Combining Geometric Inhomogeneous Random Graphs with the Stochastic Block Model

Bacheolor Thesis of

Florian Küfner

At the Department of Informatics
Institute of Theoretical Informatics

Reviewer:	Thomas Bläsius
Second reviewer:	Dorothea Wagner
Advisor:	Christopher Weyand

September 2022 – February 2023

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I hereby declare that this document has been composed by myself and describes my own work, unless otherwise acknowledged in the text. I also declare that I have read and observed the *Satzung zur Sicherung guter wissenschaftlicher Praxis am Karlsruher Institut für Technologie.*

Karlsruhe, February 2023

. . . .

(Florian Küfner)

Abstract

The structure of real-world networks is often shaped by local clustering, i.e., densely connected regions with few interconnections. Random graph models that are meant to replicate this property can be attributed to one of two approaches. Geometric models on the one hand assign a geometric position to each vertex. Vertices that lie close together are more likely to be connected. Block models on the other hand work with a partition of the vertex set where vertices in the same community have a better chance of forming an edge.

In this paper we take the geometric inhomogeneous random graphs (GIRGs) and the stochastic block model (SBM) as representatives and state a hybrid model of the two. By that we also take over the power law weights from the GIRG model which leads to an inhomogeneous degree distribution.

In the resulting hybrid model we investigate the expected number of k-sized cliques for constant k. First we prove a matching lower and upper bound for a simplified variant of our model. The proof also gives insight into the community structure of the dominating clique type. Afterwards we generalize the lower bound for our original model and discuss an experimental setup that approaches an upper bound. The setup compares the two variants of our model in terms of the number of their triangles.

Zusammenfassung

Netzwerke aus der Realität weisen häufig eine clusterartige Struktur auf. Das heißt es gibt Regionen die in sich sehr viele Kanten aufweisen, aber untereinander nur spärlich verbunden sind. Modelle von Zufallsgraphen, die diese Eigenschaft wiederspiegeln sollen, folgen zumeist einem von zwei verschiedenen Ansätzen. Geometrische Modelle weisen jedem Knoten eine Koordinate zu. Knoten mit einer geringeren Distanz werden anschließend mit einer höheren Wahrscheinlichkeit verbunden. Block-Modelle hingegen partitionieren die Knotenmenge, wobei Kanten innerhalb einer Partitionsklasse eine höhere Auftrittswahrscheinlichkeit haben als solche, die zwischen zwei Klassen verlaufen.

In diesem Paper nehmen wir die geometric inhomogeneous random graphs (GIRGs) und das stochastic block model (SBM) als Repräsentaten der Ansätze heran und betrachten ein Kreuzungsmodell der beiden. Damit übernehmen wir insbesondere auch die Gewichte des GIRG Modells. Diese sorgen dafür, dass die Knotengrade einem Potenzgesetz folgen.

In dem Modell, welches wir auf diese Weise erhalten untersuchen wir die asymptotisch erwartete Anzahl von Cliquen der Größe k für ein konstantes k. Zunächst zeigen wir eine übereinstimmende obere und untere Schranke für eine vereinfachte Variante unseres Modells. Durch den Beweis erhalten wir zusätzlich einen Einblick darin wie der dominante Cliquentyp aussieht. Entweder gehören in den dominanten Cliquen alle Knoten der selben Partitionsklasse an oder oder alle auftretenden Klassen sind unterschiedlich.

Anschließend verallgemeinern wir die untere Schranke für unser ursprüngliches Modell und stellen einen Versuchsaufbau vor, dessen Absicht es ist sich einer oberen Schranke anzunähern. Der Versuch vergleicht die beiden Varianten unseres Modells in der Anzahl ihrer Dreiecke.

Contents

1 Introduction

1.1 Motivation

In the recent past reflecting real-world networks such as social networks, citation networks, or the internet has become an important objective of graph models [KZ09 | New03]. Real-world networks form pairwise connections between entities based on the importance and the similarity of the involved entities. For example in a social network there are prominent individuals that are known by many people, and two individuals that share similar interest or live near each other are more likely to know each other.

When we take a closer look at the properties in which the similarity of two entities is measured, we notice that there are both continuous as well as discrete ones. Age and location would be examples of continuous quantities, while the hobbies of a person or the scientific field of a paper are discrete.

Both types of similarity are captured by existing random graph models. On the one hand there are random geometric graphs. They assign a point from a continuous ground space to each vertex and the edge probability of two vertices is inversely proportional to the distance between them [Pen03]. Block based models like the stochastic block model (SBM) on the other hand partition the vertex set into several clusters. Edges are assigned based on fixed a priori connection probabilities defined for each pair of clusters [HLL83].

By themselves both approaches have been explored extensively, this paper investigates a hybrid model that assigns continuous points as well as discrete cluster memberships. A similar approach was already taken by Galhotra, Mazumdar, Pal, and Saha, when they generalized the definition of geometric random graphs using the SBM [GMPS18]. What will finally differentiate our model from theirs is that in their model every vertex is stochastically equivalent which leads to a homogeneous degree distribution.

This brings us back to the other factor in real-world connections which we called the importance of entities earlier. While most vertices are of very low, constant degree, few are very highly connected. This characteristic of real-world networks is called a scale-free or inhomogeneous degree distribution.

The first random graph model to replicate this property are Chung Lu graphs. In this model power law distributed weights are assigned to all vertices and edges are more likely granted between high weighted vertices [CL02a | CL02b].

Through the incorporation of such weights other models can be modified to obtain a scale-free degree distribution as well. Examples are geometric inhomogeneous random graphs (GIRGs) [BKL19 | BKL16] which generalize random geometric graphs, or the degree-corrected stochastic block model [KN11].

In this paper we want to create a model that has an inhomogeneous degree distribution and is able to model relationships based on continuous and discrete properties. To achieve that we propose a hybrid model of GIRGs and the SBM. The GIRGs model brings in inhomogeneous weights and a continuous ground space, the SBM augments them with a discrete cluster structure.

The pattern that we chose to analyse in greater detail within our new model are constant sized cliques. Among other things cliques are a good indicator for the locality and therefore for the clustering of graphs. There are known results when comes to counting cliques in other random graph models. Bläsius, Friedrich, and Krohmer investigated the expected number of not necessarily constant sized cliques in hyperbolic random graphs which in many ways behave similar to GIRGs [BFK18]. Michielan and Stegehuis on the other hand showed the asymptotic of constant sized cliques in GIRGs and that there is a dominating clique type for most parameter settings [MS22]. The latter result will come up more often because a big part of our results builds directly upon it.

1.2 Outline

In Chapter 2 we introduce some notations that we use through out our work. In Chapter 3 we propose two variants of our hybrid model while also reiterating the definition of the GIRG model. At the end of that chapter we also discuss potential alternatives and talk about the relation to the generalized GIRG model from Bringmann, Keusch, and Lengler [BKL16]. The main part of our work is presented in Chapter 4 where we determine the asymptotic number of constant sized cliques in one variation of our model, draw conclusions for the other one, and finally discuss an experimental setup in which we compare the two versions. All results are summarized in Chapter 5.

2 Preliminaries

For $n \in \mathbb{N}$ we define $[n] := \{1, \dots n\} \subseteq \mathbb{N}$.

Further for a Set S and $k \in [|S|] \cup \{0\}$ we define $\binom{S}{k}$ to be the family with all subsets of S of exactly size k. Formally that is $\binom{S}{k} := \{S' \in 2^S \mid |S'| = k\}$. The notation is based on the binomial coefficient because the cardinality can be calculated with $|\binom{S}{k}| = \binom{|S|}{k}$.

For a graph $G = (V, E)$ and a subset of it's vertices $V' \subseteq V$ we denote the subgraph of G induced by V' as $G[V']$. The induced subgraph has an edge between vertices $u, v \in V'$ if and only if the edge uv was present in G as well.

At one point we will use the multinomial coefficient which is denoted by $\binom{n}{k_1, k_2, \dots, k_m}$ and refers to the number of different distributions of n balls into m bins of the sizes $k_1, \dots k_m$ for $\sum_{i=1}^{n} k_i = n$. The balls are distinguishable and their order within a bin is arbitrary. Thus the value of the multinomial coefficient is given by

$$\binom{n}{k_1, k_2, \dots, k_m} = \frac{n!}{k_1! \cdot k_2! \dots k_m!}.$$

The binomial coefficient can be seen as the special case of the multinomial coefficient where $m = 2$. Namely we have $\binom{n}{k} = \binom{n}{k, (n-k)}$.

3 Model

In this chapter we will propose two variants of a model that combines the concepts of both GIRGs as well as the stochastic block model. When introducing the edge probabilities we will also quickly revise the definition of the GIRG model to emphasize the similarities and differences.

In the latter part of this section we will further discuss which model is more desirable and what other variations we thought about.

3.1 Definitions

Let $n \in \mathbb{N}$ be the number of vertices and $r(n) \in \mathbb{N}$ the number of clusters. The vertex set is identified with $[n] := \{1, \ldots, n\}$ and partitioned into disjoint sets $C_1 \cup \ldots \cup C_r = V$. For $v \in V$ we also write $m_v \in [r]$ for the unique index with $v \in C_{m_v}$. The connection probability between the blocks is controlled by a symmetric matrix $D := D(n)$ of size $r \times r$. Below for vertices $u, v \in V$ the notation $D_{u,v} := D_{m_u, m_v}$ is used.

Power-law Weights. In addition to the assigned memberships there are weights $w_1, \ldots, w_n$ that are following a power law with exponent $2 < \tau < 3$. The weights can be sampled from distribution with a cumulative probability function $F = F_n : \mathbb{R} \to [0, 1]$, where F has to own the following properties. For a fixed $w_{min} > 0$ we demand $F(z) = 0$ for all $z \leq w_{min}$, and $F(z) = 1 - \Theta(z^{1-\tau})$ for all $z \geq w_{min}$. Alternatively we can use the deterministic weight function $w_v = \gamma \cdot (n/v)^{1/(\tau-1)}$ with parameter $\gamma = \Theta(1)$.

