The Power of Consensus: Random Graphs Have No Communities

Romain Campigotto
LIP6 – CNRS
Université Pierre et Marie Curie
France
Email: romain.campigotto@lip6.fr

Jean-Loup Guillaume
LIP6 – CNRS
Université Pierre et Marie Curie
France
Email: jean-loup.guillaume@lip6.fr

Massoud Seifi
LIP6 – CNRS
Université Pierre et Marie Curie
France

***Abstract*—Communities are a powerful tool to describe the structure of complex networks. Algorithms aiming at maximizing a quality function called modularity have been shown to effectively compute the community structure. However, some problems remain: in particular, it is possible to find high modularity partitions in graph without any community structure, in particular random graphs. In this paper, we study the notion of consensual communities and show that they do not exist in random graphs. For that, we exhibit a phase transition based on the strength of consensus: below a given threshold, all the nodes belongs to the same consensual community; above this threshold, each node is in its own consensual community.**

I. Introduction

Complex networks appear in various contexts such as computer science (e.g. networks of Web pages), sociology (e.g. collaborative networks), biology (e.g. gene regulatory networks). These networks can generally be represented by graphs, where nodes represent entities and edges indicate interactions between them. For example, a social network can be represented by a graph whose nodes are individuals and edges represent a kind of social relationship.

An important feature of such networks is that they are generally composed of highly interconnected sub-networks called *communities* [1], [2]. Communities can be considered as groups of nodes which share common properties and/or play similar roles within the graph. The automatic detection of such communities has attracted much attention in recent years and many community detection algorithms have been proposed (see [3] for a survey). Most of these algorithms are based on the maximization of a quality function known as *modularity* [4], which measures the internal density of communities. Modularity maximization is an **NP**-hard problem [5] and most algorithms use heuristics.

In random graphs, however, links appear independently of each other, so a strong inhomogeneity in the density of links on these graphs is not expected. Therefore, random graphs should not have communities using the previous definition. As shown in [6], it is however possible to find partitions with significantly high modularity in random networks. A good community detection algorithm should therefore be able to find communities but also to indicate their absence.

Here, we assume that, if multiple runs of a non-deterministic community detection algorithm agree that a given set of nodes belong to a community, then this set is certainly more significant than a community found by a single run. In the following, we will show that this definition of *consensual community*[1] (denoted throughout the paper by CC and CCs for the plural term) allows to make the distinction between real graphs and random graphs in terms of community structure. More precisely, we will prove that random graphs only contain trivial CCs, i.e. containing all the nodes of the graph or only a single node. We will also show there is a phase transition between these two states depending on a resolution parameter. The notion of consensual clustering has been introduced in [7] and its application to networks in [8], [9], [10].

We provide a general description of algorithms used for detecting CCs in Sect. II. We then present experimental results on artificial and real networks in Sect. III and the proof of the absence of CCs in random graphs in Sect. IV. We finally conclude in Sect. V.

II. Consensual Communities

Following the works from *E. Diday* [7] on consensual clustering of vectors, different studies have proposed to adapt this method to graphs and to combine different partitions into CCs. The common features of these methods consist in (i) compute different partitions and (ii) combine these partitions to find similarities.

Two main approaches are used to obtain different partitions. The first one consists in disturbing a given network by rewiring a small fraction of links [11] or changing slightly the weights on links [12], [13]. The second one, that we are going to use hereafter, consists in using non-deterministic algorithms to obtain different partitions. For instance, the *Louvain* method [14] (among others) can give different results depending on the order in which nodes are considered by the algorithm. This has been used in [8], [10] to compute CCs and in [9] to compute overlapping ones.

A. Definitions and Experiments

Given a graph $G = (V, E)$ with $n = |V|$ nodes, we apply $\mathcal{N}$ times a non-deterministic community detection algorithm $\mathcal{A}$ to G. At the end of each execution, each pair of nodes $(i, j) \subseteq V \times V$ is classified either in the same community or in different communities. We keep track of this in a matrix of size $n \times n$, which we denote by $P_{ij}^{\mathcal{N}} = [p_{ij}]_{n\times n}^{\mathcal{N}}$, where

[1] *Consensual commmunities* are also referred as *cores*.

p_{ij} represent the fraction of the $\mathcal{N}$ executions in which i and j were classified in the same community. Note that P is a symmetric matrix ($p_{ij} = p_{ji}$), and we set $p_{ii} = 0$. From $P^{\mathcal{N}}_{ij}$, we create a complete weighted graph $G' = (V, V \times V, W)$, where the weight of the link (i, j) is p_{ij}. Finally, given a threshold $\alpha \in [0, 1]$, we remove all links having $p_{ij} < \alpha$ from G' to obtain the *virtual graph with threshold*, G''_α. The connected components in the virtual graph G''_α obtained with a given α are called α-CCs.

