Department of Mathematics and Systems Analysis

Structure and estimation of network models with overlapping communities

Joona Karjalainen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 175/2021

Structure and estimation of network models with overlapping communities

Joona Karjalainen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall M1 of the school on 17 December 2021 at 12.

Aalto University
School of Science
Department of Mathematics and Systems Analysis

Supervising professor
Associate Professor Lasse Leskelä, Aalto University School of Science, Finland

Preliminary examiners
Assistant Professor Miklós Rácz, Princeton University, USA
Associate Professor François Caron, University of Oxford, United Kingdom

Opponent
Professor Remco van der Hofstad, Eindhoven University of Technology, The Netherlands

Aalto University publication series
DOCTORAL DISSERTATIONS 175/2021

ISBN 978-952-64-0626-8 (printed)
ISBN 978-952-64-0627-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0627-5

Unigrafia Oy
Helsinki 2021

Finland

Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Joona Karjalainen

Name of the doctoral dissertation
Structure and estimation of network models with overlapping communities

Publisher School of Science

Unit Department of Mathematics and Systems Analysis

Series Aalto University publication series DOCTORAL DISSERTATIONS 175/2021

Field of research Mathematics

Manuscript submitted 7 September 2021 **Date of the defence** 17 December 2021

Permission for public defence granted (date) 11 November 2021 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Many types of data in different fields of science can be naturally represented as networks. Social relationships in groups of people, the structure of the internet, and traffic networks can all be understood as collections of nodes and connections between them. Real-world networks often show signs of community structure, i.e., some groups of nodes are more densely connected to each other than to the rest of the nodes. Since communities may emerge through many different mechanisms, it is natural to describe these networks with statistical models where the communities are allowed to overlap. Even in the absence of obvious communities, various other types of structure are commonly observed in data. For example, the degrees of adjacent nodes tend to be correlated, and node pairs have an increased probability of being adjacent if they have common neighbors.

This dissertation is concerned with the structure of large and sparse statistical network models with overlapping communities. This structure is described using statistical quantities and distributions and their limits as the number of nodes tends to infinity. The focus is on the asymptotic behavior of subgraph frequencies, joint degree distributions of adjacent nodes, and various summary statistics. New results are proved on their convergence, and exact formulas are provided for their limits. These results lead to new estimators of the model parameters based on counting the frequencies of small subgraphs. The consistency of these estimators is proved under complete or partly incomplete data.

The results show that the models have structural similarities with many real-world networks, such as non-trivial clustering, degree correlations, and power laws. This illustrates how some empirical observations on network data can be explained with an underlying overlapping community structure.

Keywords networks, random graphs, parameter estimation, asymptotic theory, overlapping communities

ISBN (printed) 978-952-64-0626-8 **ISBN (pdf)** 978-952-64-0627-5

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2021

Pages 175 **urn** http://urn.fi/URN:ISBN:978-952-64-0627-5

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Joona Karjalainen

Väitöskirjan nimi
Päällekkäisiä yhteisöjä sisältävien verkostomallien rakenne ja estimointi

Julkaisija Perustieteiden korkeakoulu

Yksikkö Matematiikan ja systeemianalyysin laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 175/2021

Tutkimusala Matematiikka

Käsikirjoituksen pvm 07.09.2021 **Väitöspäivä** 17.12.2021

Väittelyluvan myöntämispäivä 11.11.2021 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Monia eri tieteenaloilla esiintyviä aineistoja voidaan luontevasti esittää verkostoina. Esimerkiksi ihmisten välisiä sosiaalisia suhteita, Internetin rakennetta ja liikenneverkostoja voidaan esittää kokoelmina solmuja ja niiden välisiä kytköksiä. Tosimaailman verkostoissa havaitaan usein yhteisörakennetta, eli solmut muodostavat ryhmiä, jotka ovat sisäisesti tiiviisti kytkeytyneitä, mutta kytkökset ryhmän ulkopuolisiin solmuihin ovat vähäisiä. Koska yhteisörakenne voi syntyä useiden eri mekanismien kautta, on luontevaa että verkostoa kuvaava tilastollinen malli sallii yhteisöjen päällekkäisyyden. Vaikka aineistossa ei esiintyisi selkeitä yhteisöjä, usein voidaan havaita muunlaista rakenteellisuutta. Esimerkiksi kytkettyjen solmujen asteet korreloivat, ja solmuparit ovat todennäköisemmin kytkettyjä, jos niillä on yhteisiä naapureita.

Tässä väitöskirjassa tutkitaan suurten ja harvojen päällekkäisiä yhteisöjä sisältävien tilastollisten verkostomallien rakennetta. Tätä rakennetta kuvataan tilastollisilla suureilla ja jakaumilla sekä näiden raja-arvoilla solmujen määrän kasvaessa kohti ääretöntä. Työssä analysoidaan pienten osaverkkojen lukumäärien, kytkettyjen solmujen asteiden yhteisjakaumien ja erilaisten tilastollisten tunnuslukujen asymptoottista käyttäytymistä. Näille todistetaan suppenemistuloksia ja niiden raja-arvoille annetaan täsmällisiä kaavoja. Tulosten perusteella johdetaan pienten osaverkkojen laskentaan perustuvia estimaattoreita mallien parametreille, ja niiden tarkentuvuus todistetaan täysin ja osittain havaituille aineistoille.

Tulokset osoittavat, että malleilla on rakenteellisia yhtäläisyyksiä monien tosimaailman verkostojen kanssa, kuten epätriviaaleja klusterointiominaisuuksia, asteiden korrelaatioita ja potenssilakeja. Tämä havainnollistaa sitä, kuinka eräitä verkostoaineistoista tehtyjä havaintoja voidaan selittää piilevien ja päällekkäisten yhteisöjen rakenteen avulla.

Avainsanat verkostot, satunnaisverkot, parametrien estimointi, asymptoottinen teoria, päällekkäiset yhteisöt

ISBN (painettu) 978-952-64-0626-8 **ISBN (pdf)** 978-952-64-0627-5

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2021

Sivumäärä 175 **urn** http://urn.fi/URN:ISBN:978-952-64-0627-5

Preface

This dissertation is my contribution to the theory of random graphs and network science. The work was carried out at Aalto University in the Stochastics and Statistics research group, which I joined in 2015 to write my Master's thesis on random graphs with communities. This topic turned out to be very deep indeed, and I am grateful for having had the opportunity to continue my work as a PhD student and to pursue answers to many of the questions I had at the time.

First and foremost, I wish to thank my supervisor Prof. Lasse Leskelä for his guidance, patience, and support. His insight and advice have been invaluable during my PhD project. I also appreciate that I have been entrusted with much responsibility, which has allowed me to grow as a researcher.

I am grateful to Prof. Remco van der Hofstad for agreeing to be my opponent, and my pre-examiners Prof. François Caron and Prof. Miklós Rácz for their valuable comments on my dissertation.

I wish to thank all the people who I have had the pleasure to work with. I thank my co-authors Johan van Leeuwaarden, Mindaugas Bloznelis, and Tommi Gröhn. I have learned much about mathematics and research in general during our projects.

My work has been partially supported by Emil Aaltosen Säätiö and the Magnus Ehrnrooth Foundation, for which I am grateful for. This support has given my work stability, and allowed me to focus on my research in the long term. I extend my thanks to the organizers of the numerous summer and winter schools which I have had the chance to participate during my studies.

I wish to thank my current and former colleagues in the Stochastics and Statistics group, including Niko, Matias, Paavo, and Sami. Special thanks to Hoa for the discussions on mathematics and the ups and downs of life as a PhD student. I wish to thank Prof. Kalle Kytölä, who has always been generous with his time to discuss mathematics with me.

Finally, I wish to thank my family and friends. I am grateful to my parents Leila and Jouko, who have always been supportive of my endeavours,

and my siblings Tuomas, Miika, Saana, Tuukka, Perttu, and Iina. I thank Olli-Pekka, Juho, Patrick, Jyrki, Elisa, and Päivi. Without all of you I would not be the person I am today.

Espoo, November 21, 2021,

Joona Karjalainen

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I J. Karjalainen and L. Leskelä. Moment-based parameter estimation in binomial random intersection graph models. In *Algorithms and Models for the Web Graph (WAW 2017)*, *Lecture Notes in Computer Science, volume 10519*, Toronto, Canada, pp. 1–15, June 2017.

II J. Karjalainen, J.S.H. van Leeuwaarden, and L. Leskelä. Parameter estimators of sparse random intersection graphs with thinned communities. In *Algorithms and Models for the Web Graph (WAW 2018)*, *Lecture Notes in Computer Science, volume 10836*, Moscow, Russia, pp. 44–58, May 2018.

III T. Gröhn, J. Karjalainen, and L. Leskelä. Clique and cycle frequencies in a sparse random graph model with overlapping communities. Submitted to a journal, arXiv:1911.12827, 23 pages, April 2021.

IV M. Bloznelis, J. Karjalainen, and L. Leskelä. Assortativity and bidegree distributions on Bernoulli random graph superpositions. Accepted for publication in *Probability in the Engineering and Informational Sciences,* 31 pages, August 2021.

V J. Karjalainen. A note on parameter estimation of thinned random intersection graphs. In *22nd European Young Statisticians Meeting,* Athens, Greece, pp. 51–55, September 2021.

VI M. Bloznelis, J. Karjalainen, and L. Leskelä. Normal and stable approximation to subgraph counts in superpositions of Bernoulli random graphs. Submitted to a journal, arXiv:2107.02683, 15 pages, July 2021.

Author's Contribution

In all publications, the names of the authors are listed in alphabetical order.

Publication I: "Moment-based parameter estimation in binomial random intersection graph models"

The idea of estimating the parameters in binomial random intersection graphs came from the author's numerical experiments with various network models. The author conducted the simulations, derived one of the two estimators of the parameter μ, wrote the proofs of consistency based on combinatorial arguments by Prof. Leskelä, and contributed to the writing of the final version of the article.

