10.24166/im.01.2017
Janson S, Łuczak T, Rucinski A (2000) Random Graphs. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York. https://doi.org/10.1002/9781118032718
Karjalainen J, Leeuwaarden JSHv, Leskelä L (2018) Parameter estimators of sparse random intersection graphs with thinned communities. In: International Workshop on Algorithms and Models for the Web-Graph. Springer, Cham. pp 44–58
Karonski M, Scheinerman ER, Singer-Cohen KB (1999) On random intersection graphs: The subgraph problem. Comb Probab Comput 8(1&2):131–159
Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. In: Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, "Random Graphs '93" (Poznań, 1993), vol. 6. pp 161–179. https://doi.org/10.1002/rsa.3240060204
Newman MEJ (2010) Networks. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001
Newman ME (2003a) Properties of highly clustered networks. Phys Rev E 68(2):026121
Newman ME (2003b) The structure and function of complex networks. SIAM Rev 45(2):167–256
Newman ME (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323–351
Rybarczyk K (2011) Diameter, connectivity, and phase transition of the uniform random intersection graph. Discret Math 311(17):1998–2019
Singer KB (1996) Random Intersection Graphs. ProQuest LLC, Ann Arbor. Thesis (Ph.D.)–The Johns Hopkins University. http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:9617602
Solé RV, Montoya M (2001) Complexity and fragility in ecological networks. Proc R Soc London Ser B Biol Sci 268(1480):2039–2045
Stegehuis C, van der Hofstad R, van Leeuwaarden JSH (2016a) Epidemic spreading on complex networks with community structures. Sci Rep 6. article number: 29748
Stegehuis C, van der Hofstad R, van Leeuwaarden JSH (2016b) Power-law relations in random networks with communities. Phys Rev E 94:012302. https://doi.org/10.1103/PhysRevE.94.012302
Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the internet. Phys Rev E 65:066130. https://doi.org/10.1103/PhysRevE.65.066130
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440
Yağan O (2016) Zero-one laws for connectivity in inhomogeneous random key graphs. IEEE Trans Inform Theory 62(8):4559–4574. https://doi.org/10.1109/TIT.2016.2574742
Yağan O, Makowski AM (2012) Zero-one laws for connectivity in random key graphs. IEEE Trans Inform Theory 58(5):2983–2999. https://doi.org/10.1109/TIT.2011.2181331

Publisher's Note