We will suppose hereafter that $\mathcal{N}$ is large enough, so that $P^{\mathcal{N}} = P^{\infty}$. Previous works have indeed shown a fast convergence of the $P^{\mathcal{N}}$ matrix when N grows [8], [9]. We will therefore concentrate on the α parameter, which has a strong influence on the number and size of CCs.

The non-deterministic algorithm $\mathcal{A}$ we use here is the *Louvain* algorithm [14], which is a local search method which aims at maximizing the modularity [4] function. The *Louvain* method is currently the fastest algorithm to find communities on complex networks (it takes less than five seconds on networks with more than one million of nodes and edges). It is therefore well-suited to be run many times (typically with $\mathcal{N} = 100$ or more).

B. Properties of Consensual Communities

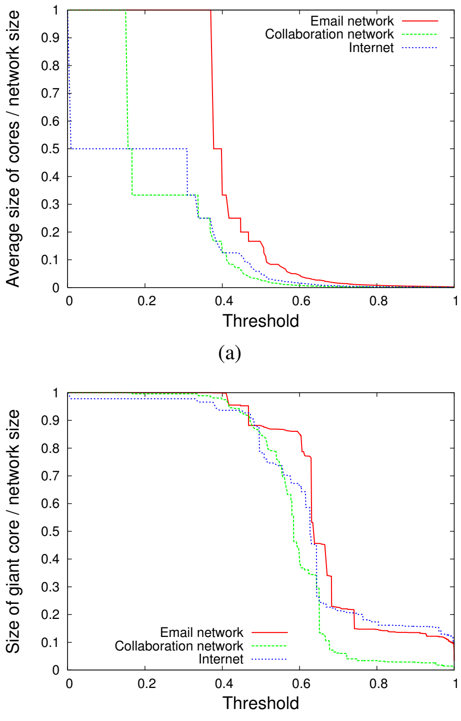
We computed CCs of complex networks of different sizes from different domains: a collaboration network [15], an email network [16] and a snapshot of the Internet (created by *M. Newman*, unpublished). As Fig. 1(a) shows, a large threshold, e.g. $\alpha = 1$, leads to tiny CCs, most of which consisting of only a single node. On the contrary, a low threshold gives a single CC (if the original graph is connected), and with $\alpha < 0.5$, we generally have a giant CC containing the majority of nodes. When the threshold increases, this giant CC will split into smaller ones. But in the Internet or email network, even with an α equal to 1, we still have a large CC containing approximately 10% of the node (see Fig. 1(b)).

This smooth decrease can also be understood through the study of the distribution of the values inside the P^{∞}_{ij} matrix. Figure 2 shows the p_{ij} distributions for three networks. We observe that if most pairs are nearly always separated and that a fair amount are always grouped together, there are also some pairs of nodes which are sometimes together and sometimes separated. This explains that significant CCs appear for a wide range of values of α.

These results show that the notion of CC can be used to detect different levels of communities. We will now show that they can also be used to show the absence of a real community structure in random graphs.

III. Consensual Communities in Random Graphs

In random graphs, all pairs of nodes have the same probability to be connected. Hence, they should not have preferential binding inducing specific and identifiable nodes groups. Therefore, we could conclude that there are no community structure in random graphs. However, several studies show that it is possible to find partitions with high modularity in random graphs [6], [17]. Indeed, the links concentration fluctuates in generated graphs, which means that subsets of nodes with a density larger than global density can appear. The phenomenon is even more pronounced in regular or quasi-regular graphs, like trees, torus or grid graphs, in which community detection algorithms can also find partitions with good modularity [18].



(a)

(b)

Fig. 1. (a) Average and (b) maximal size of CCs vs threshold α.