Publication II: "Parameter estimators of sparse random intersection graphs with thinned communities"

The idea of estimating the parameters in this model and the main strategy behind the convergence proofs came from the author. The author derived the estimators, proved their consistency, and conducted the numerical experiments. All authors contributed to the writing of the final version of the article.

Publication III: "Clique and cycle frequencies in a sparse random graph model with overlapping communities"

The idea of studying the frequencies of general cliques and cycles came from Prof. Leskelä, and the first calculations of the expected values, including the general formula for the clique frequencies, were made by Gröhn.

The author wrote the first version of the article and gave rigorous proofs for all the results with the exception of Lemma 5.8 and parts of Lemma 5.9. The moment conditions of Theorems 4.1 and 4.2 were subsequently improved jointly by the author and Prof. Leskelä.

Publication IV: "Assortativity and bidegree distributions on Bernoulli random graph superpositions"

The idea of studying the bidegree distributions originated from Prof. Leskelä. The idea of studying power laws in the bidegree distributions originated from Prof. Bloznelis.

The author wrote the proofs of Theorem 1(i), Lemma 2, and Lemma 3 with the guidance of Prof. Leskelä and Prof. Bloznelis. The author contributed to the arguments leading to Theorem 1(ii) and Theorem 2, and wrote parts of their proofs. All authors contributed to the writing of the final version.

Publication V: "A note on parameter estimation of thinned random intersection graphs"

This article is the author's personal work.

Publication VI: "Normal and stable approximation to subgraph counts in superpositions of Bernoulli random graphs"

The idea of proving normal and stable limits for clique counts originated from Prof. Bloznelis. The author was behind the main innovation leading to the proof of Lemma 1, which extends the idea from cliques to general 2-connected subgraphs.

The author and Prof. Bloznelis found the proof strategy of approximating the subgraph counts by layer-wise counts independently. The author contributed to the arguments of the proofs of Theorem 1, Remark 1, and Lemma 5. All authors contributed to the writing of the final version.

1. Introduction

When Paul Erdős and Alfréd Rényi wrote their seminal papers on random graph theory in the early 1960s [15, 16], they were not concerned whether their models described real-world networks or not. Rather, they focused on how simple model assumptions could lead to interesting and non-trivial phenomena, which could be greatly affected by small changes in the parameters. Through advances in technology, we have now gained access to large network data sets consisting of millions or billions of nodes, including data from such diverse fields as biology, economics, and social sciences. After years of studies on empirical data, it has become clear to scientists that the first random graphs, such as those studied by Erdős and Rényi, do not generally resemble real-world networks.

The word "structure" has many interpretations in different contexts. For example, social networks tend to show signs of *community structure*, i.e., people tend to form densely connected groups. This may be due to common hobbies, occupations, geographical locations, or a number of other factors. On the other hand, technological networks often show something completely different – most of the edges are concentrated around a small number of nodes called *hubs*, which leads to star-like structures. Many of the interesting structural properties of networks can be described in terms of correlations. A form of correlation between edges is described by the *clustering phenomenon*, the tendency of nodes to form triangles where two adjacent edges are found. Another type of correlation is found in networks with *assortative mixing*, where the degrees of adjacent nodes are positively correlated, i.e., neighbors of "active" nodes are often also active themselves.

The above-mentioned phenomena have intrigued researchers for decades, and a number of models have been proposed in the literature to study and explain them, such as preferential attachment models, configuration models, and stochastic block models, to name a few. Although simulating statistical network models is usually straightforward, their rigorous mathematical analysis has turned out to be far from trivial. While classical probability theory and statistics tend to focus on large numbers of independent (or weakly dependent) and identically distributed random variables,

almost nothing in networks seems to fit this framework – the edges can be dependent in infinitely many ways, and the number of observations of a network is typically limited to one. This challenge has led to new developments in, e.g., the theory of Stein's method and computational algorithms.

The general theme of this dissertation is network models with overlapping communities. They are partly motivated by social networks, where it seems obvious that the communities could be formed by many different mechanisms, and that each person can belong to one or more of them. In light of this observation, it is perhaps surprising that models with overlapping communities are not as ubiquitous in the mathematics literature as, e.g., the aforementioned preferential attachment models and configuration models. In this context, the word "community" should be understood in a broad sense. For example, clusters of densely connected nodes in food webs or protein interaction networks can be viewed as communities, although this may not be consistent with the daily usage of the word.

The models studied in this dissertation are generally large and sparse. In this context, "large" means that we study the asymptotic behaviour of statistical quantities and distributions as the number of nodes tends to infinity. Although almost anything can happen in small networks, some events become negligibly unlikely in the limit, which can reveal important information about the structure of the model. This is also the classical approach of asymptotic theory in statistics, where asymptotic results are derived with the justification that they are approximately valid for large data sets. The study of sparse networks is partly motivated by the fact that real-world networks tend to be sparse and, on the other hand, the existence of a community structure can be difficult to verify in very dense networks. From a theoretical point of view, it is interesting that provably non-trivial structural properties, such as clustering, can be found despite the sparseness.

The results of this dissertation mainly fall into three categories: theorems are proved on (i) asymptotic behaviour of subgraph counts of, e.g., cycles and cliques, (ii) asymptotic behaviour of summary statistics, such as the clustering coefficient and degree correlation coefficients, and (iii) parameter estimation. Numerical experiments are also presented to illustrate the theory. The network models vary in their complexity, and some of them have been more widely studied in the literature than others. The parameter estimators presented in the publications are based on the method of moments, and they only require computing the frequencies of certain small subgraphs in the data, and applying a simple mathematical formula. This leads to easily implementable algorithms, light computational loads, and provable accuracy in the limit. The results illustrate how certain structural properties of real-world networks could be explained with an overlapping community structure, and show that statistical inference on

these models is indeed feasible.

This remainder of this dissertation is organized as follows. Chapter 2 provides an overview of different statistical quantities and distributions used to describe the structure of networks. The literature on this topic is very vast, and the overview is limited to those concepts which are relevant for the publications. Different statistical network models are briefly discussed in Chapter 3. We present a general framework from which the models of this dissertation can be derived, and motivate their study. Chapter 4 includes much of the mathematical theory, notation, and terminology, which are often omitted or only briefly mentioned in the literature, yet important for understanding the results and proofs of this dissertation. In Chapter 5, we give a brief introduction to the moment-based approach to parameter estimation, and discuss different assumptions on graph evolution as the number of nodes tends to infinity. The publications of this dissertation and their relationships to each other are summarized in Chapter 6, and the publications themselves are placed at the end of the dissertation.

2. Structure of networks

This chapter reviews important concepts in network science and random graph theory. The definitions given in this chapter serve as an introduction to the different ways that structure can be understood in the context of networks, and motivate the study of the models in Chapter 3.

2.1 Random and non-random graphs

The word network is typically used for something that exists in the real world, such as social networks or transportation networks, which can in theory be observed and/or from which data can be collected. Real-world networks are represented as graphs, which are here considered purely mathematical objects. However, the terms graph and network are used interchangeably in many sources, without risk of confusion.

The graphs in this dissertation are *simple, undirected, and unweighted*. Such a graph G is defined by the pair $(V(G), E(G))$, where $V(G)$ is the set of nodes (typically a set of the form $\{1, 2, \dots, n\}$), and $E(G)$ is the set of edges. An edge is a pair $\{v, w\}$, where $v, w \in V(G)$ and $v \neq w$. Two nodes v and w are *adjacent* if $\{v, w\} \in E(G)$. Sometimes it is useful to describe the graph by its *adjacency matrix*. Let $V(G) = [n] := \{1, 2, \dots, n\}$. Then the adjacency matrix A is defined as

$$A_{ij} \;=\; \begin{cases} 1, & \text{if } \{i, j\} \in E(G), \\ 0, & \text{otherwise.} \end{cases}$$

The *degree* of node i, the number of its neighbors in G, is defined in terms of the adjacency matrix as

$$d_G(i) \;=\; \sum_{j=1}^{n} A_{ij}.$$

The notation $d(i)$ is often used when the graph G can be inferred from the context. If the degree of a node equals zero, we say that it is *isolated*.

A *random graph* is a random variable whose realizations are graphs. The node set is usually fixed, and only the edges are considered random. Although from the probability theoretic point of view random graphs should be considered functions, they are usually defined with an algorithm or a probability distribution

$$\mathbb{P}(G = g), \quad g \in \mathcal{G},$$

where $\mathcal{G}$ is the set of all graphs on the node set V.

2.2 Degree distributions

The degree distribution of a graph captures important information about its structure. An imbalanced distribution may suggest that a large proportion of the edges are concentrated around a small number of nodes, whereas a tightly concentrated distribution may indicate some form of homogeneity in the underlying phenomenon. Nodes with exceptionally high degrees are often called hubs. As illustrated in [13] and [41], many data sets on the structure of the Internet and the World Wide Web show imbalanced degree distributions. For applications of degree distributions to statistical inference problems, see, e.g., [14] and [53].

When G is a fixed (i.e., non-random) graph on the node set $[n]$, the degree distribution is defined as

$$f(k) \;=\; \frac{\#\{i \in [n] : \, d_G(i) = k\}}{n} \;=\; \mathbb{P}(d_G(V) = k), \quad k \in \{0, 1, 2, \ldots\},$$

where V is uniformly distributed on the set $[n]$. This is sometimes referred to as the *empirical degree distribution* to emphasize the fact that G is considered fixed. When G is a random graph, the *model degree distribution* is defined similarly,

$$\mathbb{P}(d_G(1) = k), \quad k \in \{0, 1, 2, \ldots\}.$$

This terminology of empirical and model quantities is also used in the sequel.