Geometry. Let $d \in \mathbb{N}$ be a constant. We denote the d-dimensional torus as $\mathbb{T}^d$ and imagine it as the d-dimensional cube $[0, 1]^d$ where 0 and 1 are identified in every dimension. For every vertex v we draw a position $x_v \in \mathbb{T}^d$ uniformly and independently at random. To measure the distance between two points $x, y \in \mathbb{T}^d$ we take $||x - y||$ where $||.||$ denotes the L_∞-norm on the torus. Due to the properties of the torus that is $||x - y|| = \max_{1 \leq i \leq d} \min\{|x_i - y_i|, 1 - |x_i - y_i|\}$.

Edge Probability. We denote the sum over all weights as W. Moreover there is a parameter $0 < T < 1$ called the temperature of the model.

As announced we first recapture the edge probability for GIRGs. It mainly consists of the term q_{uv} which is artificially capped to 1 to receive the actual edge probability p_{uv}.

$$q_{uv} := \left(\frac{w_v w_u / W}{||x_u - x_v||^d} \right)^{1/T}, p_{uv} := \min\{1, q_{uv}\}. \tag{3.1}$$

We can say that intuitively heavier vertices are more likely to have edges than vertices with little weight. Also two vertices with a small distance in the underlying geometry are more likely to be connected than two vertices which lie far apart.

Now we can actually go into the definition of our models which we call the B and $B2$ model. There will be a couple of definitions that expand on similar definitions made for GIRGs. The notation for the new models will be the same as for the GIRGs supplemented by the superscript B or $B2$ respectively.

The first example for this are the edge probabilities p_{uv}^B and p_{uv}^{B2}. They additionally use the cluster assignments of the vertices u and v as well as the matrix D. The term q_{uv} however remains untouched.

$$p_{uv}^B := \min\left\{1, D_{u,v} \cdot q_{uv}\right\}, \tag{3.2}$$

$$p_{uv}^{B2} := D_{u,v} \cdot \min\left\{1, q_{uv}\right\} = D_{u,v} \cdot p_{uv}. \tag{3.3}$$

Further for a graph G^B drawn from the new model that uses the edge probability p^B, we will write G or G^{B2} for the graphs on the same weights, positions and cluster assignments that instead use the edge probability p or p^{B2}, respectively. For the event that $u, v \in V$ are connected in G we write $u \sim v$. Analogically the events that u and v are connected in G^B and G^{B2} are referred to as $u \sim^B v$ and $u \sim^{B2} v$.

3.2 Discussion of Alternatives

Before jumping into any analysis of the models proposed above we want to talk about why those are the definitions we settled with and what alternatives went through our minds. The premise of all discussed models will be that they are combining GIRGs with the SBM in some way. Why this general idea is of interest was broadly talked about in the introduction section and won't be subject here.

One of the first things we notice when trying to combine the GIRG and SMB model is that they already have a lot of similarities. Both require some sort of additional vertex variables and depending on those form a probability function for the edges $p : E \longrightarrow [0, 1]$ from which the edges are sampled independently at random.

Probably the most natural way of intersecting the two models would be to generate both graphs and give the resulting graph an edge if and only if both or at least one of the generated graphs have that exact edge. The resulting edge probability for the first approach is just the product of the edge probabilities from both models. This is exactly what the $B2$ model does. The latter option of granting an edge if at least one of the two graphs includes that edge is rather unattractive on closer examination. First of all the resulting probability $1 - (1 - p_{\text{GIRG}})(1 - p_{\text{SBM}}) = p_{\text{GIRG}} + p_{\text{SBM}} - p_{\text{GIRG}} \cdot p_{\text{SBM}}$ is cumbersome to calculate. Besides that with badly chosen parameters it could happen very easily that one of the models is completely dominating the other one.

On the other hand coming from the standpoint of the GIRG model it is not unusual to have a factor inside the min-term that controls the expected average degree inside the graph. From there the idea of pulling the SBM factor inside as well isn't all that far away. This is exactly what we have done with our initial approach, the B model. In this variant the Stochastic Block Model is embedded inside the GIRG edge probability. In a sense we thereby soften the impact of the Stochastic Block model. If the connection factor lies inside the minimum it can still be balanced out by the GIRG term. On the contrary in the $B2$ model where the edge probabilities are just multiplied the resulting probability is always capped by the probability of the SBM.

If we decide to have the SBM factor inside the min-term there are still at least two options where to precisely put it. Either we could just write it in front of the q_{uv} expression from the GIRG model or we could even draw it into the parenthesis where it is raised to the power of $\frac{1}{T}$. First note that for a fixed temperature T the two options are equivalent because we can just normalize every entry of the SBM matrix D with the exponent T.

However, there comes one case to mind one might want to consider where T isn't constant. Observing the limit of the probability for $T \longrightarrow 0$ for GIRGs yields a threshold model where the probability of every edge is either 1 or 0. For the GIRG edge probability given in Equation (3.1) we receive

$$p_{uv}^{\mathrm{TH}} = \begin{cases} 1 & \text{if } w_u w_v \geq W||x_u - x_v||^d, \\ 0 & \text{otherwise.} \end{cases}$$

When applying the same procedure to the models mentioned above where one time the SBM factor is inside and the other time outside the bracket we obtain different results. If the factor is inside the brackets we find it again in the inequality that distinguishes the cases. On the other hand if in front of the brackets it either vanishes completely if greater than 1 or takes the place of the value in the case of $w_u w_v \geq W||x_u - x_v||^d$ otherwise. That again is not in the spirit of the threshold model which is supposed to grant edges deterministically for given positions and weights.

Visible in the definition made by Equation (3.2) we decided to write the factor outside the brackets. The results can be generalized for the other model without restrictions, because the threshold model does not play a role in the following considerations.

3.3 Relation to Model from Bringmann, Keusch, and Lengler

To further differentiate our new models from previous inhomogeneous geometric graph models we take a look at the fairly general model introduced by Bringmann, Keusch, and Lengler that especially includes GIRGs and hyperbolic random graphs [BKL16].

Like GIRGs the model has power law weights and a ground space from which a position is assigned to each vertex. Based on those an edge probability is formed for each vertex pair. The main difference to GIRGs is that there are no restrictions on what the ground space exactly looks like and what the belonging distance function is. Instead they demand that after fixing the geometric position of one of two vertices the marginal edge probability is asymptotically just the edge probability of Chung Lu graphs. That is the product of the weights of both vertices divided by the sum over all weights.

Special case of Inclusion First we want to take a look at a restricted case of our model which is covered by the general model. As a benefit we get a few properties that were proven for the general model. Among those is the high probability for a giant component of size $\Theta(n)$, a small diameter and short average distances between vertices [BKL16].

The restriction we lay on our model is that the matrix D only has constant entries. According to Theorem 7.3 in the paper about the general model [BKL16] every edge probability function p^* that satisfies the following equation for all $x_u, x_v \in [0,1]^d$ and a parameter $\alpha \in \mathbb{R}_{>0}$, $\alpha \neq 1$ is included in their model.

$$p_{uv}^* = \Theta\left(\min\left\{1, V(||x_u - x_v||)^{-\alpha} \cdot \left(\frac{w_u w_v}{W}\right)^{\max\{\alpha,1\}}\right\}\right),$$

where $V(r)$ denotes the volume of the r-ball around 0 where the r-ball around x is defined as $B_r(x) := \{y \in \mathbb{T}^d \mid \ ||x-y|| \leq r\}$ for $r \geq 0$.

On behalf of both models we show this equation for the probability function p^B. With $\alpha := 1/T$,

$$\begin{aligned}
p^B_{uv} &\overset{(3.2)}{=} \min\left\{1, D_{u,v} \cdot \left(\frac{w_v w_u/W}{||x_u - x_v||^d}\right)^{1/T}\right\} \\
&= \min\left\{1, D_{u,v} \cdot ||x_u - x_v||^{-d/T} \cdot \left(\frac{w_v w_u}{W}\right)^{1/T}\right\} \\
&= \min\left\{1, D_{u,v} \cdot ||x_u - x_v||^{-\alpha d} \cdot \left(\frac{w_v w_u}{W}\right)^{\alpha}\right\} \\
&\overset{\star}{=} \Theta\left(\min\left\{1, \left(\frac{\pi^{d/2}}{\Gamma(d/2+1)}||x_u - x_v||^d\right)^{-\alpha} \cdot \left(\frac{w_v w_u}{W}\right)^{\max\{\alpha,1\}}\right\}\right) \\
&= \Theta\left(\min\left\{1, V(||x_u - x_v||)^{-\alpha} \cdot \left(\frac{w_u w_v}{W}\right)^{\max\{\alpha,1\}}\right\}\right),
\end{aligned}$$

where in $\star$ we only change constants factors in front of $||x_u - x_v||^d$ and use $\alpha \geq 1$ which follows from $T \leq 1$. In the step after we apply the formula for the volume of a r-ball that uses Euler's gamma function.

Cases of Exclusion Now we also want to argue that there are instances of our models that are not included in the general model. We consider a vertex u with a fixed position and cluster membership $m_u = i$. Further we assume that the amount of clusters $r(n)$ is super constant and for almost all $j \in [r(n)]$ we have $D_{i,j} \in o(1)$. Then we see that the marginal edge probability $\mathbb{E}_{x_v,m_v}[p^B_{uv} \mid x_u, m_u]$ is in $o(\min\{1, \frac{w_u w_v}{W}\})$ and therefore does not satisfy the condition of the model from Bringmann, Keusch, and Lengler.