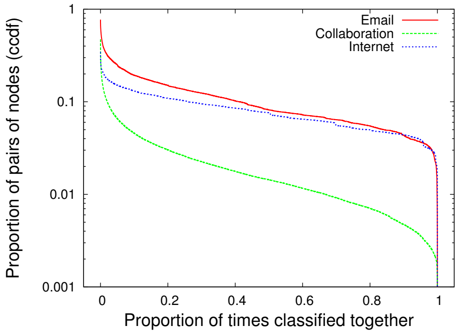
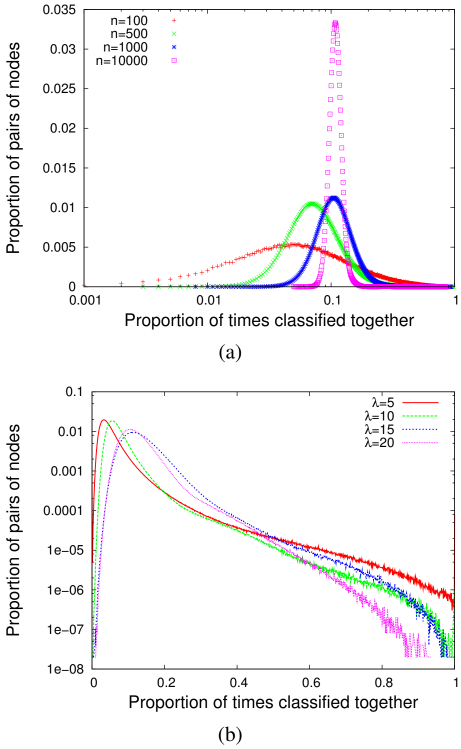


Fig. 2. p_{ij} complementary cumulative distribution for three real-world networks.

A good algorithm for community detection should indicate that communities obtained in random graphs are not real communities. We will now show that random graphs do not exhibit any non-trivial CCs. For this, we will use two different random graphs models: the classical *Erdős-Rényi* model [19] which is used to mimic the number of nodes and links only, and the configuration model [20], which also respects the full degree distribution.

A. Values of p_{ij} in Random Graphs

First of all, Fig. 3(a) and 3(b) show the distribution of p_{ij} values for an *Erdős-Rényi* random graph with different values of the number of nodes and the average degree. We observe a high concentration of p_{ij} at an average value (around 0.1 for large graphs using realistic values of the average degree) which is very different from the distributions observed on real graphs where the maximum of the distribution is at the zero value (see Fig. 2). We further observe on Fig. 3(b) that large values of p_{ij} appear. However, the concentration of values increases both with the size of the network and with the average degree and these large values are therefore less and less frequent.



(a)

(b)

Fig. 3. Distribution of the p_{ij} averaged over 100 random *Erdős-Rényi* graphs. (a) Networks with different number of nodes n and an average degree of $\lambda = 20$. (b) Networks with $n = 1{,}000$ nodes and different values of the average degree.

This concentration of values implies that even if partitions with a good modularity can be found in random graphs, these partitions are very different from one another since most pairs are classified in the same community only once every ten runs.

B. Comparison with Real Graphs

To compare more precisely real and random networks, we generated random graphs from the *Erdős-Rényi* model (resp. configuration model) that have the same size and the same average degree (resp. the same degree distribution) as two real-world networks. In Fig. 4, the *Erdős-Rényi* model shows no pair of nodes with $p_{ij} = 0$, which means that all pairs of nodes have been grouped together at least once during 1,000 runs of the *Louvain* algorithm, regardless of their position in the network. The same is observed for the configuration model.

Conversely, there is nearly no pair of nodes which are always grouped together, except for the leaves (nodes of degree 1) of the network which are always grouped with their only neighbor. This presence of nodes of degree 1 is very common with the configuration model when the degree distribution is a power-law. The same is observed for the *Erdős-Rényi* model since the real average degree is small and nodes of degree 1 are not so uncommon. This explains the small increase observed for the p_{ij} values around 1. Furthermore, as predicted by the experiments on *Erdős-Rényi* random networks (see Figs 3(a) and 3(b)), the maximum of the values is around 0.1.