Power laws

Degree distributions in real-world networks often show signs of *power-law* type behaviour, i.e., the empirical degree distribution behaves approximately as

$$\mathbb{P}(d_G(V) > k) \;\approx\; ak^{-\alpha}$$

for some constants $a > 0$ and $\alpha > 0$. This type of behaviour can be studied with a log-log plot of the pairs $(k, \mathbb{P}(d_G(V) > k))$, where a power-law distribution would show an approximately straight line in the right tail. It should be noted that since the degree distribution of any finite network

is bounded, the power law can never hold in a strict sense, but in mathematical models they can be obtained as, e.g., limits of the model degree distributions. When the distribution is truly unbounded, the parameter α, sometimes called the power-law exponent, limits the number of finite moments. Many papers have been written about power-law degree distributions in real networks as well as different random graph models, see, e.g., [1, 2, 13, 41] for empirical studies and [7, 29, 63, 68, 69, 71] for theoretical results on the topic.

2.3 Subgraph counts

One of the common ways to study the structure of a network is to search for recurring patterns, i.e., commonly appearing subgraphs (also known as *motifs* or *graphlets*). Although they do not necessarily give a clear picture of the network as a whole, small and common subgraphs can be seen as a way to describe the *local structure* of the network. For example, social networks have many more triangles than one would expect from a graph where the edges are generated independently [48]. Examples of common subgraphs have been found in networks in biochemistry, neurobiology, ecology, and engineering [45, 56]. Subgraph counts have also been used for classification [66], and to measure the similarity between different data sets [28].

The subgraph counts are defined for edges, 2-stars, and triangles by

$$N_{K_2} = \#E(G) = \sum_{1 \le i < j \le n} A_{ij}, \quad N_{S_2} = \sum_i \sum_{j<k} A_{ij}A_{ik} = \sum_i \binom{d_G(i)}{2},$$

$$N_{K_3} = \sum_{i<j<k} A_{ij}A_{jk}A_{ik}.$$

In the definition of N_{S_2} we ignore the possible edge between j and k, and so every triangle contains three 2-stars. The number of *induced* 2-stars, where this link is not allowed, is defined by the formula $\sum_i \sum_{j<k} A_{ij}A_{ik}(1 - A_{jk})$. It is worth noting that although 2-stars contain three nodes, they can be counted with $\approx n^2$ operations instead of n^3. More generally, counting k-stars (graphs on $k+1$ nodes with edges $\{1,2\}, \dots, \{1, k+1\}$) is much faster than, e.g., counting cliques. Specialized algorithms, both exact and approximate, exist for counting subgraphs efficiently, e.g., [30, 35, 57, 65].

We now give a general mathematical definition for subgraph counts.

Definition 2.1 (Subgraph, isomorphic graphs)**.** *A graph R is a subgraph of G, denoted by $R \subset G$, if $V(R) \subset V(G)$ and $E(R) \subset E(G)$. Graphs R and R' are isomorphic if there exists a bijection $\phi : V(R) \to V(R')$ such that $\{\phi(i), \phi(j)\} \in E(R')$ if and only if $\{i, j\} \in E(R)$.*

Definition 2.2 (Indicator function)**.** *The indicator function of an event B,*

denoted by $\mathbb{I}(B)$, is defined as

$$\mathbb{I}(B) \;=\; \begin{cases} 1 & \textit{if } B \textit{ is true, and} \\ 0, & \textit{otherwise.} \end{cases}$$

Definition 2.3 (General subgraph counts)**.** *Let G be a graph on n nodes. The subgraph count for R is defined by*

$$N_R = \sum_{R' \in \mathcal{G}_n(R)} \mathbb{I}(G \supset R'),$$

where $\mathcal{G}_n(R)$ is the set of R-isomorphic subgraphs of the complete graph with node set $V(G)$.

Sometimes it is useful to normalize the counts to the interval between 0 and 1, e.g., when comparing graphs of different sizes. The *subgraph density* is defined analogously as

$$\frac{1}{\#\mathcal{G}_n(R)} \sum_{R' \in \mathcal{G}_n(R)} \mathbb{I}(G \supset R').$$

It is clearly desirable for a network model to be able to produce realistic (i.e., similar to observed) subgraph counts. Analysis of the asymptotic behaviour of subgraph counts in different models has been presented in [5, 61, 62], and Publications I, II, III, and VI of this dissertation.

2.4 Clustering coefficient

The *clustering coefficient* (or *transitivity coefficient*) is one of the simplest summary statistics that can be easily computed from a graph. It answers the question "what is the probability that two of my friends are also friends with each other?", and measures the tendency to form small dense subgraphs in otherwise sparse graphs. Not surprisingly, human social networks tend to have non-negligible clustering coefficients [49, 51, 75], in the sense that the clustering coefficient is much larger than the edge density $\#E(G)/\binom{n}{2}$.

The (empirical) clustering coefficient for a fixed graph G is defined by

$$t(G) \;=\; 3\frac{N_{K_3}(G)}{N_{S_2}(G)}.$$

This definition is motivated by the fact that we may express $t(G)$ as

$$\mathbb{P}(A_{IJ} = 1 \mid A_{IK} = 1,\ A_{JK} = 1),$$

where (I, J, K) are three distinct nodes chosen uniformly at random from $V(G)$.

For a random graph G we define the *model clustering coefficient* analogously as

$$\tau(G) \;=\; \mathbb{P}(A_{12} = 1 \mid A_{13} = 1,\; A_{23} = 1).$$

It should be noted that when G is random, $t(G)$ is also random, whereas $\tau(G)$ is a constant. Moreover, it is not necessarily true that these two quantities are close to each other in any stochastic sense. If they are, then $t(G)$ can be considered a reasonable estimator for $\tau(G)$. See [9] for a theoretical study of a model where $t(G) - \tau(G)$ is not only close to zero for large n, but also normally distributed (with proper scaling).

2.5 Degree-degree correlations

Many networks show signs of *assortative mixing*, i.e., the neighbors of nodes with high degrees also tend to have high degrees. This phenomenon has been observed in human social networks, whereas the opposite seems to be true in other types of data sets, such as technological and biological networks [47, 48], i.e., they are *disassortative*. It has been pointed out by several authors that many network models do not necessarily reproduce this aspect well – the degrees of adjacent nodes tend to be nearly uncorrelated [47, 69, 70].

Bidegree distributions

We start by defining the bidegree distribution, from which different correlation measures can be derived. For a fixed graph G, the *empirical bidegree distribution* is defined by

$$f^{(2)}(s,t) \;=\; \frac{1}{2\#E(G)} \sum_{(i,j)\in E_{dir}(G)} \mathbb{I}(d(i) = s, d(j) = t),$$

where $E_{dir}(G)$ is the set of directed edges,

$$E_{dir}(G) = \{(i,j) \in V(G) \times V(G) \;:\; \{i,j\} \in E(G)\}.$$

It should be noted that the graph G is still undirected, and the set $E_{dir}(G)$ is mostly defined for notational convenience. The distribution $f^{(2)}(s,t)$ is clearly symmetrical in s and t, and both marginals are equal to the size-biased degree distribution

$$f^*(s) \;=\; \frac{sf(s)}{\sum_t tf(t)},$$

where f is the empirical degree distribution.

The *model bidegree distribution* is defined for a random graph G by

$$f_2(s,t) \;=\; \mathbb{P}(d(1) = s, d(2) = t \mid (1,2) = E_{dir}(G))$$

and, analogously, the marginals are equal to

$$f_1^*(s) \;=\; \frac{s f_m(s)}{\sum_t t f_m(t)},$$

where f_m is the model degree distribution.

Assortativity

There are many ways to measure the correlation of adjacent nodes. The *empirical assortativity* is defined as the Pearson correlation coefficient

$$\mathrm{Cor}(d(I), d(J)) \;=\; \frac{\sum_{s,t} st f^{(2)}(s,t) - (\sum_s s f^*(s))^2}{\sum_s s^2 f^*(s) - (\sum_s s f^*(s))^2},$$

where (I, J) are chosen uniformly at random from the set of all adjacent node pairs. Similarly, the *model assortativity* is defined by

$$\mathrm{Cor}^*(d(I), d(J)) \;=\; \frac{\mathbb{E}^*(d(I)d(J)) - (\mathbb{E}^*(d(I))^2}{\mathbb{E}^*(d(I))^2 - (\mathbb{E}^* d(I))^2},$$

where $\mathbb{E}^*$ denotes the conditional expectation given the event $\{(I, J) \in E_{dir}(G)\}$. It is worth noting that as with the clustering coefficient, the empirical assortativity is not necessarily a good estimator of the model assortativity even for large graphs, and this kind of approximation requires a proof.

Rank-based correlation coefficients

Some authors have criticized the use of assortativity as a measure of degree correlation. In particular, it may converge to zero in some models with obvious negative dependencies, and in some models it may fail to converge to any constant [69]. In these cases, the following rank-based correlation coefficients may be better alternatives [70, 71].

Spearman's rank correlation coefficient is based on the idea of ranking the samples from largest to smallest, and comparing these ranks instead of the numerical values of the samples. With non-continuous distributions it is possible that some of the samples have equal values. The reader is referred to [4] for information on different tie-breaking methods, and to [46] for more general theory on the topic. For a joint distribution f we use the definition

$$\rho_{Spe}(f) \;=\; \mathrm{Cor}(r_1(X^{(1)}), r_2(X^{(2)})),$$

where $(X^{(1)}, X^{(2)})$ is f-distributed and $r_i = \frac{1}{2}(f^{(i)}(-\infty, x) + f^{(i)}(-\infty, x])$, with $f^{(i)}$ denoting the i-th marginal of f. Other definitions for Spearman's rank correlation exist in the literature – the above definition corresponds to the *mid-rank* tie-breaking convention.

Kendall's rank correlation coefficient is defined similarly by

$$\rho_{Ken}(f) \;=\; \mathrm{Cor}(\mathrm{sgn}(X^{(1)} - Y^{(1)}), \mathrm{sgn}(X^{(2)} - Y^{(2)})),$$

where $\mathrm{sgn}(x) = \mathbb{I}(x > 0) - \mathbb{I}(x < 0)$, and $(X^{(1)}, X^{(2)})$ and $(Y^{(1)}, Y^{(2)})$ are mutually independent and f-distributed. It is clear from this definition that, since sgn is a bounded function, this type of correlation can be defined without the requirement of finite variance, in contrast with assortativity. The definitions of both rank correlation coefficients can be used for empirical and model quantities by choosing the bidegree distribution f accordingly.

