Universität Zürich

Master's Thesis

Connectedness and Giant Component in the Stochastic Block Model

Author:
Vanessa Rey
19-750-850

Supervisor:
Prof. Dr. Ashkan Nikeghbali
Co-supervisor:
Dr. Thomas Lehéricy

Universität Zürich[UZH]

Institute of Mathematics
Universität Zürich
December 2024

Abstract

In this Master's thesis, we develop the theory of random graphs, especially the Erdős-Rényi random graph. We introduce a random graph model with communities called the (Symmetric) Stochastic Block Model (SSBM). This model is characterized by two different connectivity probabilities of the vertices, depending on whether the vertices are in the same community or not. We show which assumptions have to be made on the connectivity probabilities such that the graph is connected. In addition, we use the theory of Poissonization of a random graph by Nicolas Curien, to poissonize this specific model and use this, to show which assumption has to hold on the connectivity probability such that the graph has a giant component.

Acknowledgements

I want to sincerely thank my supervisor Professor Dr. Ashkan Nikeghbali for giving me the opportunity to write my Master's Thesis in this field and for sparking my interest in probability theory with the various lectures I have attended with him. Further, I would like to express my sincerest gratitude to my co-supervisor Dr. Thomas Lehéricy, for his profound guidance and support during this time and for always answering my many questions along the way with much patience.

I would like to thank my friends from university who have accompanied me during this time, as well as my studies in general. A special thanks goes to Luana and Elia. It was a joy to complete this journey since the first semester with you. Going through non-physical times in Covid, back to the standard studying days and finishing a Master's degree.

I would like to express my gratitude towards my family, for their understanding and support in the background throughout my entire academic journey.

Last, but by no means least, I would like to thank Davide. Thank you for being by my side and believing in me, listening to my rants but also pushing me to peak performance and helping me with my various questions about analysis.

To everyone else that contributed in one way or the other to the success of this thesis and therefore the completion of my Master's degree, thank you. With high probability, I was having a good time.

Contents

1 Introduction **1**

2 Notation **5**

3 Preliminaries **7**

3.1 Basics of Graph Theory 7

3.2 Random Graphs 9

3.3 Stochastic Block Model 11

3.4 Average Degree in the Symmetric Stochastic Block Model 13

4 Connectedness in the Symmetric Stochastic Block Model **15**

4.1 If $\frac{a+(k-1)b}{k} < 1$ then SSBM is disconnected 15

4.2 If $\frac{a+(k-1)b}{k} > 1$, then SSBM is connected 20

5 Giant Component Poissonized **24**

5.1 Defining the Model 24

5.2 Lukasiewicz Walk 30

5.3 Relation between Poissonized Graph and Core 40

5.4 Phase Transition for the Giant Component 40

6 Conclusion and Open Problems **61**

7 Appendix **63**

References **65**

1 Introduction

A lot of real-world data sets can be represented as networks of interacting items, and it is often interesting to consider items with shared properties. The idea of clustering a random graph is to group vertices with similar properties into the same community, or to group them in a way that the vertices inside the clusters are more densely connected than across the clusters. The theory of random graphs and networks has applications in the analysis of sociological behaviour, gene expressions, recommendation systems, webpage sorting and many more (Abbe, 2018). In the field of sociology, a possible network can for example be the structure of friendships in a group of individuals. Clustering in this case is important because the probability of an edge between two individuals is much higher if the two vertices in question have another mutual acquaintance, or several. In the case of the group of friends, this means that the probability that two of your friends know one another is much higher than the probability that two people chosen randomly from the population know one another (M. E. J. Newman, Watts and Strogatz, 2002). Therefore, it makes sense to cluster the different friendships such that the probability of having an edge inside the cluster (knowing one another inside the group of friends) is higher, than the probability of having an edge between different communities (knowing someone from a different group of friends).
It is however not a necessary condition that the probability to connect vertices inside the communities is higher, this depends on what the model should represent. One of the most used random graph models with clusters is the Stochastic Block Model, which is the main topic of this thesis.

This work starts with an introduction to graph theory. A graph is a mathematical object with vertices and edges. The edges connect two vertices with each other (or are self-loops). In our case, the edges are however not determined, but chosen randomly. This gives a random graph where one can put probabilities on whether an edge is present or not. The most famous random graph model is the Erdős-Rényi Random Graph. We have as vertex set the set $\{1, 2, ..., n\}$ and for each pair of vertices $\{x, y\}$, with $x \neq y \in \{1, 2, ..., n\}$, there is an edge between the two vertices with probability p, independently of all other edges and there is no edge with probability $1 - p$. In this work, we study two different graph properties on random graphs, namely connectedness of a graph and the existence of a giant component. For a graph to be connected, any two vertices in the graph need to be connected through a series of edges, On the other side, the idea of the giant component can be applied, when connectedness does not hold. In this case, we still have a large fraction of vertices that are connected to each other inside this giant component and only a negligible part is isolated from the giant component.

The main part of this thesis is built on the Stochastic Block Model (SBM), which is a random graph model with clusters, contrary to an Erdős-Rényi random graph that does not have clusters.

This means that we group the vertices in the SBM into different communities with a certain probability, which is helpful to model real-world networks. A SBM is built by using the vertex set $\{1, 2, ..., n\}$, the number of communities k and the probability vector $p = (p_1, p_2, ..., p_k)$. The latter being the prior probability on the communities, defining how likely it is that a vertex lies in the given community. In addition, we need a $(k \times k)$-symmetric matrix W with entries in $[0, 1]$, where the entry w_{ij}, $1 \leq i, j \leq k$, defines the probability of connecting a vertex from community i to a vertex from community j. In some sense, the prior probability p defines where the vertices are (in which community) and the matrix W defines where the edges are (which vertices are connected through an edge). A special case of the SBM is if the entries of W have the same value on all of the diagonal (q_{in}, for the connections *inside* the community) and a different value off the diagonal (q_{out}, for the connections *outside* of the community) and p is uniform. Then, we speak of the *symmetric* Stochastic Block Model (SSBM). The idea here is that it does not matter in which communities the vertices lie, we are only interested if they are in the same or in different communities. Hence, the probability of connecting two vertices of different communities is the same over all clusters, and no preference is made. See Figure 1 for an illustrative example of the SSBM.

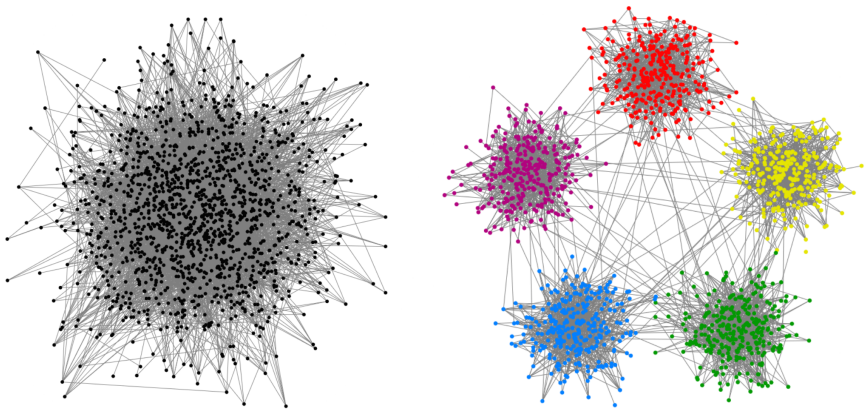


Figure 1: These two graphs represent the same graph drawn from the SBM with 1000 vertices, 5 communities, all with the same number of vertices, $q_{in} = \frac{1}{50}$ and $q_{out} = \frac{1}{1000}$. The goal of clustering in this case is to get the right graph with the communities grouped together from the scrambled graph on the left, up to some level of accuracy. (Abbe, 2018)

As an initial investigation of the SSBM, we analyse the degree distribution in the Symmetric Stochastic Block Model, meaning we study what the distribution of the number of edges at each vertex is and what the expected number of edges at a vertex is. We see that in the case where the in-community probability is $q_{in} = \frac{a}{n}$ and the across community probability is $q_{out} = \frac{b}{n}$, then the expected degree is $\frac{a+(k-1)b}{k}$, where $a, b > 0$ are fixed. On the other hand, if $q_{in} = a\frac{\ln(n)}{n}$ and

$q_{out} = b\frac{\ln(n)}{n}$, then the expected degree is $\frac{a+(k-1)b}{k}\ln(n)$.

As mentioned earlier, we are interested in studying if certain graph properties are present with probability 1, if we let the number of vertices n go to infinity. The first property we investigate, is the one of connectedness in the SSBM. It turns out that the following proposition holds.

Proposition. *We have the following two cases*

i) If $\frac{a+(k-1)b}{k} < 1$*, then the probability that* SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ *is disconnected tends to 1 as* $n \to \infty$*.*

ii) If $\frac{a+(k-1)b}{k} > 1$*, then the probability that* SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ *is connected tends to 1 as* $n \to \infty$*.*

The proposition tells us that if a vertex has on average more than $\ln(n)$ edges attached to it, the model is connected and it is, with probability tending to 1 as n goes to infinity, not connected if a vertex has on average less than $\ln(n)$ edges attached. To prove this, we use isolated vertices, i.e. vertices that have no edges attached to them. Then we show that if we assume $\frac{a+(k-1)b}{k} < 1$, the probability that there exists at least one isolated vertex goes to 1 and hence the graph cannot be connected, as there is no way to get to this isolated vertex from any other vertex.

As mentioned earlier, if connectivity does not need to hold, we can wonder if there exists a large fraction of vertices that is still connected and only a small part remains disconnected from them. This is the theory of the giant component and to prove such a statement, we use the technique by Nicolas Curien (Curien, 2024b) to poissonize the SSBM. This means that we randomise the number of vertices in the communities according to a Poisson distribution with parameter α_i for each community $i \in \{1, ..., k\}$. Then, we attach a stack of vertices to each community and connect the vertices of the stack to the vertices of the *core* (the SSBM graph) with probability q_{in} if the vertex of the stack is assigned to the same community as the vertex in the core or with probability q_{out} if they are in different communities. Using this, we can *explore* the graph and its neighbouring vertices. We start with a vertex in the stack belonging to community 1 and reveal its neighbours. Then, we remove the vertex of the stack and place the neighbouring vertices inside the stack to explore their neighbours. We analyse the distribution of this sequence of explorations and show that we can encode it in a random walk $(\underline{\mathbb{S}}_i : i \geq 0)$. The main result is that this stochastic process converges a.s. to a limit.

Proposition. *It holds that*

$$\left(\frac{\mathbb{S}_{\lfloor nt \rfloor}}{n/k}\right)_{t\geq 0} \xrightarrow[n\to\infty]{a.s.} \begin{pmatrix} 1-e^{-\frac{a+(k-1)b}{k}t}-t \\ \dots \\ 1-e^{-\frac{a+(k-1)b}{k}t}-t \end{pmatrix}_{t\geq 0}$$

in the uniform norm over t where the vector on the right is in $\mathbb{R}^k$.

With this result, we explain how one can attain the condition for the existence of a giant component in the SSBM with connection probability $q_{in} = \frac{a}{n}$ inside the communities and $q_{out} = \frac{b}{n}$ outside the communities.

Claim. *If $c = \frac{a+(k-1)b}{k} > 1$, then there exists a unique giant component roughly of size $\beta(c)n$ in the core, where $\beta(c)$ is the first positive root of $1 - e^{-\frac{a+(k-1)b}{k}t} - t$. On the other hand, if $\frac{a+(k-1)b}{k} < 1$, then the size of all components vanishes asymptotically for $n \to \infty$.*

Lastly, we give concluding remarks and also reference to open problems and further research on the topic.

2 Notation

Before starting with the preliminaries of this thesis, let us introduce some notation that is used throughout this thesis.

- We denote by $\mathbb{N} = \{0, 1, 2, 3, 4...\}$ and $\mathbb{N}^* = \{1, 2, 3, 4...\}$.
- We denote by $[n] := \{1, 2, ..., n\}$ the set including the integers from 1 to n.
- $\mathbf{1}$ is the all one vector in the respective dimension, i.e. $\mathbf{1} = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$. Respectively, $\mathbf{a}$ denotes the vector with all entries equal to a.
- Vectors are indicated by an underline: $\underline{L} = \begin{pmatrix} L_1 \\ ... \\ L_k \end{pmatrix}$.
- We use the standard little-o and big-O notations, meaning that $f_n = o(g_n)$ if $\lim_{n\to\infty} \left|\frac{f_n}{g_n}\right| = 0$ and $f_n = O(g_n)$ if $\limsup_{n\to\infty} \left|\frac{f_n}{g_n}\right| < \infty$. In particular, $f_n = o(1)$ means that f_n is vanishing as $n \to \infty$.
- Define by $\lfloor x \rfloor$ the integer value of x, i.e. $0 \leq x - \lfloor x \rfloor < 1$. Furthermore, we have that $x = \lfloor x \rfloor + \{x\}$, where $\{x\} \in [0, 1[$ is the fractional part of x.
- Let $\sigma(X_i : i \geq 0)$ be the σ-algebra generated by the random variables $X_i, i \geq 0$.
- We denote by $\text{Poi}(\lambda)$ the Poisson distribution with parameter λ, i.e. if $X \sim \text{Poi}(\lambda)$ then $\mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$ for all $k \in \mathbb{N}$.
- Similarly, we denote by $\text{Bin}(n, p)$ a Binomial random variable with n trials and success probability p, i.e. if $X \sim \text{Bin}(n, p)$, then $\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ for all $0 \leq k \leq n$.
- We write

$$\underline{X} \sim \text{Poi} \begin{pmatrix} \alpha_1 \\ ... \\ \alpha_j \\ ... \\ \alpha_k \end{pmatrix} \tag{1}$$

for a random vector with independent coordinates, X_i, where each coordinate is distributed according to a Poisson random variable with parameter α_i. We use the following notation

and the notation in (1) interchangeably.

$$\underline{X} \sim \begin{pmatrix} \text{Poi}(\alpha_1) \\ \dots \\ \text{Poi}(\alpha_j) \\ \dots \\ \text{Poi}(\alpha_k) \end{pmatrix}$$

Both denote the same random vector with independent coordinates.

- By *a.s.* we always denote almost surely, and by *i.i.d.* we mean independent and identically distributed.
- We say a stochastic process $(X_t)_{t\geq 0}$ has *càdlàg sample paths* if a.s. all sample paths are right continuous with left limits (**c**ontinue **à** **d**roite, **l**imites **à** **g**auche).
- Let the indicator function be denoted with

$$\mathbb{1}_{\{i=x\}} = \begin{cases} 1 & \text{if } i = x \\ 0 & \text{if } i \neq x. \end{cases}$$

- Lastly, in Chapter 5 we always use the same notation, which is listed here, so it can be found quickly.
  - $c := \frac{a+(k-1)b}{k}$
  - $A := (1 - \frac{a}{n})(1 - \frac{b}{n})^{k-1}$

3 Preliminaries

3.1 Basics of Graph Theory

Before we can start defining the models used in this paper, we need to build some basic knowledge on graph theory. We start with graphs in general, then we move on to random graphs and the Stochastic Block Model.

Definition 3.1 (Diestel, 2017)**.** A *graph* is a pair $G = (V, E)$ of sets such that $E \subseteq V^2$. We call the elements of V the *vertices* and the elements of E are called *edges*. The vertex set of a graph G is referred to as $V(G)$ and the edge set as $E(G)$.

By convention, we number the vertices from $1, ..., n$, thus $V(G) = \{1, 2, 3, ..., n\}$, and hence call it a *labelled* graph. Furthermore, we work with *undirected* graphs, meaning the edges have no direction, i.e. the edge $\{1, 2\}$ is the same as the edge $\{2, 1\}$. Let us also define what it means when an edge starts and ends at the same vertex.

Definition 3.2. We call a *self-loop* an edge that starts and ends at the same vertex, for example we would have a self-loop at vertex 1 if $\{1, 1\} \in E$.

Remark 3.3. If we have n vertices, then the total number of possible undirected edges (without self-loops) is $\binom{n}{2}$.

Example 3.4. Let $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{2, 5\}, \{3, 4\}\}$. Then, a possible display of the graph is given in Figure 2.

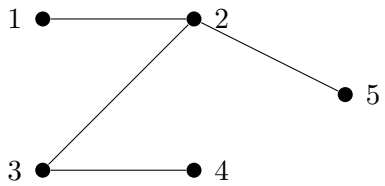


Figure 2: A labelled graph with vertices 1 to 5.

Next, let us give some definitions on the connecting edges.

Definition 3.5 (Diestel, 2017)**.** A vertex v is *incident* with an edge e if $v \in e$. Two vertices x, y of G are *adjacent* (or *neighbours*) if $\{x, y\}$ is an edge of G.

Definition 3.6 (Diestel, 2017)**.** We call a graph *complete* if all the vertices of G are pairwise adjacent.

Another important concept is the one of subgraphs.

Definition 3.7 (Roch, 2024)**.** A *subgraph* of $G = (V, E)$ is a graph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$. Implied in this definition is the fact that the edges in E' are incident with V' only. A subgraph is said to be *spanning* if $V' = V$.

Note that the preceding definitions do not exclude that there are multiple edges between two vertices of a graph. To exclude this case, we introduce the next definition.

Definition 3.8. We call a graph *simple* if it has no multiple edges between any two vertices.

We can also wonder what the number of edges at a given vertex tells us.

Definition 3.9 (Diestel, 2017)**.** The *degree* $d_G(v) = d(v)$ of a vertex v is the number of edges at v. A vertex of degree 0 is called *isolated*.

Now, let us define what it means to connect different vertices with each other.

Definition 3.10 (Roch, 2024)**.** A *walk* in G is a sequence of (not necessarily distinct) vertices $x_0 \sim x_1 \sim ... \sim x_k$. Note that we need the requirement that consecutive vertices of a walk are adjacent. If the vertices of a walk are all distinct, we call it a *path*. If the endvertices x_0, x_k of a path coincide, we call it a *cycle*. See Figure 3 for an illustration of this.

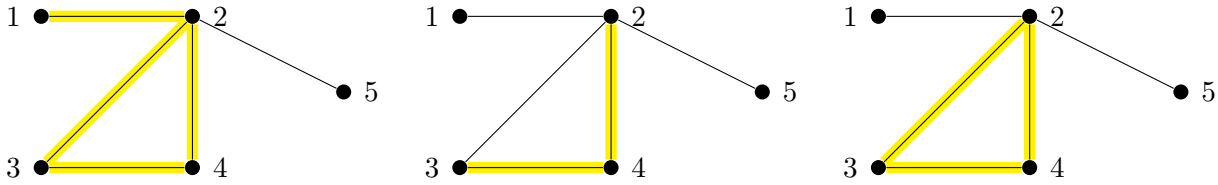


Figure 3: On the left the highlighted part is a walk, in the middle it is a path and on the right it is a cycle.

Using the theory of walks, we can define a fundamental concept of this thesis.

Definition 3.11 (Diestel, 2017)**.** A graph G is called *connected* if it is non-empty and any two of its vertices are linked by a path in G. In other words, there are no isolated vertices. See Figure 4 for an example of this. A graph which is not connected is called *disconnected*.

Figure 4: On the left we have a connected graph, while the graph on the right is disconnected (as 5 is an isolated vertex)

.

Lastly, let us introduce a specification of the graphs using cycles.

Definition 3.12 (Roch, 2024)**.** A *forest* is a graph with no cycle. A *tree* is a connected forest. A *spanning tree* of G is a subgraph which is a tree and is also spanning.

Note that a tree is exactly like one would think of a tree. It has paths that only go one way and vertices of degree one are called *leaves*, these are the ones that are at the end.

3.2 Random Graphs

Now as we have the theory on graphs, let us move on to random graphs, meaning the edges are not fixed anymore but they are built randomly. What this means exactly is described in Definition 3.13.

Definition 3.13 (Roch, 2024)**.** Let $n \in \mathbb{N}$ and $p \in [0,1]$. Set $V := [n]$. Under the *Erdős-Rényi graph model* on n vertices with density p, a random graph $G = (V, E)$ is generated as follows: For each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability p, independently of all other edges. We write $G \sim G(n, p)$.

In this model, we say that an edge is *present* with probability p and *absent* with probability $(1 - p)$.

Remark 3.14 (M. Newman, 2018)**.** Note that the model is called "Erdős-Rényi model" because it is mostly associated with the mathematicians Paul Erdős and Alfréd Rényi.

The general model is called the "Binomial Random Graph Model" and is defined as follows.

Definition 3.15 (Frieze and Karonski, 2024)**.** First, fix $0 \leq p \leq 1$. For $0 \leq m \leq \binom{n}{2}$, we assign to each graph G with vertex set $[n]$ and m edges a probability

$$\mathbb{P}[G] = p^m (1-p)^{\binom{n}{2} - m}.$$

Equivalently, we can start with an empty graph with vertex set $[n]$ and perform $\binom{n}{2}$ Bernoulli experiments inserting edges independently with probability p. Such a random graph is called a *Binomial Random Graph*.

We are often interested in what happens if we let the number of vertices go to infinity, i.e. $n \to \infty$. For this, we use Definition 3.16.

Definition 3.16. We say that a graph property $\mathscr{P}$ happens *with high probability* if $\mathbb{P}[G(n,p) \in \mathscr{P}] \underset{n \to \infty}{\longrightarrow} 1$.

Note that "a graph property" can for example be that the graph is connected or has a certain number of triangles and so on.

Let us continue to study the properties of Erdős-Rényi Random Graphs. It is observable that the probability that a graph has a certain property often "jumps" at a given value of the connectivity probability p, i.e. if p is below a certain value the graph property does not hold but if it is above this value then the graph property holds with high probability.

One of the most interesting graph properties, which is also widely studied, is the connectivity of a random graph. Theorem 3.17 gives a possible statement for this.

Theorem 3.17 (Durrett, 2007)**.** *Consider* $G = G(n, \frac{c \ln(n)}{n})$*. The probability that* G *is connected tends to* 0 *if* $c < 1$ *and to* 1 *if* $c > 1$*.*

A proof for Theorem 3.17 can be found in (Acemoglu and Ozdaglar, 2009) and the adapted version of this proof for the Symmetric Stochastic Block Model is presented in Chapter 4.