4 Cliques of Constant Size k

In this section we want to investigate the asymptotic number of k-sized cliques, for a fixed constant k, in graphs drawn from one of our new models. The following results build on a paper by Michielan and Stegehuis that is counting cliques in GIRGs [MS22].

4.1 Setting

In this consideration we assume slight simplifications that apply to both models. Instead of allowing any symmetric matrix D, we focus on the case where there are only two different factors: one for inter- and one for intra-cluster connections. While the edge probabilities of vertices in the same cluster are multiplied by 1, the probability for inter-cluster connections is diminished by the prefactor $f(n)$ with $f(n) \leq 1$. Formally the connection matrix D is given by

$$D := \begin{pmatrix} 1 & & f(n) \\ & \ddots & \\ f(n) & & 1 \end{pmatrix}.$$

Further we want to restrict $r(n)$, the number of clusters in which the vertices are distributed to be $r(n) \in \omega(1)$ i.e. to be super constant. With that we exclude the case $r(n) \in \Theta(1)$ which would represent an unpleasant special case at a later point in time. Luckily we can reduce that case to the plain GIRG model for which the number of constant sized cliques is already extensively covered, for example in [MS22].

If $r(n) \in \Theta(1)$ there always has to be a cluster that has $\Theta(n)$ vertices by pigeonhole principle. Because according to our earlier assumption the diagonal entries of the matrix D are equal to 1, the subgraph induced by the vertices of such a cluster is nothing but a GIRG with less, but still a linear amount of vertices. According to the results of [MS22] the expected number of k-sized cliques in a GIRG is polynomial in n and therefore won't change asymptotically as long as the new number of vertices stays linear in n. On the other hand the expected amount of cliques surely won't be enlarged because for every vertex pair $\{u, v\} \in \binom{V}{2}$ the inequality $p_{uv}^{B2} \leq p_{uv}^{B} \leq p_{uv}$ holds.

4.2 Definitions

For a subset $U \subseteq V$ we denote the biggest number of vertices in U that share the same cluster as $c_{\max}(U)$. Formally that is

$$c_{\max}(U) := \max_{i \in [r]} \{|C_i \cap U|\}. \tag{4.1}$$

In the following we want to look at subsets $U \subseteq V$ with $|U| = k$, in other words potential cliques of size k. Moreover we are interested in the cluster distribution within those subsets. Specifically we will distinguish them by the size of their biggest cluster i.e. the number $c_{\max}(U)$. To count the number of cliques among such subsets for $k \in \mathbb{N}$, $l \in [k]$ we define the random variables

$$\begin{aligned} N^B(k,l) :=& \left| \left\{ U \in \binom{V}{k} \,\middle|\, G^B[U] \text{ is clique} \wedge c_{\max}(U) = l \right\} \right| \\ \leq& \left| \left\{ U \in \binom{V}{k} \,\middle|\, G^B[U] \text{ is clique} \right\} \right| =: N^B(k). \end{aligned} \tag{4.2}$$

While $N^B(k)$ is just the total number of cliques of size k in G^B, the random variable $N^B(k,l)$ only counts those cliques whose biggest cluster is of size l exactly. The supplemented inequality holds trivially because the first set is a subset of the second one. For the $B2$ model both definitions are analogous.

On the same basis let $N(k)$ denote the number of cliques of size k in the graph G which uses the edge probability from the GIRG model.

4.3 Simplified Model ($B2$)

We start of with the analysis of the $B2$ model. To prove the precise asymptotic of the expected number of k-sized cliques we show a matching lower and upper bound.

4.3.1 Lower Bound

Our first goal will be to lower bound $\mathbb{E}[N^{B2}(k)]$. The rough plan is to lower bound $\mathbb{E}[N^{B2}(k,l)]$ by a product of the expected number of cliques in the GIRG model, $\mathbb{E}[N(k)]$ and a diminishing function $h(l)$ which will turn out to be

$$h(l) := r(n)^{(1-l)} \cdot f(n)^{\binom{k}{2}-\binom{l}{2}}.$$

As defined above $r(n) \geq 1$ is the total number of clusters and $f(n) \leq 1$ is the factor appearing in the probability for inter-cluster edges.

We have $N^{B2}(k) = \sum_{l\in[k]} N^{B2}(k,l)$ and $N^{B2}(k,l) \geq 0$ for $l \in [k]$ which follows naturally from the definitions in Equation (4.2). Thereby we can lower bound $\mathbb{E}[N^{B2}(k)]$ by $\mathbb{E}[N^{B2}(k,l)]$ for every $l \in [k]$. To receive the biggest and thus tightest lower bound from this we find out which $\mathbb{E}[N^{B2}(k,l)]$ grows the fastest asymptotically by finding the maximum of the function $h(l)$.

The following lemma states the maxima of $h(l)$ which we will use in the proof the of lower bound.

Lemma 4.1: *In the value domain $l \in [k]$ the function $h(l)$ is maximized by*

$$\max_{l\in[k]}(h(l)) = \begin{cases} h(1) = f(n)^{\binom{k}{2}} & \text{if } r(n) \geq f(n)^{-\frac{k}{2}}, \\ h(k) = r(n)^{1-k} & \text{if } r(n) \leq f(n)^{-\frac{k}{2}}. \end{cases} \tag{4.3}$$

Proof. To find the arguments that maximize the function $h(l)$ we use the fact that the logarithm of a function maintains the position of extremes and only changes their value. We then pretend the function is continuous and derive it to find local maxima inside the interval $(1,k)$. For each of those points the floor and the ceiling are potential candidates for maxima on $[k]$. After that we calculate the values at the edges 1 and k. Finally from all those candidates we will determine the global maximum depending on the relation of the functions $f(n)$ and $r(n)$.

As proposed we first apply the logarithm and start simplifying:

$$\begin{aligned}\log(h(l)) &= \log(r(n)^{(1-l)} \cdot f(n)^{\binom{k}{2}-\binom{l}{2}}) \\ &= (1-l)\log(r(n)) + (\binom{k}{2} - \binom{l}{2})\log(f(n)) \\ &= (1-l)\log(r(n)) + \frac{k(k-1)}{2}\log(f(n)) - \frac{l(l-1)}{2}\log(f(n)).\end{aligned}$$

Now calculating the derivation with respect to l is way easier because instead of a product we have a sum on our hands.

$$\begin{aligned}\log(h(l))\frac{d}{dl} &= \left((1-l)\log(r(n)) + \frac{k(k-1)}{2}\log(f(n)) - \frac{l(l-1)}{2}\log(f(n))\right)\frac{d}{dl} \\ &= -\log(r(n)) + 0 - \frac{2l-1}{2}\log(f(n)) \overset{!}{=} 0 \\ &\Leftrightarrow -\log(r(n)) + \frac{1}{2}\log(f(n)) = l\log(f(n)) \\ &\Leftrightarrow -\frac{\log(r(n))}{\log(f(n))} + \frac{1}{2} = l.\end{aligned}$$

This makes $\lfloor -\frac{\log(r(n))}{\log(f(n))} + \frac{1}{2} \rfloor$ and $\lceil -\frac{\log(r(n))}{\log(f(n))} + \frac{1}{2} \rceil$ potential candidates for our global maximum. Before we calculate their values we check the sign of the second derivation to find out if the point is a local minimum or maximum in the continuous extension of $\log(h(l))$. We get

$$\begin{aligned}\log(h(l))\frac{d^2}{dl^2} &= \left(-\log(r(n)) - \frac{2l-1}{2}\log(f(n))\right)\frac{d}{dl} \\ &= -\log(f(n)) = \log(\frac{1}{f(n)}) \overset{f(n)\leq 1}{\geq} 0.\end{aligned}$$

More precisely the second derivation is either non-negative or 0 in the special case that $f(n) = 1$.

Let us look at the special case first. We know there has to be at least one cluster which is why we have $r(n) \geq f(n)^{-\frac{k}{2}} = 1^{-\frac{k}{2}} = 1$. At the same time our function h shrinks down to $h(l) = r(n)^{(1-l)} \cdot 1$ and is easily maximized by $h(1) = r(n)^0 = 1$. This fits into the first case of Equation (4.3) in Lemma 4.1.

On the other hand if $f(n) < 1$, the second derivative is strictly positive and the extreme we found earlier is in fact a minimum and therefore not of further interest. So all that we are left with are the boundary points 1 and k. We calculate their values and compare them to find a decision rule depending on $r(n)$ and $f(n)$:

$$\begin{aligned}h(1) = f(n)^{\binom{k}{2}} &\leq r(n)^{(1-k)} = h(k) \\ \Leftrightarrow f(n)^{\frac{1}{2}\cdot k(k-1)} &\leq r(n)^{(1-k)} \\ \Leftrightarrow f(n) &\leq r(n)^{-\frac{2}{k}} \\ \Leftrightarrow r(n) &\leq f(n)^{-\frac{k}{2}},\end{aligned}$$

completing the proof of Lemma 4.1. ■

Under the use of Lemma 4.1 we are now able to prove the following theorem that states a lower bound for $\mathbb{E}[N^{B2}(k)]$.