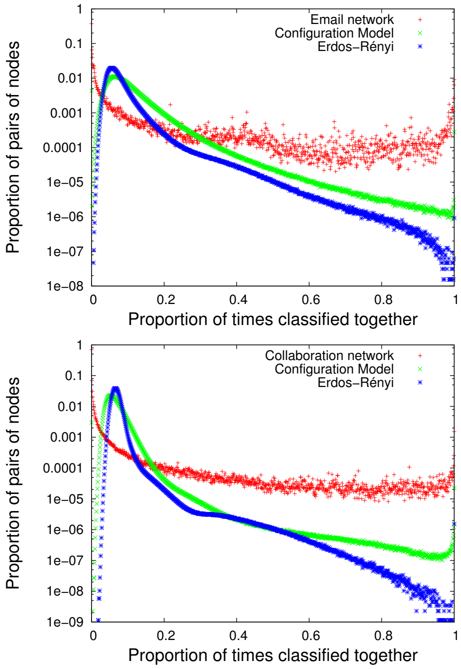


Fig. 4. p_{ij} distribution for two real-world networks together with *Erdős-Rényi* and configuration model random graphs with the same size: the email network (top) and the collaboration network (bottom).

There is two direct consequences of this distribution: (i) for very low values of the threshold, there is a single CC containing all nodes since there is no value close to zero and therefore the virtual graph contains all links, and (ii) for large values of the threshold, the virtual graph contains almost no links and therefore high threshold CCs are reduced to single nodes. Interestingly, in random networks, there is a sharp transition (see Fig. 5), at a threshold value around 0.3, which is not present in real-world networks. This phase transition cannot be directly deduced from the previous remarks and we will now use more arguments to prove its existence.

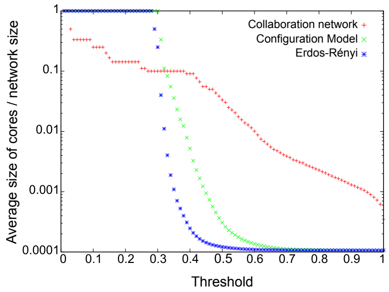


Fig. 5. Average size of CCs for a real networks and two random networks generated with the *Erdős-Rényi* and the configuration models.

IV. Existence of a Phase Transition

We recall that for a given threshold α, α-CCs are defined as connected components of the weighted graph G''_α whose adjacency matrix is P^∞, in which we have deleted weighted links with a value less than α. In random graphs, we observe that a small α gives one CC containing all the nodes of the graph. Then, after a rapid phase transition (based on the choice of α), we obtain only trivial CCs.

In the sequel, we give arguments to show the existence of this phase transition. Throughout the proof, we use extensively the fact that graphs are random and thus all connections appear independently. These assumptions can be related to classical mean field assumptions in statistical physics.

A. Values of p_{ij} for two connected nodes are highly concentrated around a mean value

Since we are considering random graphs, we can suppose that nodes (and their neighbors) in the input graph are similar. Thus, regardless of the results of the community detection algorithm used, nodes will be in expectation in the same community than a proportion p of their neighbors. Moreover, the random aspect of the graph implies this proportion p concerns neighbors which have been chosen randomly and independently for each run of the algorithm. In an equivalent way, we obtain that all p_{ij} are approximately equal to p.

Of course, this argument holds only if we assume that all elements in the graph are random. Indeed, the existence of correlations or specific properties on nodes can harm it. This is for instance the case of modularity applied on graphs having very low average degree. In particular, a node of degree 1 is always placed in the community of its unique neighbor and the above mentioned argument cannot be applied. The complete absence of correlations is therefore only valid for large networks with a sufficiently large average degree.

Figure 6 (all pairs) is an experimentation on a 10,000 nodes random *Erdős-Rényi* graph with different average degrees. We can observe that when the average degree is increasing, the effects of low degree nodes disappear and the distribution of p_{ij} is much more concentrated.

B. Values of p_{ij} for two connected nodes are higher than those of two non-connected nodes

On Fig. 6 (bottom), we can see that the distribution is composed of two distinct modes which correspond respectively to connected pairs of nodes, i.e., links, and non-connected pairs of nodes. p_{ij} values for connected nodes are higher than for non-connected nodes.

Two nodes i and j not connected and having a nonzero p_{ij} were necessarily classified at least once in the same community. As communities are necessarily connected subgraphs of the input graph, there exists a path connecting them and having only nonzero p_{uv}, for each nodes u and v belonging to the path. For instance, i and j can have a common neighbor k such that p_{ik} and p_{jk} are positive.