Remark 3.18. We can make a connection between the value of c in Theorem 3.17 and the degree of a vertex in the model. As we randomly draw edges between the n vertices with probability p, we have that each vertex has a degree distribution according to $\text{Bin}(n-1, p)$. Let d_i denote the degree of vertex i, then we have that

$$\mathbb{P}[d_i = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

by the definition of the model because we choose k vertices among the $n-1$ other vertices that will be connected to i. Consequently, these vertices have probability p and the other ones are not connected with probability $(1-p)$. Hence, the expected degree is

$$\mathbb{E}[d_i] = (n-1) \cdot p.$$

Therefore, if we have the model $G(n, \frac{c\ln(n)}{n})$, the expected degree is $(n-1) \cdot \frac{c\ln(n)}{n} \approx c\ln(n)$ for big n. This means, if the expected degree grows faster than $\ln(n)$, we have a connected graph and if it grows slower than $\ln(n)$ then we have a disconnected graph.

If the graph is not connected, we can wonder if there is a part of the graph where a lot of the vertices are connected through a path. This is the principle of the "giant component", i.e. a connected component that includes a "big" proportion of the total number of vertices. For this, let us first introduce some terminology.

Definition 3.19 (Curien, 2024a)**.** We can build an equivalence relation on the paths. We say that $x \leftrightarrow y$ if there is a path between x and y. The equivalence classes that result out of this relation are called *connected components*.

We can think of the connected components as clusters, in which it is possible to move from any vertex in the cluster to any other vertex in the cluster without leaving the cluster.

Notation 3.20 (Curien, 2024a)**.** We denote the connected component of a vertex v in $G(n,p)$ as $\mathcal{C}^v$. Furthermore, we denote by $\text{C}_1^{\max}, \text{C}_2^{\max}, \ldots$ the sizes (i.e. the number of vertices) of the connected components of $G(n,p)$ in non-increasing order. Meaning, $\text{C}_1^{\max}$ is the biggest cluster and the following are either of the same size or smaller.

Remark 3.21 (Curien, 2024a)**.** Note that a graph is connected if and only if it has only one connected component.

Now, we want to study these connected components and their sizes.

Theorem 3.22 (Curien, 2024a)**.** *For the giant connected component inside the graph $G(n, \frac{c}{n})$ it holds*

i) ***Subcritical: If $c < 1$****, then there exists a number $A > 0$ depending on $c > 0$ such that with high probability we have $\mathrm{C}_1^{\max}(G(n, \frac{c}{n})) \leq A \ln(n)$.*

ii) ***Supercritical: If $c > 1$****, then there exists a number $A > 0$ depending on $c > 0$ such that with high probability we have $\mathrm{C}_2^{\max}(G(n, \frac{c}{n})) \leq A \ln(n)$. Moreover, for the biggest connected component it holds $n^{-1}\mathrm{C}_1^{\max}(G(n, \frac{c}{n})) \underset{n\to\infty}{\to} (1 - \alpha(c))$ in probability where $\alpha(c)$ is the smallest solution in $(0, 1]$ to the equation*

$$x = e^{-c(1-x)}. \tag{2}$$

A proof for Theorem 3.22 can be found in (Curien, 2024b) and the adapted version to the Symmetric Stochastic Block Model can be found in this thesis in Chapter 5.

Remark 3.23. We can also state the supercritical case in a different but equivalent way by making the substitution $t = 1 - x$ in (2). Then, $n^{-1}\mathrm{C}_1^{\max}(G(n, \frac{c}{n})) \to \beta(c)$ in probability where $\beta(c)$ is the smallest solution in $(0, 1]$ to the equation $1 - e^{-ct} - t = 0$.

The idea of Theorem 3.22 is that if the connection probability p_n is strictly smaller than $\frac{1}{n}$, all connected components are small, meaning of size smaller than $\ln(n)$. Whereas if the connection probability is strictly bigger than $\frac{1}{n}$, then there exists a unique giant component which grows linearly in n, i.e. it has roughly size $n(1 - \alpha(c))$ (or size $n\beta(c)$), while the other component sizes are small.

Remark 3.24 (Curien, 2024a)**.** In the critical case that $p_n = \frac{1}{n}$, we have that the size of the biggest component is with high probability proportional to $n^{2/3}$.

Remark 3.25. As in Remark 3.18, we can also calculate the expected degree in this model, which is

$$\mathbb{E}[d_i] \approx n \cdot c/n = c.$$

We notice that the graph has a giant component if the expected degree is bigger than 1, and it does not have a giant component if the expected degree is smaller than 1.

Now let us continue with the main model in this thesis, the Stochastic Block Model.

3.3 Stochastic Block Model

The idea of the Stochastic Block Model is that we do not just have all n vertices and connect them, but that we can *cluster* the vertices, meaning we can build different groups with them. An example for this is that we have articles about the election in the US and we want to group them to either being democratic or republican. The edges could for example represent the references

that articles make to other articles. Hence, the general idea of clustering is to group vertices with similar properties into the same cluster/group, also called *community*. Most of the time we also want that the vertices have more connections within the community than across communities (Abbe, 2018).

Definition 3.26 (Abbe, 2018)**.** (Stochastic Block Model) Let n be a positive integer (the number of vertices), k another positive integer (the number of communities), $p = (p_1, ..., p_k)$ a probability vector on $[k]$ (the prior on the communities) and W a $(k \times k)$-symmetric matrix with entries in the interval $[0, 1]$ (the connectivity probabilities). The pair (X, G) *is drawn under* $SBM(n, k, p, W)$ if X is an n-dimensional random vector with i.i.d. components distributed under p and, conditionally on X, G is an n-vertex simple graph where vertices i and j are connected with probability W_{X_i, X_j}, independently of other pairs of vertices. We also define the *community sets* by $\Omega_i = \Omega_i(X) := \{v \in [n] : X_v = i\}$, $i \in [k]$.

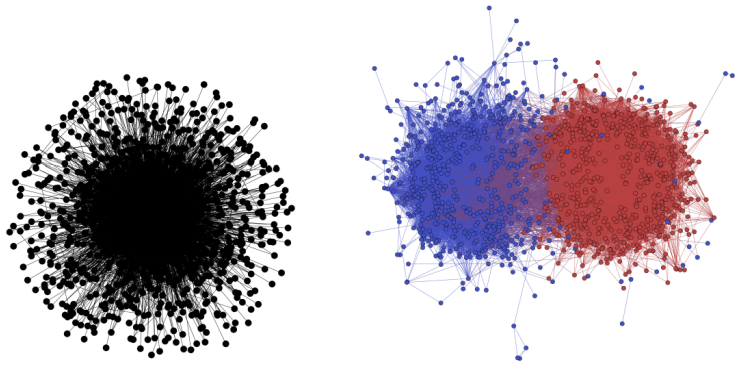


Figure 5: This figure shows the two graphs obtained by plotting the real data set of political blogs in the US. Each vertex is a blog and each edge represents the fact that one of the blogs refers to the other. The left graph is plotted with a random arrangement of the vertices, while the right one is obtained by applying some clustering algorithm on it. The blue and red colour indicate left and right leaning blogs with an accuracy of 95%. (Abbe, 2018)

Let us look at Definition 3.26 in more depth. The idea is that X is the distribution of the vertices: The component X_1 defines in which community vertex 1 lies, X_2 defines vertex 2 and so on. The probability vector p defines which communities are more likely to be chosen, and the matrix W defines with which probability we connect the vertices. Therefore, if we choose two vertices, say 1 and 2, we first need to check in which community they are, say $X_1 = 4$ and $X_2 = 1$, so vertex 1 is in community 4 and vertex 2 is in community 1. Then, we can look into our matrix at row 4 and column 1 to get the connection probability for these two vertices. Important here is to note that all vertices in community 4 have the same probability to be connected to vertices in community 1, not just the two specific vertices we choose. Typically, we assume that W scales with the number of vertices n, while the number of communities, and therefore p, does not scale with n.

Remark 3.27 (Abbe, 2018)**.** If all entries of W are the same, then the model collapses to the Erdős-Rényi model with i.i.d. labels, since the probability to connect two vertices is the same everywhere, no matter in which community they are.

Remark 3.28 (Abbe, 2018)**.** Note that by the law of large numbers, we have

$$\frac{1}{n}|\Omega_i| \underset{n\to\infty}{\longrightarrow} p_i \text{ almost surely.}$$

This is the setup for the general SBM model. To make computations easier, we are going to simplify the model by only looking at the *Symmetric* Stochastic Block Model. This is the case when p is uniform and W takes the same value on the diagonal and outside the diagonal. The idea is that we only differentiate if we make an edge **inside** the community or **across** communities. We do no longer differentiate between the different communities.

Definition 3.29 (Abbe, 2018)**.** (X, G) *is drawn under SSBM*(n, k, q_{in}, q_{out}) if W takes value q_{in} on the diagonal and q_{out} off the diagonal, and the community prior is $p = \mathbf{1}/\mathbf{k} \in \mathbb{R}^k$.

Notice that we choose each community with the same probability, so there do not need to be the same number of vertices in each community.

Now that we have the general setup of the model, we are also interested in analysing connectivity and the giant component in the SBM or SSBM. This is interesting because the Stochastic Block Model, and more general community detection, are used in different settings, for example to understand sociological behaviour, gene expressions, recommendation systems and many more. The idea when using the SBM for community detection is that we can observe a graph and then we want to recover the labels X to build the clusters (for example build the different friend groups). There are different types of recoveries for the community labels, for example *exact recovery* or *weak recovery*. However, we will not give the definitions here, but only explain why connectedness and giant component are useful for this. For exact recovery to be possible, the graph needs to be connected and for weak recovery to be possible, the graph needs to have at least a giant component (Abbe, 2018). This is why we are interested in analysing the behaviour of the SSBM (or SBM), which is done in the Chapters 4 and 5.

3.4 Average Degree in the Symmetric Stochastic Block Model

First, let us also establish the degree distribution in the Symmetric Stochastic Block Model, which is helpful later on. Let us look at the model SSBM$(n, k, a/n, b/n)$ with $q_{in} = a/n$ $q_{out} = b/n$ and $a, b > 0$. As we are in the symmetric case, we have that each community has asymptotically size $\frac{n}{k}$ (Abbe, 2018). Let d_i denote the degree of vertex i, C_i denote the community the vertex is in, $B_{in} \sim \text{Bin}(\frac{n}{k}, q_{in})$ and $B^j_{out} \sim \text{Bin}(\frac{n}{k}, q_{out})$, for $1 \leq j \leq k$ and $j \neq i$ be independent random variables. Then, we have that d_i has distribution (assuming self-loops

are also possible)

$$d_i = B_{in} + \sum_{\substack{j=1 \\ j \neq i}}^{k} B_{out}^{j}. \tag{3}$$

Hence, for the expected degree, we get

$$\mathbb{E}[d_i] = q_{in} \cdot \frac{n}{k} + \sum_{\substack{j=1 \\ j \neq i}}^{k} q_{out} \cdot \frac{n}{k} = q_{in} \cdot \frac{n}{k} + (k-1) \cdot q_{out} \cdot \frac{n}{k} = \frac{q_{in} \cdot n + (k-1) \cdot q_{out} \cdot n}{k}.$$

Replacing $q_{in} = \frac{a}{n}$ and $q_{out} = \frac{b}{n}$, which is given by our model, we get

$$\mathbb{E}[d_i] = \frac{\frac{a}{n} \cdot n + (k-1) \cdot \frac{b}{n} \cdot n}{k} = \frac{a + (k-1) \cdot b}{k}.$$

So, the expected degree for each vertex is $\frac{a+(k-1)\cdot b}{k}$ in the SSBM$(n, k, a/n, b/n)$ model.

In the SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ model we similarly get the average degree

$$\mathbb{E}[d_i] = \frac{a\frac{\ln(n)}{n} \cdot n + (k-1) \cdot b\frac{\ln(n)}{n} \cdot n}{k} = \frac{a + (k-1)b}{k} \ln(n).$$

4 Connectedness in the Symmetric Stochastic Block Model

Now, that we have all the preliminaries set, let us first analyse when the SSBM is connected. The idea is to develop a theorem similar to Theorem 3.17 but in the case of the Symmetric Stochastic Block Model SSBM$(n, k, a\frac{ln(n)}{n}, b\frac{ln(n)}{n})$. The proof of Theorem 4.1 is an adaptation of (Acemoglu and Ozdaglar, 2009).

Theorem 4.1. *We have the following two cases*

i) If $\frac{a+(k-1)b}{k} < 1$*, then* SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ *is disconnected with high probability.*

ii) If $\frac{a+(k-1)b}{k} > 1$*, then* SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ *is connected with high probability.*

Where we remember the definition of "with high probability" from Definition 3.16.

Proof. Let us start with part i).

4.1 If $\frac{a+(k-1)b}{k} < 1$ then SSBM is disconnected

We want to show if $\frac{a+(k-1)b}{k} < 1$, then $\mathbb{P}[\text{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ is disconnected$] \underset{n\to\infty}{\longrightarrow} 1$. Equivalently, we want to show that the probability that there exists at least one isolated node goes to 1.

Note that the cardinality of each community is $\frac{n}{k}$ as we are in the symmetric model.
Let I_i, for $1 \leq i \leq n$, be a Bernoulli random variable defined as

$$I_i = \begin{cases} 1 & \text{if node i is isolated} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then it holds (using $q_{in} := a\frac{\ln(n)}{n}$ and $q_{out} := b\frac{\ln(n)}{n}$) that the probability that vertex i is isolated satisfies

$$q := \mathbb{P}[I_i = 1] = (1 - q_{in})^{\frac{n}{k}}(1 - q_{out})^{(k-1)\frac{n}{k}}.$$

This holds, as we neither want to have any connections to other vertices in the same community (and also no self-loops), nor any connections to the $k - 1$ other communities.

Now, we want to approximate this probability

$$(1-q_{in})^{\frac{n}{k}} = \left((1-q_{in})^n\right)^{\frac{1}{k}} = \left(e^{n\ln(1-q_{in})}\right)^{\frac{1}{k}} = \left(e^{n\ln(1-a\frac{\ln(n)}{n})}\right)^{\frac{1}{k}} \overset{(*)}{=} \left(e^{n\left(-a\frac{\ln(n)}{n}+O\left(\frac{\ln(n)^2}{n^2}\right)\right)}\right)^{\frac{1}{k}}$$

$$= \left(e^{-a\cdot\ln(n)}e^{O\left(\frac{\ln^2(n)}{n}\right)}\right)^{\frac{1}{k}} = e^{-\frac{a\ln(n)}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)} \approx e^{-\frac{a\ln(n)}{k}}$$

where in $(*)$ we used the standard approximation $\ln(1-x) = -x + O(x^2)$ and the last approximation holds for big n.

Similarly, we get

$$(1-q_{out})^{(k-1)\frac{n}{k}} = \left((1-q_{out})^{\frac{n}{k}}\right)^{k-1} = \left(e^{-\frac{b\ln(n)}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)}\right)^{k-1} = e^{-\frac{(k-1)b\ln(n)}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)}$$

$$\approx e^{-\frac{(k-1)b\ln(n)}{k}}.$$

Hence, we can write q as

$$q = e^{-\frac{a\ln(n)}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)}e^{-\frac{(k-1)b\ln(n)}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)} = e^{\left(-\frac{a+(k-1)b}{k}\right)\ln(n)}e^{O\left(\frac{\ln^2(n)}{n}\right)} \tag{5}$$

$$= n^{-\frac{a+(k-1)b}{k}}e^{O\left(\frac{\ln^2(n)}{n}\right)} = n^{-c}e^{O\left(\frac{\ln^2(n)}{n}\right)} \tag{6}$$

with $c := \frac{a+(k-1)b}{k}$.

Next, let $X := \sum_{i=1}^n I_i$ denote the total number of isolated nodes. We want to show for $\frac{a+(k-1)b}{k} < 1$ that $\mathbb{P}[X=0] \to 0$ for $n \to \infty$.

We have

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n I_i\right] = \sum_{i=1}^n \mathbb{E}[I_i] = \sum_{i=1}^n \mathbb{P}[I_i = 1] = \sum_{i=1}^n n^{-c}e^{O\left(\frac{\ln^2(n)}{n}\right)} = n\cdot n^{-c}e^{O\left(\frac{\ln^2(n)}{n}\right)}. \tag{7}$$

If $c < 1$, we have

$$\mathbb{E}[X] \underset{n\to\infty}{\longrightarrow} \infty. \tag{8}$$

Now, we want to show that this implies $\mathbb{P}[X=0] \to 0$ as $n \to \infty$.

For this we compute

$$\mathrm{Var}(X) = \mathrm{Cov}\left(\sum_{i=1}^{n} I_i, \sum_{j=1}^{n} I_j\right) = \sum_{i=1}^{n} \mathrm{Var}(I_i) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Cov}(I_i, I_j) = n\cdot\mathrm{Var}(I_1) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Cov}(I_i, I_j) \tag{9}$$

where we used that the I_i's are identically distributed.

Let us look at the covariance in detail

$$\sum_{i=1}^{n}\sum_{j\neq i} \mathrm{Cov}(I_i, I_j) = \sum_{i=1}^{n}\left(\sum_{\substack{j\neq i\\ X_i = X_j}} \mathrm{Cov}(I_i, I_j) + \sum_{\substack{j\neq i\\ X_i \neq X_j}} \mathrm{Cov}(I_i, I_j)\right) = \sum_{i=1}^{n} A^i + B^i$$

with $A^i := \sum_{\substack{j\neq i\\ X_i = X_j}} \mathrm{Cov}(I_i, I_j)$ and $B^i := \sum_{\substack{j\neq i\\ X_i \neq X_j}} \mathrm{Cov}(I_i, I_j)$.

There are two cases, namely that the vertices i and j are in the same community, for this there are $\frac{n}{k} - 1$ possibilities and the second case is that i and j are not in the same community, for this there are $(k-1)\cdot\frac{n}{k} = n - \frac{n}{k}$ possibilities. Let us analyse the two cases further.

Case 1: Assume that the two vertices i and j are in the same community and that the vertices 1 and 2 are also in the same community. As the I_i's are identically distributed, we get

$$A^i = \sum_{\substack{j\neq i\\ X_i = X_j}} \mathrm{Cov}(I_i, I_j) = \left(\frac{n}{k} - 1\right)\mathrm{Cov}(I_1, I_2) = \left(\frac{n}{k} - 1\right)(\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]). \tag{10}$$

Next, let us calculate $\mathbb{E}[I_1 I_2]$. As $I_i \in \{0, 1\}$, we get

$$\begin{aligned}\mathbb{E}[I_1 I_2] = \mathbb{P}[I_1 = 1, I_2 = 1] &= (1-q_{in})^{\frac{n}{k}}(1-q_{out})^{(k-1)\frac{n}{k}}\cdot(1-q_{in})^{\frac{n}{k}-1}(1-q_{out})^{(k-1)\frac{n}{k}}\\ &= (1-q_{in})^{2\frac{n}{k}-1}(1-q_{out})^{2\frac{n}{k}(k-1)}.\end{aligned} \tag{11}$$

Note that we used here first the formula for vertex 1 being isolated, meaning q. Then, we used the fact that if vertex 1 is already isolated, then there will be no edge between vertex 1 and 2 for sure, so we do not need to include this.

Thus, plugging Equation (11) into (10) and using that $\mathbb{E}[I_i] = q$ we get:

$$A^i = \left(\frac{n}{k} - 1\right)\left((1-q_{in})^{2\frac{n}{k}-1}(1-q_{out})^{2\frac{n}{k}(k-1)} - q^2\right).$$

Case 2: Now assume that the two vertices are not in the same community and that the vertices

1 and 2 are not in the same community either. Then, using the possibilities stated before, we obtain

$$B^i = \sum_{\substack{j \neq i \\ X_i \neq X_j}} \mathrm{Cov}(I_i, I_j) = \left(n - \frac{n}{k}\right) \mathrm{Cov}(I_1, I_2) = \left(n - \frac{n}{k}\right) (\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]). \tag{12}$$

Again, let us calculate $\mathbb{E}[I_1 I_2]$ separately

$$\begin{aligned}\mathbb{E}[I_1 I_2] &= \mathbb{P}[I_1 = 1, I_2 = 1] = (1 - q_{in})^{\frac{n}{k}} (1 - q_{out})^{(k-1)\frac{n}{k}} \cdot (1 - q_{in})^{\frac{n}{k}} (1 - q_{out})^{(k-1)\frac{n}{k} - 1} \\ &= (1 - q_{in})^{2\frac{n}{k}} (1 - q_{out})^{2\frac{n}{k}(k-1) - 1}\end{aligned}$$

where we again used the fact that if vertex 1 is already isolated, then there will be no edge between vertex 1 and 2 for sure.