Theorem 4.2: *For a natural constant k the following lower bound is applicable:*

$$E[N^{B2}(k)] \in \begin{cases} \Omega(f(n)^{\binom{k}{2}} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \geq f(n)^{-\frac{k}{2}}, \\ \Omega(r(n)^{1-k} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \leq f(n)^{-\frac{k}{2}}. \end{cases}$$

Proof. We start by breaking down the expected value into two probabilities. For that we can use the linearity of expectation and conditional probabilities as follows

$$\begin{aligned} \mathbb{E}[N^{B2}(k,l)] &= \mathbb{E}\left[\sum_{U \in \binom{V}{k}} 1_{\{G^{B2}[U] \text{ is clique} \wedge c_{\max}(U)=l\}}\right] \\ &= \sum_{U \in \binom{V}{k}} \mathbb{E}\left[1_{\{G^{B2}[U] \text{ is clique} \wedge c_{\max}(U)=l\}}\right] \\ &= \sum_{U \in \binom{V}{k}} \mathbb{P}\left(G^{B2}[U] \text{ is clique} \wedge c_{\max}(U) = l\right) \\ &= \sum_{U \in \binom{V}{k}} \mathbb{P}\left(G^{B2}[U] \text{ is clique} \mid c_{\max}(U) = l\right) \cdot \mathbb{P}\left(c_{\max}(U) = l\right). \end{aligned} \tag{4.4}$$

We first look at the latter probability. To determine the asymptotic behavior of the probability we state both an upper and lower bound that match asymptotically.

We start with the lower bound. Note that we can only decrease the probability of an event by replacing it with a more special event. If the first l vertices of a subset $U \in \binom{V}{k}$ are all in the same cluster and all the remaining vertices are all in different, unique clusters we especially have $c_{\max}(U) = l$. Using this idea we get

$$\mathbb{P}(c_{\max}(U) = l) \geq \left(\frac{1}{r(n)}\right)^{l-1} \cdot \prod_{i=l+1}^{k} \left(1 - \frac{i-l}{r(n)}\right) \in \Theta(r(n)^{(1-l)}), \tag{4.5}$$

where we can exclude the case that there are less than $k-l+1$ clusters and the event is therefore impossible. This is because $k - l + 1 \leq k \in \Theta(1)$, while $r(n) \in \omega(1)$. The case is therefore, at least for the asymptotic probability, irrelevant. Along the same line of argumentation we get that $(1 - \frac{i-l}{r(n)}) = (1 - \Theta(\frac{1}{r(n)})) = 1 - o(1) \in \Theta(1)$.

Now to receive a upper bound for the probability we calculate the probability of a more general event namely that for a subset $U \in \binom{V}{k}$ there is a cluster of size at least l within U. The probability for this more general event we further bound up by taking the probability that the event happens for a fixed cluster times the number of clusters $r(n)$. By that we factor in every scenario where the generalized event occurs at least once. The scenarios where more than one cluster in U has size at least l we count even more often.

Overall we get

$$\begin{aligned}\mathbb{P}(c_{\max}(U) = l) &\leq \mathbb{P}(c_{\max}(U) \geq l) \\ &\leq r(n) \cdot \mathbb{P}(|C_1 \cap U| \geq l) \\ &= r(n) \cdot \sum_{i=l}^{k} \binom{k}{i} \left(\frac{1}{r(n)}\right)^{i} \left(1 - \frac{1}{r(n)}\right)^{k-i} \\ &\leq \sum_{i=l}^{k} \binom{k}{i} \left(\frac{1}{r(n)}\right)^{i-1} \\ &\leq \left(\frac{1}{r(n)}\right)^{l-1} \sum_{i=l}^{k} \binom{k}{i} \in \Theta(r(n)^{(1-l)}),\end{aligned} \tag{4.6}$$

where we use $k, l \in \Theta(1)$.

Next we turn to the first probability in Equation (4.4), it describes the event that $G^{B2}[U]$ is a clique under the assumption that the biggest cluster in U is of size l.

Note that the original edge probability of the GIRG model p_{uv} doesn't depend on the cluster memberships of the vertices. Thus we can lower bound our overall probability by arranging the clusters such that the number of inter-cluster edges is maximized. When sticking with the premise that $c_{\max}(U) = l$, an upper bound on the number of inter-cluster edges is achieved if besides the one cluster of size l every smaller appearing cluster in U is of size 1. In that setting every edge is an inter-cluster edge except those within the set of size l where all vertices are in the same cluster. By that argument the number of inter-cluster edges is at most $\binom{k}{2} - \binom{l}{2}$. Note that taking an edge as an inter- instead of an intra-cluster edge will never increase its likelihood and thus the likelihood of the clique. We call the l-sized cluster $W \subseteq U$ and obtain

$$\begin{aligned}\mathbb{P}(G^{B2}[U] \text{ is clique} \mid c_{\max}(U) = l) &= \prod_{\{u,v\} \in \binom{W}{2}} p_{uv}^{B2} \prod_{\{u,v\} \in \binom{U}{2} \setminus \binom{W}{2}} p_{uv}^{B2} \\ &\geq \prod_{\{u,v\} \in \binom{W}{2}} 1 \cdot p_{uv} \prod_{\{u,v\} \in \binom{U}{2} \setminus \binom{W}{2}} f(n) \cdot p_{uv} \\ &= \left(\prod_{\{u,v\} \in \binom{U}{2}} p_{uv} \right) \cdot f(n)^{\binom{k}{2} - \binom{l}{2}} \\ &= \mathbb{P}(G[U] \text{ is clique}) \cdot f(n)^{\binom{k}{2} - \binom{l}{2}}.\end{aligned} \tag{4.7}$$

Now we are finally putting everything back together with Equation (4.4) and get

$$\begin{aligned}\mathbb{E}[N^{B2}(k,l)] &= \sum_{U\in\binom{V}{k}} \mathbb{P}(G^{B2}[U] \text{ is clique} \mid c_{\max}(U)=l)\cdot\mathbb{P}(c_{\max}(U)=l)\\ &\overset{4.5-4.7}{\geq} \sum_{U\in\binom{V}{k}} \mathbb{P}(G[U] \text{ is clique})\cdot f(n)^{\binom{k}{2}-\binom{l}{2}}\cdot\Theta(r(n)^{(1-l)})\\ &= \sum_{U\in\binom{V}{k}} \Omega\left(\mathbb{P}(G[U] \text{ is clique})\cdot f(n)^{\binom{k}{2}-\binom{l}{2}}\cdot r(n)^{(1-l)}\right)\\ &= \Omega\left(f(n)^{\binom{k}{2}-\binom{l}{2}})\cdot r(n)^{(1-l)}\sum_{U\in\binom{V}{k}}\mathbb{P}(G[U] \text{ is clique})\right)\\ &= \Omega\left(h(l)\cdot\mathbb{E}[N(k)]\right).\end{aligned}$$

Together with $\mathbb{E}[N^{B2}(k,l)] \overset{4.2}{\leq} \mathbb{E}[N^{B2}(k)]$ and Lemma 4.1 this completes the proof of Theorem 4.2. ■

4.3.2 Upper Bound

In this section we will prove a matching upper bound to the lower bound shown above. By that we overall determine the precise asymptotic difference between the expected amount of k-sized cliques in the $B2$ model compared to the GIRG model.

Our approach is very similar to what we did for lower bound. First of all we use the linearity of expectation to rewrite the expected value into the probability that a k-vertex subset is a clique. We then split this probability up with the law of total probability along the number of inter cluster edges x in the clique. For each $x\in\{0,\dots,\binom{k}{2}\}$ we then bound the probability by a product of the clique probability in GIRGs and a diminishing function that we will call $g(x)$.

Before we state our upper bound we will show two lemmas which we need for the proof. The first lemma will help us to find an upper-bound for the number of involved clusters given a fixed number of inter-cluster edges inside a clique.

Lemma 4.3: *Let $H=(V_1\cup\ldots\cup V_r,E)$ be a complete r-partite graph on n vertices with the minimal number of edges. Then $r-1$ of the disjoint vertex sets are of size* 1.

Proof. Let $H=(V=V_1\cup\ldots\cup V_r,E)$ be a complete r-partite graph on n vertices with the minimal number of edges. For the sake of contradiction assume that there are $i,j\in[r]$ with $i\neq j$ and $|V_i|,|V_j|>1$.

Construct a new partition by moving $|V_j|-1$ vertices from V_j to V_i. Let $H'=(V,E')$ be the r-partite graph on that new vertex set partition. We now want to compare $|E'|$ with $|E|$. For every vertex in V_i we lose $|V_j|-1$ edges because all but one vertex from V_j are now in the same partition class. At the same time the one vertex left behind in V_j gains that many edges. Therefore we have $|E'|=|E|-|V_i|\cdot(|V_j|-1)+(|V_j|-1)=|E|-(|V_i|-1)\cdot(|V_j|-1)$. Because we assumed $|V_i|,|V_j|>1$ we can bound from below $|E|-(|V_i|-1)\cdot(|V_j|-1)\leq|E|-1<|E|$. This contradicts our first assumption that H had a minimal number of edges among r-partite graphs with n vertices which concludes the proof of Lemma 4.3. ■

The mentioned function $g(x)$ that diminishes the clique probability depending on the amount of involved inter-cluster edges will turn out to be

$$g(x) := r(n)^{a^+(k,x)-k}\cdot f(n)^x,$$

where $a^+(k,x) = k + \frac{1}{2} - \sqrt{k^2 - k - 2x + \frac{1}{4}}$. For the latter proof we need to determine the maximum of $g(x)$. The second lemma does just that.

Lemma 4.4: *On the interval* $x \in [0, \binom{k}{2}]$ *the function* $g(x)$ *is bounded by*

$$\max_{x\in[0,\binom{k}{2}]} g(x) \leq \begin{cases} f(n)^{\binom{k}{2}} & \text{if } r(n) \geq f(n)^{-\frac{k}{2}}, \\ r(n)^{1-k} & \text{if } r(n) \leq f(n)^{-\frac{k}{2}}. \end{cases}$$

Proof. We treat the cases separately.