Let us assume that nodes i and j have a unique common neighbor k. As the graph is purely random, we can suppose that the probability that i and k are placed in the same community is $p_{ik} = p$, and the one that k and j are in the same community is $p_{kj} = p$. We also suppose they are independent, because edges linking i, j and k can be inside as well as between different communities, without any correlation. Thus, to i and j be classified in the same community, these two events must occur simultaneously. Therefore, $p_{ij} = p_{ik} \times p_{kj} = p^2$. The independence assumption is clearly unfounded in real networks, in particular due to the existence of strong local correlation as measured by the clustering coefficient.

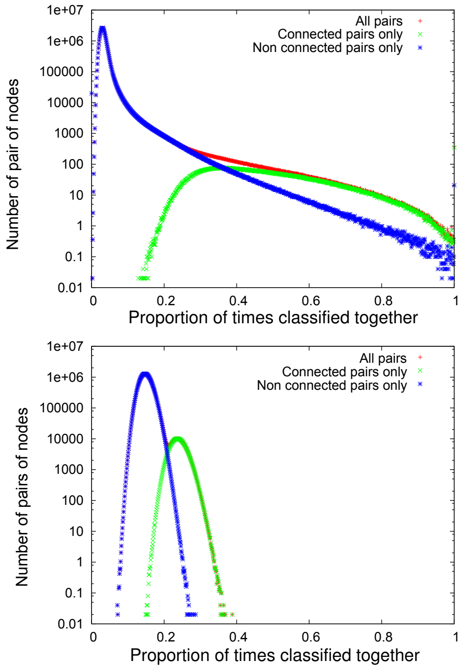


Fig. 6. p_{ij} distribution with a distinction between connected and non-connected pairs of nodes for a random graph with different average degree (5 and 100) and 10,000 nodes. The curve with all pairs is nearly completely overlapped by the other two curves, expect for average degree 5.

In the case where nodes i and j have no common neighbor but are connected with a longer path in the input graph, by using the same reasoning, we have $p_{ij} = \prod_{uv \in P} p_{uv} = p^k$, where P is a shortest path of size k linking i and j. This calculation holds if i and j have only one common neighbor.

It is easy to compute p_{ij} in the case where the two nodes have z nodes in common. We obtain $p_{ij} = 1 - (1 - p^2)^z$, that corresponds to 1 minus the probability that i and j are not linked with a common neighbor. However, if we assume that we have large graphs having low average degree, the probability of having more than one common neighbor (if we already have one) is very low. For these reasons, we can assume that values of p_{ij} are higher for connected pairs than non-connected pairs.

C. Existence of a Phase Transition

If we suppose that all connected pairs (i, j) have $p_{ij} = p$, and that non-connected nodes u and v have a lower probability of being connected, thus, for a threshold below p, only pairs of connected nodes provide connectivity, and as all connected pairs have nearly the same p_{ij}, we have only one CC containing all the nodes of the input graph (for large enough values of the average degree, the graph is connected, otherwise we have as many CCs as the number of connected components).

Conversely, since the p_{ij} distribution for connected pairs is strongly centered on the value p, any value of the threshold above p will destroy the CCs very quickly.

D. The proportion of intra-community links is equal to p

Finally, we can compute the value of this threshold. Let us assume that $k\%$ of links are intra-community links. Then, this means that for each execution of the algorithm, one node u will be put in expectation with $k\%$ of its neighbors, or equivalently each neighbor will be with the given node u for $k\%$ of the executions. This value k is thus the value of p_{ij} corresponding to the p that we have used so far.

Computing exactly the value of p is an open problem that seems to be difficult [6]. However, numerical studies (see Fig. 7) show that it decreases with the graph density but that the exact decrease pattern is quite complex.

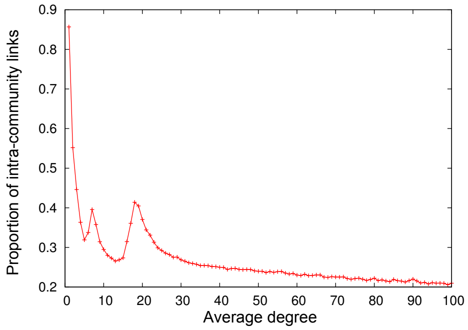


Fig. 7. Proportion of internal links for a random graph with 10,000 nodes.