3. Network models

This chapter discusses statistical network models and their asymptotic properties. We give a brief introduction to the Erdős-Rényi model, define a general class of models with overlapping communities, and discuss its special cases. Although it is common to speak of *communities* in networks, these should be understood in an abstract sense. As pointed out in [26], not only social networks, but many different kinds of networks can be viewed as bipartite structures of nodes and communities.

3.1 The Erdős-Rényi model

One of the simplest statistical network models is the Erdős-Rényi model (***ER graph***, $G(n,p)$, or simply *random graph* in some sources), where all the edges are generated independently of each other. Although named after the mathematicians Paul Erdős and Alfréd Rényi, it was first studied by Gilbert in his article from 1959 [22].

The distribution of an ER graph $G(n,p)$ can be expressed as

$$\mathbb{P}(G = g) \;=\; p^{\#E(g)}(1-p)^{\binom{n}{2}-\#E(g)}, \quad g \in \mathcal{G},$$

where $p \in (0,1)$, and $\mathcal{G}$ is the set of all graphs on the node set $V(G)$. This model, like all models considered in this dissertation, are *exchangeable*, meaning that the distribution is invariant under permutations of the node set $V(G)$.

Due to its simplicity, the ER graph is probably the most widely studied network model. Many theoretical results have been first established for ER graphs, and then extended to more complicated models. Topics on subgraph counts, connectivity, and emergence of large connected components are well covered in the literature, e.g., the books [12, 20, 33, 68].

Consider now a sequence of independent ER graphs, $(G(n,p_n))_{n\in\mathbb{N}}$. We say that a random graph *sparse*, if the mean degree remains bounded as $n \to \infty$, which in this case means that

$$\mathbb{E}(d_G(1)) \;=\; (n-1)p_n \;\not\to \infty \quad \text{as } n \to \infty.$$

Clearly, the mean degree converges to a constant if p_n is of the order n^{-1}. Since the edges are independent, many properties of ER graphs can be studied using classical statistical theory of independent and identically distributed random variables. This is not true for most network models, such as those introduced in the next section.

In many ways, the ER graph expresses the idea of a network that has "no structure". In the sparse case, the clustering coefficient tends to zero, and the degrees of adjacent nodes are uncorrelated. Since the degrees are binomially distributed, they do not model the power-law distributions found in some networks.

3.2 Models with overlapping communities

We now define classes of network models with overlapping communities, and discuss their properties. These models consist of independent and identically distributed communities, and each community can be viewed as a small ER graph. It turns out that we can obtain tractable sparse graphs with interpretable parameters that do not suffer from the trivialities of sparse ER graphs.

Let n be the number of nodes, m the number of communities, and P the community *type* distribution, i.e., the joint distribution of the size of the community size $X \in \{1, \dots, n\}$ and the within-community edge probability $Y \in [0, 1]$. The communities C_k, $k = 1, \dots, m$, are generated independently of each other as follows:

1. Generate the number of nodes, $X_k = \#V(C_k)$, and the edge probability Y_k from the distribution P.
2. Choose the node set $V(C_k)$ uniformly at random from the subsets of $[n] = \{1, \dots, n\}$ of size X_k.
3. Generate the edges between the $\binom{X_k}{2}$ node pairs of $V(C_k)$ independently with probability Y_k.

The resulting graph G is defined as the superposition of the communities:

$$V(G) \;=\; [n], \qquad E(G) \;=\; \bigcup_{k=1}^{m} E(C_k).$$

Special cases of this model include the following.

- **(Active) random intersection graphs:**

  When $X_k \sim \mathrm{Bin}(n, p)$, $p \in (0, 1)$, and $Y_k = 1$, we obtain the classical random intersection graph introduced in [40]. In this case the events that a particular node belongs to a particular community, $\{v \in V(C_k)\}$, are independent.

This model is *active* in the sense that the nodes can be thought to choose their communities (according to the binomial distribution), instead of the communities choosing their members. An edge is generated between nodes i and j if and only if the sets $\{k : i \in V(C_k)\}$ and $\{k : j \in V(C_k)\}$ intersect. This class of models is the topic of Publication I.

When m is large compared to n, this model is known to behave similarly to the ER graph [17, 60]. A generalization of the model, which allows the nodes to choose their communities in non-binomial ways, was introduced in [24]. The asymptotic structure of these kinds of models has been studied in the literature from various points of views, see, e.g., [7, 20, 61].

- **Passive random intersection graphs:**

  Another class of models introduced in [24], where we let the distribution of X_k be arbitrary and set $Y_k = 1$. In this case the events $\{v \in V(C_k)\}$ are no longer independent, but the edges are still determined by the intersections of the sets $\{k : v \in V(C_k)\}$.

  The model is passive in the sense that the nodes are chosen by the communities according to the specified distribution of X_k. Although the definition of the model seems quite simple, it models certain aspects of real networks: in particular, it allows for non-trivial clustering coefficients and power-law degree distributions even in the sparse case (when X_k is chosen suitably). Moreover, it allows for negative correlations between degrees and clustering [7], which corresponds to empirical observations from network data sets presented in [18].

- **Thinned random intersection graphs:**

  This class of models was proposed in Publication II. Let X_k be arbitrary and let Y_k be a fixed constant $q \in (0, 1)$, for all k. This model is similar to the passive random intersection graph, but allows for sparse community structures, which corresponds more closely to the intuitive idea of community.

  For example, if we observed a subgraph with nodes $\{1, \ldots, 7\}$, and all edges between them with the exception of $\{6, 7\}$, it would be natural to think that this subgraph was formed by one community. This is not possible in a passive intersection graph, where each community forms a clique. In particular, describing real networks with passive random intersection graphs may lead to overestimating the numbers of communities. Perhaps surprisingly, q can be estimated from data with moment-based methods [37].

- **Superpositions of Bernoulli random graphs:**

  This is the most general class of models, which is obtained by letting the joint distribution $(X_k, Y_k) \sim P$ be arbitrary. This is mostly motivated by that (i) two communities of equal sizes may have very different densities, and so Y_k should be allowed to be random, and (ii) large communities can be expected to be sparser than smaller ones, and so X_k and Y_k should be allowed to be correlated. Although the edges within each community are still independent (obtained through a Bernoulli trial), this model allows for fairly complex structures, e.g., mixtures of very small and very large communities. This class of models is also of theoretical importance, because it contains the previous models as special cases. Under certain assumptions on P and m, these models admits non-trivial clustering coefficients and heavy-tailed (namely, compound Poisson) degree distributions [11].

  Superpositions of Bernoulli random graphs are studied in Publications III, IV, and VI. Non-trivial clustering and assortativity, power-law degree distributions, and asymptotically normal subgraph counts are obtained under mild assumptions on the community type distribution P.

Several closely related models have been introduced in the literature. Yang and Leskovec defined and studied the *Community-Affiliation Graph Model* in [76]. This model treats communities and the edge probabilities as latent parameters, and only the edges are treated as random. The *inhomogeneous* random intersection graph [8, 63] assigns a weight to each node and community, and the probability that a node belongs to a community is a function of both weights. A model with more general distributions on the communities C_k (i.e., not necessarily ER) called *random intersection graph with communities* was introduced in [67]. Other models, such as the *Overlapping Stochastic Block Model* [43], allow edges between nodes that are not in the same community. A recent review of various block models is given in [44].

4. Mathematical preliminaries

This chapter reviews concepts and theorems in probability theory that are used in the publications of this dissertation. In particular, we discuss convergence of random variables and distributions.

The following shorthand notation is often used to denote convergence of (random) sequences. Deterministic convergence of a real-valued sequence is denoted by $a_n \to a$. For a sequence of random variables $X_1, X_2, \ldots$, convergence in probability, in distribution, and almost surely are denoted by

$$X_n \xrightarrow{p} X, \quad X_n \xrightarrow{d} X, \quad \text{and} \quad X_n \xrightarrow{a.s.} X,$$

respectively. When convergent or bounded sequences appear in equations, the following notation is often useful. All the limits are to be understood in the sense "as $n \to \infty$". Following the conventions of [32, 33], denote:

- $a_n = o(b_n)$, if $a_n/b_n \to 0$.
- $a_n = O(b_n)$, if there exist constants C and n_0 such that $|a_n| \leq Cb_n$ for $n \geq n_0$.
- $a_n = \omega(b_n)$, if $b_n = o(a_n)$.
- $a_n = \Omega(b_n)$, if there exist constants $c > 0$ and n_0 such that $a_n \geq cb_n$ for $n \geq n_0$.
- $a_n = \Theta(b_n)$, if there exist constants $c_1, c_2 > 0$ and n_0 such that $c_1 b_n \leq a_n \leq c_2 b_n$ for all $n \geq n_0$. The notation $a_n \asymp b_n$ is equivalent to $\Theta(b_n)$.
- $a_n \sim b_n$, if $a_n/b_n \to 1$. Although we also use $X \sim \mu$ to denote "X has the distribution μ", usually there is no risk of confusion.
- $a_n \ll b_n$ and $b_n \gg a_n$, if $a_n \geq 0$ and $a_n = o(b_n)$.
- $X_n = o_p(a_n)$, if $X_n/a_n \to 0$ in probability. Especially, if X_n converges to X in probability, we often write $X_n = X + o_p(1)$.
- $X_n = O_p(a_n)$, if for every $\varepsilon > 0$ there exist constants C_ε and n_ε such that $\mathbb{P}(|X_n| > C_\varepsilon a_n) \leq \varepsilon$ for every $n \geq n_\varepsilon$.

There are many ways to define convergence for a sequence of probability measures (see, e.g., [21] for a review of convergence in different metrics). In this dissertation, the most relevant of these is the following.