Case 1: $r(n) \geq f(n)^{-\frac{k}{2}}$ Remember that because we are searching for an upper bound, its always fine to make the maximum of $g(x)$ bigger than it actually is. Therefore we simplify

$$\begin{aligned} g(x) &= r(n)^{a^+(k,x)-k} \cdot f(n)^x \\ &\leq f(n)^{-\frac{k}{2}(a^+(k,x)-k)} \cdot f(n)^x \\ &= f(n)^{-\frac{k}{2}(a^+(k,x)-k)+x}, \end{aligned} \tag{4.8}$$

where the upper bound can be confusing at first, but is still correct because the exponent of the $r(n)$-term in $g(x)$ is negative. When maximizing the last expression in the above equation we just minimize the exponent because $0 \leq f(n) \leq 1$. Like in the proof of Lemma 4.1 we achieve this by first finding potential local extremes through the derivation and comparing them against the values of the boundaries 0 and $\binom{k}{2}$. We first write out the full exponent and reshape it with the intention to make it easier to derive.

$$\begin{aligned} -\frac{k}{2}(a^+(k,x) - k) + x &= -\frac{k}{2}(k + \frac{1}{2} - \sqrt{k^2 - k - 2x + \frac{1}{4}} - k) + x \\ &= -\frac{k}{4} + \frac{k}{2}(k^2 - k - 2x + \frac{1}{4})^{\frac{1}{2}} + x. \end{aligned}$$

Now we calculate the derivation we search for zeros.

$$\begin{aligned} (-\frac{k}{2}(a^+(k,x) - k) + x)\frac{d}{dx} &= \frac{k}{4}(-2)(k^2 - k - 2x + \frac{1}{4})^{-\frac{1}{2}} + 1 \overset{!}{=} 0 \\ &\Leftrightarrow \frac{k}{2(k^2 - k - 2x + \frac{1}{4})^{\frac{1}{2}}} = 1 \\ &\Leftrightarrow k = 2(k^2 - k - 2x + \frac{1}{4})^{\frac{1}{2}} \\ &\overset{(*)}{\Rightarrow} k^2 = 4(k^2 - k - 2x + \frac{1}{4}) \\ &\Leftrightarrow 8x = 3k^2 - 4k + 1 \\ &\Leftrightarrow x = \frac{3}{8}k^2 - \frac{1}{2}k + \frac{1}{8}, \end{aligned}$$

where the other direction of (*) is also valid. This can be reproduced by inserting the resulting value of x in the upper equation. Also we notice that for $k \geq 2$ the potential extreme lies inside the interval of our interest.

This now leaves us behind with three candidates for the minimum of the exponent as a function of $x \in [0, \binom{k}{2}]$. There are the borders 0 and $\binom{k}{2}$ as well as the potential local extreme $\frac{3}{8}k^2 - \frac{1}{2}k + \frac{1}{8}$. By simply plugging those candidates into the exponent we receive the value $\binom{k}{2}$ for both boundaries and $\frac{5}{8}k^2 - \frac{3}{4}k + \frac{1}{8}$ for the inner point. The comparison yields

$$\frac{5}{8}k^2 - \frac{3}{4}k + \frac{1}{8} \overset{k\geq 2}{\geq} \frac{4}{8}k^2 + \frac{1}{4}k - \frac{3}{4}k + \frac{1}{8} = \binom{k}{2} + \frac{1}{8} > \binom{k}{2}.$$

That means for $r(n) \geq f(n)^{-\frac{k}{2}}$ we now have that $g(x) \leq f(n)^{\binom{k}{2}}$ as claimed.

Case 2: $r(n) \leq f(n)^{-\frac{k}{2}}$ We now proceed in a very similar way with the case $r(n) \leq f(n)^{-\frac{k}{2}} \Leftrightarrow r(n)^{-\frac{2}{k}} \geq f(n)$. Again, as in Equation (4.8), we can upper bound $g(x)$ by an expression that has only one base, namely

$$\begin{aligned} g(x) &= r(n)^{a^+(k,x)-k} \cdot f(n)^x \\ &\leq r(n)^{a^+(k,x)-k} \cdot r(n)^{-\frac{2}{k}\cdot x}. \end{aligned} \tag{4.9}$$

This expression however is maximized by maximizing the exponent, because $r(n) \geq 1$. This can again be achieved by finding the zeros of the derivation and comparing their values to those of the boundaries.

$$\begin{aligned} a^+(k,x) - k - \frac{2}{k}x &= k + \frac{1}{2} - \sqrt{k^2 - k - 2x + \frac{1}{4}} - k - \frac{2}{k}x \\ &= \frac{1}{2} - (k^2 - k - 2x + \frac{1}{4})^{\frac{1}{2}} - \frac{2}{k}x \\ (\frac{1}{2} - (k^2 - k - 2x + \frac{1}{4})^{\frac{1}{2}} - \frac{2}{k}x)\frac{d}{dx} &= (-2) \cdot \frac{1}{2} \cdot (k^2 - k - 2x + \frac{1}{4})^{-\frac{1}{2}} - \frac{2}{k} \overset{!}{=} 0 \end{aligned}$$

We notice that after multiplying the equation with $-\frac{k}{2}$, adding 1 to both sides afterwards and squaring them, we arrive at the exact same equality as with the other derivation, which means that the zero is in fact the same, namely $\frac{3}{8}k^2 - \frac{1}{2}k + \frac{1}{8}$. Despite that the value of the exponent in that point can be different. That is why we again plug all three candidates, the potential extreme as well as the boundary points into the exponent. We receive $(1-k)$ for both edges of the interval and $\frac{3}{2} - \frac{5}{4}k - \frac{1}{4k}$ for the inner point. Because

$$\frac{3}{2} - \frac{5}{4}k - \frac{1}{4k} \overset{k\geq 2}{\leq} \frac{3}{2} - \frac{1}{2} - k = 1 - k,$$

we obtain $1-k$ as the biggest possible exponent for $x \in [0, \binom{k}{2}]$. With that follows $g(x) \leq r(n)^{1-k}$ concluding the proof of Lemma 4.4. ■

Now we are ready to state and prove the upper bound.

Theorem 4.5: *For a natural constant k the following upper bound is applicable:*

$$E[N^{B2}(k)] \in \begin{cases} \mathcal{O}(f(n)^{\binom{k}{2}} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \geq f(n)^{-\frac{k}{2}}, \\ \mathcal{O}(r(n)^{1-k} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \leq f(n)^{-\frac{k}{2}}. \end{cases}$$

Proof. For $U \in \binom{V}{k}$ with a fixed cluster distribution we call $m_{\text{inter}}(U)$ the number of inter-cluster edges inside U. This function is closely related to $c_{\max}(U)$ defined in Equation (4.1). Especially we know that $m_{\text{inter}}(U) = 0$ if and only if $c_{\max}(U) = k$ and $m_{\text{inter}}(U) = \binom{k}{2}$ if and only if $c_{\max}(U) = 1$.

As we did for the lower bound in Equation (4.4) we will break down the expected value to two probabilities that are easier to estimate. Both the linearity of expectation as well as the law of total probability come in handy.

$$\begin{aligned}\mathbb{E}[N^{B2}(k)] &= \mathbb{E}\left[\sum_{U\in\binom{V}{k}} 1_{\{G^{B2}[U] \text{ is clique}\}}\right] \\ &= \sum_{U\in\binom{V}{k}} \mathbb{E}\left[1_{\{G^{B2}[U] \text{ is clique}\}}\right] \\ &= \sum_{U\in\binom{V}{k}} \mathbb{P}\left(G^{B2}[U] \text{ is clique}\right) \\ &= \sum_{U\in\binom{V}{k}} \sum_{x=0}^{\binom{k}{2}} \mathbb{P}\left(m_{\text{inter}}(U) = x\right) \mathbb{P}\left(G^{B2}[U] \text{ is clique} \mid m_{\text{inter}}(U) = x\right).\end{aligned} \tag{4.10}$$

We first turn towards the second probability. We can draw the SBM prefactors in front of the entire product, exactly as in Equation (4.7), because they are outside the minimum in our simplified model. This time we already know the exact number of inter-cluster connections instead of having to derive them from the size of the biggest cluster in U. The remaining product over the p_{uv}-terms is nothing but the probability to form a clique in the original GIRG model. Formally that means

$$\mathbb{P}\left(G^{B2}[U] \text{ is clique} \mid m_{\text{inter}}(U) = x\right) = f(n)^x \cdot \mathbb{P}\left(G[U] \text{ is clique}\right). \tag{4.11}$$

Now we arrive at the more complicated probability that a cluster configuration causes exactly x inter-cluster edges when the k vertices of U are distributed uniformly at random among the $r(n)$ clusters. Calculating the exact function in $x \in \{0, \ldots, \binom{k}{2}\}$ is cumbersome, not only because it is by nature discontinuous but also because it has various jumps. For example there always are $r(n)$ configurations where all vertices have fallen into the same cluster and $m_{\text{inter}}(U) = 0$, but for $k \geq 3$ there is no configuration for exactly one inter-cluster edge. As a solution for this problem we will come up with a continuous function that represents an upper bound for the probability at every point $x \in \{0, \ldots, \binom{k}{2}\}$.

The amount of different clusters appearing inside a vertex set $U \in 2^V$ will be referred to as $a(U) := |\{i \in [r(n)] \mid C_i \cup U \neq \emptyset\}|$ in the remainder. The next step is to find a lower bound for $m_{\text{inter}}(U)$ depending on $a(U)$.

Because the subgraph induced by the inter-cluster edges is just a complete $a(U)$-partite graph, we can apply Lemma 4.3. The direct consequence is that for a fixed value of $a(U)$ the cluster distributions with the least amount of inter-cluster connections are those who put $k - a(U) + 1$ vertices in one cluster and only 1 in every other cluster. In this setting every edge outside this one big cluster is an inter-cluster edge, which makes $\binom{k}{2} - \binom{k-a(U)+1}{2}$ inter-cluster edges in total. Overall this yields

$$m_{\text{inter}}(U) \geq \binom{k}{2} - \binom{k - a(U) + 1}{2}.$$

We will now solve this inequality for $a(U)$. Then for a fixed $m_{\text{inter}}(U)$ we obtain an upper bound for the amount of different clusters $a(U)$ in U. This will be very helpful because as it turns out asymptotically the likelihood for a certain cluster arrangement only depends on how many clusters are involved.