Hence, plugging this into (12) and using again that $\mathbb{E}[I_i] = q$ we get

$$B^i = \sum_{\substack{j \neq i \\ X_i \neq X_j}} \mathrm{Cov}(I_i, I_j) = \left(n - \frac{n}{k}\right) \left((1 - q_{in})^{2\frac{n}{k}} (1 - q_{out})^{2\frac{n}{k}(k-1) - 1} - q^2\right).$$

Now, using the two reformulations for A and B and the fact that the I_i's are Bernoulli distributed, Equation (9) becomes

$$\begin{aligned}\mathrm{Var}(X) &= n\mathrm{Var}(I_1) + \sum_{i=1}^{n} \Bigg[\left(\frac{n}{k} - 1\right) \left((1 - q_{in})^{2\frac{n}{k} - 1} (1 - q_{out})^{2\frac{n}{k}(k-1)} - q^2\right) \\ &\quad + \left(n - \frac{n}{k}\right) \left((1 - q_{in})^{2\frac{n}{k}} (1 - q_{out})^{2\frac{n}{k}(k-1) - 1} - q^2\right) \Bigg] \\ &= nq(1 - q) + n \Bigg[\left(\frac{n}{k} - 1\right) \left(\frac{(1 - q_{in})^{2\frac{n}{k}} (1 - q_{out})^{2\frac{n}{k}(k-1)}}{1 - q_{in}} - q^2\right) \\ &\quad + \left(n - \frac{n}{k}\right) \left(\frac{(1 - q_{in})^{2\frac{n}{k}} (1 - q_{out})^{2\frac{n}{k}(k-1)}}{1 - q_{out}} - q^2\right) \Bigg] \\ &= nq(1 - q) + n \left[\left(\frac{n}{k} - 1\right) \left(\frac{q^2}{1 - q_{in}} - q^2\right) + \left(n - \frac{n}{k}\right) \left(\frac{q^2}{1 - q_{out}} - q^2\right) \right]\end{aligned}$$

$$= nq(1-q) + n\left(\frac{n}{k}-1\right)\left(\frac{q^2 - q^2 + q_{in}\cdot q^2}{1-q_{in}}\right) + n\left(n-\frac{n}{k}\right)\left(\frac{q^2 - q^2 + q_{out}\cdot q^2}{1-q_{out}}\right)$$

$$= nq(1-q) + n\left(\frac{n}{k}-1\right)\frac{q^2\cdot q_{in}}{1-q_{in}} + n\left(n-\frac{n}{k}\right)\frac{q^2\cdot q_{out}}{1-q_{out}}.$$

Now, we want to estimate this. We say $a(n) \sim b(n)$ if $\frac{a(n)}{b(n)} \to 1$ as $n \to \infty$. Since $q_{in} \to 0$ and $q_{out} \to 0$ as $n \to \infty$, it holds $q \to 0$ as $n \to \infty$. Thus,

$$\begin{aligned}
\mathrm{Var}(X) &\sim nq + \frac{n^2}{k}\frac{q^2\cdot q_{in}}{1-q_{in}} + n\left(n-\frac{n}{k}\right)\frac{q^2\cdot q_{out}}{1-q_{out}} = nq + \frac{n^2}{k}\frac{q^2\cdot q_{in}}{1-q_{in}} + n^2\left(1-\frac{1}{k}\right)\frac{q^2\cdot q_{out}}{1-q_{out}} \\
&= nq + \frac{n^2}{k}q^2\frac{q_{in}}{1-q_{in}} + \frac{n^2}{k}q^2(k-1)\frac{q_{out}}{1-q_{out}} = nq + \frac{n^2}{k}q^2\left(\frac{q_{in}}{1-q_{in}} + (k-1)\frac{q_{out}}{1-q_{out}}\right) \\
&= nq + \frac{n^2}{k}q^2\left(\frac{q_{in} - q_{in}q_{out} + (k-1)q_{out} - (k-1)q_{in}q_{out}}{(1-q_{in})(1-q_{out})}\right) \\
&\sim nq + \frac{n^2}{k}q^2(q_{in} + (k-1)q_{out} - kq_{in}q_{out}) \\
&= nq + \frac{1}{k}n^2\cdot n^{-2c}e^{O\left(\frac{\ln^2(n)}{n}\right)}\left(a\frac{\ln(n)}{n} + (k-1)b\frac{\ln(n)}{n} - k\frac{ab\ln(n)^2}{n^2}\right) \\
&= nq + \frac{1}{k}e^{O\left(\frac{\ln^2(n)}{n}\right)}\left(a\cdot n\ln(n)\cdot n^{-2c} + (k-1)b\cdot n\cdot\ln(n)\cdot n^{-2c} - kab\ln(n)^2 n^{-2c}\right) \\
&\sim nq = n\cdot n^{-c}e^{O\left(\frac{\ln^2(n)}{n}\right)} = \mathbb{E}[X]
\end{aligned}$$

where the last approximation comes from the fact that $n\ln(n)n^{-2c}$ and $\ln(n)^2\cdot n^{-2c}$ grow slower than nq as $n \to \infty$ if $c < 1$.

Therefore, we have that

$$\begin{aligned}
\mathbb{E}[X] \sim \mathrm{Var}(X) &= \mathbb{E}[(X-\mathbb{E}[X])^2] = \sum_{x\in\mathbb{N}}(x-\mathbb{E}[X])^2\mathbb{P}[X=x] \\
&= (0-\mathbb{E}[X])^2\mathbb{P}[X=0] + \sum_{x\in\mathbb{N}^*}(x-\mathbb{E}[X])^2\mathbb{P}[X=x] \geq (0-\mathbb{E}[X])^2\mathbb{P}[X=0] \\
&= \mathbb{E}[X]^2\mathbb{P}[X=0].
\end{aligned}$$

Thus, it follows by (8) that

$$\mathbb{P}[X = 0] \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \sim \frac{\mathbb{E}[X]}{\mathbb{E}[X]^2} = \frac{1}{\mathbb{E}[X]} \underset{n\to\infty}{\longrightarrow} 0.$$

We conclude that the probability that no node is isolated goes to 0 as $n \to \infty$ and hence $\mathbb{P}[\text{at least one node is isolated}] \underset{n\to\infty}{\longrightarrow} 1$. Therefore,

$$\mathbb{P}[\mathrm{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n}) \text{ is disconnected}] \underset{n\to\infty}{\longrightarrow} 1.$$

4.2 If $\frac{a+(k-1)b}{k} > 1$, then SSBM is connected

Now we want to prove case *ii*), i.e. if $\frac{a+(k-1)b}{k} > 1$, then

$$\mathbb{P}[\mathrm{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n}) \text{ is connected}] \underset{n\to\infty}{\longrightarrow} 1$$

or equivalently $\mathbb{P}[\mathrm{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n}) \text{ is disconnected}] \underset{n\to\infty}{\longrightarrow} 0$.

From (7) we have $\mathbb{E}[X] = n^{1-\frac{a+(k-1)b}{k}} e^{O\left(\frac{\ln^2(n)}{n}\right)} \underset{n\to\infty}{\longrightarrow} 0$ for $\frac{a+(k-1)b}{k} > 1$. Thus, the probability of having isolated nodes goes to 0. However, we need more to establish connectivity.

The event "graph is disconnected" is equivalent to the existence of j nodes without an edge to the remaining nodes, for some $j \leq \frac{n}{2}$. There are a few different cases, depending on whether the vertices are in the same community or not. Hence, let us pick j vertices that have no edge to the remaining $n - j$ nodes. Let us look at vertex i and say that s of the remaining $n - j$ vertices are in the same community. Note that we need $s \leq \frac{n}{k} - 1$. Thus, $n - j - s$ are in other communities. We have that

$$\mathbb{P}[\text{vertex } i \text{ is not connected to } n - j \text{ vertices} \,|\, s] = (1 - q_{in})^s (1 - q_{out})^{n-j-s}.$$

If we denote by s_i (where $0 \leq s_i \leq \frac{n}{k} - 1$ for all i) the number of vertices among the remaining $n - j$ vertices that are in the same community as vertex i, we get

$$\mathbb{P}[\{1, ..., j\} \text{ not connected to } n - j \text{ vertices} \,|\, s_1, ..., s_j] = \prod_{i=1}^{j} (1 - q_{in})^{s_i} (1 - q_{out})^{n-j-s_i}.$$

This holds as the edges are independently present and thus the probability that a vertex i is not connected to the $n - j$ remaining vertices is independent of the probability that a different vertex l is not connected to the $n - j$ remaining vertices.

With this and using a union bound, it follows,

$$\mathbb{P}[\exists j \text{ nodes not connected to the rest} \,|\, s_1, ..., s_j] \leq \binom{n}{j} \prod_{i=1}^{j} (1 - q_{in})^{s_i} (1 - q_{out})^{n-j-s_i}.$$

Using the union bound, we get,

$$\begin{aligned}
\mathbb{P}[\text{SSBM is disconnected}] &= \mathbb{P}\left[\bigcup_{j=1}^{\lfloor n/2 \rfloor} \{\exists j \text{ nodes not connected to the rest}\} \right] \\
&\leq \sum_{j=1}^{\lfloor n/2 \rfloor} \binom{n}{j} \prod_{i=1}^{j} (1 - q_{in})^{s_i} (1 - q_{out})^{n-j-s_i} \\
&= \sum_{j=1}^{\lfloor n/2 \rfloor} \binom{n}{j} \prod_{i=1}^{j} \left(1 - a\frac{\ln(n)}{n}\right)^{s_i} \left(1 - b\frac{\ln(n)}{n}\right)^{n-j-s_i}. \qquad (13)
\end{aligned}$$

Let us analyse the different summands in detail

$$\prod_{i=1}^{j} \left(1 - a\frac{\ln(n)}{n}\right)^{s_i} \left(1 - b\frac{\ln(n)}{n}\right)^{n-j-s_i} = \exp\left(\ln\left(\prod_{i=1}^{j} \left(1 - a\frac{\ln(n)}{n}\right)^{s_i} \left(1 - b\frac{\ln(n)}{n}\right)^{n-j-s_i} \right) \right)$$

$$= \exp\left(\sum_{i=1}^{j} s_i \ln\left(1 - a\frac{\ln(n)}{n}\right) + (n - j - s_i) \ln\left(1 - b\frac{\ln(n)}{n}\right) \right)$$

$$= \exp\left(n \sum_{i=1}^{j} \frac{s_i}{n} \ln\left(1 - a\frac{\ln(n)}{n}\right) + \frac{n - j - s_i}{n} \ln\left(1 - b\frac{\ln(n)}{n}\right) \right). \qquad (14)$$

Now, we observe that s_i fluctuates around the mean $\frac{n}{k}$, so replacing this we get $\frac{s_i}{n} = \frac{1}{k}$ and $n - j - s_i = n - j - \frac{n}{k} = n\left(1 - \frac{1}{k}\right) - j = n \cdot \frac{k-1}{k} - j$. Hence, Equation (14) becomes

$$\exp\left(n \sum_{i=1}^{j} \frac{1}{k} \ln\left(1 - a\frac{\ln(n)}{n}\right) + \left(\frac{k-1}{k} - \frac{j}{n}\right) \ln\left(1 - b\frac{\ln(n)}{n}\right) \right).$$

Next, we use the known inequality $\ln(1-x) \leq -x$, which holds for all x, to get

$$\exp\left(n\sum_{i=1}^{j}\frac{1}{k}\ln\left(1-a\frac{\ln(n)}{n}\right)+\left(\frac{k-1}{k}-\frac{j}{n}\right)\ln\left(1-b\frac{\ln(n)}{n}\right)\right)$$

$$\leq \exp\left(n\sum_{i=1}^{j}\frac{1}{k}\left(-a\frac{\ln(n)}{n}\right)+\left(\frac{k-1}{k}-\frac{j}{n}\right)\left(-b\frac{\ln(n)}{n}\right)\right)$$

$$= \exp\left(n\sum_{i=1}^{j}-\frac{1}{n}\frac{a\ln(n)+(k-1-\frac{jk}{n})b\ln(n)}{k}\right)$$

$$= \exp\left(n\sum_{i=1}^{j}-\frac{1}{n}\frac{a+(k-1)b}{k}\ln(n)+\frac{j}{n^2}b\ln(n)\right)$$

$$= \exp\left(nj\left(\frac{j}{n^2}b\ln(n)-\frac{a+(k-1)b}{k}\frac{\ln(n)}{n}\right)\right)$$

$$= \exp\left(j^2b\frac{\ln(n)}{n}-j\frac{a+(k-1)b}{k}\ln(n)\right).$$

So, in Equation (13) we get

$$\sum_{j=1}^{\lfloor n/2\rfloor}\binom{n}{j}\prod_{i=1}^{j}\left(1-a\frac{\ln(n)}{n}\right)^{s_i}\left(1-b\frac{\ln(n)}{n}\right)^{n-j-s_i}\leq\sum_{j=1}^{\lfloor n/2\rfloor}\binom{n}{j}\exp\left(j^2b\frac{\ln(n)}{n}-j\frac{a+(k-1)b}{k}\ln(n)\right).$$

Notice that for $j \leq \frac{n}{2}$ the term $j^2b\frac{\ln(n)}{n}$ is relatively small compared to $j\frac{a+(k-1)b}{k}\ln(n)$, so we can approximate

$$\sum_{j=1}^{\lfloor n/2\rfloor}\binom{n}{j}\exp\left(j^2b\frac{\ln(n)}{n}-j\frac{a+(k-1)b}{k}\ln(n)\right)\approx\sum_{j=1}^{\lfloor n/2\rfloor}\binom{n}{j}\exp\left(-j\frac{a+(k-1)b}{k}\ln(n)\right)$$

$$=\sum_{j=1}^{\lfloor n/2\rfloor}\binom{n}{j}n^{-\frac{a+(k-1)b}{k}j}.$$

Using Stirling's formula $j! \sim \left(\frac{j}{e}\right)^j$, which implies $\binom{n}{j} \leq \frac{n^j}{\left(\frac{j}{e}\right)^j}$ (see Corollary 7.3 in the appendix), we get

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \binom{n}{j} n^{-\frac{a+(k-1)b}{k} j} \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{n^j}{\left(\frac{j}{e}\right)^j} \cdot n^{-\frac{a+(k-1)b}{k} j} = \sum_{j=1}^{\lfloor n/2 \rfloor} n^{j(1-\frac{a+(k-1)b}{k})} e^j j^{-j}$$

$$= \sum_{j=1}^{\lfloor n/2 \rfloor} \left(\frac{e}{j}\right)^j \frac{1}{n^{j(\frac{a+(k-1)b}{k}-1)}} \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{e^2}{j} \frac{1}{n^{\frac{a+(k-1)b}{k}-1}} = \frac{e^2}{n^{\frac{a+(k-1)b}{k}-1}} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j}$$

where in the last inequality we used the fact that $\left(\frac{e}{j}\right)^j \leq \frac{e^2}{j}$ for all $j \geq 1$ (see Corollary 7.4 in the appendix) and $\frac{1}{n^{j(\frac{a+(k-1)b}{k}-1)}}$ is maximal, if j is equal to the smallest possible value (as $\frac{a+(k-1)b}{k} > 1$).

Now, we see that we have a partial sum of the harmonic series and $\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j}$ can be approximated by $\ln(\lfloor \frac{n}{2} \rfloor)$. Hence,

$$\mathbb{P}[\text{SSBM is disconnected}] \approx \frac{e^2}{n^{\frac{a+(k-1)b}{k}-1}} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j} \approx \frac{e^2 \ln(\lfloor \frac{n}{2} \rfloor)}{n^{\frac{a+(k-1)b}{k}-1}} \underset{n\to\infty}{\longrightarrow} 0$$

assuming that $\frac{a+(k-1)b}{k} > 1$ and using l'Hôpital's rule.

Thus, $\mathbb{P}[\text{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ is disconnected$] \underset{n\to\infty}{\longrightarrow} 0$, which implies that

$$\mathbb{P}[\text{SSBM}(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n}) \text{ is connected}] \underset{n\to\infty}{\longrightarrow} 1,$$

and this concludes the proof.

□

Remark 4.2. Note that in Section 3.4 we saw that the expected degree in the SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ model is $\frac{a+(k-1)b}{k} \ln(n)$. Furthermore, we saw in Theorem 3.17 that if $c < 1$, then an Erdős-Rényi graph is disconnected with high probability and if $c > 1$, then an Erdős-Rényi graph is connected with high probability where the expected degree satisfies $\mathbb{E}[d_i] = c \ln(n)$. Hence, we see that by replacing c with $\frac{a+(k-1)b}{k}$ in the SSBM, we get exactly the same statement.

5 Giant Component Poissonized

Let us now continue with the main part of this thesis. Namely, to introduce a new proof technique for the giant component in the Symmetric Stochastic Block Model. The idea for this chapter is from (Curien, 2024b). The basic idea is that we do not assume the total number of vertices per community to be fixed, but to be random according to a Poisson random variable with parameter α and then prove a theorem for the giant component in this adapted model. The last step is to prove that we can sandwich the SSBM with fixed number of vertices between two poissonized SSBMs.

5.1 Defining the Model

Let $\alpha_1 > 0, \alpha_2 > 0, ..., \alpha_k > 0$ be fixed and let $N_i \sim \text{Poi}(\alpha_i)$ be independent random variables, denoting the respective community sizes. We denote by $\underline{\alpha} := (\alpha_1, \alpha_2, ..., \alpha_k)^\top$ and by $\eta = \sum_{i=1}^k \alpha_i$ the average (or expected) number of vertices in the graph. This holds since the number of vertices in the communities are Poisson distributed and hence the parameter, α_i, is the expected number of vertices in community i. Consider $\text{SSBM}_{\text{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$, which conditionally on $N_1, N_2, ..., N_k$ is made of a classical $\text{SSBM}(\sum_{i=1}^k N_i, k, q_{in}, q_{out})$ random graph that we call the *core*.

Remark 5.1. Later, we will have that $q_{in} = a/\eta$ and $q_{out} = b/\eta$ for some $a, b > 0$.

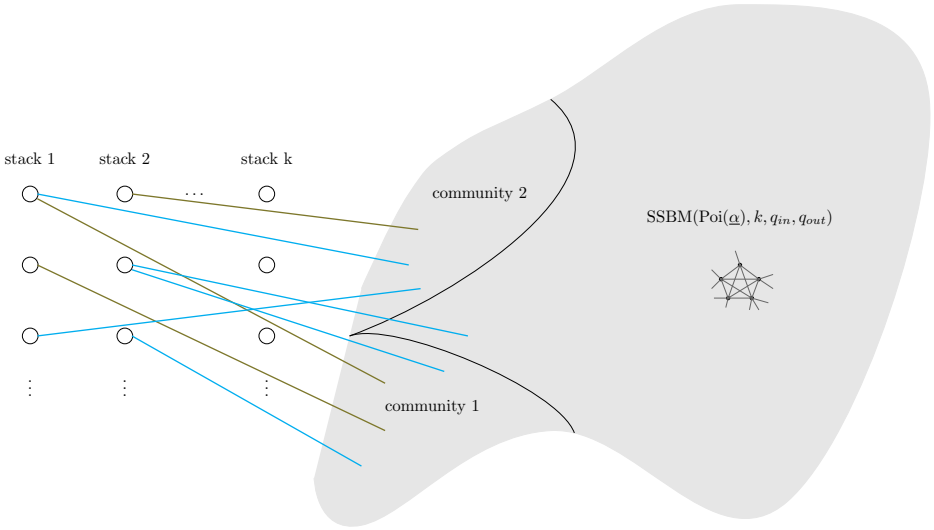


Figure 6: The stacks are the vertices on the left and the grey part represents the core. Vertices in the stacks are connected with probability q_{in} (olive line) to vertices in the same community and with probability q_{out} (blue line) to vertices of different communities. Notice that for better readability, the connections of stack k to the core are omitted.

Next, we want to define the *stacks*. The idea is that a stack is an infinitely long sequence of vertices, such that every vertex in the stack is linked to every vertex in the core ("the normal SSBM graph") independently. As we are working with k different communities, we also need k different stacks, meaning for every community we have one stack. Therefore, we have stacks $1, 2, ..., k$ of vertices which are linked to every vertex of SSBM$(\underline{\alpha}, k, q_{in}, q_{out})$ with the respective probabilities independently, i.e. if we are in stack i, then the vertices of this stack are connected with probability q_{in} to the vertices of community i in SSBM and with probability q_{out} to the vertices of SSBM in the other communities. See Figure 6 for an illustration of this.

Remark 5.2. Note that there are no edges between vertices of the stacks. There are only edges between vertices of the SSBM and the stack.

Definition 5.3. A *step of exploration* in SSBM$_{\text{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ is the following: Choose a stack i which is in community C_i, then take the vertex ρ on the top of this stack and reveal its neighbours $y_1, y_2, ..., y_L$ with $L \geq 0$, inside the core and in the respective communities. Note that we thus analyse the neighbours in each community separately. Next, we see those vertices $y_1, ..., y_L$ as new vertices of the stacks, i.e. we put the vertices of community j into stack j and erase all possible connections between $y_1, ..., y_L$ and between $y_1, ..., y_L$ and other vertices of the stack. See Figures 7 and 8 for an illustration of this.

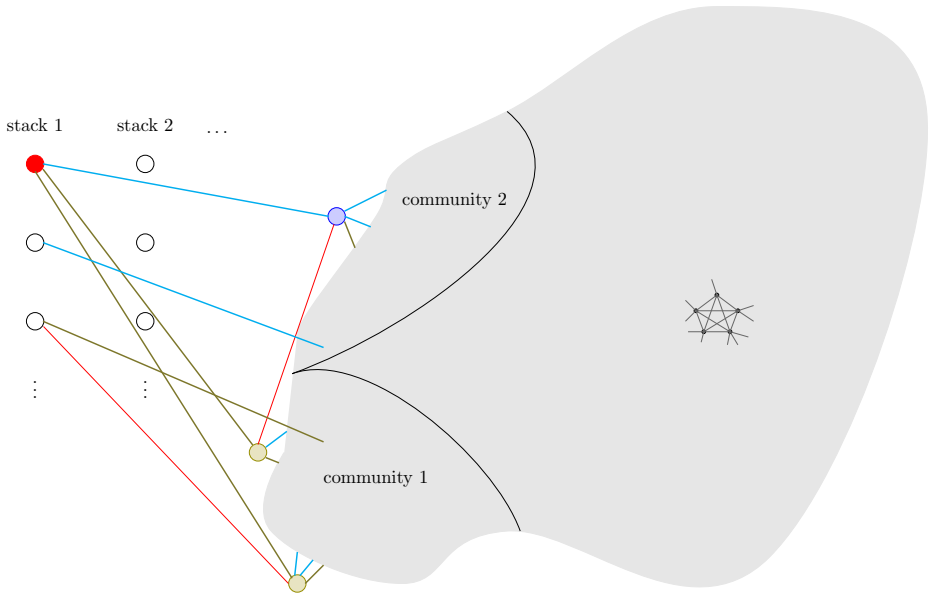


Figure 7: In this figure, we see an exploration starting with the first vertex in stack 1 (red) and revealing its neighbours in the communities. It has one neighbour in community 2 (blue) and two neighbours in community 1 (olive). As a next step, the explored vertex (red) is deleted from stack 1 and the red edges are also deleted. The two olive vertices are then put on top of the other vertices in stack 1 and the blue one is put on top of stack 2. The resulting graph is presented in Figure 8.