Definition 4.1 (Weak convergence)**.** *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on a metric space $(\mathcal{X}, d)$ with a Borel σ-algebra, and let $X \sim \mu$, $X_n \sim \mu_n$. We say that $\mu_n \to \mu$ weakly (and $X_n \to X$ in distribution) if*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X)$$

for all continuous and bounded functions $f : \mathcal{X} \to \mathbb{R}$.

In practice, the terms *weak* convergence and convergence *in distribution* are often used interchangeably. There are many equivalent ways to define this type of convergence in metric spaces. These are collected in what is known as the portmanteau theorem.

Theorem 4.2 (Portmanteau theorem)**.** *For any random elements X and $X_1, X_2, \ldots$ in a metric space S, the following conditions are equivalent.*

(i) X_n *converges to* X *in distribution;*

(ii) $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ *for all bounded Lipschitz functions f;*

(iii) $\liminf_n \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for every open set* $G \subset S$*;*

(iv) $\limsup_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for every closed set* $F \subset S$*.*

Moreover, if X and $X_1, X_2, \ldots$ are real-valued with cumulative distribution functions $F, F_1, F_2, \ldots$, then the above conditions are equivalent to $F_n(x) \to F(x)$ for all points $x \in \mathbb{R}$ at which F is continuous.

In general, weak convergence does not guarantee that $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for an unbounded function f. In particular, we often need the expected values to converge, $\mathbb{E}X_n \to \mathbb{E}X$. The relevant condition for achieving this is *uniform integrability*.

Definition 4.3 (Uniform integrability)**.** *A family of random variables $\{X_t\}_{t \in T}$ is uniformly integrable if*

$$\lim_{r \to \infty} \sup_{t \in T} \mathbb{E}|X_t| \mathbb{I}(|X_t| > r) = 0.$$

For a sequence $X_1, X_2, \ldots$ with $\mathbb{E}(|X_n|) < \infty$, $\forall n$, this is equivalent to

$$\lim_{r \to \infty} \limsup_{n \to \infty} \mathbb{E}|X_n| \mathbb{I}(|X_n| > r) = 0.$$

Uniform integrability is often applied together with truncation arguments – if we can prove a result for random variables X_n bounded by r, then there is hope that the result still holds when the truncation is removed by letting $r \to \infty$, as long as uniform integrability holds. If the random variables are also non-negative (as they often are in the context of random graphs), we can also obtain convergence of expected values by the following theorem.

Theorem 4.4 (Uniform integrability and convergence of expectations)**.** *Let $X, X_1, X_2, \ldots$ be $\mathbb{R}_+$-valued random variables with $X_n \to X$ in distribution. Then $\mathbb{E}X_n \to \mathbb{E}X < \infty$ if and only if $X_1, X_2, \ldots$ is uniformly integrable.*

Without uniform integrability, a lower bound is still obtained for non-negative random variables by the following theorem (cf. condition (iii) in the portmanteau theorem).

Theorem 4.5 ([36], Lemma 3.11)**.** *Let $X, X_1, X_2, \ldots$ be $\mathbb{R}_+$-valued random variables such that X_n converges to X in distribution. Then $\mathbb{E}X \leq \liminf_n \mathbb{E}X_n$.*

Instead of studying the relationship between two random variables X and Y, we are sometimes only interested in the relationship between their distributions. It is often useful to employ a *coupling argument*, where we define new random variables $\hat{X}$ and $\hat{Y}$ with the same distributions as X and Y, but with a different joint distribution. The relationship between $\hat{X}$ and $\hat{Y}$ can then be studied to reveal information about the distributions of X and Y.

Definition 4.6 (Coupling)**.** *Let X and Y be random variables. We say that a random variable $(\hat{X}, \hat{Y})$ is a* coupling *of X and Y, if X and $\hat{X}$ have the same distribution, and Y and $\hat{Y}$ have the same distribution. Similarly, for two probability measures μ and ν we define a coupling as a probability measure P whose marginals are μ and ν.*

A special case of a coupling is given by *Skorokhod's coupling theorem* (Anatoliy Skorokhod, [64]), which allows us to treat weakly convergent sequences as if they converged pointwise, in the following sense. This formulation is presented in the standard reference book of Billingsley, [6].

Theorem 4.7 (Skorokhod's coupling theorem)**.** *Let $\mu, \mu_1, \mu_2, \ldots$ be such that $\mu_n \to \mu$ weakly and μ has a separable support. Then there exist random variables $X, X_1, X_2, \ldots$, defined on a common probability space, such that $X_n \sim \mu_n$, $X \sim \mu$, and $X_n \to X$ pointwise.*

The *continuous mapping theorem* is central to many proofs in probability and statistics, especially in parameter estimation. It states that for a continuous function f, we may prove $f(X_n) \to f(X)$ by proving that $X_n \to X$ (with a suitable mode of convergence). This is often used when X_n converges to a constant a in probability, and an estimator is defined as a function $f(X_n)$, which is continuous at a. The following formulation is from [72]. It is important to note that here X is not necessarily a one-dimensional random variable, and so the function f may combine several quantities to form an estimator.

Theorem 4.8 (Continuous mapping theorem)**.** *Let $f : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$.*

(i) If $X_n \xrightarrow{d} X$*, then* $f(X_n) \xrightarrow{d} f(X)$.

(ii) If $X_n \xrightarrow{p} X$*, then* $f(X_n) \xrightarrow{p} f(X)$.

(iii) If $X_n \xrightarrow{a.s.} X$*, then* $f(X_n) \xrightarrow{a.s.} f(X)$.

We finish the review of classical probability theory with *Pratt's lemma* (John W. Pratt, [55]), a basic tool to verify the convergence of expected values $\mathbb{E}X_n$ by introducing a sequence of dominating random variables Y_n. This is sometimes referred to as Lebesgue's dominated convergence theorem, although Pratt's lemma does not require a single dominating random variable Y, unlike most formulations of the former.

Theorem 4.9 (Pratt's lemma, [27])**.** *Let* $X, X_1, X_2, \ldots$ *and* $Y, Y_1, Y_2, \ldots$ *be random variables with* $X_n \xrightarrow{a.s.} X$ *and* $Y_n \xrightarrow{a.s.} Y$*. If* $|X_n| \leq Y_n$ *for all* n *and* $\mathbb{E}Y_n \to \mathbb{E}Y < \infty$ *as* $n \to \infty$*, then* $\mathbb{E}X_n \to \mathbb{E}X$ *as* $n \to \infty$*. The theorem remains true when almost sure convergence is replaced by convergence in probability.*

4.1 Stable distributions

One of the most important theorems in probability theory is the central limit theorem, which states that

$$\frac{\sum_{i=1}^{n}(X_i - EX_i)}{\sqrt{n}}$$

is asymptotically normal, given that $X_1, X_2, \ldots$ are i.i.d. and have a finite variance. Clearly this results holds for a large class of different distributions of X_i, and we say that these distributions belong to the *domain of attraction* of the normal distribution.

Definition 4.10 (Domain of attraction)**.** *Let the random variables* $X_1, X_2, \ldots$ *be real-valued, independent, and identically distributed with cdf* $F(x)$*. If there exist sequences* A_n *and* B_n *such that the cumulative distribution functions of*

$$S_n \;=\; \frac{1}{B_n}\sum_{k=1}^{n} X_k - A_n$$

converge to a cdf $V(x)$ *as* $n \to \infty$*, then we say that* $F(x)$ *belongs to the domain of attraction of* $V(x)$*.*

When the finite variance condition is not satisfied, the question of existence of a limit distribution is more complicated, and there are many possible (non-normal) limit distributions with their own domains of attraction. Almost all of these are what are known as *stable distributions* (or α*-stable distributions*), distributions that are "stable" under linear combinations of i.i.d. random variables, in the following sense.

Definition 4.11 (Stable distribution)**.** *Let X, X_1, X_2 be i.i.d. random variables with cdf $F(x)$. If for all $a > 0, b > 0$ there exist $c > 0$ and $d \in \mathbb{R}$ such that*

$$aX_1 + bX_2 \overset{d}{=} cX + d,$$

where $\overset{d}{=}$ denotes equality in distribution, then $F(x)$ is stable.

The statement "almost all distributions" above is motivated by the following theorem.

Theorem 4.12. *A non-degenerate random variable X has a stable distribution if and only if there exist sequences A_n and B_n, and i.i.d. random variables $X_1, X_2, \ldots$ such that*

$$\frac{1}{B_n}\sum_{i=1}^{n} X_n - A_n \overset{d}{\to} X \quad \textit{as } n \to \infty.$$

There exist many parametrizations for stable distributions. The one used here follows the convention of [23]. Although stable distributions cannot generally be expressed analytically via a density function, their characteristic functions have simple forms. Despite the absence of an analytical density function, samples can be generated from stable distributions in a straightforward way with a transformation of uniformly distributed random variables ([52], Theorem 1.3).

Theorem 4.13 (Representations of stable distributions)**.** *Let $F(x) = \mathbb{P}(X \leq x)$ be a cdf with characteristic function $f(t) = \mathbb{E}(e^{itX})$. $F(x)$ is stable if and only if*

$$\log f(t) = i\gamma t - c|t|^{\alpha}(1 + i\beta\frac{t}{|t|}\omega(t,\alpha)),$$

where $0 \leq \alpha \leq 2$, $-1 \leq \beta \leq 1$, $\gamma \in \mathbb{R}$, and $c \geq 0$, and

$$\omega(t,\alpha) = \begin{cases} \tan(\frac{\pi}{2}\alpha) & \textit{if } \alpha \neq 1, \\ \frac{2}{\pi}\log|t| & \textit{if } \alpha = 1. \end{cases}$$

The following theorem gives conditions for verifying that a sum of i.i.d. random variables converges (in distribution) to a stable distribution. We omit the formulas for the sequences A_n and B_n here, and refer to the more explicit formulation in [52] (Theorem 3.12). The statement below follows the convention of [23], and omits the case of a normal limit.