We proceed by first simplifying the term and then solving for $a(U)$:

$$\begin{aligned}
m_{\text{inter}}(U) \geq \binom{k}{2} - \binom{k-a(U)+1}{2} &= \frac{1}{2}(k(k-1) - (k-a(U)+1)(k-a(U))) \\
&= \frac{1}{2}(k^2 - k - k^2 + 2ka(U) - a(U)^2 - k + a(U)) \\
&= \frac{1}{2}(-a(U)^2 + (2k+1)a(U) - 2k) \\
\Leftrightarrow 0 &\geq \frac{1}{2}(-a(U)^2 + (2k+1)a(U) - 2k - 2m_{\text{inter}}(U)) \\
\Leftrightarrow 0 &\leq a(U)^2 + (1-2k)a(U) + 2k + 2m_{\text{inter}}(U)).
\end{aligned}$$

The latter quadratic polynomial in $a(U)$ has two roots at the points $a^+(k, m_{\text{inter}}(U))$ and $a^-(k, m_{\text{inter}}(U))$ with

$$a^*(k,x) := k + \frac{1}{2} * \sqrt{k^2 - k - 2x + \frac{1}{4}},$$

for $* \in \{+,-\}$. Because the sign of the $a(U)^2$ term is positive the function is positive if and only if $a(U)$ is either smaller than $a^-(k, m_{\text{inter}}(U))$ or greater than $a^+(k, m_{\text{inter}}(U))$. Note that the latter criteria implies $a(U) > k$ and is therefore unfulfillable because there can be at most k different clusters among the k vertices.

Our actual insight from this is that for a fixed value of $m_{\text{inter}}(U)$ we know $a(U)$ lies within $[1, a^+(k, m_{\text{inter}}(U))] \cap \mathbb{N}$. In other words the event $m_{\text{inter}}(U) = x$ implies $a(U) \in [1, a^+(k,x)] \cap \mathbb{N}$. By this reasoning we can upper bound $\mathbb{P}(m_{\text{inter}}(U) = x)$ which is the probability we are interested in by $\mathbb{P}(a(U) \in [1, a^+(k,x)] \cap \mathbb{N})$ which is easier to calculate.

The probability space that we are moving in is still the uniformly random and independent distribution of k vertices in $r(n)$ different clusters. There are $r(n)^k$ possible events in total of those we now want to count the positive events i.e. the cluster distributions that use at most $\lfloor a^+(k,x) \rfloor$ different clusters.

First there are $\binom{r(n)}{a}$ combinations of a clusters. We may assume that the set of used clusters is already selected and they are arbitrarily ordered. For every of the a clusters we call $c_i : i \in [a]$ the number of vertices within that cluster. After choosing that function c all that is left is to distribute the k vertices into those bins of already fixed size. For counting the options in the last step we can use the multinomial coefficient. Overall we get

$$\begin{aligned}
\mathbb{P}(m_{\text{inter}}(U) = x) &\leq \mathbb{P}\left(a(U) \in [1, a^+(k,x)] \cap \mathbb{N}\right) \\
&= \sum_{a=1}^{\lfloor a^+(k,x) \rfloor} \mathbb{P}(a(U) = a) \\
&= \sum_{a=1}^{\lfloor a^+(k,x) \rfloor} \frac{\#\text{cluster assignments with } a(U) = a}{r(n)^k} \\
&= \sum_{a=1}^{\lfloor a^+(k,x) \rfloor} \frac{\binom{r(n)}{a} \cdot \sum_{c:\sum c_i = k} \binom{k}{c_1, \ldots c_a}}{r(n)^k}.
\end{aligned}$$

At this point we realize two things. On the one hand, the multinomial coefficient is computed as $\binom{k}{c_1,\ldots c_a} = \frac{k!}{c_1!\ldots c_a!} \in \Theta(1)$ because for $i \in [a]$ we have $c_i \leq k \in \Theta(1)$. On the other hand, the total number of options for c can be upper bounded by $a^k \leq k^k \in \Theta(1)$, again because every c_i is in $[k]$. Further we use Stirling's approximation and $a \leq k \in \Theta(1)$ to receive $\binom{r(n)}{a} \in \Theta(r(n)^a)$. We get

$$\begin{aligned}
\mathbb{P}\left(m_{\text{inter}}(U) = x\right) &\leq \sum_{a=1}^{\lfloor a^+(k,x)\rfloor} \frac{\binom{r(n)}{a} \cdot \sum_{c:\sum c_i=k} \binom{k}{c_1,\ldots c_a}}{r(n)^k} \\
&= \sum_{a=1}^{\lfloor a^+(k,x)\rfloor} \frac{\binom{r(n)}{a}\Theta(1)}{r(n)^k} \\
&= \sum_{a=1}^{\lfloor a^+(k,x)\rfloor} \frac{\Theta(r(n)^a)}{r(n)^k} \\
&\leq \lfloor a^+(k,x)\rfloor \cdot \frac{\Theta(r(n)^{\lfloor a^+(k,x)\rfloor})}{r(n)^k} \in \mathcal{O}\left(r(n)^{a^+(k,x)-k}\right),
\end{aligned}$$

where in the end we leave out the floor function. By doing that we can only make the term bigger. If we plug all those results into Equation (4.10) we arrive at

$$\begin{aligned}
\mathbb{E}[N^{B2}(k)] &= \sum_{U\in\binom{V}{k}} \sum_{x=0}^{\binom{k}{2}} \mathcal{O}\left(r(n)^{a^+(k,x)-k}\right) \cdot f(n)^x \cdot \mathbb{P}\left(G[U] \text{ is clique}\right) \\
&= \sum_{U\in\binom{V}{k}} \mathbb{P}\left(G[U] \text{ is clique}\right) \cdot \mathcal{O}\left(\sum_{x=0}^{\binom{k}{2}} r(n)^{a^+(k,x)-k} \cdot f(n)^x\right) \\
&= \mathbb{E}[N(k)] \cdot \mathcal{O}\left(\sum_{x=0}^{\binom{k}{2}} r(n)^{a^+(k,x)-k} \cdot f(n)^x\right) \\
&=: \mathbb{E}[N(k)] \cdot \mathcal{O}\left(\sum_{x=0}^{\binom{k}{2}} g(x)\right),
\end{aligned}$$

where in the last step we define the function $g(x)$ which we know from Lemma 4.4. Note that because the sum over x has only $\binom{k}{2} + 1 \in \Theta(1)$ terms the asymptotic behavior is dictated by the largest of those terms. So all that is left to do to arrive at the bound postulated in Theorem 4.5 is to plug the result for the maximum of g for $x \in \{0,\ldots \binom{k}{2}\}$ from Lemma 4.4. Note that the maximum over $[0, \binom{k}{2}]$ will always be at least as big as the maximum over its subset $\{0,\ldots,\binom{k}{2}\}$. ■

4.3.3 Interpretation

As already said, when put together the statements of the Theorems 4.2 and 4.5 provide the precise asymptotic of the expected number of k-sized cliques in graphs of the simplified model. Notably the way in which we proved the lower bound gives us further insight in the clique structure of the model.

By first estimating $\mathbb{E}[N^{B2}(k,l)]$ we basically counted the cliques appearing in vertex sets of different cluster structure separately. The structural variable that we distinguished the classes by was the size of the biggest cluster within the vertex set. Depending on the variable functions $f(n)$ and $r(n)$ we then found the class of vertex sets which contributes the most to the overall number of cliques. Showing afterwards that the upper bound asymptotically matches the expected cliques emerging from that dominant class alone, gives us the insight that only the cliques from that class are asymptotically significant. In the case that $r(n) > f(n)^{-\frac{k}{2}}$ the dominating class of cliques consists of vertex sets where every vertex is from a different cluster, while when $r(n) < f(n)^{-\frac{k}{2}}$ cliques with all vertices from the same cluster are the most significant.

Intuitively if there are a lot of clusters and the penalty factor for inter-cluster edges is in comparison not that big, it is not worth searching for subsets with a big common cluster which in this setting are very rare. Instead it is most promising to just look at vertex sets where every cluster is unique, which there are a lot of, and accept to only have inter-cluster edges. On the other hand if the inter-cluster penalty is large compared to a rather small amount of clusters, it is worthwhile to search for subsets where all vertices are in the same cluster to avoid inter-cluster edges completely.

A very similar result was found by [MS22] for the basic GIRG model, where the dominating clique type can either be geometric or non-geometric. Vertex sets where all the vertices lie geometrically very close together and thus are very likely to form a clique are called geometric cliques, while vertex sets with pairwise constant distance but very high weights form non-geometric cliques. Whether one or the other clique type is dominant is determined by the ratio of the constant clique size k as well as the exponent of the power law weight distribution τ. More precisely if $k > \frac{2}{3-\tau}$ the non-geometric cliques outweigh asymptotically and their expected number is $\Theta(n^{(3-\tau)k/2})$. On the other hand if $k < \frac{2}{3-\tau}$ geometric cliques dominate the overall expected number of cliques, asymptotically growing with $\Theta(n)$.

In the simplified model we analysed above we were able to split the expected number of cliques into the expected number of cliques in GIRGs and another cluster dependent factor. Therefore the two phase shifts described above are both applicable to our model and operate orthogonally to each other. If we want to state the full asymptotic of expected k-sized cliques in our model we therefore have to distinct four cases in total. Figure 4.1 illustrates the parameter space divided into the four corresponding quadrants.

4.4 Original Model (B)

As indicated earlier the main purpose of the so called simplified model was to function as an intermediate step for the analysis of our original model. Even though the attempt of totally generalizing the bounding process wasn't successful, there are still many things we can transfer back for our original model. Our main objective still is to determine the expected number of k-sized cliques for a natural constant k.

4.4.1 Lower Bound

The first thing we can get right away is a lower bound. Note that we are still in the restricted setting where $D_{uv} = 1$ if u and v are in the same cluster and $D_{uv} = f(n) \leq 1$ otherwise. Thus we have $D_{uv} \leq 1$ for all vertices $u, v \in V$. With that

$$p^B_{uv} = \min\{1, D_{uv} \cdot q_{uv}\} \geq D_{uv} \cdot \min\{1, q_{uv}\} = p^{B2}_{uv}. \tag{4.12}$$

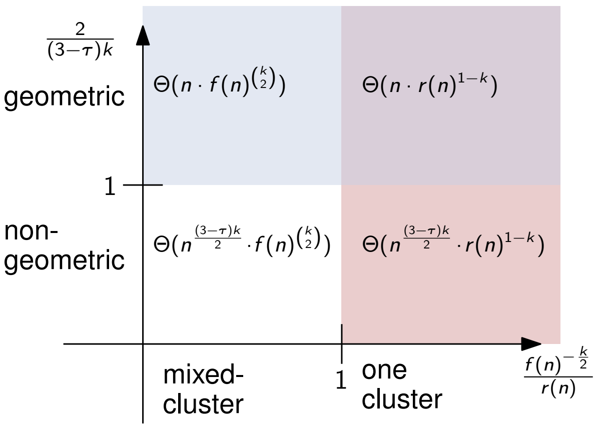
$\frac{2}{(3-\tau)k}$
geometric
$\Theta(n \cdot f(n)^{\binom{k}{2}})$
$\Theta(n \cdot r(n)^{1-k})$
1
non-geometric
$\Theta(n^{\frac{(3-\tau)k}{2}} \cdot f(n)^{\binom{k}{2}})$
$\Theta(n^{\frac{(3-\tau)k}{2}} \cdot r(n)^{1-k})$
mixed-cluster
1
one cluster
$\frac{f(n)^{-\frac{k}{2}}}{r(n)}$

Figure 4.1: Visualization of the asymptotic number of k-sized cliques across the parameter space.

In other words, every edge in our original model is at least as likely as in the simplified model. Therefore the lower bound shown in Section 4.3.1 functions as a lower bound for the expected number of k-sized cliques in the B model as well. Formally that is

$$E[N^B(k)] \in \begin{cases} \Omega(f(n)^{\binom{k}{2}} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \geq f(n)^{-\frac{k}{2}}, \\ \Omega(r(n)^{1-k} \cdot \mathbb{E}[N(k)]) & \text{if } r(n) \leq f(n)^{-\frac{k}{2}}. \end{cases} \tag{4.13}$$

Note that this bound is weaker for our original model than it is for the simplified variant. While we already succeeded finding a matching upper bound for the latter, it is yet possible that this bound isn't tight for the former.

4.4.2 Upper Bound

Finding a fitting upper bound turned out to be too difficult within the limits of this bachelor thesis. In this chapter we will on the one hand try to substantiate what makes it difficult and why we can't just generalize the approach we took with the simplified model. On the other hand we want to present and discuss some experimental results in which we compared the two models with each other.

Difficulties When we look at the upper as well as the lower bound that we were able to show for the simplified model we notice that they aren't even fully written out. In both cases the expected value $\mathbb{E}[N(k)]$ that denotes the expected amount of cliques inside the GIRG model remains inside the expression. We were always able to transform the expected value into a shape were we had the expected value for GIRGs, which is asymptotically determined by [MS22], and a multiplicative factor. Further that latter factor was always relatively easy to handle because it had no dependencies to the weights, positions or other parameters from the GIRG model.

The process of isolating those two independent factors was not that hard as long as the cluster connection factor originating from the SBM stood outside the min-term. This is exactly what makes the original model with the factor inside the minimum operation significantly harder to analyse. Even if for an edge uv we already know that it is an inter-cluster edge

	k	τ	$r(n)$	$f(n)$
Non-geometric mixed-cluster cliques	3	2.1	$n^{\frac{1}{3}}$	$n^{-\frac{3}{9}}$
Geometric mixed-cluster cliques	3	2.6	$n^{\frac{1}{3}}$	$n^{-\frac{3}{9}}$
Non-geometric one cluster cliques	3	2.1	$n^{\frac{1}{3}}$	$n^{-\frac{1}{9}}$
Geometric one cluster cliques	3	2.6	$n^{\frac{1}{3}}$	$n^{-\frac{1}{9}}$

Figure 4.2: Table of the four different parameter configurations we used for our experiments

it's still possible for the two vertices to be that heavy or that close to each other that there connection probability is still 1. In other words the connection probabilities of the two models are much more entangled in this model.

Experimental Results Note that since we were not able to show a matching upper bound we can not tell if the lower bound we derived in the previous section is tight. Even though practical experiments can neither formally confirm nor deny the tightness of the lower bound they can give a cue where to search for the correct answer. In this section we first want to propose a generic experimental setup to check the tightness of our lower bound. Further we will discuss the results of an implementation of the setup.

In order to see if the lower bound is tight we have to check if the model matches the lower bound asymptotically. This can be done by generating graphs and counting their k-sized cliques. Rather than dividing this number directly through the asymptotic term, our setup compares the two model variants with each other. Both methods are basically equivalent because we already know that the proposed asymptotic is tight for the $B2$ model.

We start out by generating the weights, positions and cluster memberships for all the vertices. From those values we can already compute p^B_{uv} and p^{B2}_{uv} for all potential edges uv. We reduce the variance of the experiment by using a coupled random experiment for each edge. That means for two vertices u, v a uniformly random value y_{uv} is drawn from the interval $[0, 1)$. For $* \in \{B, B2\}$ the edge uv in G^* is then granted if and only if $p^*_{uv} \geq y_{uv}$.

After the generation of the graphs we count the k-sized cliques for each of them. We then divide the number of cliques in the $B2$ graph through that of the B graph. Because of the coupled random events we have $u \sim^{B2} v \Rightarrow u \sim^B v$, and the number of cliques in G^B is always higher, thus we end up with a ratio between 0 and 1.

This procedure is repeated for different values of n. The quotient staying constant for variable values of n could indicate that the two models share the same asymptotic. If the diagram runs to zero for $n \longrightarrow \infty$ it could indicate that the B graph has an asymptotic that grows strictly faster than $B2$.

We performed this experiment on a small scale. Figures 4.3 to 4.6 show the average quotient for four different combinations of the inter-cluster function $f(n)$, the number of clusters $r(n)$, the clique size k and the power law exponent τ. The actual parameters are documented in Figure 4.2. We chose them such that each parameter set covers one quadrant in Figure 4.1. We considered triangles only, i.e. kept $k = 3$ because counting bigger cliques takes much longer. Each point in the diagrams is an average of about 80 quotients $\frac{N^{B2}(3)}{N^B(3)}$. The red error bars visualize the standard deviation of the sample. We aligned the average degree of graphs between the four parameter sets to make them comparable. We achieve this by multiplying the edge probability with a constant we experimentally estimated beforehand.

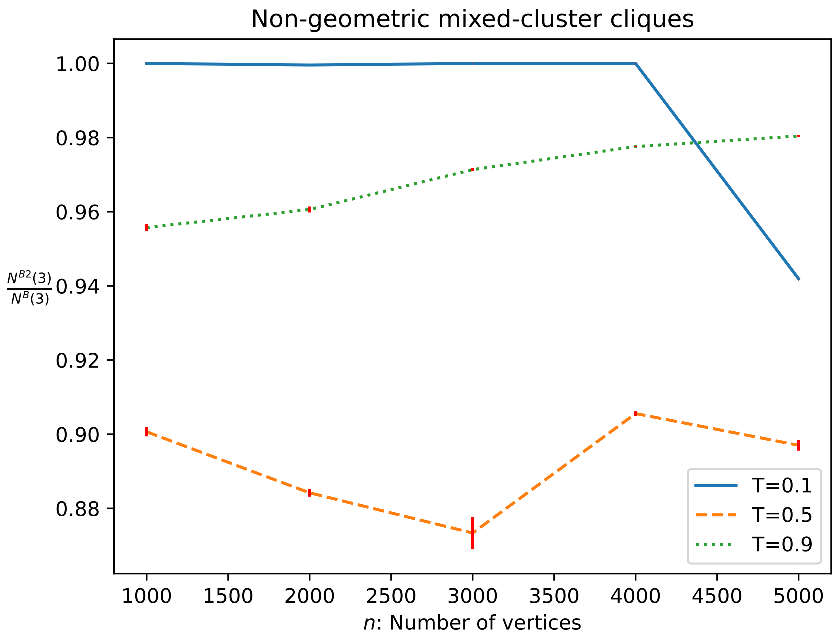

Figure 4.3: Ratio of the number of triangles in G^B and G^{B2} as a function of $n = |V|$ for different temperatures. In the parameter quadrant where non-geometric cliques with multiple clusters are dominant in $B2$.

Unfortunately the four diagrams show no clear, common trend. Some graphs tend to zero for increasing n e.g. the blue line in Figure 4.4, others indicate constant behaviour like the green line in Figure 4.4. For the two parameter sets with one-cluster cliques, the model variants behave almost the some for low temperature and lie further apart for high temperature, see Figures 4.5 and 4.6. Further experimentation is required to evaluate and interpret the observed trends.

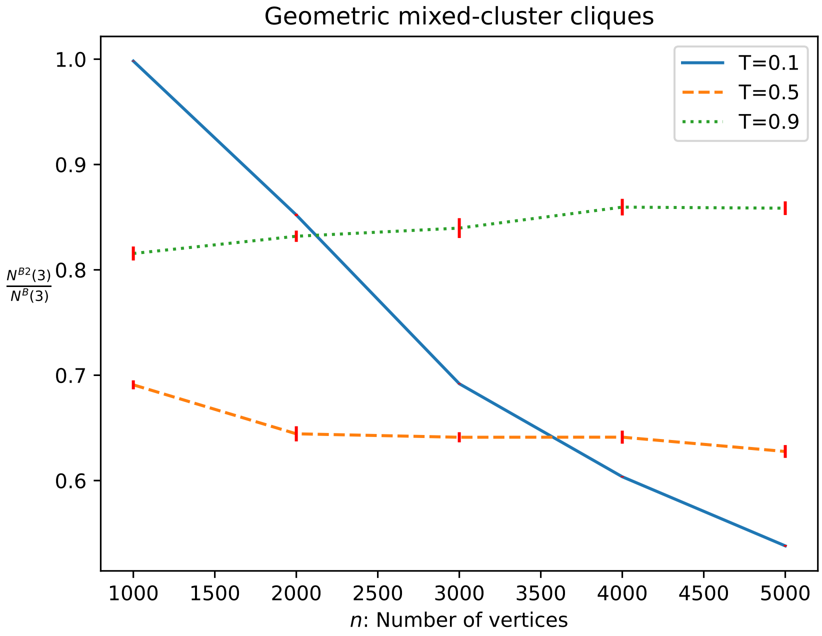

Figure 4.4: Ratio of the number of triangles in G^B and G^{B2} as a function of $n = |V|$ for different temperatures. In the parameter quadrant where geometric cliques with multiple clusters are dominant in $B2$.

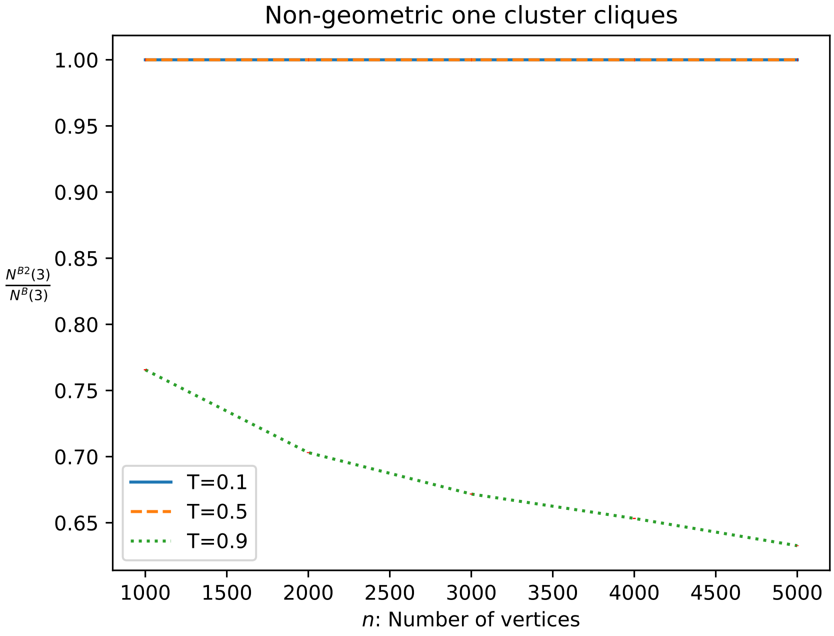

Figure 4.5: Ratio of the number of triangles in G^B and G^{B2} as a function of $n = |V|$ for different temperatures. In the parameter quadrant where non-geometric cliques with only one clusters are dominant in $B2$.

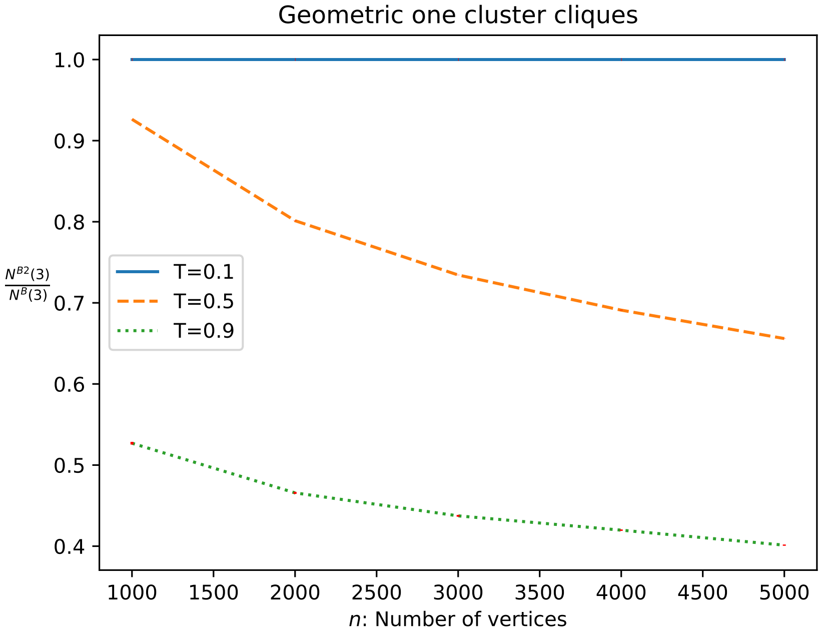

Figure 4.6: Ratio of the number of triangles in G^B and G^{B2} as a function of $n = |V|$ for different temperatures. In the parameter quadrant where geometric cliques with only one clusters are dominant in $B2$.

5 Conclusion

We discussed different approaches of defining a hybrid random graph model that is able to model continuous as well as discrete similarities between entities. The definition we came up with gives each vertex a power law weight, a point in the d-dimensional torus and a cluster membership. After stating two variants of our model, we investigated the asymptotic number of k-sized cliques for constant k in each of them. For the simplified version we were able to show a matching lower and upper bound. The resulting asymptotic showed a phase shift depending on the number of clusters and the factor diminishing the probability of inter-cluster connections.

We saw that the simplified model presents a lower bound for the other model. To approach an upper bound we suggested an experimental setup that compares the two model variants. Unfortunately we could not see a clear result in a first basic implementation yet.

With the definition of this new model many interesting questions turned up. A natural next step would be to further study the asymptotic of the second model to come up with an upper bound. In addition the investigation of k-sized cliques could be expanded to super constant values for k. With that one could also study the expected clique number of both models. Moreover, it would be interesting to see if some of the inherent GIRG characteristics, like the scale-free degree distribution or a small diameter, also apply to the hybrid models.

Bibliography

[BFK18] Thomas Bläsius, Tobias Friedrich, and Anton Krohmer. “Cliques in Hyperbolic Random Graphs”. In: *Algorithmica* Volume 80 (Sept. 2018), pp. 2324–2344. ISSN: 1432-0541. DOI: *10.1007/s00453-017-0323-3*.

[BKL16] Karl Bringmann, Ralph Keusch, and Johannes Lengler. “Average Distance in a General Class of Scale-Free Networks with Underlying Geometry”. In: *CoRR* Volume abs/1602.05712 (2016). arXiv: *1602.05712*.

[BKL19] Karl Bringmann, Ralph Keusch, and Johannes Lengler. “Geometric inhomogeneous random graphs”. In: *Theoretical Computer Science* Volume 760 (2019), pp. 35–54. ISSN: 0304-3975. DOI: *https://doi.org/10.1016/j.tcs.2018.08.014*.

[CL02a] Fan Chung and Linyuan Lu. “Connected Components in Random Graphs with Given Expected Degree Sequences”. In: *Annals of Combinatorics* Volume 6 (Nov. 2002), pp. 125–145. DOI: *10.1007/PL00012580*.

[CL02b] Fan Chung and Linyuan Lu. “The average distances in random graphs with given expected degrees”. In: *Proceedings of the National Academy of Sciences* Volume 99 (2002), pp. 15879–15882. eprint: *http://www.pnas.org/content/99/25/15879.full.pdf+html*.

[GMPS18] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. “The Geometric Block Model”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Volume 32 (Apr. 2018). DOI: *10.1609/aaai.v32i1.11905*.

[HLL83] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. “Stochastic blockmodels: First steps”. In: *Social Networks* Volume 5 (1983), pp. 109–137. ISSN: 0378-8733. DOI: *https://doi.org/10.1016/0378-8733(83)90021-7*.

[KN11] Brian Karrer and M. E. J. Newman. “Stochastic blockmodels and community structure in networks”. In: *Phys. Rev. E* Volume 83 (Jan. 2011), p. 016107. DOI: *10.1103/PhysRevE.83.016107*.

[KZ09] Michael Kaufmann and Katharina Zweig. “Modeling and Designing Real–World Networks”. In: *Algorithmics of Large and Complex Networks: Design, Analysis, and Simulation*. Edited by Jürgen Lerner, Dorothea Wagner, and Katharina A. Zweig. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 359–379. ISBN: 978-3-642-02094-0. DOI: *10.1007/978-3-642-02094-0_17*.

[MS22] Riccardo Michielan and Clara Stegehuis. “Cliques in geometric inhomogeneous random graphs”. In: *Journal of Complex Networks* Volume 10 (Feb. 2022). ISSN: 2051-1329. eprint: *https://academic.oup.com/comnet/article-pdf/10/1/cnac002/42466016/cnac002.pdf*.

[New03] M. E. J. Newman. “The Structure and Function of Complex Networks”. In: *SIAM Review* Volume 45 (2003), pp. 167–256. eprint: *https://doi.org/10.1137/S003614450342480*.

[Pen03] Mathew Penrose. *Random geometric graphs*. Vol. 5. OUP Oxford, 2003.