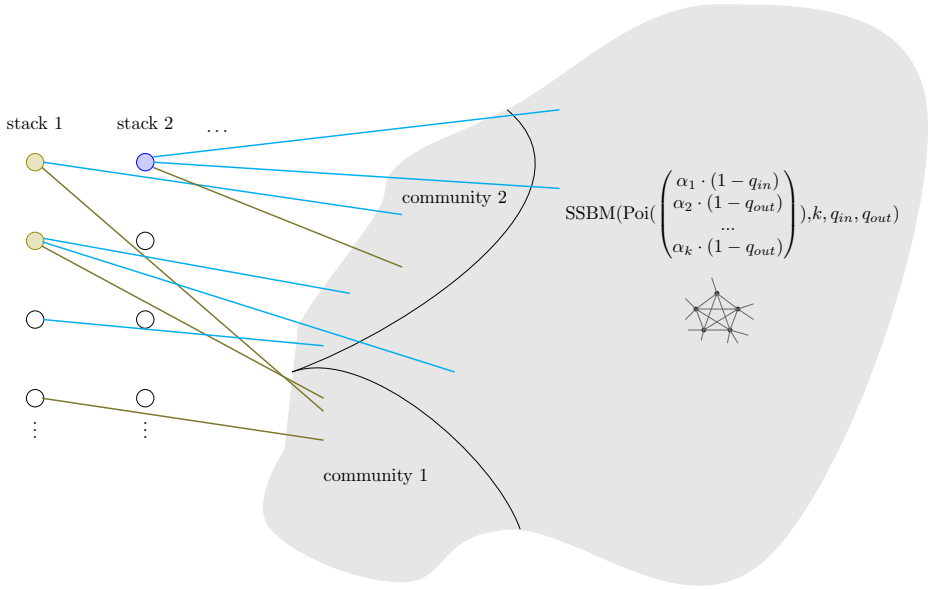


Figure 8: This figure depicts the graph after the removal of the explored vertex and with the found neighbours in the stacks. As a next exploration, we choose the blue vertex and continue in the same way as before. Notice that the resulting graph has a new parameter for the Poisson process, proven in Lemma 5.4.

Then, we get the Lemma 5.4.

Lemma 5.4 (Markov Property for $\mathrm{SSBM}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$)**.** *Let us choose a vertex ρ in stack j (which represents community j) and let $L_i \geq 0$ be the number of neighbours of ρ in the respective community i of the core, with $1 \leq i \leq k$. Then, remembering the notations for a random vector with Poisson distribution, we have that*

$$\underline{L} = \begin{pmatrix} L_1 \\ \dots \\ L_{j-1} \\ L_j \\ L_{j+1} \\ \dots \\ L_k \end{pmatrix} \sim \mathrm{Poi} \begin{pmatrix} \alpha_1 \cdot q_{out} \\ \dots \\ \alpha_{j-1} \cdot q_{out} \\ \alpha_j \cdot q_{in} \\ \alpha_{j+1} \cdot q_{out} \\ \dots \\ \alpha_k \cdot q_{out} \end{pmatrix}.$$

In particular, the coordinates are independent. Furthermore, we have that conditionally on $\underline{L}$ the community sizes in the core, made after removing ρ and placing the $\underline{L}$ neighbours in the

stack, have law

$$\text{Poi}\begin{pmatrix} \alpha_1 \cdot (1-q_{out}) \\ ... \\ \alpha_{j-1} \cdot (1-q_{out}) \\ \alpha_j \cdot (1-q_{in}) \\ \alpha_{j+1} \cdot (1-q_{out}) \\ ... \\ \alpha_k \cdot (1-q_{out}) \end{pmatrix}.$$

Before continuing to the proof of Lemma 5.4, we want to pay attention to the fact that the value q_{in} is appearing exactly in row j. In the other rows, we only have q_{out}. Therefore, q_{in} appears in the row which corresponds to the stack that we choose the starting vector in. Intuitively, $\underline{L}$ is the distribution of the adjacent vertices of ρ in the communities, so L_1 is the distribution of the neighbours of ρ in community 1, L_2 is the distribution of the neighbours in community 2 and so on.

Proof of Lemma 5.4. Let us define the vector $\underline{N'} = \underline{N} - \underline{L}$, which is the number of remaining vertices in the communities after the revelation of the $\underline{L} = \begin{pmatrix} L_1 \\ ... \\ L_k \end{pmatrix}$ neighbours of ρ and where $\underline{N} = \begin{pmatrix} N_1 \\ ... \\ N_k \end{pmatrix}$. Then, remark the following factorization:

$$\mathbb{P}[\underline{L} = \underline{l}, \underline{N'} = \underline{n}] = \mathbb{P}[\underline{L} = \underline{l}, \underline{N} - \underline{L} = \underline{n}] = \mathbb{P}[\underline{L} = \underline{l}, \underline{N} = \underline{L} + \underline{n}] = \mathbb{P}[\underline{L} = \underline{l}, \underline{N} = \underline{l} + \underline{n}]$$

$$= \mathbb{P}[\underline{L} = \underline{l} | \underline{N} = \underline{n} + \underline{l}] \cdot \mathbb{P}[\underline{N} = \underline{n} + \underline{l}]$$

where $\underline{l} = \begin{pmatrix} l_1 \\ ... \\ l_k \end{pmatrix}$ and $\underline{n} = \begin{pmatrix} n_1 \\ ... \\ n_k \end{pmatrix}$.

Now, we observe that conditionally on the number of vertices in the communities, $\underline{N}$, the number of neighbours, $\underline{L}$, is a Binomial random vector and that $\underline{N}$ is a Poisson random vector. Thus, it

follows

$$\mathbb{P}[\underline{L}=\underline{l}|\underline{N}=\underline{n}+\underline{l}]\cdot\mathbb{P}[\underline{N}=\underline{n}+\underline{l}]=\mathbb{P}\left[\begin{pmatrix}L_1\\...\\L_k\end{pmatrix}=\begin{pmatrix}l_1\\...\\l_k\end{pmatrix}\middle|\begin{pmatrix}N_1=n_1+l_1\\...\\N_k=n_k+l_k\end{pmatrix}\right]\cdot\mathbb{P}\left[\begin{pmatrix}N_1=n_1+l_1\\...\\N_k=n_k+l_k\end{pmatrix}\right]$$

$$\overset{(*)}{=}\prod_{i=1}^{k}\mathbb{P}[L_i=l_i|N_i=n_i+l_i]\cdot\mathbb{P}[N_i=n_i+l_i]$$

$$\overset{(**)}{=}\prod_{\substack{i=1\\i\neq j}}^{k}\binom{n_i+l_i}{l_i}q_{out}^{l_i}(1-q_{out})^{n_i+l_i-l_i}\cdot e^{-\alpha_i}\frac{\alpha_i^{n_i+l_i}}{(n_i+l_i)!}\cdot\binom{n_j+l_j}{l_j}q_{in}^{l_j}(1-q_{in})^{n_j+l_j-l_j}\cdot e^{-\alpha_j}\frac{a_j^{n_j+l_j}}{(n_j+l_j)!}$$

$$=\prod_{\substack{i=1\\i\neq j}}^{k}\frac{(n_i+l_i)!}{(n_i+l_i-l_i)!l_i!}q_{out}^{l_i}(1-q_{out})^{n_i+l_i-l_i}e^{-\alpha_i}\frac{\alpha_i^{n_i+l_i}}{(n_i+l_i)!}$$

$$\cdot\frac{(n_j+l_j)!}{(n_j+l_j-l_j)!l_j!}q_{in}^{l_j}(1-q_{in})^{n_j+l_j-l_j}e^{-\alpha_j}\frac{a_j^{n_j+l_j}}{(n_j+l_j)!}$$

$$=\prod_{\substack{i=1\\i\neq j}}^{k}\frac{1}{n_i!}\frac{1}{l_i!}q_{out}^{l_i}(1-q_{out})^{n_i}e^{-\alpha_i}e^{\alpha_i q_{out}}e^{-\alpha_i q_{out}}\alpha_i^{n_i}\alpha_i^{l_i}\cdot\frac{1}{n_j!}\frac{1}{l_j!}q_{in}^{l_j}(1-q_{in})^{n_j}e^{-\alpha_j}e^{\alpha_j q_{in}}e^{-\alpha_j q_{in}}\alpha_j^{n_j}\alpha_j^{l_j}$$

$$=\prod_{\substack{i=1\\i\neq j}}^{k}\left(e^{-\alpha_i(1-q_{out})}\frac{(\alpha_i(1-q_{out}))^{n_i}}{n_i!}\right)\left(e^{-\alpha_i q_{out}}\frac{(\alpha_i q_{out})^{l_i}}{l_i!}\right)$$

$$\cdot\left(e^{-\alpha_j(1-q_{in})}\frac{(\alpha_j(1-q_{in}))^{n_j}}{n_j!}\right)\left(e^{-\alpha_j q_{in}}\frac{(\alpha_j q_{in})^{l_j}}{l_j!}\right)$$

$$=\left[\prod_{\substack{i=1\\i\neq j}}^{k}\left(e^{-\alpha_i(1-q_{out})}\frac{(\alpha_i(1-q_{out}))^{n_i}}{n_i!}\right)\left(e^{-\alpha_j(1-q_{in})}\frac{(\alpha_j(1-q_{in}))^{n_j}}{n_j!}\right)\right]$$

$$\cdot\underbrace{\left[\prod_{\substack{i=1\\i\neq j}}^{k}\left(e^{-\alpha_i q_{out}}\frac{(\alpha_i q_{out})^{l_i}}{l_i!}\right)\left(e^{-\alpha_j q_{in}}\frac{(\alpha_j q_{in})^{l_j}}{l_j!}\right)\right]}_{=:A(\underline{l})}$$

where we used in $(*)$ that the components are independent. Let us give further explanations,

why the equality $(**)$ holds. Note that L_i denotes the number of neighbours of ρ in community i, which is a different community than the one ρ belongs to. Knowing that the total number of vertices in the community is $N_i = n_i + l_i$, it holds that the number of neighbours of ρ, L_i, is a Binomial random variable with $n_i + l_i$ trials and success probability q_{out} (we take l_i out of the $n_i + l_i$ total vertices that are neighbours to ρ and the remaining $n_i + l_i - l_i$ vertices should not be connected to ρ). Similarly, the case where we look for the number of neighbours inside the community of ρ follows.

Now, let us sum over all possible values to get

$$\mathbb{P}[\underline{L} = \underline{l}] = \sum_{\underline{n} \in \mathbb{N}^k} \mathbb{P}[\underline{L} = \underline{l}, \underline{N}' = \underline{n}]$$

$$= \sum_{\underline{n} \in \mathbb{N}^k} \left[\prod_{\substack{i=1 \\ i \neq j}}^{k} \left(e^{-\alpha_i(1-q_{out})} \frac{(\alpha_i(1-q_{out}))^{n_i}}{n_i!} \right) \left(e^{-\alpha_j(1-q_{in})} \frac{(\alpha_j(1-q_{in}))^{n_j}}{n_j!} \right) \right] \cdot A(\underline{l})$$

$$= A(\underline{l}) \cdot \sum_{(n_1, \ldots n_{j-1}, n_{j+1}, \ldots, n_k) \in \mathbb{N}^{k-1}} \prod_{\substack{i=1 \\ i \neq j}}^{k} \left(e^{-\alpha_i(1-q_{out})} \frac{(\alpha_i(1-q_{out}))^{n_i}}{n_i!} \right)$$

$$\cdot \sum_{n_j \in \mathbb{N}} \left(e^{-\alpha_j(1-q_{in})} \frac{(\alpha_j(1-q_{in}))^{n_j}}{n_j!} \right)$$

$$\overset{(***)}{=} A(\underline{l}) \cdot \sum_{n_1 \in \mathbb{N}} \left(e^{-\alpha_1(1-q_{out})} \frac{(\alpha_1(1-q_{out}))^{n_1}}{n_1!} \right) \sum_{n_2 \in \mathbb{N}} \ldots \sum_{n_k \in \mathbb{N}} \left(e^{-\alpha_k(1-q_{out})} \frac{(\alpha_k(1-q_{out}))^{n_k}}{n_k!} \right)$$

$$\overset{(***)}{=} A(\underline{l}) = \prod_{\substack{i=1 \\ i \neq j}}^{k} \left(e^{-\alpha_i q_{out}} \frac{(\alpha_i q_{out}))^{l_i}}{l_i!} \right) \left(e^{-\alpha_j q_{in}} \frac{(\alpha_j q_{in})^{l_j}}{l_j!} \right)$$

where in $(***)$ we used the fact that we are working with Poisson distributions.

Now, arguing again by the independence of the components, we get that

$$\mathbb{P}[L_i = l_i] = e^{-\alpha_i q_{out}} \frac{(\alpha_i q_{out}))^{l_i}}{l_i!} \sim \text{Poi}(a_i \cdot q_{out})$$

as well as

$$\mathbb{P}[L_j = l_j] = e^{-\alpha_j q_{in}} \frac{(\alpha_j q_{in})^{l_j}}{l_j!} \sim \text{Poi}(a_j \cdot q_{in}).$$

Hence,

$$\underline{L} \sim \text{Poi}\begin{pmatrix} \alpha_1 \cdot q_{out} \\ ... \\ \alpha_{j-1} \cdot q_{out} \\ \alpha_j \cdot q_{in} \\ \alpha_{j+1} \cdot q_{out} \\ ... \\ \alpha_k \cdot q_{out} \end{pmatrix}.$$

By the same reasoning as above, summing over all possible values for l_i, we also directly get that

$$\underline{N'} \sim \text{Poi}\begin{pmatrix} \alpha_1 \cdot (1 - q_{out}) \\ ... \\ \alpha_{j-1} \cdot (1 - q_{out}) \\ \alpha_j \cdot (1 - q_{in}) \\ \alpha_{j+1} \cdot (1 - q_{out}) \\ ... \\ \alpha_k \cdot (1 - q_{out}) \end{pmatrix}.$$

□

Remark 5.5. Note that the proof gives us the independence between the explorations, i.e. the number of discovered vertices in the communities after two explorations is independent of the number after one exploration.

5.2 Lukasiewicz Walk

Now, we want to analyse successive explorations of the type as described in Definition 5.3. For this, we want to fix a deterministic sequence of numbers, which indicates in which stack we pick the next vertex to explore. Let us use the sequence $(1, 2, 3, ..., k, 1, 2, 3, ..., k, 1, 2, 3...)$, meaning we choose the first vertex in stack 1, reveal its neighbours, remove the chosen vertex from the stack, then we place the found neighbours at the top of the according stacks and remove the edges between them. The next vertex we choose to explore should be the top one in stack 2. We again reveal its neighbours, remove the vertex from the stack and place the neighbours in the stacks, then choose the top one from stack 3 and so on. This procedure gives the so-called *Lukasiewicz exploration* of the graph $\text{SSBM}_{\text{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$. By Lemma 5.4 we hence get a sequence of independent Poisson random variables indicating the number of revealed neighbours

with respective distribution:

$$\begin{pmatrix} (\alpha_1 \cdot q_{in}) \\ \text{Poi}(\alpha_2 \cdot q_{out}) \\ \text{Poi}(\alpha_3 \cdot q_{out}) \\ \dots \\ \text{Poi}(\alpha_k \cdot q_{out}) \end{pmatrix}, \begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{out}(1-q_{in})) \\ \text{Poi}(\alpha_2 \cdot q_{in}(1-q_{out})) \\ \text{Poi}(\alpha_3 \cdot q_{out}(1-q_{out})) \\ \dots \\ \text{Poi}(\alpha_k \cdot q_{out}(1-q_{out})) \end{pmatrix}, \begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{out}(1-q_{in})(1-q_{out})) \\ \text{Poi}(\alpha_2 \cdot q_{out}(1-q_{in})(1-q_{out})) \\ \text{Poi}(\alpha_3 \cdot q_{in}(1-q_{out})^2) \\ \dots \\ \text{Poi}(\alpha_k \cdot q_{out}(1-q_{out})^2) \end{pmatrix}, \dots,$$

$$\begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{out}(1-q_{in})(1-q_{out})^{k-2}) \\ \text{Poi}(\alpha_2 \cdot q_{out}(1-q_{in})(1-q_{out})^{k-2}) \\ \dots \\ \text{Poi}(\alpha_{k-1} \cdot q_{out}(1-q_{in})(1-q_{out})^{k-2}) \\ \text{Poi}(\alpha_k \cdot q_{in}(1-q_{out})^{k-1}) \end{pmatrix}, \begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{in}(1-q_{in})(1-q_{out})^{k-1}) \\ \text{Poi}(\alpha_2 \cdot q_{out}(1-q_{in})(1-q_{out})^{k-1}) \\ \dots \\ \text{Poi}(\alpha_{k-1} \cdot q_{out}(1-q_{in})(1-q_{out})^{k-1}) \\ \text{Poi}(\alpha_k \cdot q_{out}(1-q_{in})(1-q_{out})^{k-1}) \end{pmatrix}, \dots$$

Remark 5.6. This sequence indicates the distribution of the found adjacent vertices. Consequently, after one step there were

$$\begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{in}) \\ \text{Poi}(\alpha_2 \cdot q_{out}) \\ \text{Poi}(\alpha_3 \cdot q_{out}) \\ \dots \\ \text{Poi}(\alpha_k \cdot q_{out}) \end{pmatrix}$$

many neighbouring vertices in the communities found. After two steps, we have found

$$\begin{pmatrix} \text{Poi}(\alpha_1 \cdot q_{out}(1-q_{in})) \\ \text{Poi}(\alpha_2 \cdot q_{in}(1-q_{out})) \\ \text{Poi}(\alpha_3 \cdot q_{out}(1-q_{out})) \\ \dots \\ \text{Poi}(\alpha_k \cdot q_{out}(1-q_{out})) \end{pmatrix}$$

many adjacent vertices in the communities (we just use again Lemma 5.4 with the new total number of vertices, which is $\alpha_i(1-q_{out})$ for $i \neq 1$ and $\alpha_1(1-q_{in})$). Note, that the system always stays the same. The rule is, we always have q_{in} in the corresponding community in which the explored vertex lies in and then afterwards we add $(1-q)$ according to the previous distribution, where q is either q_{in} or q_{out}.

Claim 5.7. *This technique gives, for a general step $mk+u$ with $1 \leq u \leq k$ and $m \geq 0$, a Poisson distribution for the adjacent vertices with parameter*

$$\underline{\alpha}^{(mk+u)} := (1-q_{in})^m (1-q_{out})^{m(k-1)+u-2} \begin{pmatrix} \alpha_1 q_{out}(1-q_{in}) \\ ... \\ \alpha_{u-1} q_{out}(1-q_{in}) \\ \alpha_u q_{in}(1-q_{out}) \\ \alpha_{u+1} q_{out}(1-q_{out}) \\ ... \\ \alpha_k q_{out}(1-q_{out}) \end{pmatrix}.$$

Proof. Let us first look at the step $mk+1$. The parameter of the Poisson distribution is then equal to

$$(1-q_{in})^m (1-q_{out})^{m(k-1)} \begin{pmatrix} \alpha_1 q_{in} \\ \alpha_2 q_{out} \\ ... \\ \alpha_k q_{out} \end{pmatrix}.$$

This follows since we visited every community exactly once per cycle and we did m cycles. Thus, we have visited every community m times (which explains the factor $(1-q_{in})^m$) and we have visited the other communities $(k-1)m$ times (which leads to the factor $(1-q_{out})^{(k-1)m}$). Now we are at the first step of the next cycle, hence we are in community 1. We can also write this as

$$(1-q_{in})^m (1-q_{out})^{m(k-1)+1-2} \begin{pmatrix} \alpha_1 q_{in}(1-q_{out}) \\ \alpha_2 q_{out}(1-q_{out}) \\ ... \\ \alpha_k q_{out}(1-q_{out}) \end{pmatrix}.$$

If we now continue to the step $mk+u$, we note that we are in community u and thus we have already visited the communities $2, ..., u-1$, so there will be an extra factor $(1-q_{in})$ for the communities $1, ..., u-1$ and $(1-q_{out})$ for the communities $u, ..., k$. From step $mk+1$ to step $mk+u$, we have made $u-1$ steps, thus we also need to add $(1-q_{out})^{u-1}$ as a factor for all the

times that we picked a vertex from another community. Collecting all this together gives us,

$$(1-q_{in})^m(1-q_{out})^{m(k-1)+1-2}(1-q_{out})^{u-1}\begin{pmatrix}\alpha_1 q_{out}(1-q_{in})\\ \dots\\ \alpha_{u-1}q_{out}(1-q_{in})\\ \alpha_u q_{in}(1-q_{out})\\ \alpha_{u+1}q_{out}(1-q_{out})\\ \dots\\ \alpha_k q_{out}(1-q_{out})\end{pmatrix}$$

$$=(1-q_{in})^m(1-q_{out})^{m(k-1)+u-2}\begin{pmatrix}\alpha_1 q_{out}(1-q_{in})\\ \dots\\ \alpha_{u-1}q_{out}(1-q_{in})\\ \alpha_u q_{in}(1-q_{out})\\ \alpha_{u+1}q_{out}(1-q_{out})\\ \dots\\ \alpha_k q_{out}(1-q_{out})\end{pmatrix}$$

which is what we wanted to show. □

To make the computations a bit easier let us define

$$A := (1-q_{in})(1-q_{out})^{k-1}. \tag{15}$$

Then, we get that

$$\underline{\alpha}^{(mk+u)} = A^m(1-q_{out})^{u-2}\begin{pmatrix}\alpha_1 q_{out}(1-q_{in})\\ \dots\\ \alpha_{u-1}q_{out}(1-q_{in})\\ \alpha_u q_{in}(1-q_{out})\\ \alpha_{u+1}q_{out}(1-q_{out})\\ \dots\\ \alpha_k q_{out}(1-q_{out})\end{pmatrix} = A^m\begin{pmatrix}\alpha_1 q_{out}(1-q_{in})(1-q_{out})^{u-2}\\ \dots\\ \alpha_{u-1}q_{out}(1-q_{in})(1-q_{out})^{u-2}\\ \alpha_u q_{in}(1-q_{out})^{u-1}\\ \alpha_{u+1}q_{out}(1-q_{out})^{u-1}\\ \dots\\ \alpha_k q_{out}(1-q_{out})^{u-1}\end{pmatrix}$$

By using Claim 5.7, we get

$$\underline{\alpha}^{(mk+u)} = A^m\underline{\alpha}^{(u)}. \tag{16}$$

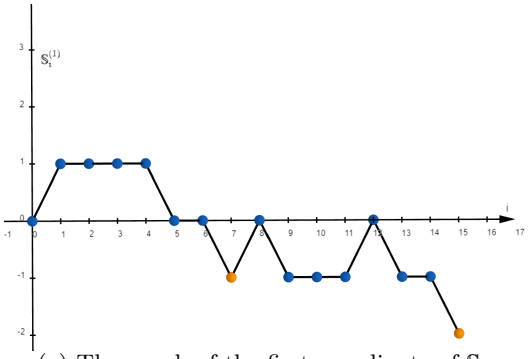
See that

$$\underline{\alpha}^{(1)} = \begin{pmatrix} \alpha_1 q_{in} \\ \alpha_2 q_{out} \\ ... \\ \alpha_k q_{out} \end{pmatrix}.$$

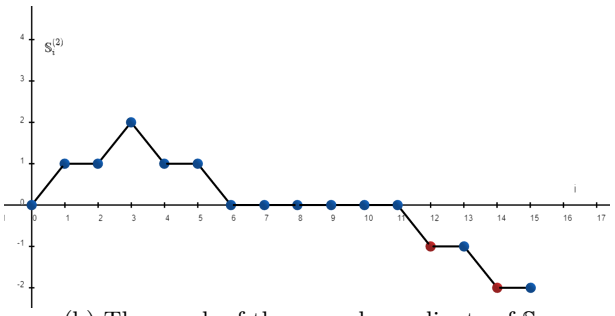
Now that we have the distribution of the neighbours at an arbitrary step, let us get back to the Lukasiewicz exploration of the graph $\mathrm{SSBM}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$. We encode this exploration in a process $(\underline{\mathbb{S}}_i : i \geq 0)$, where

$$\underline{\mathbb{S}}_i = \begin{pmatrix} \mathbb{S}_i^{(1)} \\ ... \\ \mathbb{S}_i^{(k)} \end{pmatrix}.$$

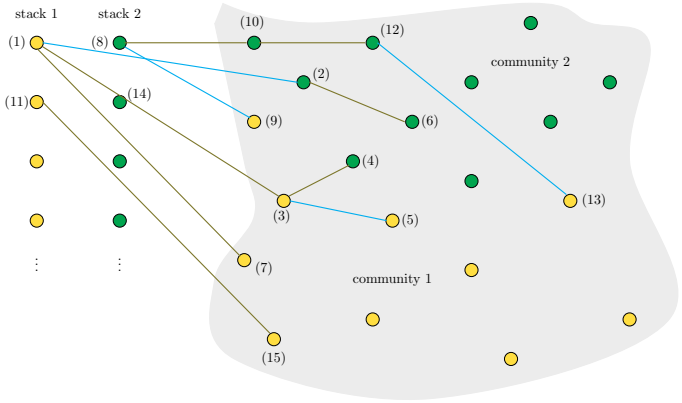
We call this the *Lukasiewicz walk*, defined by $\underline{\mathbb{S}}_0 = \mathbf{0}$ and where $\Delta\underline{\mathbb{S}}_i = \underline{\mathbb{S}}_i - \underline{\mathbb{S}}_{i-1}$ is the number of discovered neighbours in each community at step i minus the vector $\underline{e}_j = (0, 0, ..., 0, 1, 0, ..., 0) \in \mathbb{Z}^k$, which removes 1 from row j if the chosen vertex was in stack j. This is necessary as we remove each time the explored vertex from the stack. Note that every new minimal record in one of the components of $\underline{\mathbb{S}}$ corresponds to the exploration of a new connected component of a new vertex of the initial stacks. We can think of this as the vertices of the stacks being labelled $-1, -2, -3, ...$ because the Lukasiewicz walk can make arbitrary big steps in positive direction (by the number of neighbours) but only -1 in negative direction by removing the vertex from the stack. Therefore, every time the Lukasiewicz walk hits a new infimum number (meaning a new negative number) in one of the coordinates, we have exhausted all the vertices put on top of this stack before. If explored vertices from other communities do not have any neighbours in this community, then the next vertex we pick in this community will be a new one from the infinite stack, and we will therefore build a new connected component in this coordinate. We observe that this definition could lead to the interruption of the connected components, as there might be no new neighbours in one of the communities, and thus we start a new component, but there might still be more neighbours in another community and the previous component might not be exhausted yet. This means that the used method here is a simplification. See also Figure 9 for an illustration of the Lukasiewicz walk with two communities.



(a) The graph of the first coordinate of $\underline{\mathbb{S}}_i$.



(b) The graph of the second coordinate of $\underline{\mathbb{S}}_i$.



(c) Illustration of an exploration in the Symmetric Stochastic Block Model with two communities. The grey part is the random graph, while the arrays on the left are the corresponding stacks.

Figure 9: These figures depict an example of a Lukasiewicz walk in two communities. The idea is that we always look at how many neighbours a vertex in community j has in each community and then add this number minus the vector e_j to $\underline{\mathbb{S}}_i$. The numbers in brackets indicate the order in which we explore the vertices. Notice that the first orange minimal record in $\mathbb{S}_i^{(1)}$ indicates that the next time, when we have to pick a vertex from stack 1, it will be a new one from the infinite array (vertex (11)). The same holds for the red minimal record at 12 in Figure (b), the next time we have to pick a vertex in stack 2 it will be a new one (vertex (14)).

Before we can calculate $\underline{\mathbb{S}}_i$ explicitly, we need to introduce the Poisson counting process.

Definition 5.8 (Coculescu, 2023)**.** A *Poisson counting process*, also called *Poisson point process*, with rate λ is a stochastic process $(\mathcal{P}(t))_{t\geq 0}$ with càdlàg sample paths such that

i) $\mathcal{P}(0) = 0$,

ii) for some natural number n and $0 = t_0 < ... < t_n$, the increments $(\mathcal{P}(t_i) - \mathcal{P}(t_{i-1}))_{1\leq i\leq n}$ are independent,

iii) $\mathcal{P}(t) - \mathcal{P}(s) \sim \text{Poi}(\lambda(t-s))$ for all $0 \leq s \leq t$.

Remark 5.9. Note that by definition of a Poisson counting process with rate 1, we have that at each step t, $\mathcal{P}(t)$ is a Poisson random variable with parameter t.

Corollary 5.10. *A Poisson counting process $(\mathcal{P}(t))_{t\geq 0}$ is a.s. non-decreasing.*

Proof. Let us fix $0 \leq s \leq t$. Since $(\mathcal{P}(t))_{t\geq 0}$ has stationary increments and each $\mathcal{P}(t) - \mathcal{P}(s)$ is Poisson distributed, it follows by the property that a Poisson random variable is non-negative that

$$\mathbb{P}[\mathcal{P}(s) > \mathcal{P}(t)] = \mathbb{P}[\mathcal{P}(t) - \mathcal{P}(s) < 0] = 0.$$

As $\mathbb{Q}_+$ is countable, it holds

$$\mathbb{P}\left[\bigcup_{\substack{u,v\in\mathbb{Q}_+\\ u\leq v}} \{\mathcal{P}(u) > \mathcal{P}(v)\}\right] \leq \sum_{\substack{u,v\in\mathbb{Q}_+\\ u\leq v}} \mathbb{P}[\mathcal{P}(u) > \mathcal{P}(v)] = 0.$$

Therefore,

$$\mathbb{P}\left[\bigcup_{\substack{u,v\in\mathbb{Q}_+\\ u\leq v}} \mathcal{P}(u) > \mathcal{P}(v)\right] = 0.$$

Thus, it follows

$$\mathbb{P}[\forall\, u \leq v, u, v \in \mathbb{Q}_+ : \mathcal{P}(u) \leq \mathcal{P}(v)] = \mathbb{P}\left[\bigcap_{\substack{u,v\in\mathbb{Q}_+\\ u\leq v}} \mathcal{P}(u) \leq \mathcal{P}(v)\right]$$

$$= 1 - \mathbb{P}\left[\bigcup_{\substack{u,v\in\mathbb{Q}_+\\ u\leq v}} \mathcal{P}(u) > \mathcal{P}(v)\right] = 1 - 0 = 1.$$

Using that $(\mathcal{P}(t))_{t\geq 0}$ has càdlàg sample paths, we can conclude that

$$\mathbb{P}[\forall\, s \leq t : \mathcal{P}(s) \leq \mathcal{P}(t)] = 1.$$

□

Proposition 5.11. *For a Poisson counting process $(\mathcal{P}(t))_{t\geq 0}$ with rate 1 it holds*

$$\frac{\mathcal{P}(t)}{t} \xrightarrow[t\to\infty]{a.s.} 1.$$

Proof. The idea is that we want to use the Strong Law of Large Numbers (see appendix Theorem 7.1). For this, we have to write $\mathcal{P}(t)$ as a sum of i.i.d. random variables. Let us use $\mathcal{P}(t) - \mathcal{P}(t-1)$, for some $t \geq 1$, which are independent by the properties of a Poisson counting process. Note that $\mathcal{P}(t) - \mathcal{P}(t-1) \sim \text{Poi}(1)$ by definition of a Poisson counting process with rate 1. Furthermore, as the increments are Poi(1) distributed, it follows that $\mathcal{P}(t) - \mathcal{P}(t-1) \in L^1$, i.e. $\mathbb{E}[|\mathcal{P}(t) - \mathcal{P}(t-1)|] < \infty$. Now, as $\mathcal{P}(0) = 0$, we can write

$$\mathcal{P}(\lfloor t \rfloor) = \sum_{i=1}^{\lfloor t \rfloor} \mathcal{P}(i) - \mathcal{P}(i-1).$$

Using the Strong Law of Large Numbers and the fact that the expectation of a Poi(1) random variable is 1, we get

$$\frac{\mathcal{P}(\lfloor t \rfloor)}{\lfloor t \rfloor} \xrightarrow[t\to\infty]{a.s.} 1.$$

Next, note that by Corollary 5.10 it holds $\mathcal{P}(\lfloor t \rfloor) \leq \mathcal{P}(t) \leq \mathcal{P}(\lceil t \rceil)$. Hence,

$$\frac{\mathcal{P}(\lfloor t \rfloor)}{\lceil t \rceil} \leq \frac{\mathcal{P}(t)}{t} \leq \frac{\mathcal{P}(\lceil t \rceil)}{\lfloor t \rfloor}$$

Then, it holds that

$$\frac{\mathcal{P}(\lfloor t \rfloor)}{\lceil t \rceil} \xrightarrow[t\to\infty]{a.s.} 1 \quad \text{by the Strong Law of Large Numbers and}$$

$$\frac{\mathcal{P}(\lceil t \rceil)}{\lfloor t \rfloor} \xrightarrow[t\to\infty]{a.s.} 1 \quad \text{by the Strong Law of Large Numbers.}$$

Hence, we can conclude by the squeeze theorem that $\frac{\mathcal{P}(t)}{t} \xrightarrow[t\to\infty]{a.s.} 1$. □

Now, that we have defined the Poisson counting process and also some of its properties, we can write down the Lukasiewicz walk $\mathbb{S}_i$ explicitly. For this, we want to sum all discovered neighbours until step $i \geq 0$, where $i = Mk + U$ with $1 \leq U \leq k$ and $M \geq 0$, using Lemma 5.4.

$$\underline{\mathbb{S}}_i = \begin{pmatrix} \mathbb{S}_i^{(1)} \\ \mathbb{S}_i^{(2)} \\ \mathbb{S}_i^{(3)} \\ ... \\ \mathbb{S}_i^{(k)} \end{pmatrix} = \begin{pmatrix} \text{Poi}(\alpha_1 q_{in}) \\ \text{Poi}(\alpha_2 q_{out}) \\ \text{Poi}(\alpha_3 q_{out}) \\ ... \\ \text{Poi}(\alpha_k q_{out}) \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix} + \begin{pmatrix} \text{Poi}(\alpha_1 q_{out}(1-q_{in})) \\ \text{Poi}(\alpha_2 q_{in}(1-q_{out})) \\ \text{Poi}(\alpha_3 q_{out}(1-q_{out})) \\ ... \\ \text{Poi}(\alpha_k q_{out}(1-q_{out})) \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \\ ... \\ 0 \end{pmatrix} + ...$$

$$... + \text{Poi}(\underline{\alpha}^{(Mk+U)}) - \begin{pmatrix} 0 \\ ... \\ 1 \\ ... \\ 0 \end{pmatrix} = \text{Poi}(\underline{\alpha}^{(1)}) - \underline{e}_1 + \text{Poi}(\underline{\alpha}^{(2)}) - \underline{e}_2 + ... + \text{Poi}(A^M \underline{\alpha}^{(U)}) - \underline{e}_U$$

where $\underline{e}_i$ is the vector which has a 1 exactly in the i-th row.

To sum all these terms, we make a simplification. Let us assume for the moment that we only want to calculate $\underline{\mathbb{S}}_i$ with $i = Mk$ and $M > 0$. The last parameter in the Poisson distribution of the sum then is

$$\underline{\alpha}^{(Mk)} = \underline{\alpha}^{(M-1)k+k} = (1-q_{in})^{(M-1)}(1-q_{out})^{(M-1)(k-1)+k-2} \begin{pmatrix} \alpha_1 q_{out}(1-q_{in}) \\ ... \\ \alpha_{k-1} q_{out}(1-q_{in}) \\ \alpha_k q_{in}(1-q_{out}) \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_1 q_{out}(1-q_{in})^M (1-q_{out})^{M(k-1)-1} \\ ... \\ \alpha_{k-1} q_{out}(1-q_{in})^M (1-q_{out})^{M(k-1)-1} \\ \alpha_k q_{in}(1-q_{in})^{(M-1)} (1-q_{out})^{M(k-1)} \end{pmatrix}.$$

With this simplification, we know exactly how many times we were in each community, namely we visited the communities 1 to $k-1$ exactly M times and community k $M-1$ times. Hence, we have the vector $\underline{e}_k$ $M-1$ times and the other vectors $\underline{e}_i$, $i \neq k$, M times. Therefore, we get

$$\underline{\mathbb{S}}_{Mk} = \sum_{i=1}^{Mk} \text{Poi}(\underline{\alpha}^{(i)}) - M\underline{e}_1 - M\underline{e}_2 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k.$$

Now using the fact that, by Lemma 5.4, the Poisson random variables are independent, we can sum all the parameters of the Poisson random variables in a single Poisson random variable. By denoting a standard unit-rate Poisson counting process on $\mathbb{R}_+$ by $(\mathcal{P}(t) : t \geq 0)$, we can write

the Lukasiewicz walk as

$$\begin{aligned}
\underline{\mathbb{S}}_{Mk} &= \mathcal{P}\left(\sum_{i=1}^{Mk}\underline{\alpha}^{(i)}\right) - M\underline{e}_1 - M\underline{e}_2 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k \\
&= \mathcal{P}\left(\sum_{m=1}^{M-1}\sum_{u=1}^{k}\underline{\alpha}^{(mk+u)}\right) - M\underline{e}_1 - M\underline{e}_2 - ...M\underline{e}_{k-1} - (M-1)\underline{e}_k \\
&\overset{\text{by (16)}}{=} \mathcal{P}\left(\sum_{m=1}^{M-1}\sum_{u=1}^{k}A^m\underline{\alpha}^{(u)}\right) - M\underline{e}_1 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k \\
&= \mathcal{P}\left(\sum_{m=1}^{M-1}A^m\left(\sum_{u=1}^{k}\underline{\alpha}^{(u)}\right)\right) - M\underline{e}_1 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k \\
&= \mathcal{P}\left(\left(\sum_{u=1}^{k}\underline{\alpha}^{(u)}\right)\sum_{m=1}^{M-1}A^m\right) - M\underline{e}_1 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k \\
&= \mathcal{P}\left(\left(\sum_{u=1}^{k}\underline{\alpha}^{(u)}\right)\frac{1-A^M}{1-A}\right) - M\underline{e}_1 - ... - M\underline{e}_{k-1} - (M-1)\underline{e}_k,
\end{aligned}$$

where in the last equality we used the formula for the geometric series, as $0 < A < 1$, since $0 < q_{in}, q_{out} < 1$.

If we write this community-wise, we get

$$\underline{\mathbb{S}}_{Mk} = \begin{pmatrix}
\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_1^{(u)}\right)\frac{1-A^M}{1-A}\right) - M \\
\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_2^{(u)}\right)\frac{1-A^M}{1-A}\right) - M \\
... \\
\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_{k-1}^{(u)}\right)\frac{1-A^M}{1-A}\right) - M \\
\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_k^{(u)}\right)\frac{1-A^M}{1-A}\right) - (M-1)
\end{pmatrix}.$$

Before we continue to analyse the behaviour of the Lukasiewicz walk as $n \to \infty$, let us first remark how the Lukasiewicz walk and the core are linked.

5.3 Relation between Poissonized Graph and Core

Our model is defined to have stack vertices and the core, which is the graph we want to analyse for the giant component. Consequently, we need to define how the poissonized graph is related to the core such that when we analyse properties of the poissonized graph we know how to get back to the core (the "normal" SSBM graph). If we only look at the edges that are truly used in the Lukasiewicz exploration, meaning not the edges between stack vertices and revealed vertices, nor edges between revealed vertices (i.e. the red edges in Figure 7 are not considered), then we get a spanning forest

$$\mathrm{F}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out}) \subset \mathrm{SSBM}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out}).$$

Since all the vertices are still present, it is spanning and since we removed edges that contributed to a cycle, it is a forest. Notice that the Lukasiewicz walk of this forest is exactly $\underline{\mathbb{S}}$. In particular, new minimal records in one of the coordinates of $\underline{\mathbb{S}}$ are followed by the discovery of a new tree component in $\mathrm{F}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$. This holds, since every new minimal record means that we have exhausted the vertices, put on top of the stack, in this community. Therefore, the next time vertices in the other communities do not have any neighbours in this community, we pick a new vertex from the infinite array and therefore build a new tree component. If we further remove all vertices of the stack (together with the adjacent edges) we split $\mathrm{F}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ into a finer forest $\mathrm{F}'_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ which spans the core

$$\mathrm{F}'_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out}) \subset \mathrm{SSBM}(\eta, k, q_{in}, q_{out}).$$

Thus, the connected components in $\mathrm{F}'_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ are included in the connected components in $\mathrm{SSBM}(\eta, k, q_{in}, q_{out})$ and the connected components in $\mathrm{SSBM}(\eta, k, q_{in}, q_{out})$ are included in the connected components in $\mathrm{F}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$.

Therefore, if we can make statements about the connected components in $\mathrm{F}_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ and in $\mathrm{F}'_{\mathrm{Poi}}(\underline{\alpha}, k, q_{in}, q_{out})$ then it also holds for the components in $\mathrm{SSBM}(\eta, k, q_{in}, q_{out})$. This idea will be helpful for proving statements about the giant component.

Now, that we have a general term for the Lukasiewicz walk and also know how to get from the poissonized version to the core, we want to analyse the behaviour of the Lukasiewicz walk as $n \to \infty$ and see a proof for the existence of a giant component in the next subchapter.

5.4 Phase Transition for the Giant Component

To prove a theorem for the giant component, similar to Theorem 3.22, let us assume that $q_{in} = a/n$ and $q_{out} = b/n$ so that we have the model $\mathrm{SSBM}(n, k, a/n, b/n)$. Let us thus set $\alpha_i = \frac{n}{k}$ and assume that $a, b > 0$ as well as n being a multiple of k. We see that the expected

number of vertices in the graph is then $\eta = \sum_{i=1}^{k} \alpha_i = k \cdot \frac{n}{k} = n$, which is what we want. We consider Proposition 5.12.

Proposition 5.12 (Fluid Limit)**.** *It holds that*

$$\left(\frac{\mathbb{S}_{\lfloor nt \rfloor}}{n/k}\right)_{t\geq 0} \xrightarrow[n\to\infty]{a.s.} \begin{pmatrix} 1 - e^{-\frac{a+(k-1)b}{k}t} - t \\ \dots \\ 1 - e^{-\frac{a+(k-1)b}{k}t} - t \end{pmatrix}_{t\geq 0}$$

in the uniform norm over t where the vector on the right is in $\mathbb{R}^k$.

Proposition 5.12 therefore tells us that the limit in each community is the same, which allows us to analyse its asymptotic behaviour better. But before we consider this, let us continue to prove Proposition 5.12. For this we need the bounds given in Lemma 5.13 and Lemma 5.14.

Lemma 5.13. *Let $r > 0$ and $n \in \mathbb{N}$ be such that $\frac{r}{n} \leq \frac{1}{2}$, then the following bounds hold*

$$e^{-rt-\frac{7}{6}r^2\frac{t}{n}} \leq \left(1 - \frac{r}{n}\right)^{\lfloor nt \rfloor} \leq e^{-rt+\frac{7}{6}r^2\frac{t}{n}+\frac{r}{n}}.$$

Proof. Note that we have the following inequality for $0 < x < 1$, using the Taylor expansion of ln

$$\begin{aligned} |\ln(1-x)+x| &= \left|-x + \sum_{i=2}^{\infty} -\frac{x^i}{i} + x\right| = \left|\sum_{i=2}^{\infty} -\frac{x^i}{i}\right| \leq \sum_{i=2}^{\infty} \left|-\frac{x^i}{i}\right| = \frac{x^2}{2} + \sum_{i=3}^{\infty} \frac{x^i}{i} \\ &\leq \frac{x^2}{2} + \sum_{i=3}^{\infty} \frac{x^i}{3} = \frac{x^2}{2} + \frac{1}{3}\sum_{i=0}^{\infty} x^{3+i} = \frac{x^2}{2} + \frac{1}{3}x^3 \sum_{i=0}^{\infty} x^i = \frac{x^2}{2} + \frac{1}{3}x^3 \frac{1}{1-x} \\ &= \frac{x^2}{2} + \frac{x^3}{3(1-x)} \leq \frac{x^2}{2} + \frac{x^2}{3(1-x)} \leq x^2\left(\frac{1}{2} + \frac{1}{3(1-x)}\right). \end{aligned}$$

Now, we want to use this inequality with $x = \frac{r}{n}$ and we assume $\frac{r}{n} \leq \frac{1}{2}$, then we have

$$\left|\ln(1-\frac{r}{n}) + \frac{r}{n}\right| \leq \left(\frac{r}{n}\right)^2 \left(\frac{1}{2} + \frac{1}{3(1-\frac{r}{n})}\right) \leq \left(\frac{r}{n}\right)^2 \left(\frac{1}{2} + \frac{1}{3\cdot\frac{1}{2}}\right) = \left(\frac{r}{n}\right)^2 \cdot \frac{7}{6}.$$

Hence, we have that $-\frac{7}{6}\frac{r^2}{n^2} \leq \ln(1-\frac{r}{n}) + \frac{r}{n} \leq \frac{7}{6}\frac{r^2}{n^2}$ which is equivalent to $-\frac{7}{6}\frac{r^2}{n^2} - \frac{r}{n} \leq \ln(1-\frac{r}{n}) \leq \frac{7}{6}\frac{r^2}{n^2} - \frac{r}{n}$. Notice further that $\lfloor nt \rfloor \leq nt$. Therefore, we get

$$\left(1 - \frac{r}{n}\right)^{\lfloor nt \rfloor} = e^{\ln((1-\frac{r}{n})^{\lfloor nt \rfloor})} = e^{\lfloor nt \rfloor \ln(1-\frac{r}{n})} \geq e^{\lfloor nt \rfloor(-\frac{7}{6}\frac{r^2}{n^2}-\frac{r}{n})} \geq e^{nt(-\frac{7}{6}\frac{r^2}{n^2}-\frac{r}{n})} = e^{-\frac{7}{6}r^2\frac{t}{n}-rt}.$$

On the other hand, we have that

$$\left(1-\frac{r}{n}\right)^{\lfloor nt\rfloor} = e^{\lfloor nt\rfloor \ln(1-\frac{r}{n})} \leq e^{\lfloor nt\rfloor(\frac{7}{6}\frac{r^2}{n^2}-\frac{r}{n})} \leq e^{nt\frac{7}{6}\frac{r^2}{n^2}-\lfloor nt\rfloor\frac{r}{n}} = e^{\frac{7}{6}r^2\frac{t}{n}-nt\frac{r}{n}+\{nt\}\frac{r}{n}} \leq e^{\frac{7}{6}r^2\frac{t}{n}-rt+\frac{r}{n}}$$

where we used the fact that $0 \leq \{nt\} < 1$. □

Lemma 5.14. *It holds for all $|x| \leq 1$ that*

$$(1-e^{-1})|x| \leq |1-e^x| \leq (e-1)|x|.$$

Proof. Let us split the proof into two parts. Assume first that $x \geq 0$, then we have that $e^x \geq 1$ and thus $|1-e^x| = e^x - 1 \geq 0$, as well as $|x| = x$. Now, we want to show that

$$(1-e^{-1})x \leq e^x - 1 \leq (e-1)x.$$

Let us consider the two inequalities separately.
1) First, we show $(1-e^{-1})x \leq e^x - 1$. For this, we define the function $f(x) := e^x - 1 - (1-e^{-1})x$. Its derivative is $f'(x) = e^x - (1-e^{-1}) = e^x - 1 + e^{-1} > 0$, because we have that $e^x \geq 1$ on the interval $[0,1]$. Therefore, f is increasing on $[0,1]$ and it holds that $f(0) = e^0 - 1 - (1-e^{-1})\cdot 0 = 0$. Hence, we can conclude that $f(x) \geq 0$, for every $0 \leq x \leq 1$ which implies that $(1-e^{-1})x \leq e^x - 1$.
2) For the other inequality, let us define the function $g(x) := (e-1)x - e^x + 1$. Then, $g'(x) = e - 1 - e^x$. As g is continuous, it has an extremum on a compact interval, namely at point $x = \ln(e-1) \approx 0.54$. It holds that $g''(x) = -e^x \leq 0$ and hence the extremum is a maximum. On the endpoints it holds that $g(0) = -e^0 + 1 = 0$ and $g(1) = (e-1) - e + 1 = 0$. Therefore, we can conclude that $g(x) \geq 0$ for all $0 \leq x \leq 1$ which implies that $e^x - 1 \leq (e-1)x$.

Now, assume that $x < 0$, then $e^x < 1$ and $|1-e^x| = 1 - e^x$, as well as $|x| = -x$. So, we want to show

$$-(1-e^{-1})x \leq 1 - e^x \leq -(e-1)x.$$

Again, let us look at the two inequalities separately.
1) To show $1-e^x \geq -(1-e^{-1})x$, let us again define $f(x) := 1-e^x+(1-e^{-1})x$. This is a continuous function and so it has an extremum at point $x = \ln(1-e^{-1}) \approx -0.46$. Furthermore, it holds that $f''(x) = -e^x \leq 0$ for every $x \in [-1,0]$. Thus, our extremum is a maximum and for the endpoints it holds as well that $f(-1) = 1 - e^{-1} + (1-e^{-1})(-1) = 0$ and $f(0) = 1 - e^0 + (1-e^{-1})\cdot 0 = 0$. Thus, we can conclude that$f(x) \geq 0$ for all $-1 \leq x < 0$, which implies that $1-e^x \geq -(1-e^{-1})x$.
2) For the other equality, let us use the function $g(x) := -(e-1)x - (1-e^x)$. Then, we have that $g'(x) = -(e-1) + e^x = e^x + 1 - e < 0$, as we have that $e^x + 1 < 2$ for every $-1 \leq x < 1$. Therefore, g is decreasing and it holds that $g(0) = -(e-1)\cdot 0 - (1-e^0) = 0$. Hence, $g(x) \geq 0$ for all $-1 \leq x < 0$ which finally leads to $1 - e^x \leq -(e-1)x$. □

Now that we have these two lemmas, let us continue to prove the Fluid Limit, i.e. Proposition 5.12

Proof of Proposition 5.12. We want to show that

$$\left\| \frac{\mathbb{S}_{\lfloor nt \rfloor}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \cdot \mathbf{1} \right\|_\infty \xrightarrow[n\to\infty]{a.s.} 0$$

where

$$\left\| \frac{\mathbb{S}_{\lfloor nt \rfloor}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \cdot \mathbf{1} \right\|_\infty = \max_{1\le i\le k} \left(\sup_{t\ge 0} \left| \frac{\mathbb{S}^{(i)}_{\lfloor nt \rfloor}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \right| \right).$$

Hence, we need to show that each component converges to zero. To make our computations simpler, let us define $M(t) := \frac{\lfloor nt \rfloor}{k}$ such that $\lfloor nt \rfloor = M(t)k$. Then, we have that

$$\mathbb{S}^{(i)}_{\lfloor nt \rfloor} = \mathbb{S}^{(i)}_{M(t)k} = \mathcal{P}\left(\left(\sum_{u=1}^{k} \alpha_i^{(u)} \right) \frac{1 - A^{M(t)}}{1 - A} \right) - (M(t) - \mathbb{1}_{\{i=k\}}).$$

We see that

$$\begin{aligned}
&\sup_{t\ge 0} \left| \frac{\mathbb{S}^{(i)}_{\lfloor nt \rfloor}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \right| \\
&= \sup_{t\ge 0} \left| \frac{\mathcal{P}\left(\left(\sum_{u=1}^{k} \alpha_i^{(u)} \right) \frac{1-A^{M(t)}}{1-A} \right) - (M(t) - \mathbb{1}_{\{i=k\}})}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \right| \\
&= \sup_{t\ge 0} \left| \frac{\mathcal{P}\left(\left(\sum_{u=1}^{k} \alpha_i^{(u)} \right) \frac{1-A^{M(t)}}{1-A} \right)}{n/k} - \frac{\frac{\lfloor nt \rfloor}{k}}{n/k} + \frac{\mathbb{1}_{\{i=k\}}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \right| \\
&= \sup_{t\ge 0} \left| \frac{\mathcal{P}\left(\left(\sum_{u=1}^{k} \alpha_i^{(u)} \right) \frac{1-A^{M(t)}}{1-A} \right)}{n/k} - \frac{\frac{nt - \{nt\}}{k}}{n/k} + \frac{\mathbb{1}_{\{i=k\}}}{n/k} - (1 - e^{-\frac{a+(k-1)b}{k}t} - t) \right|
\end{aligned}$$

$$= \sup_{t\geq 0}\left|\frac{\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_i^{(u)}\right)\frac{1-A^{M(t)}}{1-A}\right)}{n/k} - t + \frac{\{nt\}}{n} + \frac{\mathbb{1}_{\{i=k\}}}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t}) + t\right|$$

$$\leq \sup_{t\geq 0}\left|\frac{\mathcal{P}\left(\left(\sum_{u=1}^{k}\alpha_i^{(u)}\right)\frac{1-A^{M(t)}}{1-A}\right)}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t})\right| + \sup_{t\geq 0}\left|\frac{\{nt\}}{n} + \frac{k\cdot\mathbb{1}_{\{i=k\}}}{n}\right|$$

As $0 \leq \{nt\} < 1$, we can remove the supremum in the second part and write

$$\sup_{t\geq 0}\left|\frac{\mathbb{S}^{(i)}_{\lfloor nt\rfloor}}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t} - t)\right| \leq \sup_{t\geq 0}\left|\frac{\mathcal{P}(\left(\sum_{u=1}^{k}\alpha_i^{(u)}\right)\frac{1-A^{M(t)}}{1-A})}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t})\right| + O\left(\frac{1}{n}\right). \tag{17}$$

We reduced the problem to showing the convergence of the supremum above. First, let us look at how the different parts in the parameter of the Poisson process behave if $n \to \infty$. It holds that

$$\left(\sum_{u=1}^{k}\alpha_i^{(u)}\right) = \alpha_i q_{out} + ... + \alpha_i q_{out}(1-q_{out})^{i-2} + \alpha_i q_{in}(1-q_{out})^{i-1} + \alpha_i q_{out}(1-q_{in})(1-q_{out})^{i-1}$$

$$+ \alpha_i q_{out}(1-q_{in})(1-q_{out})^{i} + ... + \alpha_i q_{out}(1-q_{in})(1-q_{out})^{k-2}$$

$$= \alpha_i\left(q_{out}\sum_{j=0}^{i-2}(1-q_{out})^j + q_{in}(1-q_{out})^{i-1} + q_{out}(1-q_{in})(1-q_{out})^{i-1}\sum_{j=0}^{k-i-1}(1-q_{out})^j\right)$$

$$= \alpha_i\left(q_{out}\frac{1-(1-q_{out})^{i-1}}{q_{out}} + q_{in}(1-q_{out})^{i-1} + q_{out}(1-q_{in})(1-q_{out})^{i-1}\frac{1-(1-q_{out})^{k-i}}{q_{out}}\right)$$

$$= \alpha_i\left(1-(1-q_{out})^{i-1} + q_{in}(1-q_{out})^{i-1} + (1-q_{in})(1-q_{out})^{i-1}(1-(1-q_{out})^{k-i})\right),$$

where we used the formula for the geometric series as $0 < q_{in}, q_{out} < 1$.

Now, using the explicit values $q_{in} = \frac{a}{n}$ and $q_{out} = \frac{b}{n}$ and the approximation $(1-x)^n = 1 - x \cdot n + O(x^2)$, we get

$$\left(\sum_{u=1}^{k} \alpha_i^{(u)}\right) = \alpha_i \left(1 - \left(1-\frac{b}{n}\right)^{i-1} + \frac{a}{n}\left(1-\frac{b}{n}\right)^{i-1} + \left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)^{i-1}\left(1-\left(1-\frac{b}{n}\right)^{k-i}\right)\right)$$

$$= \alpha_i \left(1 - \left(1-(i-1)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right) + \frac{a}{n}\left(1-(i-1)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\right.$$

$$\left. + \left(1-\frac{a}{n}\right)\left(1-(i-1)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\left(1-\left(1-(k-i)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\right)\right)$$

$$= \alpha_i \left((i-1)\frac{b}{n} - O\left(\frac{1}{n^2}\right) + \frac{a}{n} - (i-1)\frac{ab}{n^2} + \frac{a}{n}O\left(\frac{1}{n^2}\right)\right.$$

$$\left. + \left(1-(i-1)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\left((k-i)\frac{b}{n} - O\left(\frac{1}{n^2}\right) - (k-i)\frac{ab}{n^2} + \frac{a}{n}O\left(\frac{1}{n^2}\right)\right)\right)$$

$$= \alpha_i \left((i-1)\frac{b}{n} + \frac{a}{n} + O\left(\frac{1}{n^2}\right) + \left(1-(i-1)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\left((k-i)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right)\right)$$

$$= \alpha_i \left((i-1)\frac{b}{n} + \frac{a}{n} + O\left(\frac{1}{n^2}\right) + (k-i)\frac{b}{n} + O\left(\frac{1}{n^2}\right)\right.$$

$$\left. - (i-1)(k-i)\frac{b^2}{n^2} - (i-1)\frac{b}{n}O\left(\frac{1}{n^2}\right) + (k-i)\frac{b}{n}O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{n^2}\right)O\left(\frac{1}{n^2}\right)\right)$$

$$= \alpha_i \left(\frac{a + (i-1)b + (k-i)b}{n} + O\left(\frac{1}{n^2}\right)\right) = \alpha_i \left(\frac{a + (k-1)b}{n} + O\left(\frac{1}{n^2}\right)\right).$$

Here we always used the fact that $O(\frac{1}{n^3})$, as well as all higher order terms, are also of order $O(\frac{1}{n^2})$.

Next, let us analyse how $(1-A)$ evolves, using again the Binomial approximation for $(1-\frac{b}{n})^{k-1}$ as above

$$1-A = 1-\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)^{k-1} = 1-\left(1-\frac{a}{n}\right)\left(1-(k-1)\frac{b}{n}+O\left(\frac{1}{n^2}\right)\right)$$

$$= 1-\left(1-(k-1)\frac{b}{n}+O\left(\frac{1}{n^2}\right)-\frac{a}{n}+(k-1)\frac{ab}{n^2}-\frac{a}{n}O\left(\frac{1}{n^2}\right)\right) = \frac{a+(k-1)b}{n}+O\left(\frac{1}{n^2}\right).$$

Hence, if we denote by $d := a+(k-1)b$ and notice that $d>0$, we get the following for the product of the two terms

$$\left(\sum_{u=1}^{k}\alpha_i^{(u)}\right)\cdot\frac{1}{1-A} = \alpha_i\left(\frac{d}{n}+O\left(\frac{1}{n^2}\right)\right)\cdot\frac{1}{\frac{d}{n}+O\left(\frac{1}{n^2}\right)} = \alpha_i\left(\frac{d/n}{\frac{d}{n}+O\left(\frac{1}{n^2}\right)}+\frac{O\left(\frac{1}{n^2}\right)}{\frac{d}{n}+O\left(\frac{1}{n^2}\right)}\right)$$

$$= \alpha_i\left(\frac{d}{d+n\cdot O\left(\frac{1}{n^2}\right)}+\frac{n\cdot O\left(\frac{1}{n^2}\right)}{d+n\cdot O\left(\frac{1}{n^2}\right)}\right) = \alpha_i\left(\frac{d}{d+O\left(\frac{1}{n}\right)}+\frac{O\left(\frac{1}{n}\right)}{d+O\left(\frac{1}{n}\right)}\right)$$

$$= \alpha_i(1+o(1)).$$

Therefore, we can write the Lukasiewicz walk also as

$$\mathbb{S}^{(i)}_{M(t)k} = \mathcal{P}(\alpha_i(1+o(1))\cdot(1-A^{M(t)})) - (M(t)-\mathbb{1}_{\{i=k\}}).$$

To show the convergence in (17), we can equivalently show the following

$$\sup_{t\geq 0}\left|\frac{\mathcal{P}(\alpha_i(1+o(1))\cdot(1-A^{M(t)}))}{n/k}-(1-e^{-\frac{a+(k-1)b}{k}t})\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

For this, we choose $\alpha_i = \frac{n}{k}$ for all $1\leq i\leq k$, which results in

$$\sup_{t\geq 0}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k}-(1-e^{-\frac{a+(k-1)b}{k}t})\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

We show this convergence in three different cases. The first one being the convergence with $t\in[a_n,b_n]$ where $a_n := \frac{1}{c\ln(n)}$, $c := \frac{a+(k-1)b}{k}$ and $b_n := \ln(n)$. The second case concerns the convergence for $t\in[b_n,\infty)$. The third case treats the interval around 0, i.e. $t\in[0,a_n]$. If we show convergence to 0 over the supremum in these three intervals, then we have also shown uniform convergence in t to 0.

i) Let us assume that $t \in [a_n, b_n]$ and $c = \frac{a+(k-1)b}{k}$ being the expected degree in the model. We want to show that,

$$\sup_{t\in[a_n,b_n]} \left| \frac{\mathcal{P}(\frac{n}{k}(1+o(1)) \cdot (1-A^{M(t)}))}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t}) \right| \xrightarrow[n\to\infty]{a.s.} 0. \tag{18}$$

First, we bound the term $A^{M(t)}$ using Lemma 5.13. Let $A := (1-\frac{a}{n})(1-\frac{b}{n})^{k-1}$ as defined in (15). We get

$$\begin{aligned}
A^{M(t)} &= \left(\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)^{k-1}\right)^{M(t)} = \left(1-\frac{a}{n}\right)^{M(t)}\left(1-\frac{b}{n}\right)^{(k-1)M(t)} \\
&= \left(1-\frac{a}{n}\right)^{\frac{\lfloor nt\rfloor}{k}}\left(1-\frac{b}{n}\right)^{(k-1)\frac{\lfloor nt\rfloor}{k}} = \left(\left(1-\frac{a}{n}\right)^{\lfloor nt\rfloor}\right)^{\frac{1}{k}}\left(\left(1-\frac{b}{n}\right)^{\lfloor nt\rfloor}\right)^{\frac{k-1}{k}} \\
&\geq \left(e^{-at-\frac{7}{6}a^2\frac{t}{n}}\right)^{\frac{1}{k}}\left(e^{-bt-\frac{7}{6}b^2\frac{t}{n}}\right)^{\frac{k-1}{k}} = e^{-\frac{a}{k}t-\frac{7}{6}\frac{a^2}{k}\frac{t}{n}}e^{-\frac{(k-1)b}{k}t-\frac{7}{6}\frac{(k-1)b^2}{k}\frac{t}{n}} \\
&= e^{-\frac{a+(k-1)b}{k}t-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{t}{n}} = e^{-ct-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{t}{n}}.
\end{aligned} \tag{19}$$

As $t \in [a_n, b_n]$, we can further bound this from below, using the fact that e^{-x} is decreasing

$$A^{M(t)} \geq e^{-ct-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}}. \tag{20}$$

Next, let us do the same thing for an upper bound on $A^{M(t)}$, this time using the upper bound presented in Lemma 5.13

$$\begin{aligned}
A^{M(t)} &= \left(\left(1-\frac{a}{n}\right)^{\lfloor nt\rfloor}\right)^{\frac{1}{k}}\left(\left(1-\frac{b}{n}\right)^{\lfloor nt\rfloor}\right)^{\frac{k-1}{k}} \leq \left(e^{-at+\frac{7}{6}a^2\frac{t}{n}+\frac{a}{n}}\right)^{\frac{1}{k}}\left(e^{-bt+\frac{7}{6}b^2\frac{t}{n}+\frac{b}{n}}\right)^{\frac{k-1}{k}} \\
&= e^{-\frac{a}{k}t+\frac{7}{6}\frac{a^2}{k}\frac{t}{n}+\frac{a}{kn}}e^{-\frac{(k-1)b}{k}t+\frac{7}{6}\frac{(k-1)b^2}{k}\frac{t}{n}+\frac{(k-1)b}{kn}} = e^{-\frac{a+(k-1)b}{k}t+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{t}{n}+\frac{a+(k-1)b}{kn}} \\
&= e^{-ct+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{t}{n}+\frac{c}{n}}.
\end{aligned} \tag{21}$$

Here we can also use that $t \in [a_n, b_n]$ and the fact that e^x is increasing to write

$$A^{M(t)} \leq e^{-ct+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}}. \tag{22}$$

The bounds in (20) and (22) lead to the following inequality

$$1 - e^{-ct + \frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n} + \frac{c}{n}} \leq 1 - A^{M(t)} \leq 1 - e^{-ct - \frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}}.$$

Using the monotonicity of the Poisson counting process, given in Corollary 5.10, we can bound our Poisson counting process from above and below by

$$\frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-A^{M(t)})])}{n/k} - (1-e^{-ct}) \geq \frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-e^{-ct+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})])}{n/k} - (1-e^{-ct})$$

and

$$\frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-A^{M(t)})])}{n/k} - (1-e^{-ct}) \leq \frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-e^{-ct-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}})])}{n/k} - (1-e^{-ct}).$$

Let us call the lower and upper bound

$$L_n(t) := \frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-e^{-ct+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})])}{n/k} - (1-e^{-ct})$$

$$U_n(t) := \frac{\mathcal{P}(\frac{n}{k}[(1+o(1))(1-e^{-ct-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}})])}{n/k} - (1-e^{-ct}).$$

Therefore, we have that

$$\sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}[(1+o(1)(1-A^{M(t)})])}{n/k} - (1-e^{-ct})\right| \leq \max(\sup_{t\in[a_n,b_n]}\left|L_n(t)\right|, \sup_{t\in[a_n,b_n]}\left|U_n(t)\right|). \tag{23}$$

If we can show that both of these terms converge a.s. to zero as $n \to \infty$, then we have also shown the convergence in Equation (18). Let us start with the lower bound $L_n(t)$. To make the expression a little bit clearer, let us define $\lambda_n^L(t) := (1+o(1))(1-e^{-ct+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})$. Note that we have $\sup_{t\in[a_n,b_n]}\left|\lambda_n^L(t)\right| \leq 1+o(1)$, for n big enough. Hence, if we assume n to be big enough, we get

$$\sup_{t\in[a_n,b_n]}\left|L_n(t)\right| = \sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{n/k} - (1-e^{-ct})\right| = \sup_{t\in[a_n,b_n]}\left|\lambda_n^L(t)\right|\left(\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)} - \frac{1-e^{-ct}}{\lambda_n^L(t)}\right|\right)$$

$$\leq (1+o(1))\sup_{t\in[a_n,b_n]}\left(\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)} - 1 - \left(\frac{1-e^{-ct}}{\lambda_n^L(t)} - 1\right)\right|\right)$$

$$\leq (1+o(1))\left(\sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)} - 1\right| + \sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^L(t)} - 1\right|\right). \tag{24}$$

Now, let us analyse these two parts separately, starting with the part including the Poisson counting process.

Claim 5.15. *We have that*

$$\sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)}-1\right|\xrightarrow[n\to\infty]{a.s.}0.$$

Proof. For $t\in[a_n,b_n]$ we have that

$$\lambda_n^L(t)\in[(1+o(1))(1-e^{-ca_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}}),(1+o(1))(1-e^{-cb_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})].$$

If we define the interval

$$I_n:=\left[\frac{n}{k}(1+o(1))(1-e^{-ca_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}}),\frac{n}{k}(1+o(1))(1-e^{-cb_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})\right]$$

we have that $\gamma(n):=\frac{n}{k}\lambda_n^L(t)\in I_n$.
Next, let us look at how I_n evolves asymptotically. Since $e^{-x}\xrightarrow[x\to\infty]{}0$, we get that

$$\frac{n}{k}(1+o(1))(1-e^{-cb_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})=\frac{n}{k}(1+o(1))(1-e^{-c\ln(n)+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}})\xrightarrow[n\to\infty]{}\infty.$$

On the other hand, to show that

$$\frac{n}{k}(1+o(1))(1-e^{-ca_n+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n}})=\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}})\xrightarrow[n\to\infty]{}\infty,$$

we use L'Hôpital's rule. For this define

$$g(n):=\frac{1}{n}\quad\text{and}$$

$$f(n):=\frac{1}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}).$$

Then it holds that

$$g'(n)=-\frac{1}{n^2}\quad\text{and}$$

$$f'(n)=\frac{1}{k}(1+o(1))\left(-e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}\left(\frac{1}{n\ln^2(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1-\ln(n)}{n^2}-\frac{c}{n^2}\right)\right).$$

Thus,

$$\lim_{n\to\infty} \frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}) = \lim_{n\to\infty}\frac{f(n)}{g(n)} = \lim_{n\to\infty}\frac{f'(n)}{g'(n)}$$

$$= \lim_{n\to\infty}\left(n^2\cdot\frac{1}{k}(1+o(1))e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}\left(\frac{1}{n\ln^2(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1-\ln(n)}{n^2}-\frac{c}{n^2}\right)\right)$$

$$= \lim_{n\to\infty}\left(\frac{1}{k}(1+o(1))e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}\left(\frac{n}{\ln^2(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}(1-\ln(n))-c^2\right)\right) = \infty$$

because $e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}} \underset{n\to\infty}{\longrightarrow} 1$ and furthermore $\frac{n}{\ln^2(n)} \underset{n\to\infty}{\longrightarrow} \infty$ faster than $-\ln(n) \underset{n\to\infty}{\longrightarrow} -\infty$. Therefore, we see that $\frac{n}{k}\lambda_n^L(t) \in I_n$ implies that $\gamma(n) = \frac{n}{k}\lambda_n^L(t) \underset{n\to\infty}{\longrightarrow} \infty$.
By Proposition 5.11 we know that $\frac{\mathcal{P}(\gamma)}{\gamma} \underset{\gamma\to\infty}{\overset{a.s.}{\longrightarrow}} 1$, so for all $\epsilon > 0$ there exists $\gamma_0 > 0$ such that for all $\gamma > \gamma_0$ it holds that $\left|\frac{\mathcal{P}(\gamma)}{\gamma} - 1\right| < \epsilon$.
As

$$\frac{n}{k}(1+o(1))(1-e^{-c\ln(n)+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}) \underset{n\to\infty}{\longrightarrow} \infty \quad \text{and}$$

$$\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n}}) \underset{n\to\infty}{\longrightarrow} \infty,$$

there exists some $n_\epsilon > 0$ such that $\gamma_0 < \frac{n_\epsilon}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n_\epsilon)}+\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n_\epsilon)}{n_\epsilon}+\frac{c}{n_\epsilon}})$ and thus for all $n \geq n_\epsilon$ and $\gamma \in I_n$, we have that $\gamma > \gamma_0$ and so $\left|\frac{\mathcal{P}(\gamma)}{\gamma} - 1\right| < \epsilon$. Hence, also for all $n > n_\epsilon$ it holds that $\sup_{\gamma\in I_n}\left|\frac{\mathcal{P}(\gamma)}{\gamma} - 1\right| < \epsilon$. We can now conclude

$$\forall\,\epsilon > 0\,\exists\, n_\epsilon > 0 \text{ such that } \forall\, n \geq n_\epsilon \sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)} - 1\right| < \epsilon.$$

This implies

$$\sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}\lambda_n^L(t))}{\frac{n}{k}\lambda_n^L(t)} - 1\right| \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} 0. \qquad \square$$

With this we can continue to prove that the second part of (24) also converges to zero.

Claim 5.16. *It holds that*

$$\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^L(t)} - 1\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. Let us define $x_n := \frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n} + \frac{c}{n}$. We have that

$$\begin{aligned}
\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^L(t)}-1\right| &= \sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}-\lambda_n^L(t)}{\lambda_n^L(t)}\right| = \sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}-(1+o(1))(1-e^{-ct+x_n})}{\lambda_n^L(t)}\right| \\
&= \sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}-1+e^{-ct+x_n}-o(1)+o(1)e^{-ct+x_n}}{\lambda_n^L(t)}\right| \\
&\leq \sup_{t\in[a_n,b_n]}\left|\frac{e^{-ct}(e^{x_n}-1)}{\lambda_n^L(t)}\right| + \sup_{t\in[a_n,b_n]}\left|\frac{o(1)(e^{-ct+x_n}-1)}{\lambda_n^L(t)}\right| \\
&= \sup_{t\in[a_n,b_n]} e^{-ct}\frac{|e^{x_n}-1|}{|1+o(1)||1-e^{-ct+x_n}|} + \sup_{t\in[a_n,b_n]}\frac{o(1)|e^{-ct+x_n}-1|}{|1+o(1)||1-e^{-ct+x_n}|} \\
&\leq e^{-c\cdot a_n}\sup_{t\in[a_n,b_n]}\frac{|1-e^{x_n}|}{(1+o(1))|1-e^{-ct+x_n}|} + \sup_{t\in[a_n,b_n]}\frac{o(1)|1-e^{-ct+x_n}|}{(1+o(1))|1-e^{-ct+x_n}|} \\
&\leq \sup_{t\in[a_n,b_n]}\frac{|1-e^{x_n}|}{(1+o(1))|1-e^{-ct+x_n}|} + \frac{o(1)}{1+o(1)} \\
&= \sup_{t\in[a_n,b_n]}\frac{|1-e^{x_n}|}{(1+o(1))|1-e^{-ct+x_n}|} + o(1).
\end{aligned}$$

Now, recall the inequalities proven in Lemma 5.14. We can apply them here since for n big enough we have that $x_n \in [0,1]$.

$$\begin{aligned}
\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^L(t)}-1\right| &\leq \sup_{t\in[a_n,b_n]}\frac{|1-e^{x_n}|}{(1+o(1))|1-e^{-ct+x_n}|} + o(1) \\
&\leq \sup_{t\in[a_n,b_n]}\frac{(e-1)|x_n|}{(1+o(1))(1-e^{-1})|-ct+x_n|} + o(1) \\
&= \sup_{t\in[a_n,b_n]}\frac{(e-1)x_n}{(1+o(1))(1-e^{-1})(ct-x_n)} + o(1).
\end{aligned}$$

Note that the last step follows, as $-ct+x_n \leq 0$ for n big enough. We get

$$\begin{aligned}
\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^L(t)}-1\right| &\leq \frac{(e-1)(\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n})}{(1+o(1))(1-e^{-1})(c\cdot a_n - \frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}+\frac{c}{n})} + o(1) \\
&= \frac{(e-1)(\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}+\frac{c}{n})\cdot\frac{n}{\ln(n)}}{(1+o(1))(1-e^{-1})(c\cdot\frac{1}{c\ln(n)} - \frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{ln(n)}{n}+\frac{c}{n})\cdot\frac{n}{\ln(n)}} + o(1)
\end{aligned}$$

$$= \frac{(e-1)(\frac{7}{6}\frac{a^2+(k-1)b^2}{k} + \frac{c}{\ln(n)})}{(1+o(1))(1-e^{-1})(\frac{n}{\ln^2(n)} - \frac{7}{6}\frac{a^2+(k-1)b^2}{k} + \frac{c}{\ln(n)})} + o(1) \underset{n\to\infty}{\longrightarrow} 0.$$

The convergence in the last line follows as $e^{-\frac{1}{\ln(n)}} \to 1$ and $\frac{n}{\ln^2(n)} \to \infty$ for $n \to \infty$. □

Therefore, we proved with the help of inequality (24) that

$$\sup_{t\in[a_n,b_n]} \left|L_n(t)\right| \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} 0.$$

Now, let us do the same for $\sup_{t\in[a_n,b_n]}\left|U_n(t)\right|$.

Again, let us define $\lambda_n^U(t) := (1+o(1))(1-e^{-ct-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}})$. Then, it holds again that for n big enough $\sup_{t\in[a_n,b_n]}\left|\lambda_n^U(t)\right| \leq (1+o(1))$. Hence, we can estimate the upper bound as before

$$\sup_{t\in[a_n,b_n]} \left|U_n(t)\right| \leq (1+o(1)) \left(\sup_{t\in[a_n,b_n]} \left| \frac{\mathcal{P}(\frac{n}{k}\lambda_n^U(t))}{\frac{n}{k}\lambda_n^U(t)} - 1 \right| + \sup_{t\in[a_n,b_n]} \left| \frac{1-e^{-ct}}{\lambda_n^U(t)} - 1 \right| \right).$$

We can again analyse the two parts separately.

Claim 5.17. *It holds that*

$$\sup_{t\in[a_n,b_n]} \left| \frac{\mathcal{P}(\frac{n}{k}\lambda_n^U(t))}{\frac{n}{k}\lambda_n^U(t)} - 1 \right| \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} 0.$$

Proof. Note that $\lambda_n^U(t) \in \left[(1+o(1))(1-e^{-ca_n-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}}), (1+o(1))(1-e^{-cb_n-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}})\right]$. Furthermore, it holds again that

$$\frac{n}{k}(1+o(1))(1-e^{-c\ln(n)-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}}) \underset{n\to\infty}{\longrightarrow} \infty \quad \text{and}$$

$$\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}}) \underset{n\to\infty}{\longrightarrow} \infty,$$

where we again use L'Hôpital's rule to show the second convergence. Hence, we can conclude in exactly the same way as in the proof of Claim 5.15. □

On the other hand, it also holds that

Claim 5.18.

$$\sup_{t\in[a_n,b_n]} \left| \frac{1-e^{-ct}}{\lambda_n^U(t)} - 1 \right| \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. Let us define $y_n := -\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}$, then by doing the same computations as above, we get

$$\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^U(t)}-1\right| = \sup_{t\in[a_n,b_n]} e^{-ct}\frac{\left|1-e^{-y_n}\right|}{(1+o(1))|1-e^{-ct+y_n}|} + \sup_{t\in[a_n,b_n]}\frac{o(1)\left|1-e^{-ct+y_n}\right|}{(1+o(1))|1-e^{-ct+y_n}|}$$

$$\leq e^{-c\cdot a_n}\sup_{t\in[a_n,b_n]}\frac{\left|1-e^{-y_n}\right|}{(1+o(1))|1-e^{-ct+y_n}|}+o(1)$$

$$\leq e^{-\frac{1}{\ln(n)}}\sup_{t\in[a_n,b_n]}\frac{(e-1)|y_n|}{(1+o(1))(1-e^{-1})|-ct+y_n|}+o(1),$$

where we used the inequalities presented in Lemma 5.14 in the last line. Now, we notice that $-ct+y_n = -ct-\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}\leq 0$, hence

$$\sup_{t\in[a_n,b_n]}\left|\frac{1-e^{-ct}}{\lambda_n^U(t)}-1\right| \leq e^{-\frac{1}{\ln(n)}}\sup_{t\in[a_n,b_n]}\frac{(e-1)(-y_n)}{(1+o(1))(1-e^{-1})(ct-y_n)}+o(1)$$

$$= e^{-\frac{1}{\ln(n)}}\frac{(e-1)(-y_n)}{(1+o(1))(1-e^{-1})(c\cdot a_n-y_n)}+o(1)$$

$$= e^{-\frac{1}{\ln(n)}}\frac{(e-1)\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n}}{(1+o(1))(1-e^{-1})(c\cdot a_n+\frac{a^2+(k-1)b^2}{k}\frac{b_n}{n})}+o(1)$$

$$= e^{-\frac{1}{\ln(n)}}\frac{(e-1)\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n}\cdot\frac{n}{\ln(n)}}{(1+o(1))(1-e^{-1})(\frac{1}{\ln(n)}+\frac{a^2+(k-1)b^2}{k}\frac{\ln(n)}{n})\cdot\frac{n}{\ln(n)}}+o(1)$$

$$= e^{-\frac{1}{\ln(n)}}\frac{(e-1)\frac{a^2+(k-1)b^2}{k}}{(1+o(1))(1-e^{-1})(\frac{n}{\ln^2(n)}+\frac{a^2+(k-1)b^2}{k})}+o(1)\xrightarrow[n\to\infty]{}0,$$

by the same arguments as in Claim 5.16 and using the fact that $a,b>0$. □

Hence, we proved that $\sup_{t\in[a_n,b_n]}\left|U_n(t)\right|\xrightarrow[n\to\infty]{a.s.}0$. In conclusion, we therefore showed that both upper bounds in (23) converge to zero and therefore we also showed the convergence in (18)

$$\sup_{t\in[a_n,b_n]}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k}-(1-e^{-\frac{a+(k-1)b}{k}t})\right|\xrightarrow[n\to\infty]{a.s.}0.$$

ii) Now, we need to show the same convergence, but for $t\in[b_n,\infty)$ and $b_n=\ln(n)$. For this let us first remark that for $nt\geq 2$ it holds $\frac{nt}{2}\leq\lfloor nt\rfloor$, the proof for this can be found in Lemma

7.5 in the appendix. Using this and the known inequality $(1-x) \leq e^{-x}$, we can approximate

$$\left(1-\frac{a}{n}\right)^{\lfloor nt \rfloor} \leq (e^{-\frac{a}{n}})^{\lfloor nt \rfloor} = e^{-\frac{a\lfloor nt \rfloor}{n}} \leq e^{-\frac{a}{n}\frac{nt}{2}} = e^{-\frac{a}{2}t} \quad \text{and}$$

$$\left(1-\frac{b}{n}\right)^{\lfloor nt \rfloor} \leq e^{-\frac{b}{2}t}.$$

Thus, we can bound $A^{M(t)}$ from above by

$$A^{M(t)} = \left(1-\frac{a}{n}\right)^{M(t)} \left(1-\frac{b}{n}\right)^{(k-1)M(t)} = \left(1-\frac{a}{n}\right)^{\frac{\lfloor nt \rfloor}{k}} \left(1-\frac{b}{n}\right)^{\frac{k-1}{k}\lfloor nt \rfloor} \leq (e^{-\frac{a}{2}t})^{\frac{1}{k}} (e^{-\frac{b}{2}t})^{\frac{k-1}{k}}$$

$$= e^{-\frac{a+(k-1)b}{2k}t} = e^{-\frac{c}{2}t}.$$

As e^{-x} is decreasing, we have that $e^{-\frac{c}{2}t} \leq e^{-\frac{c}{2}b_n}$ for all $t \in [b_n, \infty)$. Hence, $1-e^{-\frac{c}{2}t} \geq 1-e^{-\frac{c}{2}b_n}$ and using again the monotonicity of the Poisson process in Corollary 5.10, we get

$$\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-A^{M(t)})\right)}{n/k} - (1-e^{-ct}) \geq \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})\right)}{n/k} - (1-e^{-ct}). \quad (25)$$

On the other hand, we have that $\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-A^{M(t)})\right) \leq \mathcal{P}\left(\frac{n}{k}(1+o(1))\right)$ as $A^{M(t)} \geq 0$, since we assume that $\frac{a}{n}, \frac{b}{n}$ are probabilities. Therefore, we get

$$\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-A^{M(t)})\right)}{n/k} - (1-e^{-ct}) \leq \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\right)}{n/k} - (1-e^{-ct}). \quad (26)$$

Thus, let us show that the supremum over the interval $[b_n, \infty)$ of the two bounds in (25) and (26) converges to zero in a similar fashion as in case i). Let us start with the expression in (26). We note that

$$\sup_{t\in[b_n,\infty)} \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\right)}{n/k} - (1-e^{-ct}) \right| \leq \sup_{t\in[b_n,\infty)} \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\right)}{n/k} - 1 \right| + \sup_{t\in[b_n,\infty)} \left| e^{-ct} \right|$$

$$= (1+o(1)) \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\right)}{\frac{n}{k}(1+o(1))} - \frac{1}{1+o(1)} \right| + e^{-c\cdot b_n}$$

$$\leq (1+o(1)) \left(\left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\right)}{\frac{n}{k}(1+o(1))} - 1 \right| + \left| \frac{1}{1+o(1)} - 1 \right| \right) + e^{-c\ln(n)} \xrightarrow[n\to\infty]{a.s.} 0$$

because the Poisson process converges a.s. to 1 since the argument goes to infinity and furthermore it also holds $e^{-c\ln(n)} \underset{n\to\infty}{\longrightarrow} 0$.

On the other hand, we have for the expression in (25) that

$$\sup_{t\in[b_n,\infty)}\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})\right)}{n/k}-(1-e^{-ct})\right|$$

$$\leq \sup_{t\in[b_n,\infty)}\left|(1+o(1))(1-e^{-\frac{c}{2}b_n})\right|\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})\right)}{\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})}-\frac{1}{(1+o(1))(1-e^{-\frac{c}{2}b_n})}\right|$$

$$+\sup_{t\in[b_n,\infty)}\left|e^{-ct}\right|$$

$$=\left|(1+o(1))(1-e^{-\frac{c}{2}b_n})\right|\left(\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})\right)}{\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})}-1\right|+\left|\frac{1}{(1+o(1))(1-e^{-\frac{c}{2}b_n})}-1\right|\right)$$

$$+e^{-c\ln(n)}.$$

Note that as $e^{-\frac{c}{2}\ln(n)}\underset{n\to\infty}{\longrightarrow}0$, it follows that

$$\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})\right)}{\frac{n}{k}(1+o(1))(1-e^{-\frac{c}{2}b_n})}-1\right|\underset{n\to\infty}{\overset{a.s.}{\longrightarrow}}0 \quad \text{as well as}$$

$$\left|\frac{1}{(1+o(1))(1-e^{-\frac{c}{2}b_n})}-1\right|\underset{n\to\infty}{\overset{a.s.}{\longrightarrow}}0.$$

Thus, we have that both bounds (25) and (26) converge in the supremum to 0, which, by the same argumentation as in case i), gives us

$$\sup_{t\in[b_n,\infty)}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k}-(1-e^{-\frac{a+(k-1)b}{k}t})\right|\underset{n\to\infty}{\overset{a.s.}{\longrightarrow}}0.$$

iii) Lastly, let us look at the case $t\in[0,a_n]$ with $a_n=\frac{1}{c\ln(n)}$. We have that,

$$\sup_{t\in[0,a_n]}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k}-(1-e^{-\frac{a+(k-1)b}{k}t})\right|$$

$$\leq \sup_{t\in[0,a_n]}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k}\right|+\sup_{t\in[0,a_n]}\left|1-e^{-\frac{a+(k-1)b}{k}t}\right|.$$

Therefore, we want to show the convergence to zero of the two summands. Remember the bound in expression (19) which gives us

$$A^{M(t)} \geq e^{-ct} e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{t}{n}} \geq e^{-ct} e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}}.$$

Note that $e^{-ct} \geq e^{-c\cdot a_n}$ for all $t \in [0, a_n]$ and thus $1-e^{-ct} \leq 1-e^{-c\cdot a_n}$. Using the monotonicity of the Poisson counting process (Corollary 5.10), we get

$$\sup_{[0,a_n]} \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-A^{M(t)})\right)}{\frac{n}{k}} \right| \leq \sup_{[0,a_n]} \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-ct}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}})\right)}{\frac{n}{k}} \right|$$

$$= \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-c\cdot a_n}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}})\right)}{\frac{n}{k}} \right|$$

$$= \left| (1+o(1))(1-e^{-c\cdot a_n}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}}) \right| \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-c\cdot a_n}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}})\right)}{\frac{n}{k}(1+o(1))(1-e^{-c\cdot a_n}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{a_n}{n}})} \right|$$

$$= \left| (1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}}) \right| \left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}})\right)}{\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}})} \right|.$$

Now, by the same argument as in Claim 5.15 and using L'Hôpital's rule, we get that

$$\left| \frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}})\right)}{\frac{n}{k}(1+o(1))(1-e^{-\frac{1}{\ln(n)}-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}})} \right| \xrightarrow[n\to\infty]{a.s.} 1.$$

On the other hand, we see

$$\left| (1+o(1))(1-e^{-\frac{1}{\ln(n)}}e^{-\frac{7}{6}\frac{a^2+(k-1)b^2}{k}\frac{1}{cn\ln(n)}}) \right| \xrightarrow[n\to\infty]{} 0.$$

Combining the two, we have

$$\sup_{[0,a_n]}\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))(1-A^{M(t)})\right)}{\frac{n}{k}}\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

Lastly, we note that by Lemma 5.14

$$\sup_{t\in[0,a_n]}\left|1-e^{-ct}\right| \leq \sup_{t\in[0,a_n]}(e-1)|-ct| = \sup_{t\in[0,a_n]}(e-1)c\cdot t = (e-1)\cdot c\cdot a_n$$

$$= (e-1)\frac{1}{\ln(n)} \xrightarrow[n\to\infty]{} 0.$$

Therefore, we conclude

$$\sup_{t\in[0,a_n]}\left|\frac{\mathcal{P}\left(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)})\right)}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t})\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

As the statement holds in all the three cases, we have shown that

$$\sup_{t\in\mathbb{R}_+}\left|\frac{\mathbb{S}^{(i)}_{\lfloor nt\rfloor}}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t})\right| = \sup_{t\in\mathbb{R}_+}\left|\frac{\mathcal{P}(\frac{n}{k}(1+o(1))\cdot(1-A^{M(t)}))}{n/k} - (1-e^{-\frac{a+(k-1)b}{k}t})\right| \xrightarrow[n\to\infty]{a.s.} 0,$$

which concludes the proof.

□

Thus, Proposition 5.12 states that the asymptotic behaviour of the Lukasiewicz walk divided by the expected number of vertices in the communities is similar to the function $t \to 1-e^{-\frac{a+(k-1)b}{k}t}-t$. In Figure 10, we see graphs of the function $t\to 1-e^{-ct}-t$ for different values of $c=\frac{a+(k-1)b}{k}$. We observe that if $c>1$, there exists a second non-zero root and if $c<1$, then the function does not intersect the x-axis multiple times.

This observation leads to the following claim, where we say $X_n = o_{\mathbb{P}}(n)$ if $X_n/n \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

Claim 5.19. *Let $N_i \sim \mathrm{Poi}(n)$ for each $1\leq i\leq k$. If $c=\frac{a+(k-1)b}{k}>1$, then there exists a unique giant component of size $\beta(c)n+o_{\mathbb{P}}(n)$ in the core $\mathrm{SSBM}(\sum_{i=1}^k N_i, k, \frac{a}{n}, \frac{b}{n})$, where $\beta(c)$ is the first positive root of $t\to 1-e^{-\frac{a+(k-1)b}{k}t}-t$, and all other components are of size $o_{\mathbb{P}}(n)$. While if $\frac{a+(k-1)b}{k}<1$, then asymptotically all components are of size $o_{\mathbb{P}}(n)$.*

We do not prove Claim 5.19, but give an idea on how it could be done following the proof of Corollary 3 in (Curien, 2024b).

Figure 10: Plot of the graph of the function $t \to 1 - e^{-ct} - t$ for different values of c.

Idea of proof. The goal is to prove this using the sandwiching of Section 5.3 and show similar statements for the graph $\mathrm{F}_{\mathrm{Poi}}$ and $\mathrm{F}'_{\mathrm{Poi}}$. If we look again at Figure 10, we see that if $c > 1$, we have a second root, which is non-zero. Let us call this $\beta(c)$. If $c \leq 1$ such a root does not exist, therefore we set $\beta(c) = 0$. We argued in the paragraph before Definition 5.8 that every time our Lukasiewicz walk hits a new minimal record in one of the coordinates, afterwards we explore a new connected component. The sizes of the connected components in $\mathrm{F}_{\mathrm{Poi}}$ are therefore given by the length of the excursions of $\mathbb{S}$ above the running infimum process in each coordinate, $\hat{\mathbb{S}}_l^{(i)} = \inf\{\mathbb{S}_j^{(i)} : 0 \leq j \leq l\}$, for $1 \leq i \leq k$. We denote by $(E_m^{(i)} : m \geq 1)$ those excursions ranked in decreasing order. Notice that the excursion lengths above the running infimum of the function $t \to 1 - e^{-ct} - t$ are given by $(\beta(c), 0, 0, ...)$. Using the Fluid Limit in Proposition 5.12 one can show that

$$\left(\frac{E_m^{(i)}}{n/k} : m \geq 1 \right) \xrightarrow[n\to\infty]{a.s.} (\beta(c), 0, 0, ...)$$

for the uniform norm over t, in each coordinate. In the case that $c > 1$, it holds that $\beta(c) > 0$ and therefore we get that we have one component growing in size proportional to n in each coordinate. Nevertheless, one needs to check that the components being in each coordinate big implies they are overall big. In addition, the removal of the stack vertices may split the giant component of $\mathrm{F}_{\mathrm{Poi}}(\underline{n}, k, \frac{a}{n}, \frac{b}{n})$ into several smaller components, therefore there need to be further verifications that the component is still giant in the core. For this, one could check that the vertex in the stack has, in probability, a bounded number of neighbours as $n \to \infty$, such that there are only a bounded number of connected components in the core opened. The case, if $c < 1$ then there does not exist a giant component, is left unproven. □

Lastly, to deduce a theorem similar to Theorem 3.22, we need to go from the Poissonized SSBM

to the normal one. This can be done by defining $N_- \sim \text{Poi}(\frac{n-n^{7/12}}{k})$ and $N_+ \sim \text{Poi}(\frac{n+n^{7/12}}{k})$. Then, we have Lemma 5.20.

Lemma 5.20. *It holds that*

i) $\mathbb{P}\left[\text{Poi}\left(\frac{n-n^{7/12}}{k}\right) \leq \frac{n}{k}\right] \underset{n\to\infty}{\longrightarrow} 1,$

ii) $\mathbb{P}\left[\text{Poi}\left(\frac{n+n^{7/12}}{k}\right) \geq \frac{n}{k}\right] \underset{n\to\infty}{\longrightarrow} 1.$

Proof. i) We use Chebyshev's inequality which tells us that if $\mathbb{E}[X^2] < \infty$ and $s > 0$, then

$$\mathbb{P}[\big|X - \mathbb{E}[X]\big| \geq s] \leq \frac{\text{Var}(X)}{s^2}.$$

Notice that $\text{Var}(\text{Poi}(\lambda)) = \lambda$ and $\mathbb{E}[\text{Poi}(\lambda)] = \lambda$, thus it follows, using the fact that $(\text{Poi}(\lambda) - \lambda \geq s) \subset (\big|\text{Poi}(\lambda) - \lambda\big| \geq s)$, that

$$\mathbb{P}[\text{Poi}(\lambda) - \lambda \geq s] \leq \frac{\lambda}{s^2} \iff \mathbb{P}[\text{Poi}(\lambda) \geq \lambda + s] \leq \frac{\lambda}{s^2}.$$

Now taking $\lambda = \frac{n-n^{7/12}}{k}$ and $s = \frac{n^{7/12}}{k} + 1$, we get

$$\mathbb{P}\left[\text{Poi}\left(\frac{n-n^{7/12}}{k}\right) \geq \frac{n-n^{7/12}}{k} + \frac{n^{7/12}}{k} + 1\right] = \mathbb{P}\left[\text{Poi}\left(\frac{n-n^{7/12}}{k}\right) \geq \frac{n}{k} + 1\right] \leq \frac{\frac{n-n^{7/12}}{k}}{(\frac{n^{7/12}}{k} + 1)^2}$$

$$= \frac{\frac{n-n^{7/12}}{k}}{\frac{n^{7/6}}{k^2} + 2\frac{n^{7/12}}{k} + 1} = \frac{n - n^{7/12}}{\frac{n^{7/6}}{k} + 2n^{7/12} + k} \leq \frac{n}{\frac{n^{7/6}}{k}} = k \cdot n^{1-\frac{7}{6}} = k \cdot n^{-\frac{1}{6}} \underset{n\to\infty}{\longrightarrow} 0.$$

Hence, we conclude that $\mathbb{P}\left[\text{Poi}\left(\frac{n-n^{7/12}}{k}\right) \leq \frac{n}{k}\right] \underset{n\to\infty}{\longrightarrow} 1.$

ii) For this convergence, we use the bound $\mathbb{P}[X \leq \lambda - x] \leq e^{-\frac{x^2}{2\lambda}}$ for $X \sim \text{Poi}(\lambda)$, which can be found in (Cannone, 2019, Fact 6). It holds that

$$\mathbb{P}\left[\text{Poi}\left(\frac{n+n^{7/12}}{k}\right) \geq \frac{n}{k}\right] = 1 - \mathbb{P}\left[\text{Poi}\left(\frac{n+n^{7/12}}{k}\right) \leq \frac{n}{k} - 1\right].$$

Taking $\lambda = \frac{n+n^{7/12}}{k}$ and $x = \frac{n^{7/12}}{k} + 1$, we get

$$\mathbb{P}\left[\text{Poi}\left(\frac{n+n^{7/12}}{k}\right) \leq \frac{n}{k} - 1\right] \leq e^{-\frac{(\frac{n^{7/12}}{k}+1)^2}{2(\frac{n+n^{7/12}}{k})}}.$$

Furthermore, looking at the exponent, we get

$$\frac{(\frac{n^{7/12}}{k}+1)^2}{2(\frac{n+n^{7/12}}{k})} = \frac{\frac{n^{7/6}}{k^2}+2\frac{n^{7/12}}{k}+1}{2\frac{n}{k}+2\frac{n^{7/12}}{k}} = \frac{n^{7/6-1}+2kn^{7/12-1}+k^2n^{-1}}{2k+2kn^{7/12-1}} = \frac{n^{1/6}+2kn^{-5/12}+k^2n^{-1}}{2k+2kn^{-5/12}} \underset{n\to\infty}{\longrightarrow} \infty.$$

Hence,

$$\mathbb{P}\left[\text{Poi}\left(\frac{n+n^{7/12}}{k}\right) \leq \frac{n}{k}-1\right] \leq e^{-\frac{(\frac{n^{7/12}}{k}+1)^2}{2(\frac{n+n^{7/12}}{k})}} \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore, we conclude that

$$\mathbb{P}\left[\text{Poi}(\frac{n+n^{7/12}}{k}) \geq \frac{n}{k}\right] \underset{n\to\infty}{\longrightarrow} 1.$$

□

Using these two facts, we can build the following inclusions with the probabilities $q_{in} = a_n$, $q_{out} = b_n$, where we assume that $a_n \cdot n \underset{n\to\infty}{\longrightarrow} a$ and $b_n \cdot n \underset{n\to\infty}{\longrightarrow} b$

$$\text{SSBM}(\underline{N}_-, k, a_n, b_n) \subset \text{SSBM}(n, k, a_n, b_n) \subset \text{SSBM}(\underline{N}_+, k, a_n, b_n)$$

which hold with high probability if $n \to \infty$ since $N_- \cdot a_n \underset{n\to\infty}{\longrightarrow} a$ as $\frac{N_-}{n} \underset{n\to\infty}{\longrightarrow} 1$. We have that $\text{SSBM}(\underline{N}_-, k, a_n, b_n)$ is a Poissonized SSBM with parameter $\alpha_i = N_-$ and similarly $\text{SSBM}(\underline{N}_+, k, a_n, b_n)$. Hence, if we assume that $\frac{a+(k-1)b}{k} > 1$, then both of these random graphs have a unique giant component of size approximately $\beta(c)n$ and all other components having negligible size as $n \to \infty$. Therefore, the same has to hold for the graph that is sandwiched in the middle between the two Poissonized graphs. To conclude, we thus have Proposition 5.21.

Proposition 5.21. *(Giant component in SSBM) If* $\frac{a+(k-1)b}{k} > 1$, *then* $\text{SSBM}(n, k, a/n, b/n)$ *has with high probability a giant component that grows asymptotically in size proportional to* n, *while if* $\frac{a+(k-1)b}{k} < 1$, *then* $\text{SSBM}(n, k, a/n, b/n)$ *only has components of sizes that vanish if* $n \to \infty$. *We are referring to Definition 3.16 for the term "with high probability".*

6 Conclusion and Open Problems

In this Master's thesis, we developed first the theory of random graphs and explained a few fundamental properties, such as the connectivity and the existence of a giant component. Furthermore, we introduced the Stochastic Block Model as a variant of random graphs using communities. This model is characterised to have different connectivity probabilities, determined by the different communities the vertices are in. In this thesis, we only worked with the Symmetric Stochastic Block Model (SSBM), which assumes that there are only two different probabilities, i.e. we connect vertices in the same community with probability q_{in} and vertices that are in different communities with q_{out}. We saw that the Stochastic Block Model collapses back to the Erős-Rényi Random Graph if we assume the probability to be the same in and across communities. If we assume that the probability to have edges inside the community is $\frac{a}{n}$ and the probability to have edges across communities is $\frac{b}{n}$, then we saw that the expected average degree in the SSBM$(n, k, \frac{a}{n}, \frac{b}{n})$ is $\frac{a+(k-1)b}{k}$ and in the SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ the expected average degree is $\frac{a+(k-1)b}{k}\ln(n)$. Furthermore, we analysed the criterion for connectedness and concluded that if $\frac{a+(k-1)b}{k} > 1$ then the SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ is connected whereas if $\frac{a+(k-1)b}{k} < 1$ then the SSBM$(n, k, a\frac{\ln(n)}{n}, b\frac{\ln(n)}{n})$ is disconnected.

In the main part of the thesis, we used the technique of Nicolas Curien in (Curien, 2024a) to poissonize the SSBM. Consequently, we assumed the number of vertices in each community to be distributed according to a Poi(α_i) random variable and to have a stack with infinitely many vertices linked to each community. We stated a Markov Property for the poissonized SSBM, which helped us to see the distribution of the neighbouring vertices. We figured out that the distribution of the adjacent vertices after an arbitrary step $mk + u$, with $1 \leq u \leq k$ and $m \geq 0$, is a Poisson distribution with parameter

$$\underline{\alpha}^{(mk+u)} := (1-q_{in})^m (1-q_{out})^{m(k-1)+u-2} \begin{pmatrix} \alpha_1 q_{out}(1-q_{in}) \\ \dots \\ \alpha_{u-1} q_{out}(1-q_{in}) \\ \alpha_u q_{in}(1-q_{out}) \\ \alpha_{u+1} q_{out}(1-q_{out}) \\ \dots \\ \alpha_k q_{out}(1-q_{out}) \end{pmatrix}.$$

Then, we defined the concept of the Lukasiewicz exploration to analyse the behaviour of the different components of the poissonized SSBM. We chose one vertex in the stack and determined the number of neighbours of this vertex in each community. Then, we removed the vertex from the stack and put the adjacent vertices into the according stacks. We saw that we can encode the Lukasiewicz exploration into the Lukasiewicz walk, which is a stochastic process denoted by

$(\underline{\mathbb{S}}_i : i \geq 0)$. With this walk, we could observe the size of the components, as a new infimimum in one of the coordinates of $\underline{\mathbb{S}}_i$ results in the building of a new connected component. We saw that the Lukasiewicz walk divided by the number of vertices in each community tends asymptotically to the function $t \to 1 - e^{-\frac{a+(k-1)b}{k}t} - t$ in each coordinate, i.e.

$$\left(\frac{\underline{\mathbb{S}}_{\lfloor nt \rfloor}}{n/k}\right)_{t\geq 0} \xrightarrow[n\to\infty]{a.s.} \begin{pmatrix} 1 - e^{-\frac{a+(k-1)b}{k}t} - t \\ \dots \\ 1 - e^{-\frac{a+(k-1)b}{k}t} - t \end{pmatrix}_{t\geq 0}$$

in the uniform norm over t. Lastly, we argued why this Fluid Limit helps to find a giant component in the case $\frac{a+(k-1)b}{k} > 1$.

In summary, this Master's thesis is a first approach to poissonize the Symmetric Stochastic Block Model and is by far not complete. There still remain a few open problems, for example one could prove the existence of a giant component rigorously, using the theory developed in this thesis.

Claim. *Let $N_i \sim \text{Poi}(n)$ be independent random variables, for each $1 \leq i \leq k$. If $c = \frac{a+(k-1)b}{k} > 1$, then there exists a unique giant component of size $\beta(c)n + o_{\mathbb{P}}(n)$ in the core $\text{SSBM}(\sum_{i=1}^k N_i, k, \frac{a}{n}, \frac{b}{n})$, where $\beta(c)$ is the first positive root of $t \to 1 - e^{-\frac{a+(k-1)b}{k}t} - t$, and all other components are of size $o_{\mathbb{P}}(n)$. While if $\frac{a+(k-1)b}{k} < 1$, then asymptotically all components are of size $o_{\mathbb{P}}(n)$.*

Another open problem is to analyse the critical case for the giant component, i.e. if $\frac{a+(k-1)b}{k} = 1$. One could hope to find a similar statement as in the Erdős-Rényi Random Graph (Remark 3.24).

Conjecture. *In the critical case that $\frac{a+(k-1)b}{k} = 1$, we have that the size of the biggest component is with high probability proportional to $n^{2/3}$.*

In this thesis, the simplification that the in-community probability is $q_{in} = \frac{a}{n}$ and the across-community probability is $q_{out} = \frac{b}{n}$ was used. The technique should also work with a more general setting on the communities. Let $q_{in} = a_n$ and $q_{out} = b_n$ where we assume that asymptotically $a_n \cdot n \underset{n\to\infty}{\longrightarrow} a$ and $b_n \cdot n \underset{n\to\infty}{\longrightarrow} b$. Adapting the formulas to this setting could also be a continuation of this thesis.

Furthermore, the technique of Poissonization can be used to give a different proof for the connectivity threshold, for this see also (Curien, 2024a), in which the proof for the Erdős-Rényi model can be found and which should be adaptable in the style of this thesis. One can hope that the Poissonization gives a rather short proof for the connectedness in the SSBM case.

For further reading on the topic of the Stochastic Block Model and also applications using the theory of connectedness and giant component in the SSBM, I can recommend (Abbe, 2018). In this paper, applications, such as algorithms to recover the initial labelling of the vertices in the communities if one only observes a given graph, are presented.

7 Appendix

Theorem 7.1 (Strong Law of Large Numbers)**.** *Let $(\xi_i)_{i\geq 1}$ be an i.i.d. family of random variables such that $\xi_i \in L^1$, $\mathbb{E}[\xi_i] = \mu$ and $X = \sum_{i=1}^{n} \xi_i$. Then, it holds that*

$$\frac{X}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

Theorem 7.2 (Stirling's formula)**.** *We have that for $j \in \mathbb{N}^*$*

$$j! \sim \sqrt{2\pi j}\left(\frac{j}{e}\right)^j$$

and

$$j! > \sqrt{2\pi j}\left(\frac{j}{e}\right)^j e^{\frac{1}{12j}-\frac{1}{360j^3}}.$$

Corollary 7.3. *It follows that*

$$\binom{n}{j} \leq \frac{n^j}{\left(\frac{j}{e}\right)^j}$$

for $j \leq n \in \mathbb{N}^*$.

Proof. Using the bound presented in Theorem 7.2 and the facts that for $j \geq 1$ it holds that

$$12 \leq 360j^2 \iff 12j \leq 360j^3 \iff \frac{1}{12j} \geq \frac{1}{360j^3} \iff \frac{1}{12j} - \frac{1}{360j^3} \geq 0 \iff e^{\frac{1}{12j}-\frac{1}{360j^3}} \geq 1$$

and

$$j \geq \frac{1}{2\pi} \iff \sqrt{2\pi j} \geq 1$$

we get

$$j! \geq \sqrt{2\pi j}\left(\frac{j}{e}\right)^j e^{\frac{1}{12j}-\frac{1}{360j^3}} \geq \left(\frac{j}{e}\right)^j.$$

Thus,

$$\binom{n}{j} = \frac{n!}{j!(n-j)!} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-j+1)}{j!} \leq \frac{n^j}{\left(\frac{j}{e}\right)^j}.$$

□

Corollary 7.4. *For $j \in \mathbb{N}^*$ it holds*

$$\left(\frac{e}{j}\right)^j \leq \frac{e^2}{j}.$$

Proof. Note that for $j = 1$ the inequality is satisfied. Hence, assume $j \geq 2$. Then, we have that

$$\left(\frac{e}{j}\right)^j \leq \frac{e^2}{j} \iff e^{j-2} \leq j^{j-1} \iff \frac{j-2}{j-1} \leq \ln(j).$$

Let us show that the last inequality is true. Define $f(j) := \ln(j) - \frac{j-2}{j-1}$, then we have

$$f'(j) = \frac{1}{j} - \frac{1 \cdot (j-1) - 1 \cdot (j-2)}{(j-1)^2} = \frac{1}{j} - \frac{1}{(j-1)^2}.$$

Now, it holds that for $j = 2$, $f(2) = \ln(2) - \frac{2-2}{2-1} = \ln(2) > 0$ and for $j \geq 3$ we have that $j^2 - 3j + 1 > 0$ and thus $j^2 - 2j + 1 > j$, which gives

$$\frac{1}{j} > \frac{1}{j^2 - 2j + 1} \iff \frac{1}{j} > \frac{1}{(j-1)^2}.$$

Thus, we have that $f'(j) > 0$ for all $j \geq 3$ and hence f is increasing for $j \geq 3$. We conclude that $f(j) \geq f(3) > 0$ for all $j \geq 3$ and therefore $\frac{j-2}{j-1} \leq \ln(j)$ for all $j \geq 2$. □

Lemma 7.5. *It holds that for $t > 0$ and $n \in \mathbb{N}$ such that $nt \geq 2$*

$$\frac{nt}{2} \leq \lfloor nt \rfloor \leq nt.$$

Proof. The right-hand side follows by the definition of the floor function. For the left-hand side, let us remark that

$$nt \geq 2 \iff 2nt \geq 2 + nt \iff 2nt - 2 \geq nt \iff nt - 1 \geq \frac{nt}{2}.$$

To conclude, we note that $\lfloor nt \rfloor = nt - \{nt\} \geq nt - 1$. □

References

Abbe, Emmanuel (Mar. 2018). 'Community Detection and Stochastic Block Models: Recent Developments'. In: *Journal of Machine Learning Research* 18, pp. 1–86. URL: https://jmlr.org/papers/v18/16-480.html (visited on 09/12/2024).

Acemoglu, Daron and Asu Ozdaglar (2009). *Networks Lecture 4: Erdös-Renyi Graphs and Phase Transitions*. URL: https://economics.mit.edu/sites/default/files/inline-files/Lecture%204%20-%20Erdos-Renyi%20Graphs%20and%20Phase%20Transitions.pdf (visited on 31/08/2024).

Cannone, Clément (2019). *A short note on Poisson tail bounds*. URL: http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf (visited on 21/11/2024).

Coculescu, Delia (2023). *Lecture Notes on Mathematical Finance II*. Universität Zürich.

Curien, Nicolas (2024a). *A random Walk among random Graphs*. URL: https://www.imo.universite-paris-saclay.fr/~nicolas.curien/enseignement.html (visited on 29/11/2024).

— (2024b). 'Erdős-Rényi Poissonized'. In: *Comptes Rendus. Mathématique* 362, pp. 649–656. URL: https://comptes-rendus.academie-sciences.fr/mathematique/articles/10.5802/crmath.578/ (visited on 06/12/2024).

Diestel, Reinhard (2017). *Graph Theory*. Fifth Edition. Springer-Verlag. DOI: 10.1007/978-3-662-53622-3.

Durrett, Rick (2007). *Random Graph Dynamics*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511546594.

Frieze, Alan and Michal Karonski (2024). *Introduction to Random Graphs*. Cambridge University Press. URL: https://www.math.cmu.edu/~af1p/BOOK.pdf (visited on 23/08/2024).

Newman, M. E. J., D. J. Watts and S. H. Strogatz (2002). 'Random graph models of social networks'. In: *PNAS* 99, pp. 2566–2572. URL: https://www.pnas.org/doi/10.1073/pnas.012582999.

Newman, Mark (2018). *Random graphs*. Vol. 1. Oxford University Press. DOI: 10.1093/oso/9780198805090.003.0011.

Roch, Sébastien (2024). *Modern Discrete Probability: An Essential Toolkit*. Cambridge University Press. DOI: 10.1017/9781009305129.