Theorem 4.14 (Generalized central limit theorem)**.** *Let $X_1, X_2, \ldots$ be real-valued i.i.d. random variables with cdf $F(x)$. For $F(x)$ to belong to the domain of attraction of a stable distribution with $0 < \alpha < 2$, it is necessary and sufficient that*

(i) it holds that

$$\frac{F(-x)}{1 - F(x)} \to \frac{c_1}{c_2} \quad \textit{as } x \to \infty,$$

where $c_1 + c_2 > 0$ and $|c_1 - c_2| \leq c_1 + c_2$,

(ii) for every constant $k > 0$

$$\frac{1 - F(x) + F(-x)}{1 - F(kx) + F(-kx)} \to k^\alpha \quad \textit{as } x \to \infty.$$

The parameters of the stable distribution depend on α*,* c_1*, and* c_2*. The sequences* A_n *and* B_n *depend also on* $F(x)$*.*

4.2 Wasserstein metrics

The *Wasserstein* metrics (or *Wasserstein distances*, introduced by Leonid Vaserstein [73]) measure the distance between two probability distributions. Convergence in a Wasserstein metric is a stronger notion than convergence in distribution, and in particular, it implies the convergence of certain moments. It has applications in optimal transport theory [74], and is often used together with Stein's method to prove asymptotic distributions of sums of dependent random variables, see e.g. [25, 58, 59].

Definition 4.15 (Wasserstein space)**.** *Let* $(\mathcal{X}, d)$ *be a Polish metric space, and let* $p \geq 1$*. The Wasserstein space of order* $p \geq 1$ *is defined as*

$$P_p(\mathcal{X}) \ := \ \left\{ \mu \in P(\mathcal{X}) : \int_{\mathcal{X}} d(x_0, x)^p \, \mu(dx) < \infty \right\},$$

where $P(\mathcal{X})$ *is the set of Borel probability measures on* $\mathcal{X}$*, and* x_0 *is arbitrary.*

Definition 4.16 (Wasserstein distance)**.** *Let* μ *and* ν *be probability measures in* $P_p(\mathcal{X})$*. The Wasserstein-p distance between* μ *and* ν *is defined as*

$$W_p(\mu, \nu) \ = \ \inf\{\mathbb{E}(d(X, Y)^p)^{1/p} : \ X \sim \mu, \ Y \sim \nu\},$$

where the infimum is taken over all couplings of X *and* Y*.*

The function W_p above is called the Wasserstein-p metric. Naturally, for probability measures $\mu, \mu_1, \mu_2, \ldots$ we say that $\mu_n \to \mu$ in W_p, if $W_p(\mu_n, \mu) \to 0$ as $n \to \infty$.

Theorem 4.17. *Let* $X_n \sim \mu_n$*,* $X \sim \mu$ *be real-valued random variables. Assume that* $\mu_n \to \mu$ *in* W_p*. Then* $\mu_n \to \mu$ *weakly and* $\mathbb{E}(|X_n|^p) \to \mathbb{E}(|X|^p)$*.*

The following theorem is sometimes useful for verifying convergence in W_2 for two-dimensional random vectors. A more general version of the theorem (with $p \in [1, \infty)$ and a general Polish space) is presented in [74] (Definition 6.8 and Theorem 6.9).

Theorem 4.18. *Let* $\mu, \mu_1, \mu_2, \ldots \in P_2(\mathbb{R}^2)$*. Then* $\mu_n \to \mu$ *in* W_2 *if and only if*

$$\int \phi(x, y) d\mu_n(x, y) \to \int \phi(x, y) d\mu(x, y)$$

for all continuous functions ϕ *with* $|\phi(x, y)| \leq C(1 + x^2 + y^2)$ *for some* $C \in \mathbb{R}$*.*

4.3 Probabilistic inequalities

We now summarize probabilistic inequalities that are not as well known as, e.g., Markov's inequality or Chebyshev's inequality (which are omitted in this presentation), but which are utilized in the publications of this dissertation. The first inequality is one of many exponential inequalities for binomial random variables.

Theorem 4.19 ([11], Lemma A.7)**.** *Let X be $\mathrm{Bin}(n,p)$-distributed with mean $\mu = np$. Then*

$$\mathbb{P}(X = r) \;\leq\; e^{-\frac{s^2}{2(r+s)}},$$

for all $s > 0$ and all integers r such that $|r - \mu| \geq s$.

Janson's inequality, like the previous inequality, gives an upper bound for probabilities concerning sums of Bernoulli-distributed random variables, but in this case the random variables are allowed to be dependent. For the purposes of this dissertation, X can be considered to be a subgraph count, which is a sum of dependent indicator functions, e.g., due to the fact that different subgraphs may overlap.

Theorem 4.20 (Janson's inequality [31, 33])**.** *Let $\{J_i\}_{i\in Q}$ be a set of independent random indicator variables and let $\{Q(\alpha)\}_{\alpha\in A}$ be a family of subsets of the index set Q. Define $I_\alpha = \prod_{i\in Q(\alpha)} J_i$ and $X = \sum_{\alpha\in A} I_\alpha$. Assume that the index set A is finite. Then for all $0 \leq t \leq \mathbb{E}X$,*

$$\mathbb{P}(X \leq \mathbb{E}X - t) \;\leq\; e^{-\frac{t^2}{2\bar{\Delta}}},$$

where $\bar{\Delta} = \mathbb{E}(X) + \sum_{A\neq B} \sum_{A\cap B\neq\emptyset} \mathbb{E}(I_A I_B)$.

The last inequality is related to ER graphs. This inequality by Janson, Oleszkiewicz, and Ruciński [34] gives an upper bound for the upper tails of subgraph counts (one of the longest standing problems in random graph theory, see e.g. [15], the paper by Erdős and Rényi from 1960). Here we denote $v_H = |V(H)|$ and $e_H = |E(H)|$ for a graph $H = (V(H), E(H))$.

Theorem 4.21. *Let G be any graph with maximum degree Δ_G, and let X_G be the corresponding subgraph count in the ER graph $G(n,p)$. For any $t > 1$ there exists a constant $c(t,G)$ such that for all $n \geq v_G$ and $p \in (0,1)$*

$$\mathbb{P}(X_G \geq t\mathbb{E}X_G) \;\leq\; e^{-c(t,G)M_G^*(n,p)},$$

where

$$M_G^*(n,p) \;=\; \begin{cases} \Theta(1), & \text{if } p \leq n^{-1/m_G}, \\ \Theta(\min_{H\subset G} \psi_H^{1/\alpha_H^*}), & \text{if } n^{-1/m_G} \leq p \leq n^{-1/\Delta_G}, \\ \Theta(n^2 p^\Delta), & \text{if } p \geq n^{-1/\Delta_G}, \end{cases}$$

where $m_G := \max_{H\subset G} e_H/v_H$, $\psi_H = n^{v_H} p^{e_H}$, and α_H^* is the fractional independence number ([34], Appendix A), for which it holds that $\alpha_H^* \leq v_H - 1$.

5. Parameter estimation in the sparse regime

This chapter consists of an overview of concepts related to parameter estimation in the network models of Chapter 3. These concepts are closely related to Publications I, II, and V, and also serve as motivation for Publication III.

We focus on consistent estimators, i.e., ones that converge in probability to the true value as the number of nodes tends to infinity. Clearly, consistency can depend on how the model evolves as a function of n. For example, many of the theorems of this dissertation assume that sufficiently many moments of the community size distribution converge to finite and non-zero numbers.

5.1 Modeling large networks

A large network is modeled by considering a sequence of independent random graphs $(G_n)_{n\in\mathbb{N}} = (G_1, G_2, \ldots)$. The parameters (and distributions) of the model are allowed to depend on n, but we usually omit this in notation and write, e.g., $m = m_n$ and $P = P_n$. The properties of a large (but finite) graph can be thought to be approximated by the limits (when they exist) of the quantities and distributions obtained from G_n as $n \to \infty$. The question of how m and P_n should depend on n is not a trivial one. As we noted in the previous chapter, if m is too large, certain random intersection graph models become equivalent with with the ER graph, which clearly makes the models uninteresting for describing community structure. On the other hand, if m is too small, we obtain graphs that are almost empty as almost none of the nodes belong to any communities.

Recall that graphs are called sparse if the mean degree remains bounded as the size of the graph grows to infinity. In this dissertation, we are mostly interested in regimes where the mean degree $\mathbb{E}(d_G(i))$ converges to a non-zero finite limit. This is partly motivated by that many real networks appear to be sparse [50], but also by the fact that we wish to fit the model to data, where the average degree will always be non-zero and finite. This

regime turns out to be interesting in many ways. For example, we also obtain non-trivial clustering coefficients [38, 39] , degree distributions [11], and assortativity coefficients [10].

Consider the general superposition model with an arbitrary layer type distribution P. The mean degree of a node is found by a simple calculation:

$$\begin{aligned}\mathbb{E}(d_{G_n}(i)) &= \mathbb{E}\left(\sum_{j\neq i}\mathbb{I}\Big(\{i,j\}\in E(G_n)\Big)\right)\\ &= (n-1)\mathbb{P}\Big(\{1,2\}\in E(G_n)\Big)\\ &= (n-1)\left(1-\prod_{k=1}^{m}\Big(1-\mathbb{P}(\{1,2\}\in E(C_k)\Big)\right),\end{aligned}$$

where $\mathbb{P}(\{1,2\}\in E(C_k))$ equals $\mathbb{E}[\mathbb{P}(\{1,2\}\in E(C_1))\,|\,\{1,2\}\subset V(C_1)]$ by the law of total probability, and so

$$\prod_{k=1}^{m}\Big(1-\mathbb{P}(\{1,2\}\in E(C_k)\Big) = \Big(1-\mathbb{E}\Big(\frac{X_1(X_1-1)Y_1}{n(n-1)}\Big)\Big)^m.$$

Denoting $P_{21}:=\mathbb{E}(X_1(X_1-1)Y_1)$, an application of the binomial theorem gives

$$\mathbb{E}(d_{G_n}(i)) = (n-1)\sum_{t=1}^{m}\binom{m}{t}(-1)^{t+1}\left(\frac{P_{21}}{n(n-1)}\right)^t.$$

There are many ways to obtain a non-trivial constant limit for the mean degree. If we view the communities C_k as a real underlying structure, rather than a purely mathematical construction, then it seems natural to assume that m/n is bounded. The mean degree is then approximately $(m/n)P_{21}$, e.g., in the following cases:

- Assume only that $(m/n)P_{21}\to c\in(0,\infty)$, and that the rest of the terms in the previous sum are negligible.
- Assume that $m/n\to\beta\in(0,\infty)$ and $P_{21}\to p_{21}\in(0,\infty)$, but do not assume anything else about the distribution P. This is one of the approaches taken in Publication II.
- Assume that $m/n\to\beta$ and that $P\to P_\infty$ (weakly) for some distribution P_∞ with $(P_\infty)_{21}\in(0,\infty)$, and that the random variables $X^{(n)}(X^{(n)}-1)Y^{(n)}$ are uniformly integrable. This is the approach taken in, e.g., Publication IV.
- Assume that $m/n\to\beta$, and that there is a (non-trivial) limiting random vector (X_∞,Y_∞), and that $(X^{(n)},Y^{(n)})\sim P_n$ has the same distribution as $(\min\{X_\infty,n\},Y_\infty)$. This is the approach taken in Publication VI.