V. Conclusion

We have shown here that CCs allow to distinguish graphs with a real community structure from graphs where this structure arises from fluctuations. To do so, we have shown that CCs in random graphs are trivial, containing either all the nodes of the graph or one node each.

Some future works remain to further understand the absence of non-trivial CCs in random graphs. First, it is necessary to compute the exact value of the threshold as a function of the parameters (size and average degree) of the *Erdős-Rényi* graphs. For graphs generated from the configuration model, the task is more difficult since there are many degree one nodes for which the modularity function requires that they are placed in the community of their only neighbour. Such local correlations are harder to take into account.

It would be interesting to analyze this phenomenon on others random graphs families such as *Watts-Strogatz* model, *Barabási-Albert* model, etc.

On a more general perspective, the computational issue has to be addressed. Indeed, computing CCs on large graphs can be hard even with very fast underlying algorithms such as *Louvain*, and different techniques should be used, such as local computations for instance.

Another perspective would be to make a similar study on regular graphs, in which we know that it does not exist community structures. In particular, for regular grids and torus, previous studies have shown that a high modularity partition can be found, but the regularity of such network naturally allows many different partitions which are simply translations of any partition. Intuitively, it means that many high quality partitions can be found and that should not exist.

Acknowledgment

This work is partially supported by the *DynGraph ANR-10-JCJC-0202* project of the French National Research Agency.

References

[1] M. Girvan and M. Newman, “Community structure in social and biological networks,” *Proc. of the National Academy of Sciences*, vol. 99, no. 12, 2002.

[2] C. Senshadhri, T. G. Kolda, and A. Pinar, “Community structure and scale-free collections of Erdős-Rényi graphs,” *Physical Review E*, vol. 85, no. 056109, 2012.

[3] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[4] M. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical Review E*, vol. 69, no. 2, 2004.

[5] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On finding graph clusterings with maximum modularity,” in *Graph-Theoretic Concepts in Computer Science*, 2007.

[6] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, “Modularity from fluctations in random graphs and complex networks,” *Physical Review E*, vol. 70, no. 2, 2004.

[7] E. Diday, “The dynamic clusters method and optimization in non-hierarchical clustering,” *Optimization Techniques*, pp. 241–258, 1973.

[8] M. Seifi, J.-L. Guillaume, I. Junier, J.-B. Rouquier, and S. Iskrov, “Stable community cores in complex networks,” in *3rd Int. Workshop on Complex Networks*, Melbourne, Florida, 2012.

[9] Q. Wang and E. Fleury, “Uncovering overlapping community structure,” in *2nd Int. Workshop on Complex Networks*, 2010, pp. 176–186.

[10] A. Lancichinetti and S. Fortunato, “Consensus clustering in complex networks,” *Scientific Reports*, vol. 2, no. 336, 2012.

[11] B. Karrer, E. Levina, and M. Newman, “Robustness of community structure in networks,” *Physical Review E*, vol. 77, no. 4, 2008.

[12] M. Rosvall and C. Bergstrom, “Mapping change in large networks,” *PloS one*, vol. 5, no. 1, 2010.

[13] D. Gfeller, J. Chappelier, and P. De Los Rios, “Finding instabilities in the community structure of complex networks,” *Physical Review E*, vol. 72, no. 5, 2005.

[14] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[15] M. Newman, “The structure of scientific collaboration networks,” *Proc. of the National Academy of Sciences*, vol. 98, no. 2, 2001.

[16] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, “Self-similar community structure in a network of human interactions,” *Physical Review E*, vol. 68, no. 6, 2003.

[17] J. Reichardt and S. Bornholdt, “Statistical mechanics of community detection,” *Physical Review E*, vol. 74, no. 1, 2006.

[18] F. de Montgolfier, M. Soto, and L. Viennot, “Asymptotic modularity of some graph classes,” in *ISAAC*, 2011, pp. 435–444.

[19] P. Erdős and A. Rényi, “On random graphs,” *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[20] E. A. Bender and E. R. Canfield, “The asymptotic number of labeled graphs with given degree sequences,” *Journal of Combinatorial Theory A*, vol. 24, pp. 296–307, 1978.