One way to view the assumption $m/n \to \beta$ is that it controls the way certain subgraphs are formed. For example, consider a triangle on the vertex set $\{1, 2, 3\}$. If the number of communities is small, then as $n \to \infty$, the edges of the triangle are most likely formed by a single community, which leads to a simple asymptotic formula for the expected triangle count. However, as shown in [39], this does not extend to all subgraphs – it is possible that, asymptotically, most triangles are formed within communities, but, e.g., large numbers of 2-stars may still be formed by two different communities.

Other ways to model large graphs exist in the literature – e.g., notions of convergence in terms of subgraph counts have been considered in [54], see also the approach in [42].

5.2 Moment-based estimators

We now give an overview of how moment-based estimators can be derived and how their consistency can be established. The binomial random intersection graph is completely defined by the parameters n, m, and p, but since the number of nodes n is considered known, the unknown parameters are m and p. In the general superposition model, the layer type distribution P is completely arbitrary, and parameter estimation can only be considered if we assume that P comes from some parametric family. For clarity, we only consider the binomial random intersection graphs in this section.

We focus on consistent estimators. Recall that if $\theta \in \mathbb{R}$ is a parameter of the model, we say that $\hat{\theta}$ is a consistent estimator of θ, if $\hat{\theta}$ is a function of the data (i.e., the graph G, but not, e.g., m) that satisfies

$$\mathbb{P}(|\hat{\theta}(G) - \theta| > \varepsilon) \ \to \ 0 \quad \text{as } n \to \infty$$

for any $\varepsilon > 0$. According to this definition, $\hat{\theta}(G) = 0$ is a consistent estimator of p in any model where $p \to 0$, but clearly fails to say anything meaningful about the network. The approach we take in Publication I is to choose a different set of parameters, as follows. Define $\lambda \in (0, \infty)$ as the limit of the mean degree as $n \to \infty$, and $\mu \in (0, \infty)$ as the limit of the mean number of communities that include node 1. When we require that m/n and np converge to constants as $n \to \infty$, it turns out that

$$\frac{m}{n} \ \to \ \mu^2/\lambda \quad \text{and} \quad np \ \to \ \lambda/\mu. \tag{5.1}$$

The constants μ and λ can then be estimated, and can be thought to describe the graph well (together with the model) for a large n.

A simple calculation shows that the number of ways that the nodes can be assigned to communities is 2^{nm}. It is not computationally feasible to evaluate the probability of obtaining a graph g from a model with

parameters m and p, even for small values of n and m. This means that a naive maximum likelihood estimator is not practical, as it does not only require evaluating this probability, but also maximizing it. Bayesian estimation methods typically rely on evaluating this probability as well.

Moment-based parameter estimation has received some attention in the literature as a practical alternative [3, 5, 19]. In this approach, we (i) derive expected values of chosen random quantities (e.g., subgraph counts) from the theoretical model, (ii) solve the parameters from these equations as functions of the expected values, and (iii) replace the expected values by the observed quantities. This is illustrated in the following example based on Publication I.

Example 5.1 (Heuristic derivation of a moment-based estimator)**.** *In the binomial random intersection graph satisfying* (5.1) *it holds that*

$$\begin{aligned} \mathbb{E}N_{K_2} &\approx \frac{1}{2}n^2\mu^2 m^{-1}, \\ \mathbb{E}N_{S_2} &\approx \frac{1}{2}n^3\mu^3(1+\mu)m^{-2}. \end{aligned} \tag{5.2}$$

Treating these approximations as equalities, we obtain

$$\frac{\mathbb{E}N_{S_2}}{(\mathbb{E}N_{K_2})^2} = 2\frac{1}{n}(1+\mu^{-1}).$$

Solving for μ gives

$$\hat{\mu} = \left(\frac{n}{2}\frac{\mathbb{E}N_{S_2}}{\mathbb{E}N_{K_2}^2} - 1\right)^{-1},$$

and replacing $\mathbb{E}N_*$ by the empirical subgraph counts N_* yields the estimator

$$\hat{\mu} = \left(\frac{n}{2}\frac{N_{S_2}}{N_{K_2}^2} - 1\right)^{-1}. \tag{5.3}$$

The consistency of the estimator $\hat{\mu}$ is established by showing that the approximations of $\mathbb{E}N_*$ are sufficiently accurate, $\mathbb{E}N_*$ is sufficiently close to N_* for K_2, S_2, and by applying the continuous mapping theorem (Theorem 4.8). Note that evaluating $\hat{\mu}$ only requires counting the numbers of edges and 2-stars, which is easily done and requires only $O(n^2)$ operations. A shortcoming of this approach is that the estimate $\hat{\mu}$ is not necessarily contained in the parameter space – it is possible, especially for a small n, that $\hat{\mu} < 0$ for some data set, although these values of μ are excluded by the definition of the model. It is worth noting that different estimators can be derived for the same parameter by choosing different quantities for the moment equations (5.2). This choice may be motivated, e.g., by the necessary assumptions to ensure consistency and the time required for computing the quantities from data.

Parameter estimation with incomplete data

Sometimes it is useful to be able to estimate parameters using only a part of the data. Some part of the the data may simply not available, or some of the observations may be unreliable. On the other hand, computations with a subset of the data are naturally faster than with the full data set.

Consider first a case where the full data set G is available. One may choose a number of nodes $n_0 < n$ and a node set of size n_0, and run the computations on the subgraph $G^{(n_0)}$ induced by these nodes. To avoid a biased data set, the nodes should be chosen independently of G. If the full data set is not available and a particular node set is chosen by necessity, it may be unclear whether $G^{(n_0)}$ is representative of the whole network. Since the mechanism by which the observations become noisy or missing may not be known, and which may or may not be independent of the true network, the mathematical justifications may be considered heuristic.

As an example, consider estimating μ as in Example 5.1. Denoting the edge and 2-star counts in $G^{(n_0)}$ by $N_*(G^{(n_0)})$, the estimator (5.3) becomes

$$\hat{\mu} \;=\; \left(\frac{n_0}{2} \frac{N_{S_2}(G^{(n_0)})}{N_{K_2}(G^{(n_0)})^2} - 1 \right)^{-1} .$$

This estimator turns out to be consistent provided that n_0 is sufficiently large, namely, $n_0/n^{2/3} \to \infty$ as $n \to \infty$ [38].

6. Summaries of the articles

Publication I. We study the binomial random intersection graph, which can be used as a parsimonious model of large and sparse networks. We propose moment-based parameter estimators and derive the expression of the asymptotic clustering coefficient as a function of the parameters. The estimators only require computing the numbers of links, 2-stars, and triangles. To our knowledge, this is the first paper to discuss parameter estimation in the case where the number of communities is of the same order as the number of nodes. We prove the consistency of these estimators with partial data, i.e., if only a part of the nodes (and all the edges between them) are observed. The performance of the estimators is illustrated with numerical experiments on simulated data, which show reasonable accuracy with small values of n. The proofs utilize the approximate densities of unions of 2-stars and triangles, for which we give explicit formulas, and which may be of independent interest.

Publication II. We propose a network model which can be viewed as a variation of the passive random intersection graph. The passive RIG inherently contains large numbers of cliques, which may not be desirable when modeling graphs with a weaker community structure. In our model, the sparsity of the communities can be tuned via a thinning parameter. We derive the asymptotic formulas for the degree variance and clustering coefficient with arbitrary community size distributions under moment conditions. We also give formulas for the probabilities for finding a specific 2-star or triangle in the graph, and show that the expected numbers of edges, 2-stars, and triangles are close to their expected values as n tends to infinity. Using these results, we derive parameter estimators for the case of binomial community sizes and prove their consistency when using partial data. The results are illustrated with real-world data sets.

Publication III. We study a network model with overlapping communities, where the community sizes and strengths are allowed to be random and correlated. The model is motivated in particular by the fact that in real networks, larger communities can be expected to be sparser than smaller ones. Our main result is that under certain conditions, the observed subgraph counts are close to their expected values in a stochastic sense. Using combinatorial arguments, we derive the asymptotic formulas for the expected numbers of k-cycles and k-cliques for all k.

Publication IV. This article concerns the same model as Publication III. In real networks it is often observed that the degrees of adjacent nodes are correlated. For example, if a person has a large number of friends in a social network, then their friends often have large numbers of friends as well. Our main result is that the joint degree distribution of adjacent nodes converges to a limit distribution, assuming that the number of communities is of the same order as the number of nodes, and that the layer type distribution converges to a limit together with suitable cross-moments. This happens in the sense of weak convergence and, with stronger assumptions, also in the sense of Wasserstein-2 metric. We obtain the asymptotic formula of the model assortativity and the convergence of Spearman's and Kendall's rank correlation coefficients. Finally, we show that the asymptotic joint degree distributions show power-law type behaviour in certain settings with an inverse relationship between the layer sizes and strengths.

Publication V. This short article generalizes the parameter estimation approach of Publication II to models with non-binomial community sizes. We discuss the estimation of the thinning parameter and the asymptotic community size distribution π. We assume that π belongs to a family of single-parameter distributions, where each distribution can be identified by the ratio of its third and second factorial moments, $(\pi)_3/(\pi)_2$. It is shown that the parameters may be estimated using the subgraph counts of links, 2-stars, and triangles also in this setting. The intuition is that the ratio of the factorial moments can be estimated from the data (without estimating the actual factorial moments), and in many cases we have a simple and continuous relation between $(\pi)_3/(\pi)_2$ and the parameter. Pareto-mixed Poisson distributions and Zipf-type distributions are considered as examples. The consistency of the estimators is based on the results of Publication III.

Publication VI. This article concerns the same superposition model as Publication III, with the assumption that the community size distributions are truncated versions of the limit distribution. We prove the asymptotic normality of counts of 2-connected subgraphs under moment conditions for the limiting layer type distribution. Cycles and cliques are considered as special cases. We also obtain convergence to α-stable distributions for 2-connected and balanced graphs under certain conditions which, to our knowledge, is the first result of its kind in the literature of sparse affiliation network models.

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] L.A.N. Amara, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. In *The Structure and Dynamics of Networks*, pages 207–210. Princeton University Press, 2011.

[3] C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B. Statistical Methodology*, 74(1):3–35, 2012.

[4] I. L. Amerise and A. Tarsitano. Correction methods for ties in rank correlations. *Journal of Applied Statistics*, 42(12):2584–2596, 2015.

[5] P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

[6] P. Billingsley. *Convergence of probability measures*. Wiley, second edition, 1999.

[7] M. Bloznelis. Degree and clustering coefficient in sparse random intersection graphs. *The Annals of Applied Probability*, 23(3):1254–1289, 2013.

[8] M. Bloznelis and J. Damarackas. Degree distribution of an inhomogeneous random intersection graph. *Electronic Journal of Combinatorics*, 20:P3, 2013.

[9] M. Bloznelis and J. Jaworski. The asymptotic normality of the global clustering coefficient in sparse random intersection graphs. In *15th International Workshop on Algorithms and Models for the Web-Graph (WAW)*, pages 16–29. Springer, 2018.

[10] M. Bloznelis, J. Karjalainen, and L. Leskelä. Assortativity and bidegree distributions on Bernoulli random graph superpositions. In *17th Workshop on Algorithms and Models for the Web Graph (WAW)*, pages 68–81. Springer, 2020.

[11] M. Bloznelis and L. Leskelä. Clustering and percolation on superpositions of Bernoulli random graphs, 2020. arXiv:1912.13404.

[12] B. Bollobás. *Random graphs*. Cambridge University Press, second edition, 2001.

[13] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *The Structure and Dynamics of Networks*, pages 183–194. Princeton University Press, 2011.

[14] A. Channarond, J.-J. Daudin, and S. Robin. Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601, 2012.

[15] P. Erdős and A. Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[16] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

[17] J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen. Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models. *Random Structures & Algorithms*, 16(2):156–176, 2000.

[18] I. Foudalis, K. Jain, C. Papadimitriou, and M. Sideri. Modeling social networks through user background and behavior. In *8th International Workshop on Algorithms and Models for the Web-Graph*, pages 85–102. Springer, 2011.

[19] O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77(380):835–840, 1982.

[20] A. Frieze and M. Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.

[21] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[22] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[23] B. V. Gnedenko and A. N. Kolmogorov. Limit distributions for sums of independent random variables. *Addison-Wesley*, 1968.

[24] E. Godehardt and J. Jaworski. Two models of random intersection graphs and their applications. *Electronic Notes in Discrete Mathematics*, 10:129–132, 2001.

[25] L. Goldstein and G. Reinert. Stein's method for the beta distribution and the Pólya-Eggenberger urn. *Journal of Applied Probability*, 50(4):1187–1205, 2013.

[26] J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

[27] A. Gut. *Probability: A graduate course*. Springer, 2013.

[28] F. Hormozdiari, P. Berenbrink, N. Pržulj, and S. C. Sahinalp. Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution. *PLoS Computational Biology*, 3(7):e118, 2007.

[29] E. Jacob and P. Mörters. Spatial preferential attachment networks: Power laws and clustering coefficients. *The Annals of Applied Probability*, 25(2):632–662, 2015.

[30] S. Jain and C Seshadhri. A fast and provable method for estimating clique counts using Turán's theorem. In *Proceedings of the 26th International Conference on the World Wide Web*, pages 441–449. Association for Computing Machinery, 2017.

[31] S. Janson. Poisson approximation for large deviations. *Random Structures & Algorithms*, 1(2):221–229, 1990.

[32] S. Janson. Probability asymptotics: Notes on notation. *arXiv:1108.3924*, 2011.

[33] S. Janson, T. Łuczak, and A. Ruciński. *Random graphs*. John Wiley & Sons, 2011.

[34] S. Janson, K. Oleszkiewicz, and A. Ruciński. Upper tails for subgraph counts in random graphs. *Israel Journal of Mathematics*, 142(1):61–92, 2004.

[35] M. Jha, C. Seshadhri, and A. Pinar. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In *Proceedings of the 24th international conference on World Wide Web*, pages 495–505. Association for Computing Machinery, 2015.

[36] O. Kallenberg. *Foundations of modern probability*. Springer, second edition, 2002.

[37] J. Karjalainen. A note on parameter estimation of thinned random intersection graphs. In *22nd European Young Statisticians Meeting*, pages 51–55. Panteion University of Social and Political Sciences, 2021.

[38] J. Karjalainen and L. Leskelä. Moment-based parameter estimation in binomial random intersection graph models. In *14th Workshop on Algorithms and Models for the Web Graph (WAW)*, pages 1–15. Springer, 2017.

[39] J. Karjalainen, J. S. H. van Leeuwaarden, and L. Leskelä. Parameter estimators of sparse random intersection graphs with thinned communities. In *15th Workshop on Algorithms and Models for the Web Graph (WAW)*, pages 44–58. Springer, 2018.

[40] M. Karoński, E. R. Scheinerman, and K. B. Singer-Cohen. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8(1-2):131–159, 1999.

[41] D. Krioukov, M. Kitsak, R. S. Sinkovits, D. Rideout, D. Meyer, and M. Boguñá. Network cosmology. *Scientific Reports*, 2(1):1–6, 2012.

[42] V. Kurauskas. On local weak limit and subgraph counts for sparse random graphs, 2015. arXiv:1504.08103.

[43] P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.

[44] C. Lee and D. J. Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[45] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[46] J. Nešlehová. On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98(3):544 – 567, 2007.

[47] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[48] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[49] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[50] M. E. J. Newman. *Networks — An Introduction*. Oxford University Press, 2010.

[51] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[52] J. Nolan. *Stable distributions: Models for heavy-tailed data*. Birkhauser, 2003.

[53] S. Ouadah, S. Robin, and P. Latouche. Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases. *Scandinavian Journal of Statistics*, 47(1):156–181, 2020.

[54] S. Petti and S. Vempala. Approximating sparse graphs: The random overlapping communities model. arXiv: 1802.03652, 2018.

[55] J. W. Pratt. On interchanging limits and integrals. *The Annals of Mathematical Statistics*, 31(1):74–77, 1960.

[56] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[57] M. Rahman, M. A. Bhuiyan, and M. Al Hasan. Graft: An efficient graphlet counting method for large graph analysis. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2466–2478, 2014.

[58] M. Raič. Normal approximation by Stein's method. In *Proceedings of the 7th Young Statisticians Meeting*, pages 71–97, 2003.

[59] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[60] K. Rybarczyk. Equivalence of a random intersection graph and $G(n, p)$. *Random Structures & Algorithms*, 38(1-2):205–234, 2011.

[61] K. Rybarczyk and D. Stark. Poisson approximation of counts of induced subgraphs in random intersection graphs. *Discrete Mathematics*, 340(9):2183–2193, 2017.

[62] A. Röllin. Kolmogorov bounds for the normal approximation of the number of triangles in the Erdős-Rényi random graph. *Probability in the Engineering and Informational Sciences*. To appear, arXiv:1704.00410.

[63] Y. Shang. Degree distributions in general random intersection graphs. *The Electronic Journal of Combinatorics*, 17:R23, 2010.

[64] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290, 1956.

[65] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. DOULION: Counting triangles in massive graphs with a coin. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 837–846. Association for Computing Machinery, 2009.

[66] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1307–1318. Association for Computing Machinery, 2013.

[67] V. Vadon, J. Komjáthy, and R. van der Hofstad. A new model for overlapping communities with arbitrary internal structure. *Applied Network Science*, 4(1):42, 2019.

[68] R. van der Hofstad. *Random graphs and complex networks*, volume 1. Cambridge University Press, 2016.

[69] R. van der Hofstad and N. Litvak. Degree-degree dependencies in random graphs with heavy-tailed degrees. *Internet Mathematics*, 10(3-4):287–334, 2014.

[70] P. van der Hoorn and N. Litvak. Convergence of rank based degree-degree correlations in random directed networks. *Moscow Journal of Combinatorics and Number Theory*, 4(4):427–465, 2014.

[71] P. van der Hoorn and N. Litvak. Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet Mathematics*, 11(2), 2015.

[72] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[73] L. N. Vaserstein. Markov processes on a countable product space, describing large systems of automata. *Problemy Peredachi Infomatsii*, 5(3):64–73, 1969.

[74] C. Villani. *Optimal transport: Old and new*. Springer, 2009.

[75] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[76] J. Yang and J. Leskovec. Structure and overlaps of ground-truth communities in networks. *ACM Transactions on Intelligent Systems and Technology*, 5(2), 2014.

Aalto-DD 175/2021

ISBN 978-952-64-0626-8 (printed)
ISBN 978-952-64-0627-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Mathematics and Systems Analysis
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS