JAGIELLONIAN UNIVERSITY

KRAKÓW, POLAND

FACULTY OF PHYSICS, ASTRONOMY
AND
APPLIED COMPUTER SCIENCE

MARIAN SMOLUCHOWSKI INSTITUTE OF PHYSICS

Static and dynamic properties of selected stochastic processes on complex networks

Jeremi K. Ochab

A thesis submitted for the degree of Doctor of Philosophy
supervised by prof. dr hab. Zdzisław Burda

KRAKÓW 2013

UNIwersytet Jagielloński

KRAKÓW

Wydział Fizyki, Astronomii

I

Informatyki Stosowanej

Instytut Fizyki im. Mariana Smoluchowskiego

Statyczne i dynamiczne własności wybranych procesów stochastycznych na sieciach złożonych

Jeremi K. Ochab

Praca doktorska wykonana pod kierunkiem
prof. dra hab. Zdzisława Burdy

KRAKÓW 2013

Wydział Fizyki, Astronomii i Informatyki Stosowanej
Uniwersytet Jagielloński

Oświadczenie

Ja niżej podpisany *Jeremi Kazimierz Ochab* (nr indeksu: 300) doktorant Wydziału Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Jagiellońskiego oświadczam, że przedłożona przeze mnie rozprawa doktorska pt. „*Statyczne i dynamiczne własności wybranych procesów stochastycznych na sieciach złożonych*” jest oryginalna i przedstawia wyniki badań wykonanych przeze mnie osobiście, pod kierunkiem prof. dr. hab. *Zdzisława Burdy*. Pracę napisałem samodzielnie.

Oświadczam, że moja rozprawa doktorska została opracowana zgodnie z Ustawą o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dziennik Ustaw 1994 nr 24 poz. 83 wraz z późniejszymi zmianami).

Jestem świadom, że niezgodność niniejszego oświadczenia z prawdą ujawniona w dowolnym czasie, niezależnie od skutków prawnych wynikających z ww. ustawy, może spowodować unieważnienie stopnia nabytego na podstawie tej rozprawy.

Kraków, dnia 26.07.2013

.....
podpis doktoranta

Abstract

This thesis is concerned with the properties of a number of selected processes taking place on complex networks and the way they are affected by structure and evolution of the networks. What is meant here by 'complex networks' is the graph-theoretical representations and models of various empirical networks (e.g., the Internet network) which contain both random and deterministic structures, and are characterised among others by the small-world phenomenon, power-law vertex degree distributions, or modular and hierarchical structure. The mathematical models of the processes taking place on these networks include percolation and random walks we utilise.

The results presented in the thesis are based on five thematically coherent papers. The subject of the first paper is calculating thresholds for epidemic outbreaks on dynamic networks, where the disease spread is modelled by percolation. In the paper, known analytical solutions for the epidemic thresholds were extended to a class of dynamically evolving networks; additionally, the effects of finite size of the network on the magnitude of the epidemic were studied numerically. The subject of the second and third paper is the static and dynamic properties of two diametrically opposed random walks on model highly symmetric deterministic graphs. Specifically, we analytically and numerically find the stationary states and relaxation times of the ordinary, diffusive random walk and the maximal-entropy random walk. The results provide insight into localisation of random walks or their trapping in isolated regions of networks. Finally, in the fourth and fifth paper, we examine the utility of random walks in detecting topological features of complex networks. In particular, we study properties of the centrality measures (roughly speaking, the ranking of vertices) based on random walks, as well as we conduct a systematic comparative study of random-walk based methods of detecting modular structure of networks.

These studies thus aimed at specific problems in modelling and analysis of complex networks, including theoretical examination of the ways the behaviour of random processes intertwines with the structure of complex networks.

Streszczenie

Niniejsza praca doktorska dotyczy własności wybranych procesów losowych zachodzących na sieciach złożonych oraz sposobów w jaki wpływa na nie struktura i zmienność w czasie tychże sieci. Przez „sieci złożone” rozumiem zapożyczone z teorii grafów modele różnorodnych sieci rzeczywistych (np. sieci internetowej), których struktury powstały w sposób częściowo losowy, a częściowo deterministyczny. Sieci takie charakteryzują się m.in. tak zwanym efektem małego świata, potęgowym rozkładem krotności wierzchołków czy modułową i hierarchiczną strukturą. Używane przeze mnie modele matematyczne procesów zachodzących na sieciach to perkolacja i błędzenia przypadkowe.

Wyniki przedstawione w niniejszej pracy doktorskiej opierają się na pięciu spójnych tematycznie artykułach naukowych. Przedmiotem pierwszego z nich jest obliczanie progowego prawdopodobieństwa wybuchu epidemii mającej miejsce na sieci dynamicznej, przy czym rozprzestrzenianie się choroby modelowane jest za pomocą perkolacji. W artykule tym znane wyniki analityczne dot. takich progów zostały rozszerzone na sieci podlegające ewolucji czasowej. Dodatkowo, numerycznie badano efekt skończonego rozmiaru sieci na wielkość powstałej epidemii. Przedmiotem drugiego i trzeciego artykułu są statyczne i dynamiczne własności dwóch diametralnie różnych błędzeń losowych na deterministycznych grafach o dużej symetrii. Analitycznie i numerycznie wyznaczone zostały stany stacjonarne i czasy relaksacji zwykłego błędzenia przypadkowego odpowiadającego dyfuzji oraz błędzenia maksymalizującego entropię. Wyniki te pozwalają lepiej zrozumieć efekt lokalizacji i uwięzienia błędzenia przypadkowego w odseparowanych częściach sieci. Wreszcie w czwartym i piątym artykule analizowana jest użyteczność błędzeń losowych w wykrywaniu pewnych topologicznych cech sieci złożonych. W szczególności, zbadano w nich własności tzw. miar centralności (ogólnie rzecz ujmując odpowiadających rankingom wierzchołków sieci) opartych o błędzenia losowe. Przeprowadzono również systematyczne porównanie opartych o błędzenia przypadkowe metod wykrywania w sieciach struktur modułowych.

Powyższe badania miały więc na celu podjęcie problemów modelowania i analizy sieci złożonych, a zwłaszcza teoretycznej analizy powiązań pomiędzy budową tych sieci i zachowaniem zachodzących na nich procesów losowych.

Za towarzyszące mi ciągle proste słowa:

Znajduj w świata głośnym szumie

to, co piękne, zacny Kumie.

Śmiej się, śmiechu innych szukaj,

I przenigdy nie trać ducha.

dziękuję Kasi Gizickiej.

Contents

List of papers	5
Introduction	7
Thesis overview	9
1 Graphs and random graphs	13
1.1 Terminology and basics	14
1.2 Graph representations	16
1.3 Structural quantities	19
1.4 Random graphs	25
2 Percolation	31
2.1 Basic concepts	32
2.2 Percolation of graphs	34
2.3 Percolation on small-world networks	37
2.4 Note on epidemic modelled by percolation	39
Study I	42
3 Random walks	45
3.1 Basics of Markov chains	46
3.2 Types of random walks	49
3.2.1 Generic random walk	49
3.2.2 Maximal-entropy random walk	49
3.2.3 Other random walks on graphs	51
3.3 Mean first-passage times	53

3.3.1	Correlation between MFPT and stationary states	54
3.3.2	MFPT and modular graph structure	56
Study II		61
Study III		63
4	Centrality measures	65
4.1	Basic properties	66
4.2	Types of centralities	67
4.3	Example applications of centrality measures	72
5	Community detection	77
5.1	What is a community?	78
5.2	Graphs with community structure	82
5.3	Community detection algorithms	84
5.3.1	Spectral methods	85
5.3.2	Random walks	86
5.3.3	Modularity optimisation	88
5.4	Centrality measures in community detection	90
5.5	Percolation in community detection	92
Study IV		94
Study V		96
	Conclusions	99
	Acknowledgements	103
	Bibliography	105
	Paper I	115
	Paper II	122
	Paper III	134
	Paper IV	147

Paper V	154
Statements of coauthorship	164

List of papers

The papers this thesis is based on are numbered and referred to with Roman numbers:

- I. J.K. Ochab, P.F. Góra, *Shift of percolation thresholds for epidemic spread between static and dynamic small-world networks*,
Eur. Phys. J. B **81**, 373–379 (2011)
- II. J.K. Ochab, Z. Burda, *Exact solution for statics and dynamics of Maximal Entropy Random Walk on Cayley trees*,
Phys. Rev. E **85**, 021145 (2012)
- III. J.K. Ochab, *Maximal Entropy Random Walk: solvable cases of dynamics*,
Acta Phys. Pol. B **43**, 1143 (2012)
- IV. J.K. Ochab, *Maximal-entropy random walk unifies centrality measures*,
Phys. Rev. E **86**, 066109 (2012)
- V. J.K. Ochab, Z. Burda, *Maximal entropy random walk in community finding*,
Eur. Phys. J-Spec. Top. **216**, 73-81 (2013)

Introduction

The development of human civilisation is increasingly data-driven. Not only the results of scientific experiments are stored as digital data, but also the traces of our daily activities as phone calls, purchases, or travels. The information which, however tritely, emerges from that data is: everything is connected [1]. It is not only our mobiles and computers that are linked; it is people, cities, and economies; it is organs, cells, and molecules; finally, but not exhaustively, it is books, words, and ideas that are related.

These networks of relationships have been a subject of intensive, systematic studies, both theoretical and empirical, for the last fifteen years, although the first attempts to model them in the present manner are probably due to social sciences [2] in late 1940s. The models firstly involve representing the topology of the network mathematically in the form of a graph. The graphs, however, turned out to be complex: they can be overwhelmingly big, they are to much extent random, but also contain significantly non-random structures, and some of their characteristics are distributed according to power laws instead of normal or Poisson distributions.

The second, highly nontrivial task in modelling a network involves representing the process that actually takes place on it, e.g., flow of money or spread of information. The range of theoretical approaches, often proposed by the physics community, includes percolation, spin models, diffusion and random walks, flow of electrical currents, or synchronisation of coupled oscillators. These can mimic

a variety of transport processes, e.g., traffic in urban or Internet networks, intracellular transport, spread of diseases, opinions, or memes.

Thirdly, depending on the time scales present in a studied system, the model graphs can either have a static, thermalised architecture or be allowed to grow or evolve. The alteration of the network can happen simultaneously with any of the processes listed above. Evolution of the network and the extrinsic process can even be coupled, e.g, so as to imitate people abstaining from social contact if they or their acquaintances have been infected.

According to the above scheme, this thesis is concerned with the properties of model processes taking place on graphs and the way they are affected by structure and evolution of complex networks. The topics explored include: modelling with percolation disease spread on a class of dynamically evolving networks, and in particular extending analytical solutions for the thresholds of epidemic outbreak and numerical studies on finite-size effects of outcomes of the epidemic; studying the static and dynamic behaviours of two diametrically opposed random walks on model deterministic graphs, specifically, their times of reaching stationary states or trapping in isolated regions; finally, applying the knowledge of their properties to reveal the complex structures of graphs.

These studies thus aimed at specific problems in modelling and analysis of complex networks, ultimately reaching the subject of community detection, which means finding groups of well-connected modules in the networks. This field of research may be considered as a developing methodology of data analysis that can be employed in basic research disciplines, including among others systems biology [3], neurosciences [4], social sciences [5], or literary studies [6]. The knowledge of a network's structure and its interaction with a given dynamical process is crucial also in applications. Among countless examples one could mention telecommunications (e.g., redesigning routing protocols in the Internet [7]), policy planning (e.g., modifications to traffic in urban street networks [8],

epidemic control and prevention [9], crisis management), social analysis (crime investigation [10], clustering of population with respect to the language spoken [11]), or data mining (WWW search [12], analysing target groups in web business). Although my own studies can be regarded rather as basic research, they were conducted with some of the above applications in mind as possible future research paths.

Thesis overview

In terms of structure and content this thesis may be regarded as merely an introduction to the detailed calculations and results of papers I-V. It is composed in such a way that each chapter prepares the ground to continue smoothly to the summary of a given study and to the original paper, while at the same time it allows to locate the study on the map of the specific field. The chapters are arranged in order of increasing breadth of knowledge needed to embrace the context of subsequent papers, which at the same time to much extent reproduces my own exploration of the discipline and the progress of my research. The increase in depth, on the other hand, takes place only within chapters in order to lead to more specific findings. Such text structure, designed to ultimately focus on the papers I-V, results in a constant struggle between conciseness and completeness of the presented material. This delicate balance can also be observed in the bibliography, where I refer to standard textbooks, vast reviews, and groundbreaking papers, as well as to some specialist articles that are narrower in scope, but relevant to my research; the most specific papers are sometimes left out from the bibliography of the thesis, but are included in the respective studies I-V.

The beginning chapter on *random graphs* is in fact extremely rudimentary, inasmuch as its first two sections merely define and name certain graph-theoretical concepts, including standard graph representations as matrices or lists; neverthe-

less, it has its purpose, since at the interface of several disciplines terminology is bound to blend and blur, and requires clarification. The chapter goes on to introduce a small number of measurable quantities that are indicative of structural properties of graphs, and are most frequently encountered in studies on complex networks. Finally, the essentials of *random* graphs are presented.

Whereas the first chapter exposes the *model structural backbone* of complex systems, the second chapter explains the properties and possible applications of what can be perceived as a basic *model process – percolation*. In the first section, this procedural perspective is used to describe the formation of giant connected components in the random graphs introduced earlier. While this provides some general context, the rest of the chapter is almost exclusively developed for the sake of the Study I. Specifically, Section 2.3 explicates the generating function technique used to find percolation thresholds in a particular type of small-world networks, and Section 2.4 outlines the connection between percolation and epidemic modelling. The chapter concludes with a summary of Study I, which extends analytical solutions for percolation thresholds to a class of small-world networks with dynamically rewired links, and provides numerical insights into finite-size effects for epidemic spread in such networks.

The subsequent chapter elaborates on another family of processes that can model transport or transmission of information, namely *random walks*. It is far beyond the scope of this humble doctoral thesis to attempt at covering a topic whose history is more than a century old, let alone encompassing it in just one chapter. The presented perspective is thus severely restricted to discrete time random walks on graphs. After introducing general definitions and properties of Markov chains and random walks, the scope is further narrowed down to selected random walks utilised in the studies on complex networks. Among other quantities characterising random walks described in the chapter, mean first-passage time matrix has become of much use in my research. For this reason,

I discuss it at length, and allow myself to present some additional, unpublished observations concerning its connection to structural properties of networks on which a random walk takes place. The summaries of Study II and III are appended to that chapter: the former concerns analytical solution for stationary states and relaxation times of the selected random walks on Cayley trees; the latter further explores numerically the dynamic behaviour of random walks on some other highly symmetric graphs. These two studies allowed to gain some mathematical intuition on behaviour of random walks on model graphs, and motivated extension of my research to examine intertwining of this behaviour with the structure of complex networks.

This is the dominant topic explored in the next two chapters. The first one reviews a variety of ways the importance of a node or a link in a network can be computed. The general term coined for this quantified importance is the *measure of centrality*. As discussed in the chapter, what the centrality precisely means depends on the research problem, on what the network represents, and on the specific processes inhabiting the system. Consequently, the relation between centrality and random walks – the model processes of choice – is highlighted. Study IV investigates this relation based on the knowledge provided in Chap.3, which allowed to unify some of the centralities in a common framework.

Due to the very close connection of the study to *community detection* in networks, however, its summary is presented only at the end of the final chapter. The chapter recalls some typical attempts at defining what a community is, and then describes several ways the construction of random graphs from Chap.1.4 can be modified to include the modular structures. Only then, the central subject of how to detect such structures is summarised, with the special attention paid to the application of random walks, but also with some classical methods outlined for comparison. The methods are additionally linked to the concepts introduced in the previous chapters, namely centrality measures and percolation. The chapter

concludes with the summaries of Study IV and V. While the former has been already mentioned, the latter is exclusively focused on the comparative study of the performance of community detection methods utilising different random walks.

The entire thesis thus briefly covers the subjects of modern mathematical techniques for modelling networks and various stochastic processes that take place on them, and for methods of complex networks analysis, which make use of static and dynamic properties of these processes.

Chapter 1

Graphs and random graphs

This chapter serves as an introduction to the rudimentary concepts providing a mathematical framework describing the structure of complex networks. The first section aims at setting conventions regarding notation and terminology, as well as reviewing several definitions concerning among others basic types of graphs, degrees, or paths that may come in handy in the subsequent chapters. In a similar manner, the second section briefly describes the ways of representing graphs as matrices, which are extensively used in analytical studies. Next, in the third section, I present an overview of the quantities that are most often analysed, both theoretically and experimentally, in complex network research. These quantities, such as vertex degree distributions, average path lengths, and clustering coefficients, are used in the primary characterisation of networks. The last section already provides the first intuitions about what complex networks actually are, since it introduces Erdős-Rényi ensemble of random graphs and the configuration model, which serve as the null models of complex networks.

The material of this chapter has been selected to recall only the concepts needed to provide foundation for research summarised in Studies I-V. As I am aware of how fragmentary this information is, I give references to a general introduction to graph theory [13] and a much more comprehensive source [14]

(accessible online for free). A broader view on modern network science can be found in [15–17], which present more physical approach, and thus much closer to my understanding of complex networks.

1.1 Terminology and basics

In this section, I review some basic terminology and graph-theoretical concepts. This mainly aims at establishing a common language with the readers from different disciplines. It can also serve as a very brief introduction (as far as definitions are concerned) to graph theory and what is called now "network science" to readers with no background in the disciplines.

Suppose we want to represent mathematically a set of entities (these might be cell phones, power plants, or genes) which can be pairwise connected to each other by another set of entities (e.g., by phone calls, power lines, or protein interactions). Usually, the former, finite and non-empty set is denoted by V and called the **vertices** (also **nodes**, or **sites**); the latter, finite set is denoted by E and called the **edges** of a graph (also **links**, or **bonds**). While the elements of the first set can be represented by labels, e.g., $u, v \in V$, the elements of the other set are unordered pairs of labels $e = \{u, v\} \in E$. These two sets together are said to represent a **simple graph** $G(V, E)$, which can otherwise be represented by drawing a diagram with dots (vertices) and lines (edges) connecting them. The graph can be also symbolised by just its name G , while the sets of vertices and edges of that graph can be denoted as $V(G)$ and $E(G)$, respectively. Two graphs are called, G and H , *isomorphic*, $G \sim H$, if and only if there exists a bijection $\phi : V(G) \longrightarrow V(H)$ such that $\forall u, v \in V(G) : ((u, v) \in E(G) \iff (\phi(u), \phi(v)) \in E(H))$. In other words, such a function, called *isomorphism*, only relabels the graph's vertices.

One of the basic quantities characterising a vertex v in a simple graph is the **degree** $k(v)$, which is the number of the vertex's neighbours, or equivalently,

the number of edges the vertex is incident with. If it is meaningful for the connections to be directed, e.g., we want to distinguish person A calling B from person B calling A, the edges are ordered pairs of vertices $e = (u, v) \in E$, and the corresponding graph is called a simple **directed graph** or a **digraph**. In digraphs, the degree separates into **in-degree** $k^{\text{in}}(v)$ and **out-degree** $k^{\text{out}}(v)$, which are the number of edges $(\cdot, v)$ pointing to v and the edges $(v, \cdot)$ coming out of v , respectively.

If it is meaningful for the connections to have a certain weight w , e.g., we want to describe the load of a power line, the edges can be denoted by $(\{u, v\}, w_{uv}) : u, v \in V, w_{uv} \in \mathbb{R}$, and the corresponding graph is called a **weighted graph**. In weighted graphs, the degree $k(v)$ remains an integer number of neighbours, but it can be generalised to **strength**, which is the sum of weights of edges incident to the given vertex $s(v) = \sum_u w_{u,v}$.

The simple, weighted, and directed graphs are thus three basic ways of selecting and representing the information about connections between some entities of interest. As regards the structure of graphs, I would like to enumerate still a few other types.

We call G such that all vertices have the same degree $k(v) = K$ a **K-regular graph**. For example, a square two-dimensional grid is a 4-regular graph, because each of its vertices has four neighbours, or physically speaking, it has the coordination number four. Grids, however, are only special cases of regular graphs.

We call G such that edges between all pairs of vertices exist a **complete graph** K_N . As the name suggests, for the total of $|V| = N$ vertices a complete graph has the maximal possible number of edges $N(N-1)/2$. A complete subgraph of a graph is often called a **clique**.

We call G such that its vertices can be divided into two disjunctive sets V_1, V_2 ($V = V_1 \cup V_2$) for which *only* edges $u, v, u \in V_1, v \in V_2$ exist, a **bipartite graph**.

Thus, the sets V_1 and V_2 are connected between each other, but neither of them is connected internally.

If the graphs are to represent, e.g., communication systems, it is natural to define the mathematical concepts corresponding the pathways of information transmission. On a graph these pathways are best described in terms of a sequence of vertices and edges $(v_1, e_1, v_2, e_2, \dots, v_{t-1}, e_{t-1}, v_t)$, in which none of the vertices is visited more than once, and the edges $e_i = \{v_i, v_{i+1}\}$ connect the consecutive vertices. We call such a sequence a **path**. The above path has a *length* t and *ends* v_1 and v_t . If $t \geq 3$ and $v_1 = v_t$, we call such a sequence a **cycle**.

Sometimes, if the pathways are to model, e.g., a particle wandering on some physical structure, the assumption that no vertices nor edges are visited twice may be rejected. In such a case, we call the sequence of vertices and edges a **walk**, and if $v_1 = v_t$, the walk is **closed**.

Based on those concepts, one more graph type can be defined that will be referred to further in the thesis: a **tree**, which we call a graph containing no cycles, and which is *connected* (i.e., between any two vertices there exists a path connecting them).

Whereas there is a whole taxonomy of many more different graphs, I restrict myself to only the tiny fraction of that bestiary that was used in the Studies I-V.

1.2 Graph representations

Before I go on further, a note is needed on possible representations of graphs. The most straightforward, and the most analytically manageable representations have a matrix form. For the number of vertices $|V| = N$, the **adjacency matrix** $\mathbf{A}$ has the size $N \times N$ and its elements take values

$$A_{uv} = \begin{cases} 1, & \text{if } (u, v) \in E \\ 0, & \text{if } (u, v) \notin E. \end{cases} \quad (1.1)$$

In case of simple undirected graphs, the matrix is symmetric, since the edges are unordered vertex pairs. If the graph is directed, $\mathbf{A}$ can be unsymmetric. If the graph is weighted, it is worthwhile to distinguish between the binary adjacency matrix $\mathbf{A}$, as defined above, and the real matrix $\mathbf{W}$, whose elements take values given by the respective edge weights w_{uv} . The latter is called a **weight matrix**.

The (in- or out-) degrees and strengths of vertices can be naturally computed with the use of adjacency or weight matrices: $k^{\text{out}}(v) = \sum_{u \in V} A_{vu}$, $k^{\text{in}}(v) = \sum_{u \in V} A_{uv}$, $s(v) = \sum_{u \in V} W_{vu}$. Similarly, the numbers of walks between any two vertices can be easily obtained from the adjacency matrix: as A_{vu} represents one step along the edge $\{v, u\}$, $\sum_{w \in V} A_{vw} A_{wu} = (\mathbf{A}^2)_{vu}$ is the number of walks of length 2 from v to u , and generally, $(\mathbf{A}^t)_{vu}$ is the number of walks of length t from v to u .

For the total number of edges $|E|$, another approach to representing a graph is to construct an **incidence matrix** $\mathbf{B}$ of size $|V| \times |E|$. The rows and columns of this matrix correspond to vertices and edges respectively, so that the elements tell whether a given vertex and edge are incident (i.e. whether the vertex is any of the two endpoints of the edge). In the case of directed graphs, the elements of the oriented incidence matrix take values

$$B_{ve} = \begin{cases} 1, & \text{if } e = (v, u) \\ -1, & \text{if } e = (u, v) \\ 0, & \text{if otherwise.} \end{cases} \quad (1.2)$$

In the case of undirected graphs (hence, *unoriented incidence matrix*) one should take the absolute value so that only binary values are allowed.

The last matrix representation is called **unnormalised Laplacian matrix** (also **Kirchoff's matrix**) $\mathbf{L}$, and can be obtained from the incidence matrix

$$\mathbf{L} = \mathbf{B}\mathbf{B}^T, \quad (1.3)$$

though the usual equivalent definition is

$$L_{uv} = \begin{cases} k(v), & \text{if } u = v \\ -1, & \text{if } \exists e = (u, v) \\ 0, & \text{if } (u, v) \notin E, \end{cases} \quad (1.4)$$

or in matrix notation

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \quad (1.5)$$

where $\mathbf{D}$ is a diagonal matrix with $D_{vv} = k(v)$. Also the **normalised Laplacians** [18]

$$\mathbf{L}_{\text{sym}} = \mathbf{D}^{1/2} \mathbf{L} \mathbf{D}^{-1/2}, \quad (1.6)$$

$$\mathbf{L}_{\text{RW}} = \mathbf{D}^{-1} \mathbf{L} \quad (1.7)$$

are often used, where the second one is related to the Generic Random Walk discussed in Chap. 3.2.1. These representations allows to describe and analyse graphs with the use of linear algebra, in particular, spectral methods (see for instance [19]). Especially for the Laplacian matrix a number of spectral properties have been found and applied to study, e.g., synchronisation [20, 21] diffusion [18] or graph partitioning [22] (the last application is discussed in Chap. 5.3.1 in the context of detection of modular structure of graphs).

As far as the data storage and computational complexity is concerned, the matrices implemented as arrays are at a disadvantage, since, naively, they use $O(N^2)$ memory and take $O(N)$ time to list the neighbours of a vertex (see, e.g., [23]). Alternatively, a structure of a graph can be stored in the form of an **adjacency list**, which loosely speaking is a collection of lists, each for one vertex, containing the vertices' neighbours. This allows to store the data more space-efficiently, and list the neighbours of a vertex in time proportional to the degree of the vertex. The disadvantage of adjacency lists is a slower time for testing if a given edge exists, depending linearly or at best logarithmically on the degree

of the edge's endpoints. I refer to these algorithmic details, only because they came in handy for the computational part of my work, including comparison of community detection methods; the issues of computational complexity, however, are not discussed further on.

1.3 Structural quantities

In the study of complex networks, vertex degrees are can be regarded as the most fundamental property of a graph, or at least the first to be measured. For a graph of size N , we call the set $\{N_k\}_{k=0,1,\dots,k_{\max}}$ corresponding to the numbers of vertices having a given degree k the **degree sequence** of a graph. In practice, we will also often call $P(k) = N_k/N$ the **degree distribution**. More formally, however, the degree distribution is a probability distribution defined for an ensemble of random graphs, and a degree sequence is just one set of numbers drawn from that distribution that describes a particular instance of a graph found in the ensemble.

The simplest degree distribution is $P(k') = \delta(k' - k)$, which can describe the degrees of any k -regular graph. Below, we will introduce random graph ensembles having, e.g., Poisson degree distribution. Nevertheless, in reality we often encounter what is called **scale-free** networks, which by definition have power-law degree distributions

$$P(k) = \frac{C}{k^\gamma}, \quad \gamma > 0, \quad k \in [0, k_{\max}]. \quad (1.8)$$

In such cases, even though the low-degree vertices are very numerous, there still is a small chance that a vertex with a degree several orders of magnitude greater appears in the graph. It should be stressed that these are not true power-laws in the sense that they always have a finite cut-off $k_{\max} < N$.

These distributions have properties which may make them rather tricky to measure. Firstly, their mean

$$\langle k \rangle = \int_{k_{\min}}^{\infty} kP(k)dk \propto \int_{k_{\min}}^{\infty} k^{-\gamma+1}dk \quad (1.9)$$

diverges for $\gamma \leq 2$. Similarly, their second moment $\langle k^2 \rangle$ diverges for $\gamma \leq 3$. This might pose a problem, since most of the empirical networks have the exponent in the range $\gamma \in [2, 3]$ (see, e.g., Table 2.1 in [24], or Table 3.1 in [17]).

In experiments, of course, the respective moments of the degree distribution have to be finite, since we are able to observe only a network of a finite size. Still, the mean degree of a power-law does not meaningfully characterise the network in the sense that given $\langle k \rangle$ the deviation is still very large and we are likely to observe nodes with degrees several order of magnitude larger, the so called *hubs*. This is one of the blueprints of a scale-free behaviour. Consequently, measuring the whole distribution involves gathering data spanning several orders of magnitude of the observed quantity. Even if this task is manageable, problems may still arise due to large fluctuations in the tail of the distribution.

Among others for these reasons other distributions can be easily mistaken for the power laws, e.g., log-normal distribution $P(k) \propto \exp \left[-\frac{(\ln k - \langle k \rangle)^2}{2\sigma^2} \right]$, a stretched exponential $P(k) \propto \exp \left[-\left(\frac{k}{k_0}\right)^\beta \right]$, or a power law with an exponential cut-off

$$P(k) \propto k^{-\gamma} \exp \left(-\frac{k}{k_0} \right), \quad (1.10)$$

where k_0 is a characteristic degree value above which the probability falls off to zero very rapidly. We mention the last distribution because it is reproduced also in some complex networks models due to constraints on the network's growth [25]. Estimates on natural cut-offs found in uncorrelated networks are discussed in [26–28].

The type of the degree distribution is also decisive in the networks robustness to random attacks [29]. By attacks we mean randomly removing nodes from the network: if only a small fraction of other nodes is disconnected from the

network, the most of the system is still able to communicate; if the network splits into several large clusters, the system fails. These ideas can be precisely defined in the language of percolation (see Chap. 2.2). Suffice to say that for Erdős-Rényi graphs defined in the next section the fraction of removed nodes needed for the percolating cluster to be disconnected is $1 - \frac{1}{\langle k \rangle}$ [cf. equation (2.9)]. This number for $\langle k \rangle = 2$ yields 50%, while for the power-law degree distributed Internet network with $\gamma = 2.5$ more than 90% of nodes need to be destroyed (because it is rather improbable to randomly hit a hub). The effect additionally strongly depends on the exponent γ , so clearly the degree distribution is vital.

Much more than that, it has been shown that in scale-free networks percolation threshold [30] and the threshold for epidemic spreading [31] is absent, which in the latter case means that any non-zero spreading rate (i.e., a ratio of infection and curing rates) results in a finite fraction of nodes invaded by an epidemic outbreak. These results are further elaborated on in the next chapter.

Since many studies are concerned with communication and information transmission, one of the key properties of a network is the distribution of distances between pairs of locations in it. What is called the **average path length** (in fact, the **average intervertex distance** averages only the shortest paths) can be defined as

$$l = \frac{1}{N(N-1)} \sum_{v \neq u} d(u, v), \quad (1.11)$$

where $d(u, v)$ is the distance (i.e., the length of the shortest path) between vertices u and v .

In a square 2D lattice, which one could expect to find in some real networks based on geographical locations, the quantity scales as $l \propto \sqrt{N}$; analogously, in a 3D cubic lattice it is $l \propto \sqrt[3]{N}$. Still, in most of the real networks, the scaling is logarithmic with respect to the number of vertices

$$l \propto \ln(N), \quad (1.12)$$

which is called the **small-world property** [32], also referred to as *six degrees of separation*. It means that information can be passed with the help of very few middlemen; e.g., it should take only around 1.3 more intermediate persons to deliver a message between any two people in the world than between any two people in Poland, even though the ratio of the respective populations is around 182.

In fact, the logarithmic scaling of distances is expected in any infinite dimensional networks, such as Cayley trees and growing trees. In lattices, this can be obtained by introducing a small number of shortcuts between random locations. However, the effect of a small world can be even stronger in scale-free networks with the degree distribution exponent in the range $\gamma \in (2, 3)$, where the scaling is $l \propto \ln(\ln N)$, and as a result they are called *ultra-small worlds* [33, 34].

Another one of the most important properties of simple random graphs, which we take as null models for complex networks, is the fact that they are uncorrelated. By this we mean lack of correlation between the degrees of neighbouring vertices. In order to restate this condition more formally, it is worthwhile to find the distribution of degrees of the nearest neighbours in such networks. To that end, imagine we take at random an edge of the graph and we move along it to reach one of its ends. The probability of thus encountering a vertex of degree k is proportional to the number of such vertices, and so to $P(k)$, as well as to the number of edges we can use to arrive at them k . Together with the normalising constant the distribution is given by

$$Q(k) = k \frac{P(k)}{\langle k \rangle}, \quad (1.13)$$

which leads to the mean degree of a nearest neighbour

$$\langle k \rangle_{\text{NN}} = \frac{\langle k^2 \rangle}{\langle k \rangle}. \quad (1.14)$$

Consequently, the probability of encountering at the two ends of an edge vertices of degrees k and k' is given by a factorised joint distribution

$$P(k, k') = kP(k)k'P(k')/\langle k \rangle^2. \quad (1.15)$$

If we fix the degree distribution in a given network to have the particular form $P(k) \sim k^{-\gamma}$, then some correlations actually have to appear. This is due to the fact that hubs, i.e., the few vertices with the greatest degrees, should on average have multiple connections between themselves. Whereas for $\gamma > 3$ the average is smaller than 1 and no multiple edges form, for $\gamma < 3$ there have to be either loops (i.e., edges pointing to oneself) or multiple edges. The only mechanism for the network to remain a simple graph (with no loops or multiple edges) is to introduce correlations (see Appendix F in [17]).

In real networks, the correlations may take the form of clustering vertices together. Especially in social networks we expect that two friends of ours know each other; as a result a triangle forms between us and the two friends. High density of such triangles is one of the hallmarks of complex networks, and is one of the first tests to be performed when analysing a network. This idea is quantified in several ways by what we call the *clustering coefficient*.

The simplest definition is that of the **global clustering coefficient** (GCC), often written as

$$C = \frac{3 \times \text{number of triangles}}{\text{number of paths of length 2}}, \quad (1.16)$$

so that the number of existing triangles is compared to the number of triangles that it is possible to form in the network.

Perhaps a slightly more popular in the literature of the subject is the definition of the **local clustering coefficient** (LCC) [32] for a node v of degree $k(v)$

$$C(v) = \frac{2t(v)}{k(v)[k(v) - 1]}, \quad (1.17)$$

where $t(v)$ is the number of triangles formed by the vertex v and pairs of its neighbours. Since at maximum there may be $k(v)[k(v) - 1]/2$, the value of LLC is bounded $0 \leq C(v) \leq 1$. To obtain a quantity corresponding to the whole network, as GCC, the local coefficients have to be averaged

$$\bar{C} = \frac{1}{N} \sum_{v \in V} C(v), \quad (1.18)$$

yielding **mean local clustering coefficient** (MLCC). (I use bar instead of angle brackets to differentiate an average over one given network from an average over an ensemble of networks.) In general, GCC and MLCC produce slightly different but strongly correlated results. Using one or the other depends mostly on their analytical tractability, e.g., the numerator in (1.16) can be computed as $\text{Tr}(\mathbf{A}^3)$.

These concepts have been also generalised to weighted and directed networks. There is however no unique way to do that, and several definitions have been proposed [35, 36]. I provide the reader only with the most widespread definition for weighted graphs by Barrat et al. [37]

$$C_W(v) = \frac{1}{s(v)[k(v) - 1]} \sum_{u,w} \frac{W_{vu} + W_{vw}}{2} A_{vu} A_{vw} A_{uw}, \quad (1.19)$$

where $s(v)$ is the strength of a vertex, $\mathbf{A}$ is the adjacency matrix, and $\mathbf{W}$ the weight matrix.

In a similar manner to LCC, [38] have proposed the edge clustering coefficient

$$C(u, v) = \frac{t(u, v)}{\min[(k(u) - 1, k(v) - 1)]}, \quad (1.20)$$

where $t(u, v)$ is the number of triangles that contain the edge (u, v) , and the denominator counts the maximal possible number of triangles that edge could have.

The idea is that edges within communities tend to share more triangles than the edges bridging communities. In community detection algorithms the low-clustering edges can therefore be pruned, leaving the communities – much alike high-betweenness edges, as discussed in Chap. 5.4.

While the triangle is only the shortest cycle, the length of the cycle can be used as a parameter allowing to generalise the idea, and interpolate between local and global network properties.

1.4 Random graphs

In this section, we briefly introduce the basic random graph models: Erdős-Rényi random graphs (together with the binomial model) and the configuration model. These models often serve as null models to be compared with the empirical data, or if needed in algorithms of network analysis. They serve as a basis for benchmark graphs with modular structure described in Chap. 5.2.

The classic **Erdős-Rényi (ER)** model takes a random graph of N labelled nodes and m edges chosen randomly from the set of all $N(N-1)/2$ possible edges [39, 40]. All the $\binom{N(N-1)/2}{m}$ possible graphs form a probability space with each of the graphs being equiprobable.

The ER model is closely related to the *binomial model* or *Gilbert model* [38], in which we take N vertices, and fix to p the probability that a given edge exists. The expected number of edges then is $E(m) = p[N(N-1)/2]$. Hence, the probability that a given labelled graph is obtained yields $p^m(1-p)^{\binom{N}{2}-m}$. It can be shown that the degree distribution of such graphs is approximately

$$P(k) \approx \binom{N-1}{k} p^k (1-p)^{N-1-k} \quad (1.21)$$

and hence for large N it approaches the Poisson distribution

$$P(k) \simeq e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \quad (1.22)$$

When one recalls the results on degree distributions of empirical networks, they are most often power laws. This already shows that ER random graphs for some purposes may be insufficient as null models in network analysis. The ensemble

has, however, the virtue that a number of properties can be exactly calculated for it.

For instance, the clustering coefficient of an ER graph can be obtained by a very simple reasoning. By construction the probability that any pair of vertices is connected is fixed and equal to p . Now, given a vertex and two of its neighbours, the probability that the neighbours are connected is precisely the same p , since in binomial model each edge is formed independently. Hence, the clustering coefficient

$$\bar{C} = C(v) = p = \frac{\langle k \rangle}{N-1}. \quad (1.23)$$

The properties of interest include also the diameter of the graph d or the average path length l . For the regime of connectedness that we are interested in one can show that if $\langle k \rangle \simeq pN \geq c \ln(N)$ for some constant c , then almost surely the diameter of the graph takes one of few possible values around $\frac{\ln N}{\ln(pN)} \approx \frac{\ln N}{\ln \langle k \rangle}$ [41]. It can be expected that the average path length scales similarly, and thus the random graphs have the small-world property.

As far as the topology of the graphs appearing in the binomial model $G(N, p)$ is concerned, it can be shown how increasing p leads to the emergence of more and more extended subgraphs. One of the results is that for the probability $p(N) = cN^{-k/l}$, where c is a positive number, almost every graph contains a subgraph with k nodes and l edges [42]. In particular, for appearance trees and cycles of all sizes the critical probability $p(N) \propto N^{-1}$. This result foreshadows our discussion of a giant connected component appearing in percolation problems in Chap. 2.2.

As noted above, we would like to construct random graphs closer to the real networks in terms of their degree distributions. In fact, the ER ensemble can be generalised into the so called **configuration model** [43], so that the random graphs reproduce almost any degree distribution we demand. Such random

labelled graphs, as they are called in graph theory, form a statistical ensemble whose members are equiprobable. The ensemble is comprised of all the graphs with a given degree distribution $\{N_k\}_{k=0,1,2,\dots}$, where $N_k = NP(k)$ is the number of vertices of degree k . This results in maximally random graphs with a given degree distribution [44, 45].

The construction procedure for a graph from this ensemble, sometimes called *Molloy-Reed construction* [46, 47], is as follows:

- fix the number of vertices N ,
- for each vertex $v \in \{1, 2, \dots, N\}$ draw the number $k(v)$ of *half-edges* (according to the degree distribution $\{N_k\}$) and attach one of their ends to the vertex,
- randomly, pairwise join the remaining ends of half-edges.

The half-edges are often also called *stubs*. Each run of the procedure results in a possibly different graph with the same degree distribution. It is noteworthy, however, that not every run of the procedure ends up in a simple graph; it might be the case that a loop (an edge connecting a vertex to itself) or multiple edges between the same pair of vertices are formed.

In order to calculate certain quantities for graphs from configuration model, it is useful to first determine the probability p_{uv} of an edge existing between a pair of vertices u, v of given degrees $k(u), k(v)$. To estimate it, first note that having chosen one of the $k(v)$ half-edges belonging to v the probability it becomes incident to u is $k(v) / \sum_{w \neq u, v} k(w) \simeq k(v) (\langle k \rangle N)^{-1}$. The probability of the half-edge and u *not* connecting $k(v)$ times is $(1 - k(v) (\langle k \rangle N)^{-1})^{k(v)}$. Finally, the probability that at least one connection takes place is

$$p_{uv} = 1 - \left(1 - \frac{k(v)}{\langle k \rangle N}\right)^{k(v)} \simeq \frac{k(u)k(v)}{\langle k \rangle N}. \quad (1.24)$$

This probability is what we need to calculate the clustering coefficient of the network, since the LCC is equivalent to the probability that given two neighbours of a node there exists a link between them [45]. However, the degree distribution

$P(k)$ is not enough to appropriately average over the neighbours' degrees. We want to use the distribution $Q(k)$ (1.13) describing degree of a vertex that has been arrived at along one of its edges. Now, the mean local clustering coefficient can be calculated [45]

$$\bar{C} = \sum_{\{k\}} \sum_{\{q\}} \frac{(k-1)(q-1)}{\langle k \rangle N} Q(k)Q(q) = \frac{1}{N} \frac{(\langle k^2 \rangle - \langle k \rangle)^2}{\langle k \rangle}, \quad (1.25)$$

where the sums run over the whole degree sequence. Similar results for global clustering coefficient in uncorrelated networks are discussed in [48, 49].

Let us note that the clustering coefficient can now be computed not only for Poissonian distribution of degrees, as for ER graphs, but also for power-law distributions. In such a case, the value of the clustering is much higher, and the difference between ER model and a scale-free configuration model can reach several orders of magnitude (depending on the network size).

Lastly, for the sake of comparison with the graph types mentioned earlier, one can estimate the average intervertex distance l [16]. Given the mean number of nearest neighbours $\langle k \rangle$ and the mean number of second nearest neighbours $\langle k^2 \rangle - \langle k \rangle$, their ratio gives the mean branching coefficient $B = \frac{\langle k^2 \rangle}{\langle k \rangle} - 1$. As a result, at the t -th step away from the starting vertex one can reach $\langle k \rangle B^{t-1}$ more nodes. Consequently, the average intervertex distance is approximately $l \approx \ln N / \ln B$ [44]. The value of l can be made even more precise and include the additive constant [34, 50].

To conclude, in this chapter I have introduced basic graph-theoretical definitions and terminology that will be of use in the following chapters in which certain model networks are analysed. I have also given an overview of the primary quantities, such as degree distribution, average path length, or clustering coefficients, which are needed to characterise and analyse complex networks. Finally, I have briefly described the Erdős-Rényi and configuration model ensembles of random

graphs, which are fundamental to modelling networks with modular structure and to the methods of community detection utilising the ensembles as null models.

Chapter 2

Percolation

Percolation theory can be thought of, as the name suggests, as a mathematical description of a liquid permeating a porous medium. The down-to-earth intuition is that in a rock there are pores that can be filled with the liquid, and that the liquid can leak between the pores through tiny cracks. One of the important questions is how porous the rock has to be for the liquid to leak from one end of the rock to the other. Such wording might appeal to oil mining companies, but a little change of imagery can do justice to physicists or chemists interested in polymerisation and gelation of macromolecules [51, 52], where the filled pores become molecules and the leaky cracks become molecular bonds. When the density of bonds increases the molecules can form larger and larger macromolecules, and ultimately a solid-like gel. Similarly, this image can be translated to a network of vertices and edges that can be permeated by some process. An example of such processes to which percolation theory can be applied is disease spread on networks that can lead to an epidemic, which is the subject of Study I summarised at the end of this chapter.

Before the above illustrations are framed in a mathematical formalism, let me remark that the theory of percolation has been a fruitful area of study particularly with respect to analysis of critical phenomena. Although this will come in handy,

the focus of this brief chapter is on issues connected to Study I as percolation of graphs and the generating function formalism used to solve percolation problems in a class of small-world networks. A comprehensive introduction can be found in [53].

2.1 Basic concepts

It is conceptually easiest to define percolation on a two-dimensional square lattice of size $N = L^2$. First, the vertices and edges of that underlying lattice are unoccupied. Next, we allow the vertices to be occupied with a given probability $p \in [0, 1]$; if two such occupied vertices are neighbours on the underlying lattice they become connected; a group of thus connected vertices forms a *cluster*. We call this process **site percolation**. If instead it is the edges of the lattice that we allow to be occupied with probability p , and we connect the edges that are incident (i.e., they share one of their ends), the respective process is called **bond percolation**. One can also introduce *directed* percolation (where not only edge, but also its direction is randomly drawn), *site-bond* percolation (where both vertices and edges are occupied, in general with different probabilities) and other percolation types.

If the occupation probability p is low, the sites are either isolated or form small connected separate clusters; although if the probability is high enough, the vertices can build up a large cluster comparable in size to the whole lattice. We call such a cluster, whose size $S \sim N$ is finite in the thermodynamic limit $N \rightarrow \infty$, a **percolating/percolation cluster** or **component**. In the context of graphs, the name **giant connected component** is often used. Technically, the definition of a percolating cluster involves existence of a path connecting opposite boundaries of the lattice, which however is equivalent to the one given above in the limit of infinite system size.

The regimes in which percolating cluster is or is not present are separated by a phase transition occurring for a critical value $p = p_c$, which we call **percolation threshold**. The value of p_c depends on the type and dimension of the lattice, as well as on the type of percolation, and can be found in textbooks [53]. For reference, in connection to Study I, we only invoke the exact value $p_c = 1/2$ for bond percolation on a 2D square lattice. As a rule, the value is smaller the higher the number of nearest neighbours (e.g., 6 for 2D triangular lattice as compared with 4 for square), and the higher the dimension ($p_c = 1$ for 1D and $p_c = 0$ for infinite dimensional lattice).

The order parameter for this transition may be defined as the probability $P_\infty = S/N$ that a randomly chosen vertex belongs to the giant component, which is zero for $p < p_c$ as there is no giant component, and greater than zero for $p > p_c$. In particular, close to the critical point it behaves in an analogous way to magnetization in the Ising model of ferromagnetism (e.g., see [54])

$$P_\infty \propto (p - p_c)^\beta, \quad (2.1)$$

where $\beta > 0$ is one of the critical exponents of the phase transition. The other exponents can be defined in a similar manner (cf. Chap. 1.3 in [54]) of which, however, for the sake of brevity we report only on the following two:

$$n_s(p_c) \propto s^{-\tau}, \quad (2.2)$$

which describes the distribution of cluster sizes s (where $Nn_s(p)$ is the number of clusters of size s , excluding the percolating cluster), and

$$\xi \propto |p - p_c|^{-\nu} \text{ for } p \longrightarrow p_c, \quad (2.3)$$

which describes the behaviour of *correlation length* near the threshold. ξ is the characteristic length of correlation function $g(\vec{r}) \propto \exp(-r/\xi)$ that describes the probability of a site at position $\vec{r}$ from an occupied site to belong to the same finite cluster.

The characteristic length is important for practical reasons, since computers can simulate only systems having finite linear size L . This results in any quantity behaving in the thermodynamic limit as $X \propto |p - p_c|^{-\chi} \propto \xi^{-\chi/\nu}$ to change its behaviour for finite L and exhibit *finite-size scaling*

$$X \propto \begin{cases} \xi^{\chi/\nu}, & \text{for } L \gg \xi \\ L^{\chi/\nu}, & \text{for } L \ll \xi. \end{cases} \quad (2.4)$$

Since it might pose problems to find the value of ξ , our focus is on the second line of the proportionality. This effect of dependence on lattice size can be seen in Study I with respect to percolation thresholds for epidemic spread.

Unlike the percolation threshold, the critical exponents do not depend on the lattice topology, but only on its dimension d (for any $d \geq 6$ the behaviour of the system is already the same as for infinite dimension; the respective exponents are called mean-field, and $d_c = 6$ is called the critical dimension). To find some of the critical exponents renormalisation techniques can be employed (see, e.g., Chap. 5.8 in [54]).

2.2 Percolation of graphs

Instead of thinking about percolation in terms of clusters forming on square, triangular, or other d -dimensional lattices, one can take as the underlying network a complete graph with N nodes. The result of bond percolation on such a graph is an instance of an Erdős-Rényi graph taken from the $G(N, p)$ ensemble, where p is the probability of occupying a given edge. A natural question to pose from the percolation perspective is what is the value p_c at which the connected component of size $S \propto N$ appears [40, 55].

For $N \rightarrow \infty$ this question can be rephrased as: what is the condition on which we can travel across a graph, so that each time we leave a vertex via a

different edge than we have entered it. The condition states

$$\sum_k kQ(k) \geq 2, \quad (2.5)$$

where $Q(k)$ has been given in (1.13), and let us recall that it describes the degree distribution of vertices at the end of a randomly chosen edge. This is equivalent to demanding the average number of second nearest neighbours z_2 to be greater than the number of nearest neighbours z_1 .

Hence, the average nearest-neighbour degree for random graphs to have a giant connected component [56] is

$$\langle k \rangle_{\text{NN}} = \frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2. \quad (2.6)$$

For the binomial model of random graphs, whose degree distribution is Poissonian (1.22), the second moment of $P(k)$ is $\langle k^2 \rangle = \langle k \rangle + \langle k \rangle^2$, which means the percolation transition takes place at

$$\langle k \rangle = 1, \text{ that is } p_c = \frac{1}{N-1}. \quad (2.7)$$

This result is identical to the classical result for Bethe lattice (i.e., infinite Cayley tree, as explained in Study II) which is regarded as an infinite dimensional system (further explanation can be found, e.g., in [53] or Chap. 4.3 in [17]). This means that ER graphs below and around percolation threshold are tree-like (i.e., they contain almost no loops). The emergence of the giant connected component is accompanied by divergence of the average size of a finite cluster to which a randomly chosen vertex belongs [44]

$$\langle s \rangle = \frac{\langle k \rangle^2}{2\langle k \rangle - \langle k^2 \rangle} + 1. \quad (2.8)$$

Under the assumption that the random networks are locally tree-like (i.e., they contain no finite cycles) it is possible to put into work the elegant technique of generating functions. Although for somewhat different calculations, this formalism will be used in the next Section. Therefore, here I will focus only on some

specific results that can be obtained with it. Beside the tree approximation, we assume no degree-degree correlations, as stated in (1.13)-(1.15). Given a network satisfying the above assumptions, the question we ask is at which point it *falls apart* if a random fraction of $1 - p$ vertices or edges is removed (note that it is a reverse problem to formation of giant connected component).

The answer is once again given by the ratio of the numbers of first z_1 and second z_2 nearest neighbours [30, 57]

$$p_c = \frac{z_1}{z_2} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \quad (2.9)$$

It can be deduced from this equation that if the second moment of the degree distribution is infinite the percolation threshold is zero, and the giant connected component cannot be destroyed by random vertex removal. (As a side note, it can be easily destroyed by removal of high-degree nodes [29, 58].) As already pointed out in Chap. 1.3, it is the case for power-law distributions with exponent $\gamma \leq 3$. To be more specific, the summary of results for scale-free network [16, 59] is as follows:

- for $4 < \gamma$ ($\langle k^3 \rangle$ is finite), $S \propto p - p_c$
- for $3 < \gamma < 4$ ($\langle k^2 \rangle$ is finite), $S \propto (p - p_c)^{1/(\gamma-3)}$
- for $\gamma = 3$ ($\langle k^2 \rangle$ is divergent), $p_c = 0$ and $S \propto p \exp(-2/(p\langle k \rangle))$
- for $2 < \gamma < 3$ ($\langle k^2 \rangle$ is divergent), $p_c = 0$ and $S \propto p^{1+1/(3-\gamma)}$.

These results also depend on the minimal and maximal (cut-off) degree of the distribution (see Chap. 4.3 in [17]), since the divergence of the second moment is valid only in the limit $N \rightarrow \infty$. For $\gamma \leq 3$, equation (2.9) can be used with the assumption, however, that the cut-off of $P(k)$ is $k_{\max} \sim N^{1/2}$ (for derivation of the cut-offs for different exponents see [26–28]), which yields

$$\begin{aligned} p_c &\sim 1/\ln N & \text{for } \gamma = 3 \\ p_c &\sim N^{-(3-\gamma)/2} & \text{for } 2 < \gamma < 3. \end{aligned} \quad (2.10)$$

It is important to see that the finite size of the network considerably increases percolation thresholds, even though in thermodynamic limit they are zero. Al-

though in Study I the degree distributions of the model small-world networks are approximately Poissonian, what ought to be stressed is the existence of finite-size dependence of percolation thresholds. As remarked in Sec. 2.4 this might be of practical significance to epidemic models.

This brief selection of results proves percolation theory useful in application to random graphs. In Chap. 5.5, we additionally discuss how percolation has been applied to community detection. Below we extend this overview by asking a technically similar question but slightly different in terms of interpretation: not how the network behaves, but how a process behaves on the network.

2.3 Percolation on small-world networks

In [60] the authors study bond percolation on two-dimensional small-world networks. These are an extension of the Watts-Strogatz model [32], which first was able to account for the small-world property. They are constructed by modifying a regular d -dimensional hypercubic lattice, so that either a number of edges is *rewired* (i.e., one of their ends is randomly changed) or *added* to the lattice by connecting random pairs of vertices. In such networks there are $dL^d = dN$ bonds belonging to the underlying d -dimensional square lattice, where L is the linear system size. Additional parameter may be introduced allowing to connect further neighbours along the principal axes; for simplicity, however, we consider only lattices where nodes are linked solely to their nearest neighbours (in geometrical sense), and hence which are $2d$ -regular graphs. To such a system, $dN\phi$ additional edges (called *shortcuts*) are added, where typically $0 < \phi \ll 1$.

The percolation problem, with p being the probability of occupying a given edge, can then be solved with the use of generating functions technique, as presented in [61]. The function

$$H(z) = \sum_{n=1}^{\infty} P(n)z^n \quad (2.11)$$

generates the probabilities $P(n)$ that a randomly chosen vertex belongs to a connected cluster of n nodes other than the percolating cluster. Since $P(n)$ is a probability distribution, its normalisation results in $H(1) = 1$ below the percolation threshold and $H(1) = 1 - S$ above it, where S is the size of the percolating cluster.

The distribution $P(n)$ corresponds to the whole small-world network; we can similarly define probabilities $P_0(n)$ for the underlying lattice (without the added edges). Then, a cluster in the small-world network may consist of several clusters on the underlying lattice connected by the shortcuts. If the probability that a cluster of size n on the lattice has exactly m shortcuts emanating from it is given by $P(m|n)$, the generating function (2.11) satisfies

$$H(z) = \sum_{n=1}^{\infty} P_0(n) z^n \sum_m P(m|n) [H(z)]^m. \quad (2.12)$$

The equation holds in the thermodynamic limit or if the shortcuts do not form loops.

The probability $P(m|n)$ can be expressed by a simple combinatorial formula

$$P(m|n) = \binom{2dN\phi p}{m} \left[\frac{n}{N} \right]^m \left[1 - \frac{n}{N} \right]^{2dN\phi p - m} \quad (2.13)$$

given the total of $2dN\phi p$ ends of the occupied edges and the probability n/N that an end is found in a given cluster of size n . Substituting it into (2.12) and summing over m , in the limit $N \rightarrow \infty$ one obtains

$$H(z) = \sum_n P_0(n) \left[z e^{2d\phi p(H(z)-1)} \right]^n, \quad (2.14)$$

or equivalently

$$H(z) = H_0 \left(z e^{2d\phi p(H(z)-1)} \right), \quad (2.15)$$

where $H_0(z) = \sum_n P_0(n) z^n$. Remembering that $H'(1) = \sum_n n P(n) = \langle n \rangle$ and $H'_0(1) = \langle n \rangle_0$ are the average cluster sizes on the small-world network and the lattice, respectively, we can conclude that $\langle n \rangle$ diverges when

$$2dp_c\phi = \frac{1}{\langle n \rangle_0}, \quad (2.16)$$

which marks the percolation transition.

The authors [60] were able to calculate $H'_0(1)$ with the use of Padé approximants and consequently find the relation between the density of shortcuts ϕ and the percolation threshold p_c .

2.4 Note on epidemic modelled by percolation

The above results are further developed in the Study I to include network dynamics. It is important to note at this point that bond percolation can be translated into SIR model of epidemic spread [62], in which the probability of infection can be expressed in terms of the percolation probability p (see [63] for different models; SIR will be the focus of Study I).

The **Susceptible-Infectious-Recovered** (SIR) model at its simplest consists in three stages of disease transmission in discrete time:

- (i) a vertex is susceptible, i.e., it represents a healthy person who can be infected by a neighbour with probability T (called *transmissibility*)
- (ii) if the vertex has been infected, each turn for the total duration of l time steps it can infect each of its neighbours with probability T
- (iii) if l time steps of being infectious have passed, the vertex is removed and can neither infect others nor be infected again (this can be interpreted, e.g., as death or immunisation).

To start a simulation one has to initially infect a random vertex while all the other vertices are susceptible. The epidemic ends when all infectious vertices have become removed. The total number of removed vertices is then called the size of the outbreak. The transmissibility T can be simply related to the percolation probability p by

$$T = \sum_{t=1}^l p(1-p)^{t-1} = 1 - (1-p)^l \quad (2.17)$$

in the discrete case. The percolation thresholds p_c can therefore be understood in terms of thresholds of infectiousness leading to epidemic outbreaks. The equivalence between SIR and bond percolation has, however, some caveats, as discussed in [64].

Since there is a vast body of theoretical results on percolation, it is worth noting which quantities are more weighty than others from the standpoint of epidemic modelling. Firstly, assuming the model of the network is appropriate, percolation thresholds p_c allow to predict whether there is a risk of epidemic outbreak. Secondly, the size of percolating cluster P_∞ corresponds to the social or economic costs connected with the epidemic. Thirdly, since the real networks tend to be relatively small (e.g., the patients and staff of a hospital), the finite-size effects do play a significant role. Following this train of thought, of practical interest are the results on scaling of the maximum and mean outbreak sizes and durations for certain degree distributions [65].

Somewhat connected to this issue is the distribution of cluster sizes in vicinity of percolation threshold, which describes smaller outbreaks, even though the giant component does not appear. The last issue can be observed in simulations of epidemic models, in which at each run of the simulation only one cluster can form; more clusters would require multiple initial infection sites. The distribution of sizes is thus gathered by rerunning the simulation and changing the position of the initial infection. This is technically slightly different from standard percolation simulations, in which the bonds or sites are all occupied at the same time, and the resulting lattice contains a number of clusters.

In this chapter, after introducing the basic ideas of percolation on lattices, I have discussed several basic results in percolation theory that had been used in the research on complex networks. They concern the critical mean degrees needed to form a connected random graph, but also the critical fraction of edges

needed to destroy a random graph. These results are of importance for instance in the context of robustness of complex networks to random attacks. They also illustrate the ideas and problems relevant to my own research presented in this thesis. From this standpoint, it is important to bear in mind the dependence of percolation thresholds on the finite size of the network. On the other hand, the basic analytical technique implicitly used in Study I is the generating functions formalism described in Sec. 2.3. All these tools have been employed to mathematically describe epidemic outbreaks simulated with the use of the SIR model of disease spread. The results of this study are summarised below.

Study I

The first study to be presented is perhaps the most focused one in terms of applications: its general goal is to examine how epidemic spread, a dynamic stochastic process on its own, is affected by the dynamics of the network. It is on purpose that I have not included in Chap. 1 Barabási-Albert networks [66], as their evolution involves network growth. We preferred to adopt small-world networks of Watts-Strogatz type [32] with two-dimensional underlying lattice, as discussed in Sec.2.3, which is justifiable for systems concerning cultivation or farming. Such a model has a constant number of vertices and edges, but naturally to rewire the *shortcuts*, i.e., the additional random edges that make the network a small-world, without changing the overall topology of the network. Rewiring the shortcuts during the epidemic conveniently lets us extend the existing analytical results of Sec.2.3. As many earlier works our study draws on the equivalence of bond percolation and SIR model of epidemic spread [62]. The problem with analytical modelling of disease spread on a dynamic network, however, is that the dynamic SIR process is mapped onto a conceptually static problem of percolation: we are given a static lattice in which some edges are or are not present. Thus, it is not entirely straightforward how to incorporate dynamics into percolation.

In the paper, we were able to predict analytically lowering of the percolation thresholds for epidemic spread resulting from the dynamics only. Although the dependence on dynamics was analytically tractable, mathematically simple to derive, and stayed in good agreement with simulations, it was in a sense expected and intuitive. What seems more surprising to us from today's perspective is the result of the numerical studies of finite-size effects. More precisely, we measured numerically the dependence of the average size of the epidemic (that is the percentage of population that is infected during an outbreak) on the size of the underlying lattice. While the finite-size effects on regular lattice clearly obeyed (2.4) and were large (in terms of shifting the percolation threshold or equivalently

raising the epidemic size for a fixed transmissibility), on small-world network (no matter static or dynamic) the effects were much smaller and the transition not so sharp (cf. Figures 6 and 7 in paper I). The difference is significant for practical reasons, for it is ultimately the size of the outbreak that determines the social or economic costs of an epidemic. Though at present, the data-driven models of dynamic networks are much richer (see [67] for a review on temporal networks), the epidemic models remain conceptually the same, and the general observations above might still remain valid.

Chapter 3

Random walks

It has been more than a century since Einstein [68, 69] and Smoluchowski [70] gave an explanation of Brownian motion and laid foundations of what is now called the theory of stochastic processes. These advances further lead to defining a model process in which a particle's movements are discretised both in time and space, and it was Pólya who first considered random walk on lattices [71]. In this thesis I restrict myself to such discrete random walks only, although it should be noted that also continuous-time random walks exist or the Wiener process in continuous time and space. These stochastic processes are of immense importance in modelling such microscopic phenomena as diffusion of molecules, transport processes in noisy media, or thermal fluctuations of polymer configurations. They can also describe processes ranging from DNA transcription, to animals' foraging strategies, and to stock price changes.

In the study complex networks random walks (RWs) are used as a proxy of various transport processes, but can also be used in methods analysing network topologies. In this Chapter, I introduce basic properties of RWs in the framework of Markov chains; later, I define several particular types of RWs useful in analysis of complex networks; finally, I discuss at length one of the quantities related to random walks, namely mean first-passage times. The particular RW types

include most notably generic random walk, corresponding to ordinary diffusion, and maximal-entropy random walk, whose properties on graphs are compared in Studies II-V. This is only a selection of topics concerned with random walks on graphs that I used in the Studies; a general introduction to Markov chains and random walks is presented, e.g., in Chap. 11-12 of [72]; more information, especially on topics covering relaxation and mixing times that I refer to can be found in [73].

3.1 Basics of Markov chains

Given a set of events $V = \{v_1, v_2, \dots, v_N\}$, and for each v_i given the set of transition probabilities $P_{ij} \geq 0$ from v_i to v_j (including v_i itself), we call them a Markov chain. The matrix of elements $P_{ij} \forall i, j = 1, \dots, N$ is called the **stochastic, transition, or Markov matrix**, with its rows normalised to one $\forall i : \sum_j P_{ij} = 1$ (which is then called *row-stochastic*).

For any time t , we define a probability vector $\vec{\pi}(t) = (\pi_1(t), \dots, \pi_N(t))^T$ describing the probability distribution over the set of events V . If the initial distribution is $\vec{\pi}(0)$, the distribution after t steps can be obtained from $\vec{\pi}(t)^T = \vec{\pi}(0)^T \mathbf{P}^t$. The probability vector defined as the solution to the equation

$$\vec{\pi}^T = \vec{\pi}^T \mathbf{P}, \quad (3.1)$$

which we call a **stationary state** or **steady state** vector, may be regarded as the probability distribution after infinite time. Let us note that by virtue of Frobenius-Perron theorem for irreducible non-negative matrices this vector is unique. In general, the limiting distribution may not be equal to the stationary state $\vec{\pi}(t) \not\rightarrow \vec{\pi}$ when $t \rightarrow \infty$. For instance, in bipartite graphs it is possible to reach states $\vec{\pi}_1 \neq \vec{\pi}_2$ such that $\vec{\pi}_2^T \mathbf{P} = \vec{\pi}_1^T$, $\vec{\pi}_1^T \mathbf{P} = \vec{\pi}_2^T$, which means the system switches between one and the other, never reaching the stationary distribution. One can define, however, an effective stationary state by averaging the two. Such

situations may appear depending on the choice of the initial distribution $\vec{\pi}(0)$ and can be similarly devised, e.g., in directed networks.

If the events V are identical to the vertices of a graph $G(V, E)$ and we assume that the edges of the graph determine the allowed transitions, then $P_{ij} \leq A_{ij}$, where $\mathbf{A}$ is the adjacency matrix of the graph. For all practical purposes, in this thesis we will call the sequence $\{\vec{\pi}(t)\}_{t=0,1,\dots,\infty}$ a discrete-time **random walk** defined by the stochastic matrix $\mathbf{P}$.¹

In connected graphs random walks are *ergodic* (or *irreducible*) Markov chains, i.e., it is possible to go from any vertex to any other vertex, or more formally $\forall i, j \exists t : (\mathbf{P}^t)_{ij} > 0$ (mark the strong inequality). Additionally, if the order of the quantifiers is reversed $\exists t \forall i, j : (\mathbf{P}^t)_{ij} > 0$, the Markov chain is *regular*, which means that it is possible to get from i to any j in exactly t steps. The latter is not true for example on bipartite graphs, since then one cannot get to any vertex in the same part of the bipartition for any odd number of steps, and to any vertex in the other part of the graph for any even number of steps. Furthermore, we call the chain *reversible* if and only if

$$\forall i, j : \pi_i P_{ij} = \pi_j P_{ji}, \quad (3.2)$$

where the above equation is called *detailed balance condition*. Such a condition ensures that starting from stationary state one cannot distinguish the chain moving forwards or backwards. In this thesis we will deal exclusively with reversible Markov chains. They have a special property that will come in handy in Study IV that their stochastic matrix can be symmetrised

$$\mathbf{S} = \Pi^{1/2} \mathbf{P} \Pi^{-1/2}, \quad (3.3)$$

where Π is a diagonal matrix with $\Pi_{ii} = \pi_i$.

Such symmetry allows to construct spectral representation of $\mathbf{P}$

$$P_{ij} = \pi_i^{-1/2} \pi_j^{1/2} \sum_{\alpha} \Lambda_{\alpha} \Psi_{\alpha i} \Psi_{\alpha j}, \quad (3.4)$$

¹For a more precise and restricted definition, see for example [72].

where $\vec{\Psi}_\alpha$ are eigenvectors of $\mathbf{S}$ associated with the eigenvalues $1 = |\Lambda_0| \geq |\Lambda_1| \geq \dots \geq |\Lambda_{N-1}|$. Remembering that upon taking increasing powers of the stochastic matrix, we come closer and closer to the stationary state, and that taking those powers in spectral decomposition leads to taking increasing powers $\Lambda_\alpha^t = \exp(t \ln \Lambda_\alpha)$, we can define the *asymptotic rate of convergence to stationary state*, often called *relaxation time*

$$\tau = 1 / \ln \Lambda_1 \quad (3.5)$$

(see Chapter 3-4 in [73] for precise definitions of relaxation time, mixing time, average hitting time, and similar concepts).

A particular class of Markov chains are what we call *absorbing* Markov chains. These are chains that contain an event v_i for which $P_{ii} = 1$, i.e., it is impossible to leave it. Such Markov chains might be helpful in constructing mean first-passage time matrices defined in Sec. 3.3. They are also somewhat similar to random-walk models of trapping physical particles [74], for which the quantities such as probability of survival $S(t)$ or return $R(t)$ can be defined, meaning the probability that a random walker is not absorbed after time t and was able to return to initial point after t . These concepts are of more interest to solid state physicists.

In our considerations all random walks are discrete-time, and are defined on finite connected graphs (predominantly undirected). It is also noteworthy that the stochastic matrix $\mathbf{P}$ is constant in time, which in some more general models may not apply.

3.2 Types of random walks

3.2.1 Generic random walk

What we call the ordinary or **generic random walk (GRW)** is defined by the stochastic matrix

$$P_{ij} = \frac{A_{ij}}{k(i)}, \quad (3.6)$$

where $k(i) = \sum_j A_{ij}$ denotes the node degree. It should be noted that in the literature on complex networks the term 'random walk' in fact means this particular definition. The factor $1/k(i)$ in the above definition corresponds to the *uniform probability* of selecting one of $k(i)$ neighbours of the node i . Such a choice maximises the entropy of nearest neighbour selection. The stationary probability distribution of GRW is given by $\pi_i = k(i)/\sum_j k(j) = k(i)/2|E|$.

The choice of uniform transition probabilities corresponds to the standard Einstein-Smoluchowski-Pólya random walk that describes ordinary diffusion. This random walk was first studied on infinite d -dimensional lattices by Pólya, who addressed the problem of recurrence of the random process. For historical reasons I recall his result [71] that the probability of a random walker (which we simply call a particle moving according the random walk) returning to the place where it started its random motion is equal to 1 only in dimensions $d = 1$ and $d = 2$, whereas for $d > 2$ the return probability is smaller than one.

3.2.2 Maximal-entropy random walk

Another type of random walk that is a subject of my research is the **maximal-entropy random walk (MERW)** [75, 76], also called *Ruelle-Bowens random walk* [77]. Instead of maximising the entropy of nearest-neighbour selection, as GRW does, MERW is defined by the condition of maximising the entropy of a set of trajectories (or walks, in graph-theoretical terms) with a given length and end-points; this is a global principle similar to the least action principle. Thus, for

any given length and end-points the stochastic matrix of MERW should render all the trajectories equiprobable. This leads to the following unique matrix

$$P_{ij} = \frac{A_{ij}}{\lambda_0} \frac{\psi_{0j}}{\psi_{0i}}, \quad (3.7)$$

where λ_0 is the largest eigenvalue of the adjacency matrix $\mathbf{A}$, and ψ_{0i} is the i -th element of the corresponding eigenvector $\vec{\psi}_0$. By virtue of the Frobenius-Perron theorem all elements of this vector are of the same sign, because the adjacency matrix $\mathbf{A}$ is irreducible.

That the stochastic matrix (3.7) ensures the equiprobability of paths can be seen by taking a walk $\gamma_{a_0 a_\tau} = (a_0, a_1, \dots, a_\tau)$ of length τ . The probability of visiting this sequence of vertices starting at a_0 and the finishing at a_τ is

$$P(\gamma_{a_0 a_\tau}) = P_{a_0 a_1} P_{a_1 a_2} \cdots P_{a_{\tau-1} a_\tau} = \frac{1}{\lambda_0^\tau} \frac{\psi_{0a_0}}{\psi_{0a_\tau}}, \quad (3.8)$$

which depends on the number of steps and the two end-points only. Hence, the intermediate vertices do not play any role, and all the walks having the same length and end-points are equally probable.

The stationary state of MERW is given by Shannon-Parry measure [78]

$$\pi_i = \psi_{0i}^2, \quad (3.9)$$

which can be interpreted as the probability of finding a particle, described quantum-mechanically as a wave function ψ_{0i} , in the ground state of the operator $-\mathbf{A}$ [75, 76]. This analogy can be drawn further to include a particle propagator $\psi_i G(\mu)_{ij} \psi_j = \pi_i \sum_{t=0}^{\infty} (\mathbf{P}^t)_{ij}$, where

$$\mathbf{G}(\mu) = \sum_{t=0}^{\infty} e^{-\mu t} \mathbf{A}^t, \quad (3.10)$$

where μ is the chemical potential in a grand-canonical ensemble of trajectories with given end-points, with $G(\mu)_{ij}$ acting as a partition function.

In the spectral representation (3.4), the stochastic matrix of MERW takes the form

$$P_{ij} = \frac{\psi_{0j}}{\psi_{0i}} \sum_{\alpha} \frac{\lambda_{\alpha} \psi_{\alpha i} \psi_{\alpha j}}{\lambda_0}, \quad (3.11)$$

where the eigenvalues and eigenvectors of the symmetrised matrix $\mathbf{S}$ (3.3) have been expressed in terms of the eigenvalues λ_α and eigenvectors of $\vec{\psi}_\alpha$ of the adjacency matrix $\mathbf{A}$:

$$\Lambda_\alpha = \frac{\lambda_\alpha}{\lambda_0}, \quad \vec{\Psi}_{\alpha i} = \vec{\psi}_{\alpha i}, \quad (3.12)$$

in particular, $\Lambda_0 = 1$ and $\Lambda_1 = \lambda_1/\lambda_0$, which leads to the relaxation time (3.5) given by $\tau = 1/\ln(\lambda_1/\lambda_0)$.

It is noteworthy that the two types of random walk, GRW (3.6) and MERW (3.7), coincide on k -regular graphs. Nevertheless, in general, their behaviour is very disparate, including their stationary states and dynamics.

3.2.3 Other random walks on graphs

In some applications it is undesirable that a random walker becomes trapped in an isolated part of a graph, e.g., in one of several components. To alleviate that problem the PageRank algorithm [12] takes advantage of a random walk with the so-called *teleportation*

$$P_{ij} = \alpha \frac{A_{ij}}{k(i)} + \frac{1 - \alpha}{N} \quad (3.13)$$

where $\alpha \in [0, 1]$ is the parameter responsible for the teleportation. This means that at each step the random walker can move to any vertex in the graph with probability $1 - \alpha$. It may be seen as adding to the original network a complete graph with some small edge weights. Practically, it allows the random walker to explore even very isolated areas of the graph and, if the graph is disconnected, to move between components. It is said that the parameter α originally had the value 0.85, and it seems that the values between 0.6 – 1 keep the desired balance. Let us note, that $\alpha = 1$ simply reproduces GRW. I discuss stationary state properties of PageRank random walk in relation to mean first-passage times in Sec. 3.3.1 and refer to it also in the context of centrality measures in Chap. 4.

An elegant generalisation of MERW that also allows teleportation is discussed in [79]. The adjacency matrix can be redefined $\tilde{A}_{ij} \rightarrow A_{ij} \exp(U_{ij})$, where U_{ij} can be thought of as an energy of transition along the edge (i, j) . Taking $U_{ij} = 0$ for the existing edges and $U_{ij} = -U_0 < 0$ for the rest, we introduce a small probability of teleportation between any two non-adjacent vertices by redefining $\tilde{P}_{ij} \rightarrow P_{ij} \exp(U_{ij})$ the stochastic matrix (3.7). This results in maximising not the entropy, as before, but the free-energy of paths. The stationary state is once again given by the eigenvectors of the modified adjacency matrix $\tilde{\mathbf{A}}$.

Apart from these, so called **biased random walks** are sometimes considered. The name comes from the fact that the transition probabilities are correlated with a chosen property of vertices or edges. Such biased RWs are contrasted to GRW, whose the transition probabilities are independent of any properties of the target vertices, and thus are considered unbiased. From this point of view, MERW can also be classified as a biased random walk.

The most common example of such RWs is the bias connected to the degree of vertices, as in the case of

$$P_{ij} = \frac{A_{ij}k(j)^\alpha}{\sum_l A_{jl}k(l)^\alpha}, \quad (3.14)$$

where α can take both positive or negative values depending on whether high or low-degree nodes should be preferred [80]. In a similar manner, a bias based on any centrality measure (see Chap. 4) can be introduced [81]

$$P_{ij} = \frac{A_{ij} \exp(\beta c(j))}{\sum_l A_{il} \exp(\beta c(l))}, \quad (3.15)$$

where $c(j)$ is for instance betweenness centrality (4.7) of a vertex j . Likewise, edge centralities can be used.

The principal idea behind these random walks is that real processes, as transport or search in networks, can exhibit a bias themselves. A simple example might be a car driver who wishes to avoid high betweenness routes. Such RWs can be used, e.g., to model and design routing protocols which would prevent traffic

congestion in communication networks. By tuning the bias to a selected graph property one might also think of utilising biased random walks to the analysis of structure of complex networks, including community detection.

3.3 Mean first-passage times

The **mean first-passage time (MFPT)**, also called the average or mean **first hitting time**, is a quantity offering relevant insights into various processes [82], such as kinetics of chemical and biochemical reactions, disease spread, animals' foraging strategies, or target information search. It is extensively used in Studies IV and V, but some of my own observations are presented also here in Sec. 3.3.2 and 3.3.1 below.

First passage time is a concise term for the time it takes a random walker starting from an initial vertex i to reach (or hit) a target final vertex f for the first time. We are not interested in what happens to the random walker after the first hit, therefore, in calculating this quantity the random walk can be treated as absorbing at the vertex f . The matrix, whose element M_{if} encodes the expected (i.e., mean) value of this time for the ordered vertex pair (i, f) , will be called MFPT matrix and denoted by $\mathbf{M}$. It ought to be stressed that this matrix is not symmetric in general; for instance, it might take less time to get from a peripheral node in the graph to a central one than the other way round. Hence, it is worthwhile to remember which matrix index refers to the initial vertex and which to the final one.

The MFPT matrix can be very neatly constructed from the stochastic matrix $\mathbf{P}$ and the stationary state π with the use of the *fundamental matrix* [83]

$$\mathbf{Z} = (\mathbf{1} - \mathbf{P} + \vec{\mathbf{e}}\vec{\pi}^T)^{-1} , \quad (3.16)$$

where $\mathbf{1}$ is the identity matrix, and $\vec{\mathbf{e}} = (1, 1, \dots, 1)^T$. The MFPT matrix is then given by

$$\mathbf{M} = (\mathbf{E}\mathbf{Z}_{\text{dg}} - \mathbf{Z})\Pi^{-1}, \quad (3.17)$$

where $\mathbf{E}$ is a matrix of all ones, $\mathbf{Z}_{\text{dg}}$ is a diagonal matrix with elements $(\mathbf{Z}_{\text{dg}})_{ii} = Z_{ii}$, and similarly $\Pi_{ii} = \pi_i$, as introduced in (3.3).

On expanding $(\mathbf{1} - \mathbf{P})^{-1}$ into Maclaurin series $\mathbf{1} + \mathbf{P} + \mathbf{P}^2 + \dots$ it can be noted that the fundamental matrix is defined so as to contain all the powers of the stochastic matrix $\mathbf{P}$, which enables to average the first passage times over all walk lengths. For the expansion to exist, the matrix $\mathbf{1} - \mathbf{P}$ has to be invertible and its eigenvalues have to lie within a unitary circle. The matrix, however, is non-invertible, as the largest eigenvalue of the stochastic matrix is $\Lambda_0 = 1$. The correction $\vec{\mathbf{e}}\vec{\pi}^T$ subtracts the zeroth mode, and guarantees a well-defined inversion. Instead of the fundamental matrix it is possible to use also other so called *generalized inverses*, the formalism of which is summarized in [84].

3.3.1 Correlation between MFPT and stationary states

As, I have indicated above the PageRank random walk (3.13) is in the limiting case $\alpha = 1$ identical to GRW. In fact, close to that value the teleportation parameter α can be treated as perturbation to GRW. What follows is presentation of how the information contained in the MFPT matrix of a random walk might be used to approximate the stationary state of the random walk with teleportation by the stationary state of the random walk without teleportation (let us call it shortly the pure random walk). A part of the calculations are performed in analogy to the note by Grolmusz [85].

For $\mathbf{P}'$ denoting the stochastic matrix of any random walk, the random walk with teleportation in matrix notation is defined by

$$\mathbf{P} = \alpha\mathbf{P}' + (1 - \alpha)\vec{\mathbf{e}}\vec{\mathbf{v}}^T, \quad (3.18)$$

where $\vec{\mathbf{e}} = (1, 1, \dots, 1)^T$ and $\vec{\mathbf{v}} = \vec{\mathbf{e}}/N$ is the personalisation vector. The form of $\vec{\mathbf{v}}$ might differ depending on the *a priori* information, e.g., in the context of web search it describes the user preferences.

Let us now denote by $\vec{\pi}(0)$ the stationary state of the pure random walk and by $\vec{\pi}(\alpha)$ the stationary state of the walk with a given teleportation α . Our aim is to express the stationary state $\vec{\pi}(\alpha)$ with the use of only the quantities corresponding to the pure random walk:²

$$\vec{\pi}(\alpha)^T - \vec{\pi}(0)^T = (1 - \alpha)[\vec{\mathbf{v}} - \vec{\pi}(0)]^T(\mathbf{1} - \alpha\mathbf{P}')^{-1}. \quad (3.20)$$

On recalling the construction of MFPT matrix, for $1 - \alpha \ll 1$ we can approximate $(1 - \alpha\mathbf{P}')^{-1} \approx \mathbf{Z} + c(\alpha)\vec{\mathbf{e}}\vec{\pi}^T$, where $\mathbb{R} \ni c(\alpha) \xrightarrow{\alpha \rightarrow 1} \infty$. The precise form of c is not important since it cancels out, leaving

$$[\vec{\pi}(\alpha) - \vec{\pi}(0)]_f \approx (1 - \alpha) \left[Z_{ff} - \pi_f(0)(1 + \frac{\bar{m}_f}{N}) \right], \quad (3.21)$$

where $\bar{m}_f = \sum_i M_{if} = \sum_i \frac{Z_{ff} - Z_{if}}{\pi_f(0)}$ is the sum over the f -th column of the MFPT matrix, i.e., over all the initial vertices. Finally, one obtains

$$[\vec{\pi}(\alpha) - \vec{\pi}(0)]_f \approx (1 - \alpha)\pi_f(0) \left[\vec{\pi}_f(0) - \frac{\vec{\mathbf{e}}}{N} \right] \cdot \vec{M}_{*f}, \quad (3.22)$$

where $\vec{M}_{*f}$ is the f -th column of the MFPT matrix. This result means that the stationary state of *any* random walk with small teleportation, as defined in (3.18), is equal to the stationary state of a RW without teleportation plus a term proportional to the mean first-passage time of reaching a given node.

² Equation (3.20) can be obtained by the following series of transformations

$$\begin{aligned} \vec{\pi}(\alpha)^T \mathbf{P} &= \vec{\pi}(\alpha)^T \\ \vec{\pi}(\alpha)^T [\alpha \mathbf{P}' + (1 - \alpha)\vec{\mathbf{e}}\vec{\pi}^T] &= \vec{\pi}(\alpha)^T \\ \vec{\pi}(\alpha)^T (\mathbf{1} - \alpha \mathbf{P}') &= (1 - \alpha)\vec{\mathbf{v}}^T \\ \vec{\pi}(\alpha)^T &= (1 - \alpha)\vec{\mathbf{v}}^T (\mathbf{1} - \alpha \mathbf{P}')^{-1}. \end{aligned} \quad (3.19)$$

Fig. 3.1 illustrates how the above equation works for GRW and PageRank. The stationary states of GRW (i.e., for PageRank with $\alpha = 1$) are proportional to vertex degrees, and thus each horizontal line corresponds to the group of vertices having the same degree. In the subsequent figures, one can see how these lines increase their slope. This is the result of (3.22): for each given line, the nodes with higher total MFPT value, $\vec{e} \cdot \vec{M}_{*f} = \sum_i M_{if}$, are lifted higher than those with smaller total MFPT. This effect can be intuitively phrased in terms of teleportation allowing random walkers to explore the isolated parts of a graph (i.e., those with smaller $\vec{M}_{*f}$).

3.3.2 MFPT and modular graph structure

The properties of MFPT might be helpful in modelling traffic dynamics in communication networks with the use of biased random walks [80]. In (3.17) the multiplication by Π^{-1} causes the columns of MFPT matrix to be proportional to the inverse of stationary state $M_{if} \sim \pi_f^{-1}$. For some random walks this approximation can be improved [80], but the information in a given column is averaged, and thus the whole matrix is flattened to just a vector. In fact, a large amount of structural information on the network, possibly indicative of the community structure, is lost.

Thus, instead of averaging the columns of MFPT matrix M_{*f} (which tells how much time it takes to get *to* a final vertex f from anywhere in the network) we average rows M_{i*} (which is the mean time to get *from* an initial vertex i to anywhere in the network), that is we calculate $\bar{M}_i = \sum_f M_{if}$. Histograms of these values for GRW and MERW in a sample LFR benchmark graph (see Chap. 5.2) have a multimodal structure. In Fig. 3.2, I have roughly separated the consecutive peaks with the vertical strokes.

In order to better visualise the structure of these histograms, we plot the values on a logarithmic scale after subtracting the maximal value $\bar{M}_{\text{MAX}} = \max_i \{\bar{M}_i\}$.

Figure 3.1: Log-log plot of the stationary probability of PageRank $\vec{\pi}(\alpha)$ versus total mean first-passage time $\sum_i M_{if}$ of reaching a vertex f for GRW on a sample network with power-law degree distribution. The case of $\alpha = 1$ is GRW, whose stationary states are proportional to degrees, hence the quantisation. Lowering α makes $\pi(\alpha)$ increase more, if the corresponding MFPT is larger, as expected from (3.13).

Figure 3.2: The histograms show the number of vertices i having a given MFPT value of going from the vertex to anywhere in the network; $\bar{M}_i = \sum_f M_{if}$ and $\bar{M}_{MAX} = \max_i \{\bar{M}_i\}$. MFPT is calculated for (left) MERW and (right) GRW. The vertical strokes are placed roughly in the minima: (left) 10000, 24000, 33500, 39000, 47000, 70000, 80000, 99000, 114000, and (right) 620, 1200, 1610, 1730, 1850, 2400, 3100.

In these histograms one can define intervals of values $\bar{M}_i$ containing significantly separated peaks. Based on the intervals, we divide the vertices of the network. A sample results for GRW are shown in Fig. 3.3, where the vertices are coloured according to the built-in community assignment.

That the average values of rows of the MFPT matrix can indicate such non-trivial structure to our best knowledge has not been shown before; usually, it was the recurrence times that were studied. This result motivated us to further examine the dynamic properties of random walks and their application to analysis of complex networks. Another interesting result connected to mean first-passage times for a special type of biased random walks is shown below.

To close this chapter, I have defined discrete-time random walks on graphs in terms of Markov chains, together with some of the primary quantities that characterise them, as stochastic matrices, stationary states, or relaxation times.

Figure 3.3: The graphs show a sample LFR network (see Chap. 5.2) with 16 built-in communities. The coloured vertices are the ones whose $\bar{M}_i$ lie in the given interval as determined from Fig. 3.2 for MERW. Each colour corresponds to the community assignment of a given vertex. As can be seen, the vertices in the given intervals often belong to just one, or only a few communities, but are not dispersed over the whole network.

Selected types of random walks, whose properties I compare in Studies II-V, have been briefly described in Sec. 3.2. In practice, the choice of a given random walk type depends on the process that is modelled, or the features of the network one wants to detect. In Sec. 3.3, mean first-passage times have been discussed at length, with particular focus on some properties that may be used in the context of centrality measures and community detection in complex networks.

Study II

In this analytical paper, we address the questions how static and dynamic properties of GRW and MERW differ on a simple model graph: a Cayley tree. This is a tree (see Chap. 1.1) constructed in the following way: the first vertex has the degree k (the *root* of the tree), all its neighbours (called jointly the first *generation* of the tree) also have the same degree k , which in the second generation there are $k(k-1)$ vertices all of which also have degree k , and so on, until the last generation (the so called *leaves*) which have no further neighbour and their degree is just 1. In the limit of infinite number of vertices (and generations) the Cayley tree becomes a Bethe lattice. Such a model thus has many symmetries; however, we also extend it a little further and break the symmetry at the root, where we allow the degree to be arbitrarily varied.

Figure 3.4: A Cayley tree with the branching number $k = 2$, except of the root $r = 3$. The last generation $g = 3$ of nodes is comprised of the leaves.

Although such a model is deterministic, and consequently has a number of symmetries, it allows to locally approximate random graphs which are known to be tree-like.

In our paper, we were able to find analytically the eigenspectrum of the adjacency matrix of Cayley trees with the parameters being the root degree r , the degree of other vertices k , and the number of generations of the tree G . Thanks to this result, we derived the exact stationary state of MERW given

by the principal eigenvector of the adjacency matrix (3.9). Interestingly, the stationary state may be localised or not depending on the degrees r and k . These results, compared with the stationary state of GRW, are visualised in an online interactive demonstration [86].

The solution to the eigenproblem of the adjacency matrix also allowed further to reveal differences in the dynamics of the two random walks. Their dynamic behaviour, and in particular their relaxation times, are governed by the second largest eigenvalues Λ_2 of their stochastic matrices, similarly as in (3.5). Knowing the eigenspectrum of $\mathbf{A}$, we were also able to find Λ_2 for the stochastic matrix of GRW. The results show that while GRW relaxes to the stationary state in time proportional to the size of the network $\tau_1 \sim N$ (as expected for a diffusion), the relaxation time of MERW is much faster $\tau_1 \sim \ln N$. The times can additionally be faster depending on the initial probability distribution of a random walk, which is due to the symmetries. These effects are visualised in another online interactive demonstration [87], allowing to vary the initial conditions, branching degrees of the tree, etc.

In a more recent work by Goltsev et al.[88], in this spirit, i.e., using approximate solutions for Cayley trees, the authors were able to calculate when an epidemic on a scale-free network can localise. The localisation is defined in terms of the participation ratio of the principal eigenvector of the adjacency matrix, which corresponds to the stationary state of MERW that we have calculated exactly.

Study III

The third paper included in this thesis is a continuation and extension of Study II, so that it summarises the above results in an orderly manner, but more importantly its aim is to attempt solving static and dynamic behaviour of the two random walks, GRW and MERW, on another model graph: in this case a ladder graph.

Figure 3.5: A ladder graph with 28 nodes and periodic boundary conditions marked by half-edges on both sides. The graph has two intact regions separated by two gaps of size $g = 3$.

This graph is comprised of two one-dimensional rings linked together, so that a vertex v_i is connected to its neighbours $v_{i\pm 1}$, but also to the mirror vertex on the other ring v'_i . Thus, it is a 3-regular graph, but in principle, due to the mirror symmetry, it is a quasi-one-dimensional system. When some edges connecting the two rings are deleted, the equation for the stationary state of MERW is the same as tight-binding equation with repelling potential introduced by the deleted edges. The behaviour of a two dimensional system of that kind can be illustrated by the online demonstration [89]. The system is interesting, since random walkers (moving according to MERW) to be trapped for a longer time in regions between two such deleted edges.

To be more precise, in a graph where two equally-sized regions of that type exist, the relaxation time of GRW is proportional to N^2 as in a normal diffusion and is independent of the number of edges taken out in between these two regions (let us call it g for *gap*). For MERW, however, the relaxation time primarily

depends exponentially on the gap size $\tau_1 \propto \exp(g)$, and so the more isolated the intact regions are, the (exponentially) harder it is for a random walker to move between them. These results were obtained numerically, since, as far, exact analytical treatment has failed. Further modifications of the model were also studied, where the relative sizes of the intact regions and the gaps were varied, however such multiparameter behaviour remains not fully understood.

The results on the relaxation times of such ladder graphs in some way represent the tendency of a given random walk to be trapped in well-connected subgraphs. From the perspective of the analysis of complex networks they provide tentative arguments for using one random walk rather than the other as a model stochastic process taking place on the network. It is especially tempting to check such trapping behaviour on real networks and use it in community detection (see Chap. 5), which we partly tried to assess in Study V.

Chapter 4

Centrality measures

The aim of my research is to analyse processes taking place on networks (which has been done in Studies I-III), as well as analyse networks with the use of these processes (which is the subject of Studies IV-V). The critical assumption in the first of these problems is that the network topology affects the behaviour of the process, and in the second that once the behaviour is known it can be traced back to the network's topology. In a sense, the process is used in a conceptually similar way as a test particle in gravity or particle physics which allows to learn about the geometry of a given field.

Below, I introduce a wide class of quantities called *centrality measures* for which the test process can be a random walk on a graph. As the name suggests, a centrality measure attributes to a vertex or an edge the value of how central it is with respect to the rest of the graph. Such measures are utilised in ranking web pages or in analysis of communication, transport, and social networks. The selection of centralities in Sec. 4.2 includes only the ones related to random walks, which are analysed in Study IV, or, for comparison, the ones that are most widely known, as betweenness and closeness.

What 'central' exactly means, depends on the particular application one has in mind. It might be simply understood as a central position in the sense of some

physical or geographical distance or perhaps as a superior economic or social status. It might be useful to think of centralities as values according to which the vertices or edges can be ranked; one such ranking that is frequently referred to is the list of web pages found by a web search engine. Such rankings can be obtained by simulating a web surfer, or as a matter of fact a random walker similar to PageRank (3.13). To what extent centralities based on different random walks are similar or different has been the topic of Study IV, to which this chapter might be regarded as an introduction. However, due to the numerous interpretations and applications, there is an entire menagerie of centrality measures. For this reason, this chapter is focused predominantly on the ones defined with the use of random walks.

4.1 Basic properties

As far as terminology is concerned, the *centrality measure* is also often referred to as *centrality index*, *importance*, or just centrality. Its minimal defining condition is as follows: given a graph $G = (V, E)$ and the set $X = V$ or $X = E$, let us call a real-valued function $c : X \rightarrow \mathbb{R}$ which is invariant to graph isomorphisms a centrality index. This function is called elsewhere a structural index [90].

There are several properties a centrality index is expected to have. Firstly, since we want to rank the vertices according to their importance, centrality, influence or some other intuitive concepts, we should be able to compare the value of centrality index for *any* two vertices in the graph. Since $c(\cdot)$ is a real-valued function this property is provided by 'less than' relation in $\mathbb{R}$. Additionally, we would rather the centrality did not have the same values for different vertices or edges; in other words, at best we expect that $c(\cdot)$ be a one-to-one function. Of course, in some cases it is an unreasonable requirement, e.g., in complete graphs,

where all the vertices and edges are structurally equivalent. Vertices or edges that are structurally different, however, should be discriminated.

In decision-making processes such grading, which tells which vertex has a higher status as well as which vertices are equivalent, might serve practical purposes. The question of which centrality index can best preserve the structural differences is discussed, e.g., in very recent papers [91, 92].

The structural equivalence and grading, however, depend on the process and the network, to which an appropriate centrality index should be selected. For instance, in *vertex-transitive* graphs all vertices are indistinguishable, i.e., for all pairs of vertices u and v there exists an automorphism $f : V(G) \rightarrow V(G)$ such that $f(u) = v$ and $f(v) = u$. Or in other words, relabelling the vertices produces the same graph. If a graph is not vertex-transitive, some centrality measures may still fail to distinguish some of its vertices [91]. For example, the degree centrality yields the same value for all vertices in a regular graph, regardless of any other features (e.g., belonging to cycles of different lengths). On the other hand, the classic centralities (degree, eigenvector, betweenness, closeness) yield the same result for all vertices in *walk-regular graphs*, which we call all graphs having the property that for all vertices in the graph $v \in V$ the number of closed walks $(v, v_1, \dots, v_{k-1}, v)$ of any length $k \in \mathbb{N}$, where $v_1, \dots, v_{k-1} \in V$, is the same.

These are some limiting cases, which show that for a given problem centrality measures should be chosen with care, as their discriminating power may differ.

4.2 Types of centralities

I do not intend to give a comprehensive catalogue of all the centrality measures that have appeared in the literature. However, a reader unfamiliar with the topic might benefit from recalling some basic definition and terminology. I do therefore provide a short list of classic centralities together with a few less known which are of interest in the context of the studies presented in Study IV. A more

comprehensive catalogue can be found in [90]. In the following, the symbolic names of the centralities will take the form c_N , with the abbreviation of the name in the subscript. The various centralities are defined either with the use of purely structural, graph theoretic concepts as paths and cycles or with the use of some signal or transport processes like random walks or electrical current flows. Below, we do not divide them into those categories, but rather try to present them in the order of their importance and/or similarity to each other.

Let us begin with the simplest structural index one could think of (first introduced in [93]), namely

Degree centrality

$$c_D(v) = k(v), v \in V. \quad (4.1)$$

As simple as it is, it requires a comment: for different applications in-, out-degree, or strength could be taken instead. All of them are clearly local properties and do not bear much information on a vertex's influence on the rest of the graph. Even so, they can be described in terms of a process propagating over the whole network.

In this case, the stationary state of the generic random walk (3.6) is a normalised version of the degree centrality. In general, one can define what we call

Stationary state centrality

$$c_{SS}(v) = \pi_v, \quad \text{where } \vec{\pi} \text{ is a solution to} \quad (4.2)$$

$$\vec{\pi}^T \mathbf{P} = \vec{\pi}^T, \quad (4.3)$$

that is $\vec{\pi}$ is the stationary state of a random walk defined by the stochastic matrix $\mathbf{P}$. Thus, assuming an opinion or information spreads on the network according to GRW, the influence of a node on the network indeed is just the number of neighbours it has. Due to this identity, the name *stationary state centrality* is hardly ever used in the literature. The cause is also that rarely have other

random walks than GRW been used. The centrality, nevertheless, is a function of the stochastic matrix as well $c_{SS}(v; \mathbf{P})$, and other random walks can be used as well.

In fact, a centrality measure based on (3.13) is now almost a household name - the *PageRank*. It is still mostly based on local information spread and yields very similar results to GRW, although they are not quantised into integers (see Fig. 3.1).

Interestingly, another random walk - MERW (3.7) - has a stationary state $\pi_i = \psi_{0i}^2$, where $\vec{\psi}_0$ is the normalised principal eigenvector of the adjacency matrix, which leads us to

Eigenvector centrality [94]

$$c_{EV}(v) = \psi_{0v}, \quad \text{where } \vec{\psi}_0 \text{ is a solution to} \quad (4.4)$$

$$\mathbf{A}\vec{\psi}_0 = \lambda_0\vec{\psi}_0. \quad (4.5)$$

This is a global measure, in the sense that the knowledge of the whole network, represented by $\mathbf{A}$, is needed to calculate the centrality of one node. The random walk interpretation allows to perceive it as a result of an iterative process of information propagation, in which the consecutive steps have probabilities proportional to the centrality of the next node on the way, and the probability of the whole walk from one node to another is weighted by λ_0^l , where l is the walk's length (note the difference between a walk and a path, as defined in Chap. 1.1).

Measuring centrality can thus be understood in terms of counting walks, which are appropriately weighted. Random walks provide only some ways of doing that. For comparison, we invoke the definition of

Estrada's subgraph centrality [95]

$$c_{Est}(v) = \left(\sum_{l=0}^{\infty} \frac{\mathbf{A}^l}{l!} \right)_{vv} = (e^{\mathbf{A}})_{vv}, \quad (4.6)$$

where the walks are weighted by factorials of their length and only *closed walks* are taken into account (hence the index vv in the definition).

More classical approach does not even weigh the walks, it only counts them. However, the set of walks is restricted to the shortest paths only. Let $s(a, b) > 0$ be the number of shortest paths between vertices a and b and let $s(a, b|v)$ be the number of these paths passing through the vertex v .

Betweenness centrality [96] is then defined as

$$c_B(v) = \sum_{\substack{a, b \in V \\ a \neq v \neq b}} \frac{s(a, b|v)}{s(a, b)}. \quad (4.7)$$

Analogously, it can be defined for edges, so that the ratio counts the shortest paths between vertices but they traverse a given edge $e = \{v, w\} \in E$. On assumption that transport processes use only the shortest paths, the betweenness centrality corresponds to the *load* of a vertex or an edge. The process also assumes that a vertex tries to send to other vertices as many packages as there are shortest paths.

The sets of paths can be restricted in other ways, e.g., including only the shortest paths not longer than some value, or including also the paths that are longer than the shortest ones by some value.

If the communication or transport process is stochastic, i.e., the sender (nor any other intermediary node) does not choose any specific route for the package, it has to perform some kind of random walk on the network to finally reach the addressee. For GRW, we show what is known as

Random-walk betweenness [97]. First, we construct a matrix

$$(P_t)_{ij} = P_{ij}, \text{ if } j \neq t, \quad (4.8)$$

with the t -th column zeroed: $(P_t)_{it} = 0, \forall i$, which means that the random walk is absorbing at the node t (random walkers cannot come out of this node). Next, we also remove the t -th row, for it does not affect the other transitions. Now, the probability that the random walker reaches node u starting from s in r steps is: $(\mathbf{P}_t^r)_{us}$. To obtain the probability that it also traverses a given edge incident

to u , the result has to be multiplied by $k(u)^{-1}$. Summing over all possible walk lengths $r = 0, \dots, \infty$ yields the mean number of times the walk has passed an edge incident to u : $k(u)^{-1}((\mathbf{I} - \mathbf{P}_t)^{-1})_{us}$. Finally, this gives

$$\mathbf{V} = \mathbf{D}_t^{-1} \cdot (\mathbf{I} - \mathbf{P}_t)^{-1} \cdot \hat{\mathbf{e}}_s = (\mathbf{D}_t - \mathbf{A}_t)^{-1} \cdot \hat{\mathbf{e}}_s, \quad (4.9)$$

where $\hat{\mathbf{e}}_s$ is a vector with s -th element equal 1, and the rest 0. The betweenness for the given source and target nodes s and t is

$$b^{(st)}(v) = \frac{1}{2} \sum_j A_{vj} |V_v^{st} - V_j^{st}|, \text{ for } v \neq s, t, \quad (4.10)$$

$$b^{(st)}(s) = b^{(st)}(t) = 1, \quad (4.11)$$

and after averaging over all source-target pairs

$$c_{\text{RW}}(v) = \frac{2}{N(N-1)} \sum_{s < t} b^{(st)}(v). \quad (4.12)$$

For GRW, as shown in [97] this centrality is identical to the *current-flow betweenness*. The latter models the network as an electrical grid with the edges having unit resistors, and the current flowing between nodes s and t according to Kirchhoff's law. We are not aware of any studies of random-walk betweenness for other random walks, the above centrality, however, can be easily generalised.

Technically somewhat similar is the centrality based on mean first-passage times introduced in Chap. 3, which is called

Markov centrality or **Random-walk closeness centrality** [98, 99]. It is simply defined as

$$c_M(v) = \frac{N}{\sum_{u \in V} M_{uv}}, \quad (4.13)$$

where $\mathbf{M}$ is the MFPT matrix from (3.17). Thus, this centrality measures inverse total distance between a given vertex and the rest of the network, understood in terms of the average distance a stochastic particle traverses to reach its destination.

In some transportation problems, called facility location problems, minimising the deterministic distances is more reasonable. More precisely, minimising the total distance to all nodes in the network. It is connected to

Closeness centrality

$$c_C(v) = \frac{1}{\sum_{u \in V} d(u, v)}, \quad (4.14)$$

where $d(u, v)$ is the distance between the two nodes. The centrality was discussed in analysis of social networks [100], where the maximum-closeness person could be considered the most central and therefore influential. However, since the interpretation of $d(u, v)$ in such cases is not straightforward, the closeness centrality might be better suited for problems in which it represents a physical distance.

Some other centralities based on geographical distances are mentioned below. The clustering coefficients can be also interpreted as centrality measures, both the vertex and edge clustering [see equations (1.17)-(1.20)].

4.3 Example applications of centrality measures

As regards the geographical structures, one has to take into account also the real Euclidean distances. In [8] Latora et al. used closeness, betweenness, and two other centralities: straightness and information centrality, in analysis of street networks of several cities.

Given the graph, in which vertices correspond to crossings, edges to the streets, and edge weights to the street lengths, one might want to find the most effective transportation routes. The idea of **straightness** is to take into account the degree to which the routes are straight, which supposedly eases transport and reduces the time of travel. For this purpose its definition

$$c_S(v) = \frac{1}{N-1} \sum_{u \neq v} \frac{d_{\text{Eucl}}(u, v)}{d(u, v)}, \quad (4.15)$$

includes the Euclidean distance $d_{\text{Eucl}}(u, v)$, so that the more winding the route is, the smaller the ratio $d_{\text{Eucl}}(u, v)/d(u, v)$.

Information centrality on the other hand is a relative change in efficiency of the graph after node removal [101]:

$$c_I(v) = \frac{e(G) - e(G \setminus \{v\})}{e(G)}, \text{ where} \quad (4.16)$$

$$e(G) = \frac{1}{N} \sum_{u \in G} c_S(u).$$

It appears that the betweenness and information centrality are able to detect the primary communication routes in the urban street networks. Moreover, the distributions of information centrality values indicates whether a city was planned or self-organised (exponential vs. power-law cumulative distribution).

To show a specific example of facility location problems, we refer to the geographical model of the Internet (in the sense of routers) as an evolving network with the new nodes attaching preferentially [102]. For a given new node v , the preference relies on finding

$$u : \min_{u < v} (\alpha d_{\text{Eucl}}(u, v) + c(u)). \quad (4.17)$$

That is the new node v connects to the node that minimises the trade-off between the closest geographical location and the operation costs due to communication delays. One of candidate functions for $c(u)$ is the closeness centrality.

We have already mentioned that betweenness centrality can be understood as load communication or transportation networks. In particular, Goh et al. [103] consider the distribution of this centrality over nodes of a network as the load of routers in data packet transmission. One can imagine that overloading of power grids, computer and routing networks, or urban networks can cause failure, blackouts, traffic jams, congestion, etc. Centrality measures can be applied in modelling such failure cascades in the framework of percolation models, as reviewed briefly in [16].

The cascade-like failures in networks due to attacks and capacity overload was studied in [104].

The simple assumption was that a limiting load

$$L(v) = (1 + \alpha)c(v) \tag{4.18}$$

could be assigned to the nodes in an undamaged network, with $\alpha \geq 0$ a tolerance parameter for an excess load the node can take. The cascading failure can then be modelled as an iterative procedure in which

- (i) a vertex is removed, and betweenness is recalculated (as a result of some shortest paths disappearing and others being redirected)
- (ii) the overloaded nodes are removed

and the procedure is repeated until there are no more overloaded nodes. Such process allows to study the extent of damages in the network, to search for vulnerable nodes in it, etc.

The function $c(v)$ is an initial load that in the absence of empirical data can be modelled by centrality measures. Motter [105] first used for that purpose closeness, degree, and other centralities. He found out that networks are more sensitive to removal of nodes with larger centralities, which can trigger global cascades.

Similar models but with overloaded links, and consequently betweenness playing the role of load, were studied in [106] or in [107] where the Kirchoff's equations of electrical currents were solved for (similarly as in calculation of current-flow betweenness).

Another large branch of applications is the analysis of the Internet (either on the level of autonomous systems or of routers) and the World Wide Web (i.e., the network of web pages) [108]. The research problems connected to the Internet are somewhat similar to designing, e.g., urban networks for car traffic, which can be considered an analogue for routing traffic of information packets. One of the basic quantities studied in this context, in order to understand how to avoid jamming, is betweenness [100]. The analysis of WWW, on the other hand, is mainly focused on ranking the web pages according to their importance (and ultimately also the

user's preferences) by search engines. The most popular algorithms of that kind are Google's PageRank [12] and Hubs & Authorities (HITS) [109], the former being a random walk stationary state centrality, and the latter a centrality based on eigenvectors of $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A} \mathbf{A}^T$ matrices.

In applications, the sensitivity of a centrality index to the presence of complete subgraphs or similar structures in a network may be helpful in detecting artefacts, e.g., so-called link farms in WWW which aim at deceiving ranking of pages in search engines. This issue is discussed in [79], where web page ranking centralities (PageRank, HITS, and stationary state centrality based on MERW) are compared.

Lastly, centrality measures can also form a basis for community detection algorithms, or can enhance them. This particular application is briefly reviewed in Chap. 5.4, after the outline of standard problems and methods of community detection has been presented. It is also at the end of the chapter on community detection that Study IV is summarised, since some of the concepts it utilises include similarity matrices used in community detection methods. The main concern of this study is, nevertheless, the relation between various centralities based on random walks that have been collectively introduced in Sec. 4.2 of the present chapter.

Chapter 5

Community detection

Real networks, or rather real systems represented in a simplified way as networks, be it biological, social or other, tend to have a non-random structure governed by its function. On the one hand, knowledge of the structure may allow to understand or predict the behaviour of the system (i.e., processes taking place in it and its own evolution); on the other, one may try to deduce the function of the elements of the system from the structure. That is why beside looking at the properties of individual vertices, as in the case of centrality indices, massive research has been conducted to reveal the modular structure of networks. In social sciences, the *modules* have for us an intuitive meaning of communities, groups of friends, business interests, etc., and have already been researched for decades [5]. Such studies can allow analysing social changes, targeting groups of customers, or understand the spread of ideas. The interest of biologists in these topics is relatively new [110] and concerns, e.g., metabolic networks, gene regulatory networks, or protein interaction networks. Knowledge of modules in such networks can help discover joint functions of groups of genes or discern the functional modules of proteins taking part in cellular processes.

Because of these and many other applications the discipline has been dynamically developing for the last decade. The results of this research is thoroughly

reviewed in [111]. Below, the intuitions about what a community is are made slightly more precise; next, dominant models of graphs with community structure are presented; and finally, a selection of community detection algorithms is given, together with their connection to percolation, random walks, and centralities. This choice of topics is intended to provide context for the Studies IV and V. Although the former concerns to a large extent centrality measures discussed in the previous chapter, it also makes use of the similarity matrices defined in Sections 5.1 and 5.3.2.

5.1 What is a community?

There are numerous attempts at developing a rigorous definition of a **community** (also called **modules** or **clusters**), some of which provide a formal method for constructing artificial networks with known communities built in (see Sec. 5.2) or a well-defined quality function that can be optimised by community detection algorithms (see Sec. 5.3). Nevertheless, there is no definition universally agreed upon. On the contrary, what is meant by the term 'community' might depend on what the network represents or what is the intended application of the information on community structure.

For that reason, only a few possible choices are presented so that the initial intuitions can be made more precise, and so that the most representative or influential ideas are covered. We begin with a several definitions based on the *local* graph properties. The easiest way to start is to consider as a community a *maximal* complete subgraph H of a graph G also called a clique [112]. The condition of maximality is essential, for we do not expect for instance every 3 out of 5 friends to constitute a separate community, provided the only knowledge we have is that they know each other. Similar definitions based on the concept of a maximal subgraph with a given property have been used, e.g., in which the diameter is bounded by n (*n-clan*)[113] or where each vertex is adjacent

to at least k other vertices in the subgraph (k -core)[114]. These restrictions can be understood as a kind of minimal conditions for connectivity within the community.

However, one has to take into consideration the connections between the examined subgraph and the rest of the graph, which can be done as follows [37]. Let us take a subgraph $H \subset G$ and a node $v \in H$. The degree of the node can be split $k(v) = k(v; H) + k(v; G \setminus H)$, where the parts denote the number of links pointing to vertices in H and outside H , respectively (usually they are denoted by k_v^{in} and k_v^{out} , which however might be confused with the in- and outdegrees of a node in a directed network; for this reason we prefer the notation above). Formally $k(v; H) = \sum_{u \in H} A_{uv}$, and $k(v; G \setminus H) = \sum_{u \notin H} A_{uv}$.

Then, we call the subgraph H

a **Community in the strong sense** if and only if

$$\forall v \in V(H) : k(v; H) > k(v; G \setminus H). \quad (5.1)$$

This means that each node in a strong community has more connections with it than with the rest of the graph.

We call H

a **Community in the weak sense** if and only if

$$\sum_{v \in V(H)} k(v; H) > \sum_{v \in V(H)} k(v; G \setminus H). \quad (5.2)$$

This means that the total number of internal connections of a whole community must be greater than the total number of its external connections. These local definitions can serve well to set conditions for the structure of artificial benchmark graphs with built-in communities, some of which are described in Sec. 5.2. They should be treated with caution, however, when used with respect to some deterministic, especially regular graphs, where they can lead to rather unintuitive results.

On the other hand, global quantities can be found that indicate whether there is any community structure in the network. Such quantities can be *fitness measures* (or *quality functions*), which means that the closer their values are to the optimum, the more distinct communities can be found in a given graph. These functions can be framed in terms of the deviation from a *null model*, i.e., a model which is known to have no community structure, as the ER graph ensemble or the like. Of course, the closer the null model to the examined network the better, since then the deviation can be attributed to the modular structure only. For this reason, the most popular null model is based on randomising the original network, so that the edges are randomly rewired, while keeping the expected degree of each vertex constant. The quality function based on this idea is called **modularity**[115].

For a given partition of the graph into communities, each node v has an assigned membership C_v to one of these communities. The modularity can then be written down as

$$Q = \frac{1}{2m} \sum_{u,v} A_{uv} - p_{uv} \delta(C_u, C_v) = \frac{1}{2m} \sum_{u,v} A_{uv} - \frac{k(u)k(v)}{2m} \delta(C_u, C_v), \quad (5.3)$$

where $m = |E|$ is the total number of edges, p_{uv} is the probability of forming an edge between vertices u and v , which in the second equality has been substituted with the value obtained for uncorrelated networks (1.24), and $\delta(C_u, C_v)$ is one if u and v are members of the same community, and zero otherwise. It should be noted that in fact it is the choice of p_{uv} that sets the null model against which the given network is tested. Due to the delta function we can sum over clusters instead of nodes

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right], \quad (5.4)$$

where n_c is the number of communities in the examined partition, l_c is the number of inter-cluster edges, and $k_c = \sum_{v \in c} k(v)$ is all the degrees in c summed. The two terms in the sum reflect the difference between the fraction of edges in the

original graph and in the null model, where the *expected* degrees of vertices are preserved. The more non-random modular structure a given graph has, the larger the modularity. Thus, given a graph which fixes $\mathbf{A}$ and the degree sequence, the goal of community detection algorithms utilising modularity is to find a partition of the graph that maximises its value (see Sec. 5.3.3) by changing the number of clusters n_c and the assignments C_v .

Another way of assigning vertices to the same community is based on the **similarity** between pairs of vertices: then, the groups of most similar vertices naturally form a community. This can be better understood if the network is embedded in a *metric space*. The central concept for metric spaces is the **distance** function $d(u, v) \geq 0$, whose defining properties are: the coincidence axiom $d(u, v) = 0 \iff u = v$, symmetry $d(u, v) = d(v, u)$, and triangle inequality $d(u, w) \leq d(u, v) + d(v, w)$. Thus intuitively, if vertices are close to each other, they are similar, and if they are distant, they are dissimilar. As a consequence, in the study of complex networks distance-like measures are often called a **dissimilarity**. The dissimilarity is not a distance, because it does not fulfil the coincidence axiom. In what follows, it can be noted that the dissimilarity d_{uv} might be null even though u and v are different vertices. In such cases, the vertices can be called structurally equivalent.

An example comes from image segmentation problems [116], where in the simplest case a number of points on a two-dimensional plane is clustered into two separate groups. Then, the Euclidean distance (or countless others) between any two points can be used as a basis for clustering algorithms. Although such visual information can be mapped into a weighted graph, the converse may not be possible.

Nonetheless, for simple graphs dissimilarities based solely on the adjacency matrix $\mathbf{A}$ can be constructed, e.g., [117]

$$d_{uv} = \sqrt{\sum_{w \neq u, v} (A_{uw} - A_{vw})^2}, \quad (5.5)$$

where the similarity is understood as having the same set of neighbours. Beside such straightforward cases, one can resort to counting paths between vertices, e.g., [118]

$$d_{uv} = \sum_{t=0}^{\infty} \frac{(\mathbf{A}^t)_{uv}}{t!}, \quad (5.6)$$

which in turn views the similarity as a property associated with all other vertices. This idea is close to utilising random walks, in particular powers of their stochastic matrices, mean-first passage times, commute times, and related quantities, which will be described in Sec. 5.3 and which have been analysed in Studies IV and V. It can be observed that most of the vertex-similarity definitions are in fact matrix analogues of some centrality measures listed in Chap. 4.

5.2 Graphs with community structure

As far, graph-theoretical properties of communities in empirical networks are still rather skimpy. We know, however, that for many classes of networks the average path length within communities is typically very small, $l < 3$. It grows approximately logarithmically for small community sizes, until about $n \leq 10$. Above this size the intervertex distance either quickly saturates or the growth becomes even slower. The distribution of sizes of the communities is broad, with the tail that can be fitted with a power-law (see, e.g., [119]).

Since neither ER nor the configuration model has the community structure built in, numerous modified models have been developed to account for this crucial characteristic. The ones presented here are regarded as the classical benchmarks utilised to test algorithms of community detection. The modular structure in such benchmarks is predefined, so that the results of the algorithms can be compared with the planted partition.

One of the simplest ways to produce a random graph model with community structure is the planted l -partition model [120]. In general, it assumes that the

vertices are divided into l groups of equal size. Each pair of vertices within one group is connected with probability p_{in} , while vertices belonging to different groups are linked with probability p_{out} . For $p_{\text{in}} > p_{\text{out}}$ the edge density within groups (intra-cluster) is greater than between them (inter-cluster), and so the whole graph has a community structure. Each group is then a subgraph of ER type (specifically, binomial model graph with $p = p_{\text{in}}$), and consequently the degree distribution is Poissonian, the average shortest path length scales as $\ln n$ within each of them as well.

A special case of this model is Girvan-Newman benchmark [121], where $n = 128$ and $l = 4$ and $\langle k \rangle = 16$. The last condition makes the probabilities p_{in} and p_{out} dependent on each other: $p_{\text{in}}n/l + p_{\text{out}}n(l-1)/l = 16$, where the first term is the mean number of intra-cluster connections z_{in} of a vertex, and the second of inter-cluster connections. For example, $z_{\text{in}} = 14$ results in dense, well-separated clusters, as only the remaining 2 edges per vertex can form bridges between communities. For $p_{\text{in}} = p_{\text{out}}$, or equivalently $z_{\text{in}} \approx 4$, the benchmark becomes a random graph. As simple as it might seem, detection of the transition between graph having community structure and devoid of it is nontrivial. Suffice it to say that community detection algorithms fail to find the correct clusters below $z_{\text{in}} < 8$. This interesting problem is raised again in the conclusions.

Further modifications can be introduced similarly as with the transition from ER to configuration model. Very successful in this respect is the Lancichinetti-Radicchi-Fortunato (LFR) benchmark [122], which not only allows for a power-law degree distribution, but also a power-law distribution of community sizes. The parameter that steers the level of inter-cluster connections is called the *mixing parameter* $\mu \in [0, 1]$, defined as the fraction of edges of a vertex that lead outside of its community.

The construction procedure is similar to the configuration model, as given below:

- fix the number of vertices n
- draw the community sizes from the community size distribution
- for each vertex $v \in \{1, 2, \dots, n\}$ draw the number $k(v)$ of half-edges (according to the degree distribution $P(k)$)
- for each vertex v attach $(1 - \mu)k(v)$ half-edges to it
- for each community randomly, pairwise join the remaining ends of the half-edges
- for each vertex v attach $\mu k(v)$ half-edges to it
- randomly, pairwise join the remaining ends of the half-edges between the vertices of different communities.

The benchmark allows for manipulating the average degree, the power-law exponents of degree and community size distributions, the minimal and maximal sizes of the distributions, and the mixing parameter. The benchmark has been further developed to include directed and weighted networks.

5.3 Community detection algorithms

We give a very brief overview of selected types of community detection methods, particularly the ones connected to random walks. There are many classical data mining methods used for data clustering, as hierarchical clustering algorithms [123] partitional clustering [124] (of which k -means clustering is the most well-known one) graph bipartitioning [125] and many other [126] that, however useful they may be, are beyond the scope of this thesis. Nevertheless, since the main stochastic processes I have researched was MERW, whose definition involves eigendecomposition of the adjacency matrix, it is worthwhile to see what information is stored in the spectral properties of matrices representing graphs. Hence, we begin with a note on spectral methods, and go on with random-walk based methods, and modularity optimisation. Additionally, sections 5.4 and 5.5

describe some other popular methods based on the concepts of centrality and percolation that we reviewed in previous chapters.

5.3.1 Spectral methods

These methods utilise eigenvalues and eigenvectors of matrices representing graphs (see Chap. 1.2). Most often either the Laplacian $\mathbf{L}$ or adjacency $\mathbf{A}$ matrix is used. This is because $\mathbf{L}$ has a number of interesting properties: it always has at least one zero eigenvalue associated with the eigenvector $(1, 1, \dots, 1)$. The number of zero eigenvalues match the number of connected components in the graph. This suggests that the eigenvalues which are close to zero and are visibly separated from the spectrum correspond to fairly well-defined communities, which can be found by examining the associated eigenvectors.

That is why the smallest non-zero eigenvalue (*Fiedler value* or *spectral gap*) and the corresponding eigenvector (*Fiedler vector*) [127] has been used for graph bipartitioning [128] that is cutting the graph into two subgraphs. By repeating the procedure iteratively a set number of times, a partition into several communities can be obtained. Such spectral bisection method [129] consists in finding the minimum cut R (i.e., the minimum number of edges connecting the two groups of vertices)

$$R = \frac{1}{4} \vec{s}^T \mathbf{L} \vec{s} = \sum_i a_i^2 \lambda_i, \quad (5.7)$$

where one minimises over the partition assignment vector $\vec{s}$, and the coefficients a_i are scalar projections $a_i = \vec{s}^T \vec{v}_i$ on the eigenvectors $\vec{v}_i$ of the Laplacian matrix. The solution in the real domain is obtained by the Fiedler vector $\vec{v}_2$, since the Fiedler value is the smallest one. However, the partition assignments s_i take values either 1 or -1 depending on whether the vertex i belongs to one or the other group. In such case, it is sufficient to choose $s_i = \text{sgn}(v_{2i})$, where sgn is the signum function, which is fairly close to the minimum.

Similar techniques can be used with the adjacency or weight matrices [130] not without some caveats, however, for which one can consult [131]. The spectral bisection method has been further developed into more sophisticated spectral clustering algorithms [116, 132, 133] that take into account k eigenvectors, which serve for embedding the graph vertices in a k -dimensional Euclidean space and performing k -means clustering (one of the classic data clustering techniques [124]), that groups the vertices into k clusters.

Interestingly, these methods are linked to random walks, since $\mathbf{L}_{\text{RW}} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{P}$ [see equations (1.4)-(1.6)] where $\mathbf{P}$ is the stochastic matrix of GRW (3.6). As a result it has been proven that if the cut R for a bipartition is properly normalised, it is equal to the probability that a random walker moves from one of the clusters to the other [134]. Unfortunately, we are not aware of any analogous studies on other types of RW.

5.3.2 Random walks

The heuristic arguments behind the random-walks based clustering methods are that since the number of paths connecting vertices in the same community is higher than for vertices belonging to different communities, random walkers are expected to wander longer within the clusters, and are less likely to leave them.

This idea has been elegantly formalised in [135], where the probabilities of moving from one cluster to another in time t are encoded in the clustered autocovariance matrix

$$\mathbf{R}(t) = \mathbf{H}^T(\Pi\mathbf{P}^t - \vec{\pi}^T\vec{\pi})\mathbf{H}, \quad (5.8)$$

where $\mathbf{P}$ is the stochastic matrix of a RW and $\vec{\pi}$ is its stationary state, Π is a diagonal matrix with $\Pi_{ii} = \pi_i$, and $\mathbf{H}$ is a $N \times c$ membership matrix, whose element $H_{vi} = 1$ if vertex v is in cluster i , and 0 otherwise. For a certain time scale t trace of matrix $\mathbf{R}(t)$ is minimised, and then optimal partitions are found by maximising the result.

Earlier studies usually adopted two other approaches: either based on dissimilarity matrices or expansion techniques discussed below (part of this discussion is also included in Study IV and is the main topic of Study V). One of the more effective algorithms is due to [136], where the dissimilarity matrix is defined in a similar manner to (5.5) as

$$d(t)_{uv} = \sqrt{\sum_w \frac{[(\mathbf{P}^t)_{uw} - (\mathbf{P}^t)_{vw}]^2}{\pi_w}}, \quad (5.9)$$

where the random walk defined by stochastic matrix $\mathbf{P}$ and stationary state $\vec{\pi}$ proceeds with t steps from the vertex u and v . The dissimilarity between the two vertices thus in a way measures the difference between how they can see the rest of the network as the RW evolves. The choice of the time t is rather arbitrary, but typically small so that the RW does not come too close to the stationary state. In the same spirit [137–139] use the whole MFPT matrix

$$d_{uv} = \frac{1}{N-2} \sqrt{\sum_{w \neq u,v} [M_{uw} - M_{vw}]^2}, \quad (5.10)$$

and hence the view the vertices have on the rest of the network is time independent (in short, $\mathbf{M}$ sums all powers of $\mathbf{P}$). The clustering methods utilising the above dissimilarity matrices then proceed with the agglomerative hierarchical clustering algorithms of choice. Other particular definitions have also appeared in the literature (see, e.g., [140, 141]), but the above ones can be regarded as the most representative.

The other approach might be called expansion and filtering the edges. Let us say that we track a process of information spread taking place on the network, and each time a piece of information is passed across a given edge, we change the weight of that edge - the more frequent the transmission the larger the weight. After some time the weights can be normalised and the smallest weights can be discarded by the filter. We assume that it is the inter-cluster edges that are filtered out, leaving us with disconnected components corresponding to the communities we have been looking for.

The idea was used for instance in [142], but perhaps a more elegant method was presented in [143]:

- (i) the RW performs t steps (rather small number, so as not to reach the stationary state), which means we take $\mathbf{T} = \mathbf{P}^t$
- (ii) $\mathbf{T}$ is inflated, i.e., each of its elements is raised to some power
- (iii) the rows (if the matrix is row-stochastic) are normalised, so that $\mathbf{T}$ is row-stochastic as well and we can go back to step (i).

Such a process reaches a stationary state with zeros and a single 1 in each row (unless some symmetries are involved). In a row corresponding to v , the position of 1 indicates what is the attracting vertex of v . The set of vertices with the same attractor form a community.

Among other noteworthy methods utilising RWs the Infomap [144] is perhaps the most successful. It strongly differs from the standard approaches presented above; generally speaking, the method treats partition into communities as means of compressing the information needed to describe a process on the network (in this case PageRank random walk). It should be noted, however, that in all these methods GRW is the predominant random walk to be used, not least because of its computational simplicity; the other, biased random walks are quite rare.

5.3.3 Modularity optimisation

Modularity (5.3) has become by far the most popular quality function used in community detection. It was first used in the divisive algorithm by Girvan and Newman [115], which is described in Sec. 5.4 below, to choose the best partition out of those produced by the algorithm. Later on, a multitude of different algorithms have emerged and have been refined, so as to find the most accurate approximation of the optimal modularity value in the least time. Since in fact the idea is rather simple and the null model associated with modularity has already been explained, only briefly a couple of such approaches are mentioned.

The first is a greedy agglomerative clustering method [145], in which the N vertices of the graph are clustered together one by one; each time the vertex to be merged with existing clusters (including single vertices) is chosen so as to maximise the modularity of such subsequent partition. In the end, the partition with the highest modularity from among those that appeared during the process is chosen.

Another classic technique used to optimise modularity is the simulated annealing [146], in which the set of all possible partitions constitutes a space explored by the algorithms. Consecutive visited states (partitions) are chosen with a standard Monte Carlo probability of transition: 1 if the difference of the quality function in the two states is positive $\Delta Q > 0$, and $\exp(\beta\Delta Q)$ if $\Delta Q < 0$, the parameter β being, physically speaking, the inverse temperature. The stationary state of such process is the optimum of the quality function, in this case modularity [147].

Finally, also spectral algorithms can be used by substituting in (5.7) the Laplacian matrix with the modularity matrix [148, 149]

$$B_{uv} = A_{uv} - \frac{k(u)k(v)}{2|E|}. \quad (5.11)$$

As earlier, the procedure follows by choosing a bipartition vector $\vec{s}$ so that its elements ± 1 reproduce the signs of the eigenvector v_1 associated with the largest positive eigenvalue of $\mathbf{B}$.

Whatever the precise method of optimising modularity is, it suffers from several problems. Firstly, the number of high-modularity partitions reaching values very close to the global maximum grows exponentially with the number of modules present in the network [150]. This means that the true maximum is impossible to find even in fairly small graphs, since there are too many high-scoring local maxima in the modularity landscape to be checked. Moreover, even though the algorithms reach values close to the global maximum, different sampling heuristics can lead to significantly different modular structures of the same network. Secondly, in random graphs (e.g., ER type) partitions can be found

to score high modularity values [147, 151], while no communities can be present by definition. This effect is due to fluctuations in distribution of edges in the graph ensemble, but in real networks one cannot tell if the community is meaningful or has appeared as a fluctuation only. Lastly, modularity optimisation has a resolution limit [152], which may result in clustering together a number of small communities (compared to the whole graph), regardless of the density of their inter-cluster connections. Due to these reasons methods based on modularity should be cross-checked against algorithms of other kinds. These problems also raise questions regarding significance of the community detection results as such, which we discuss in Conclusions to the thesis.

5.4 Centrality measures in community detection

Vertex centralities reduce and extract, in this way or the other, the information contained in the adjacency matrix, stochastic matrices, dissimilarity matrices or similar entities (which can be seen explicitly, e.g., in the case of eigenvector or Markov centralities). The full structural information contained in an $N \times N$ matrix is reduced to a vector of N numbers; the structure is lost.

The methods of community detection often use the above matrices to answer more complex (or just more dimensional) questions - to find the modular structure of a network. In a sense, some of them may be considered counterparts of the centrality measures. One of the primary approaches to community detection, however, utilised the residual information stored in centralities [115]. The hierarchical, divisive, clustering algorithm by Girvan and Newman consists in iteratively repeating the following steps:

- (i) calculating edge betweenness,
- (ii) removing the edges with the highest scores,
- (iii) going back to step (i).

After a number of removals, the network splits into components, and the consecutive partitions can be represented in the form of a dendrogram. The process continues until a given number of components is obtained, and the best partition is chosen so as to maximise modularity. For the purpose of step (i) the shortest-paths, random-walk and current-flow edge betweenness were used with similar results.

The choice of centrality, nevertheless, may depend on the specific problem to be solved. Thanks to the application in community detection, numerous modifications to the edge betweenness were introduced [153], e.g., the set of shortest paths was further restricted to *non-redundant* ones only (i.e. each vertex can be used only once as a beginning or an end of a shortest path), which enhanced the performance of the algorithm. The centralities have also been extended to weighted graphs.

Several innovations utilising centrality measures were also introduced to enhance community detection methods to find overlap between communities. For instance, in [154] the algorithm by Girvan and Newman is extended as follows. It consists in an alternative step to edge removal, namely a vertex split: if a vertex v can be split into two, v_1 and v_2 , with the edges incident to it split into two sets, so that the yet virtual edge (v_1, v_2) has more shortest paths traversing it than any real edge, the split should be performed; else, a real edge should be removed in the standard way. The algorithm thus utilises a modified version of shortest-path betweenness, called *split betweenness*, which allows to find the vertices which belong at the same time to several communities.

Feeding some other algorithms with vertex centrality values, e.g., [155] which is based on finding ground states of a frustrated ferromagnetic Ising spin model, allows to restrict the set of pairs of frustration centres (which are supposed to be community centres) from n^2 to $O(1)$. This greatly reduces the computation time. Similar enhancements, involving degree, betweenness or PageRank, were

introduced to some other algorithms [156, 157]. which we provide only examples of.

5.5 Percolation in community detection

In [158] Derenyi et al. generalise the concept of site percolation to k -clique percolation, in which adjacent k -cliques (i.e., ones that share $k - 1$ vertices) play the role of adjacent vertices. Such a definition of adjacency allows to introduce paths and connectedness in the usual manner (see Chap. 1.1). The giant connected component, in the sense of k -clique adjacency, then appears in ER graphs at p equal to

$$p_c(k) = [(k - 1)N]^{-1/(k-1)}, \quad (5.12)$$

which recovers the classical percolation threshold $p_c = 1/N$ for $k = 2$, that is k -cliques representing edges.

We mention this result, as it has been used to define what is called a k -clique community: a union of all k -cliques which are connected in the sense of k -clique adjacency. This allows to obtain overlapping community structure and to develop an algorithm for its detection [159] (called Clique Percolation Method).

I conclude this chapter with the summaries of Studies IV and V below. They are both comparative in character, and the systematic control of the numerical comparison is ensured by the use of LFR benchmark graphs introduced briefly in Sec. 5.2. This allows to produce results for network topologies close to the ones observed in real networks without risking too much bias caused by very definite structures. The core quantities discussed in this chapter that appear in the Studies are dissimilarity matrices of the kind given in (5.5)-(5.5) and (5.9)-(5.10), which encode the structural distance relations between vertices of a given graph, and from which it is moderately easy to extract the structural information.

Although the Studies focus on methods based on random walks, it is not without purpose that we decided to include in the above digest of community detection algorithms also the ones based on spectral decompositions and modularity. While the former is grounded in the classical graph theory, and is popular in image analysis, the latter has become a standard quality function and can be used as auxiliary means to other methods; both, they can serve as a point of reference for other methods.

Study IV

In Chap. 4.2 it has been indicated that a number of centrality indices, including random-walk betweenness, degree or eigenvector centrality, Markov centrality and others, are defined with the use of a random walk (which is by default GRW). Some other centralities can be obtained from the dissimilarity matrices used also in community detection, e.g., those given in (5.6) or (5.9), since they are all based on concepts of walks, shortest or weighted paths, etc. In these contexts, several observations regarding MERW has lead me to believe that some of the above approaches can be unified or at least some analogies between them can be drawn.

The aim of the paper was therefore to compare, both analytically and numerically, a number of centrality indices and a couple of dissimilarity matrices. If these quantities could be shown to be related, I thought, one could assume that the structural information about the network and the results they produce are largely equivalent. Thus, the idea was primarily reductive and synthetic.

In the paper, both MERW and GRW were shown to belong to a special class of random walks whose stochastic matrix can be symmetrised as in (3.3). This allowed me to draw analogies between one of the dissimilarity matrices (5.9) (proposed in [136]) and the MFPT matrix (3.17). In particular, it appears that the fundamental matrix $\mathbf{Z}$ (3.16), which is used to construct MFPT matrix, is in fact an unsymmetric version a particle propagator $\mathbf{G}$ (3.10) associated with a given random walk. The latter is in turn closely related to what we call a heat kernel $e^{\beta\mathbf{A}}$, proposed as a dissimilarity matrix (5.6) in [118].

I further studied the connection between several random-walk based centralities, including the eigenvector centrality (4.4), which is associated with the propagator mentioned above, as well as centralities based on powers of a RW's stochastic matrix, stationary states of RWs and Markov centrality (4.13). I compared the resulting values of centrality on a set of benchmark graphs (LFR

benchmarks were used, since at present they seem to reflect most closely the structure of real networks). The results of comparison in the form of a dendrogram (Fig. 2 in IV) show how close the different centralities, understood as vectors of real numbers, are to each other. In particular, it can be seen that centralities based on MERW give consistently distinct results from all the other methods, while they are very similar to each other. Centralities associated with GRW are slightly less similar to each other and mix with some other methods.

In sum, I was able to unify some approaches to calculating centralities, which I hope will help to synthesise part of the existing knowledge on complex networks analysis. Additionally, the choice of a random walk was once again shown to be important in the methods of analysis. More generally, the connection between spectral properties of adjacency matrices, counting paths, calculating mean first-passage times or other quantities associated with random walks is an argument for one mathematically favoured structural information on networks that could be extracted irrespective of the method one assumes. I raise this argument again in the Conclusions to the thesis, where I discuss my ongoing and prospective research.

Study V

Random-walk based methods are one of the successful branches of community detection. Both their strength and weakness at the same time is the fact that the resulting partition may depend on the chosen RW: on the one hand, this allows to model a given process as closely as possible with the appropriate RW, and thus obtain results relevant to the particular system under study; on the other hand, there exists a certain preconception (not necessarily a misconception) that all community detection methods should give similar results approximating some ideal graph partition. The latter is precisely the case, if one assesses the methods on a set of benchmark graphs constructed so as to have a certain predefined community structure known to us.

With these thoughts in mind, in our paper we aimed at extending the existing RW-based methods of community finding with a stochastic process other than just the classic GRW. Since in the previous studies (II-IV) we have shown that MERW has both static and dynamic properties strongly contrasting it with GRW, as well as by definition it contains some spectral information (which, as discussed in Sec. 5.3.1, is one of the basic means of graph partitioning), we substituted one RW for the other in several community detection methods. The aim was to systematically compare their performance. To that end we used the popular benchmark graphs by Lancichinetti, Fortunato, and Radicchi [122], briefly introduced earlier in Sec. 5.2. One thus has to bear in mind that the partitions generated by the algorithms were compared to the communities defined in terms of local quantities (inter- and intra-cluster degrees).

The comparison was performed on several algorithms, some of which had not been previously checked on benchmark graphs. We spotted one algorithm, based on the dissimilarity matrix (5.10), that performed below expected standards; other algorithms from [136, 142] and one adapted from [118] were found to do reasonably well. In the best algorithms, the two random walks gave comparable

results with no clear preference for any of them. These are only some practical conclusions on which method or which RW is better or worse with respect to the given set of benchmark graphs. With the benefit of hindsight, however, one could raise further questions, namely, is there really one favoured graph partition irrespective of the random walk or do the algorithms which we call 'good' ones somehow suppress the RW-dependent information.

Conclusions

In the course of my doctoral studies I have explored the topics of epidemiological modelling and percolation, static and dynamic properties of random walks, and the analysis of structure of complex networks; the problems I have tackled were both analytical and numerical, and although my studies have remained to a great extent theoretical in nature, the emphasis has shifted towards applications. In Paper I, I studied disease spread modelled by percolation on Watts-Strogatz type of small-world networks; beside numerical study of finite-size effects, known analytical results have been extended to networks with dynamically rewired edges, and showed lowering of epidemic thresholds. In Papers II and III, I calculated stationary states and relaxation times of the generic and maximal-entropy random walks on two model network types: on Cayley trees the results are analytical, and on ladder graphs they are numerical; they show that the two random walks have very disparate properties, with MERW being able to localise and be trapped in parts of the network. In Papers IV and V, I study the relation between a number of dissimilarity matrices based on random walks; the results of the former paper unify some of the random-walks based centralities; in the latter paper, I conduct a systematic comparison between the generic and maximal-entropy random walks utilised in community detection methods.

At this point of the thesis, I believe it is worthwhile to look a little bit farther ahead and outline the perspective of future research. The advances in the discipline of complex networks are very dynamic, and the new models and

empirical studies are brought closer and closer to each other; now, it is not only the undirected, directed, and weighted networks that are simulated or examined, but also temporal networks or multiplex¹ networks. These new entities often require qualitatively new methods of treatment and ways of defining processes like random walks [160]. Similarly, the range of questions raised has extended, so that for example the communities to be detected can have internal hierarchical structure or can overlap [161].

Even though these novel paths of research have already emerged, there are still questions that are only partly answered even for undirected graphs. One of them, which is of my personal interest and can be regarded as a natural continuation of the studies presented in this thesis, is the relation between existence of a statistically significant community structure in a network and the possibility of any algorithm to detect it. This problem has been initiated in [162], where a phase transition from detectable to undetectable community structure has been observed, and which recently has been discussed [163–165] by examining spectral properties of the modularity matrix. The problem can be approached also from the random-walk centred viewpoint of this thesis, leading to the conclusion that quantities constructed from stochastic and mean first-passage time matrices can detect the transition irrespective of the random walk chosen. Moreover, this behaviour seems to be attributable simply to the shape of spectrum of the adjacency matrix of a graph, and can be observed for as simple quantities as clustering coefficients (1.16)-(1.19).

As a final commentary, in my view the studies presented in the thesis were governed by two ideas: one is simplicity – however paradoxical it may sound in the context of science of complex networks – which resulted in using unsophisticated but, hopefully, effective mathematical techniques; the other is similarities, which

¹Briefly, in multiplex networks, for a given set of vertices there exist multiple layers of connections.

I strove to find between different concepts and approaches, and which lead to, once again hopefully, a meaningful synthesis of knowledge.

Acknowledgements

I owe a deep debt of gratitude to my thesis advisor, professor Zdzisław Burda, for his kindness, constant support, and, to put it simply – goodness. He has helped me out many a time during the four years of doctoral studies; he has always encouraged me to follow my own research interests; he has shared with me his knowledge and experience. I would also like to thank dr hab. Paweł F. Góra who was my first mentor, and has ignited my interest in complex systems.

I am sincerely thankful to Bartek Waclaw for his strong commitment and mentoring during our joint research. I am grateful to prof. Wolfhard Janke and the whole CQT group from Universität Leipzig for their hospitality and friendliness during almost a year of my stays in Leipzig. Warm and heartfelt thanks go to a couple of Leipzigers thanks to whom I felt there at home.

Many sincere thanks are due to Marcin Zagórski and several other Ph.D. Students with whom, during the four years in the Institute of Physics, I have shared: thesis advisors, rooms, floors, courses, conferences, and time spent on many friendly conversations.

Lastly, my love and thanks go to my Brother and Sister, who have supported me in many ways, to Kasia, and to many Friends indeed.

The project operated within the Foundation for Polish Science International Ph.D. Projects Programme cofinanced by the European Regional Development Fund covering, under Agreement No.MPD/2009/6, the Jagiellonian University International Ph.D. Studies in Physics of Complex Systems.

Bibliography

- [1] A.-L. Barabasi, *Linked: How Everything Is Connected to Everything Else and What It Means* (Plume, 2003).
- [2] A. Bavelas, “A mathematical model for group structures,” *Human organization* **7**, 16–30 (1948).
- [3] A.-L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nature Reviews Genetics* **5**, 101–113 (2004).
- [4] E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems,” *Nature Reviews Neuroscience* **10**, 186–198 (2009).
- [5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Vol. 8 (Cambridge University Press, 1994).
- [6] P. M. Carron and R. Kenna, “Universal properties of mythological networks,” *Europhysics Letters* **99**, 28002 (2012).
- [7] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, “Efficient routing on complex networks,” *Phys. Rev. E* **73**, 046108 (2006).
- [8] P. Crucitti, V. Latora, and S. Porta, “Centrality measures in spatial networks of urban streets,” *Phys. Rev. E* **73**, 036125 (2006).
- [9] M. Camitz, *Computer Aided Infectious Disease Epidemiology - Bridging to Public Health*, Ph.D. thesis, Karolinska Institutet, Solna, Sweden (2010).
- [10] J. J. Xu and H. Chen, “Crimenet explorer: a framework for criminal network knowledge discovery,” *ACM Trans. Inf. Syst.* **23**, 201–226 (2005).
- [11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
- [12] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer Networks and ISDN Systems* **30**, 107 – 117 (1998), proceedings of the Seventh International World Wide Web Conference.
- [13] R. J. Wilson, *Wprowadzenie do teorii grafów* (Wydawnictwo Naukowe PWN, Warszawa, 2007) [In Polish].
- [14] R. Diestel, *Graph theory* (Springer-Verlag, Heidelberg, 2010).

-
- [15] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Review of Modern Physics* **74**, 47–97 (2002).
 - [16] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, “Critical phenomena in complex networks,” *Rev. Mod. Phys.* **80**, 1275–1335 (2008).
 - [17] A. Fronczak and P. Fronczak, *Świat sieci złożonych: Od fizyki do Internetu* (Wydawnictwo Naukowe PWN, Warszawa, 2009) [In Polish].
 - [18] F. R. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics, Vol. 92 (American Mathematical Society, Providence, USA, 1997).
 - [19] P. V. Mieghem, *Graph Spectra for Complex Networks* (Cambridge University Press, New York, NY, USA, 2011).
 - [20] M. Barahona and L. M. Pecora, “Synchronization in small-world systems,” *Phys. Rev. Lett.* **89**, 054101 (2002).
 - [21] T. Nishikawa, A. E. Motter, Y.-C. Lai, and F. C. Hoppensteadt, “Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize?” *Phys. Rev. Lett.* **91**, 014101 (2003).
 - [22] A. Pothen, “Graph partitioning algorithms with applications to scientific computing,” in *Parallel Numerical Algorithms*, ICASE/LaRC Interdisciplinary Series in Science and Engineering, Vol. 4, edited by D. Keyes, A. Sameh, and V. Venkatakrishnan (Springer Netherlands, 1997) pp. 323–368.
 - [23] C. E. Leiserson, R. L. Rivest, C. Stein, and T. H. Cormen, *Introduction to algorithms* (The MIT press, 2001).
 - [24] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” *Physics Reports* **424**, 175–308 (2006).
 - [25] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, “Classes of small-world networks,” *Proceedings of the National Academy of Sciences* **97**, 11149–11152 (2000).
 - [26] Z. Burda, J. D. Correia, and A. Krzywicki, “Statistical ensemble of scale-free random graphs,” *Phys. Rev. E* **64**, 046118 (2001).
 - [27] Z. Burda and A. Krzywicki, “Uncorrelated random networks,” *Phys. Rev. E* **67**, 046118 (2003).
 - [28] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, “Correlations in interacting systems with a network topology,” *Phys. Rev. E* **72**, 066130 (2005).
 - [29] R. Albert, H. Jeong, and A.-L. Barabasi, “Error and attack tolerance of complex networks,” *Nature* **406**, 378–382 (2000).
 - [30] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, “Resilience of the internet to random breakdowns,” *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
 - [31] R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
 - [32] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature* **393**, 440–442 (1998).

-
- [33] R. Cohen and S. Havlin, “Scale-free networks are ultrasmall,” *Phys. Rev. Lett.* **90**, 058701 (2003).
 - [34] A. Fronczak, P. Fronczak, and J. A. Holyst, “Average path length in random networks,” *Phys. Rev. E* **70**, 056110 (2004).
 - [35] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, “Intensity and coherence of motifs in weighted complex networks,” *Phys. Rev. E* **71**, 065103 (2005).
 - [36] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, “Generalizations of the clustering coefficient to weighted complex networks,” *Phys. Rev. E* **75**, 027105 (2007).
 - [37] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, “Defining and identifying communities in networks,” *Proceedings of the National Academy of Sciences of the USA* **101**, 2658–2663 (2004).
 - [38] E. N. Gilbert, “Random graphs,” *The Annals of Mathematical Statistics* **30**, pp. 1141–1144 (1959).
 - [39] P. Erdős and A. Rényi, “On random graphs,” *Publicationes Mathematicae Debrecen* **6**, 290–297 (1959).
 - [40] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–61 (1960).
 - [41] F. Chung and L. Lu, “The diameter of sparse random graphs,” *Advances in Applied Mathematics* **26**, 257 – 279 (2001).
 - [42] B. Bollobás, *Random graphs*, Vol. 73 (Cambridge university press, 2001).
 - [43] B. Bollobás, “A probabilistic proof of an asymptotic formula for the number of labelled regular graphs,” *European Journal of Combinatorics* **1**, 311–316 (1980).
 - [44] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Phys. Rev. E* **64**, 026118 (2001).
 - [45] M. E. J. Newman, “Random graphs as models of networks,” in *Handbook of graphs and networks*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Berlin, 2003) pp. 35–68.
 - [46] M. Molloy and B. Reed, “A critical point for random graphs with a given degree sequence,” *Random Structures & Algorithms* **6**, 161–180 (1995).
 - [47] M. MOLLOY and B. REED, “The size of the giant component of a random graph with a given degree sequence,” *Combinatorics, Probability and Computing* **7**, 295–305 (1998).
 - [48] Z. Burda, J. Jurkiewicz, and A. Krzywicki, “Network transitivity and matrix models,” *Phys. Rev. E* **69**, 026106 (2004).
 - [49] Z. Burda, J. Jurkiewicz, and A. Krzywicki, “Perturbing general uncorrelated networks,” *Phys. Rev. E* **70**, 026106 (2004).
 - [50] A. Fronczak, P. Fronczak, and J. A. Holyst, “How to calculate the main characteristics of random uncorrelated networks,” in *AIP Conf. Proc.* **776** (2005).

- [51] P. J. Flory, “Molecular size distribution in three dimensional polymers. i. gelation1, ii. trifunctional branching units, iii. tetrafunctional branching units,” *Journal of the American Chemical Society* **63**, 3083, 3091, 3906 (1941).
- [52] W. H. Stockmayer, “Theory of molecular size distribution and gel formation in branched chain polymers,” *Journal of Chemical Physics* **11**, 45–55 (1943).
- [53] A. Aharony and D. Stauffer, *Introduction to percolation theory* (Taylor and Francis, London, 1994).
- [54] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman, *Zjawiska krytyczne: wstęp do teorii grupy renormalizacji* (Wydawnictwo Naukowe PWN, Warszawa, 1998) [In Polish].
- [55] R. Solomonoff and A. Rapoport, “Connectivity of random nets,” *The bulletin of mathematical biophysics* **13**, 107–117 (1951).
- [56] M. Molloy and B. Reed, “A critical point for random graphs with a given degree sequence,” *Random Structures & Algorithms* **6**, 161–180 (1995).
- [57] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Network robustness and fragility: Percolation on random graphs,” *Phys. Rev. Lett.* **85**, 5468–5471 (2000).
- [58] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, “Breakdown of the internet under intentional attack,” *Phys. Rev. Lett.* **86**, 3682–3685 (2001).
- [59] R. Cohen, D. ben Avraham, and S. Havlin, “Percolation critical exponents in scale-free networks,” *Phys. Rev. E* **66**, 036113 (2002).
- [60] M. E. J. Newman, I. Jensen, and R. M. Ziff, “Percolation and epidemics in a two-dimensional small world,” *Phys. Rev. E* **65**, 021904 (2002).
- [61] C. Moore and M. E. J. Newman, “Exact solution of site and bond percolation on small-world networks,” *Phys. Rev. E* **62**, 7059–7064 (2000).
- [62] P. Grassberger, “On the critical behavior of the general epidemic process and dynamical olation,” *Mathematical Biosciences* **63**, 157 – 172 (1983).
- [63] I. Nåsell, “Stochastic models of some endemic infections,” *Mathematical Biosciences* **179**, 1 – 19 (2002).
- [64] E. Kenah and J. M. Robins, “Second look at the spread of epidemics on networks,” *Phys. Rev. E* **76**, 036113 (2007).
- [65] E. Ben-Naim and P. L. Krapivsky, “Size of outbreaks near the epidemic threshold,” *Phys. Rev. E* **69**, 050901 (2004).
- [66] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science* **286**, 509–512 (1999).
- [67] P. Holme and J. Saramäki, “Temporal networks,” *Physics Reports* **519**, 97–125 (2012), temporal Networks.
- [68] A. Einstein, “Zur theorie der brownschen bewegung,” *Annalen der Physik (Leipzig)* **324**, 371–381 (1906).

-
- [69] A. Einstein, “Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen,” *Annalen der Physik* (Leipzig) **322**, 549–560 (1905).
 - [70] M. Smoluchowski, “Zur kinetischen theorie der brownischen molekularbewegung und der suspensionen,” *Annalen der Physik* (Leipzig) **326**, 756–780 (1906).
 - [71] G. Pólya, “Über eine aufgabe der wahrscheinlichkeitsrechnung betreffend die irrfahrt im straßennetz,” *Mathematische Annalen* **84**, 149–160 (1921).
 - [72] C. Grinstead and L. Snell, *Introduction to probability* (American Mathematical Society, Rhode Island, NJ, 2000).
 - [73] D. Aldous and J. Fill, *Reversible markov chains and random walks on graphs*. (2002) [Monograph in preparation].
 - [74] J.-M. Luck, *Systemes désordonnés unidimensionnels* (Commissariat de l’Energie Atomique, 1992) [In French].
 - [75] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, “Localization of the maximal entropy random walk,” *Phys. Rev. Lett.* **102**, 160602 (2009).
 - [76] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, “The various facets of random walk entropy,” *Acta Physica Polonica B* **41**, 949 (2010).
 - [77] D. Ruelle, *Thermodynamic formalism: the mathematical structure of equilibrium statistical mechanics* (Cambridge University Press, 2004).
 - [78] W. Parry, “Intrinsic markov chains,” *Transactions of the American Mathematical Society* **112**, 55–66 (1964).
 - [79] J.-C. Delvenne and A.-S. Libert, “Centrality measures and thermodynamic formalism for complex networks,” *Phys. Rev. E* **83**, 046117 (2011).
 - [80] A. Fronczak and P. Fronczak, “Biased random walks in complex networks: The role of local navigation rules,” *Phys. Rev. E* **80**, 016107 (2009).
 - [81] V. Zlatić, A. Gabrielli, and G. Caldarelli, “Topologically biased random walk and community finding in networks,” *Phys. Rev. E* **82**, 066109 (2010).
 - [82] S. Redner, *A guide to first-passage processes* (Cambridge University Press, 2001).
 - [83] J. G. Kemeny and J. L. Snell, *Finite markov chains*, Vol. 210 (Springer-Verlag New York, 1976).
 - [84] J. J. Hunter, “Generalized inverses, stationary distributions and mean first passage times with applications to perturbed markov chains,” *Res. Lett. Inf. Math. Sci.* **3** (2002).
 - [85] V. Grolmusz, “A note on the PageRank of undirected graphs,” arXiv:1205.1960v2 [cs.DS] **3** (2012).
 - [86] J. K. Ochab, “Stationary states of maximal entropy random walk and generic random walk on cayley trees,” Wolfram Demonstration Project, <http://demonstrations.wolfram.com> (2012), [Online interactive presentation].
 - [87] J. K. Ochab, “Dynamics of maximal entropy random walk and generic random walk on cayley trees,” Wolfram Demonstration Project,

- <http://demonstrations.wolfram.com> (2012), [Online interactive presentation].
- [88] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes, “Localization and spreading of diseases in complex networks,” *Phys. Rev. Lett.* **109**, 128702 (2012).
 - [89] B. Waclaw, “Generic random walk and maximal entropy random walk,” Wolfram Demonstration Project, <http://demonstrations.wolfram.com> [Online interactive presentation].
 - [90] D. Koschützki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, “Centrality indices,” in *Network Analysis*, Lecture Notes in Computer Science, Vol. 3418, edited by U. Brandes and T. Erlebach (Springer Berlin Heidelberg, 2005) pp. 16–61.
 - [91] E. Estrada, “Open problem: the discriminant power of the subgraph centrality and other centrality measures,” *Discrete Applied Mathematics*. To appear. (Preprint arXiv:1305.6836 [cs.SI]) (2013).
 - [92] M. P. Rombach and M. A. Porter, “Discriminating power of centrality measures,” arXiv:1305.3146 [cs.SI] (2013).
 - [93] C. H. Proctor and C. P. Loomis, “Analysis of sociometric data,” in *Research methods in social relations*, Vol. 2, edited by M. Jahoda, M. Deutsch, and S. W. Cook (Dryden Press, NewYork, 1951) pp. 561–586.
 - [94] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” *The Journal of Mathematical Sociology* **2**, 113–120 (1972).
 - [95] E. Estrada and J. A. Rodríguez-Velázquez, “Subgraph centrality in complex networks,” *Phys. Rev. E* **71**, 056103 (2005).
 - [96] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry* **40**, pp. 35–41 (1977).
 - [97] M. J. Newman, “A measure of betweenness centrality based on random walks,” *Social Networks* **27**, 39 – 54 (2005).
 - [98] S. White and P. Smyth, “Algorithms for estimating relative importance in networks,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’03 (ACM, New York, NY, USA, 2003) pp. 266–275.
 - [99] J. D. Noh and H. Rieger, “Random walks on complex networks,” *Phys. Rev. Lett.* **92**, 118701 (2004).
 - [100] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, “Large-scale topological and dynamical properties of the internet,” *Phys. Rev. E* **65**, 066130 (2002).
 - [101] V. Latora and M. Marchiori, “Efficient behavior of small-world networks,” *Phys. Rev. Lett.* **87**, 198701 (2001).
 - [102] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou, “Heuristically optimized trade-offs: A new paradigm for power laws in the internet,” in *Automata, Languages and Programming*, Lecture Notes in Computer Science, Vol. 2380, edited by P. Widmayer, S. Eidenbenz, F. Triguero, R. Morales, R. Conejo, and

-
- M. Hennessey (Springer Berlin Heidelberg, 2002) pp. 110–122.
- [103] K.-I. Goh, B. Kahng, and D. Kim, “Universal behavior of load distribution in scale-free networks,” *Phys. Rev. Lett.* **87**, 278701 (2001).
 - [104] A. E. Motter and Y.-C. Lai, “Cascade-based attacks on complex networks,” *Phys. Rev. E* **66**, 065102 (2002).
 - [105] A. E. Motter, “Cascade control and defense in complex networks,” *Phys. Rev. Lett.* **93**, 098701 (2004).
 - [106] Y. Moreno, R. Pastor-Satorras, A. Vázquez, and A. Vespignani, “Critical load and congestion instabilities in scale-free networks,” *EPL (Europhysics Letters)* **62**, 292–298 (2003).
 - [107] J. . H. Bakke, A. Hansen, and J. Kertész, “Failures and avalanches in complex networks,” *EPL (Europhysics Letters)* **76**, 717–723 (2006).
 - [108] R. Pastor-Satorras and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach* (Cambridge University Press, Cambridge, 2007).
 - [109] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *J. ACM* **46**, 604–632 (1999).
 - [110] B. H. Junker and F. Schreiber, *Analysis of biological networks*, Vol. 2 (Wiley-Interscience, New York, USA, 2008).
 - [111] S. Fortunato, “Community detection in graphs,” *Physics Reports* **486**, 75–174 (2010).
 - [112] R. D. Luce and A. D. Perry, “A method of matrix analysis of group structure,” *Psychometrika* **14**, 95–116 (1949).
 - [113] R. J. Mokken, “Cliques, clubs and clans,” *Quality & Quantity* **13**, 161–173 (1979).
 - [114] S. B. Seidman, “Network structure and minimum degree,” *Social Networks* **5**, 269–287 (1983).
 - [115] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical Review E* **69**, 026113 (2004).
 - [116] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**, 888–905 (2000).
 - [117] R. S. Burt, “Positions in networks,” *Social Forces* **55**, 93–122 (1976).
 - [118] E. Estrada and N. Hatano, “Communicability in complex networks,” *Physical Review E* **77**, 036111 (2008).
 - [119] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, “Characterizing the community structure of complex networks,” *PLoS ONE* **5**, e11976 (2010).
 - [120] A. Condon and R. M. Karp, “Algorithms for graph partitioning on the planted partition model,” *Random Structures and Algorithms* **18**, 116–140 (2001).
 - [121] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences of the USA* **99**, 7821–7826 (2002).

-
- [122] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E* **78**, 046110 (2008).
 - [123] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The elements of statistical learning*, Vol. 1 (Springer, Berlin, Germany, 2001).
 - [124] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1 (University of California Press, Berkeley, USA, 1967) p. 14.
 - [125] B. Kernighan and S. Lin, “An efficient heuristic procedure for partitioning graphs,” *Bell System Technical Journal* (1970).
 - [126] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*, Vol. 20 (Society for Industrial and Applied Mathematics, Philadelphia, USA, 2007).
 - [127] M. Fiedler, “Algebraic connectivity of graphs,” *Czechoslovak Mathematical Journal* **23**, 298–305 (1973).
 - [128] M. Fiedler, “A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory,” *Czechoslovak Mathematical Journal* **25**, 619–633 (1975).
 - [129] E. R. Barnes, “An algorithm for partitioning the nodes of a graph,” *SIAM Journal on Algebraic Discrete Methods* **3**, 541–550 (1982).
 - [130] W. E. Donath and A. J. Hoffman, “Lower bounds for the partitioning of graphs,” *IBM Journal of Research and Development* **17**, 420–425 (1973).
 - [131] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing* **17**, 395–416 (2007).
 - [132] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**, 888–905 (2000).
 - [133] A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, “On spectral clustering: Analysis and an algorithm,” (MIT Press, Cambridge, USA, 2002) pp. 849–856.
 - [134] M. Meila and J. Shi, “A random walks view of spectral segmentation,” in *AI and STATISTICS (AISTATS) 2001* (2001).
 - [135] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, “Stability of graph communities across time scales,” *Proceedings of the National Academy of Sciences of the USA* **107**, 12755–12760 (2010).
 - [136] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” in *Computer and Information Sciences - ISCIS 2005*, Lecture Notes in Computer Science, Vol. 3733, edited by p. Yolum, T. Güngör, F. Gürgen, and C. Özturan (Springer Berlin Heidelberg, 2005) pp. 284–293.
 - [137] H. Zhou, “Distance, dissimilarity index, and network community structure,” *Physical Review E* **67**, 061901 (2003).

-
- [138] H. Zhou, “Network landscape from a brownian particle’s perspective,” *Physical Review E* **67**, 041908 (2003).
 - [139] H. Zhou and R. Lipowsky, “Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities,” in *Computational Science - ICCS 2004*, Lecture Notes in Computer Science, Vol. 3038, edited by M. Bubak, G. Albada, P. Sloot, and J. Dongarra (Springer Berlin Heidelberg, 2004) pp. 1062–1069.
 - [140] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di, “Community detection by signaling on complex networks,” *Physical Review E* **78**, 016115 (2008).
 - [141] E. Weinan, T. Li, and E. Vanden-Eijnden, “Optimal partition and effective dynamics of complex networks,” *Proceedings of the National Academy of Sciences of the USA* **105**, 7907–7912 (2008).
 - [142] D. Harel and Y. Koren, “On clustering using random walks,” in *FST TCS 2001: Foundations of Software Technology and Theoretical Computer Science*, Lecture Notes in Computer Science, Vol. 2245, edited by R. Hariharan, V. Vinay, and M. Mukund (Springer Berlin Heidelberg, 2001) pp. 18–41.
 - [143] S. M. van Dongen, *Graph clustering by flow simulation*, Ph.D. thesis, Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht, Netherlands (2000).
 - [144] M. Rosvall and C. T. Bergstrom, “Maps of random walks on complex networks reveal community structure,” *Proceedings of the National Academy of Sciences of the USA* **105**, 1118–1123 (2008).
 - [145] M. E. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E* **69**, 066133 (2004).
 - [146] S. Kirkpatrick, D. Gelatt, Jr., and M. P. Vecchi, “Optimization by simulated annealing,” *science* **220**, 671–680 (1983).
 - [147] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, “Modularity from fluctuations in random graphs and complex networks,” *Physical Review E* **70**, 025101 (2004).
 - [148] M. E. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences of the USA* **103**, 8577–8582 (2006).
 - [149] M. E. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical Review E* **74**, 036104 (2006).
 - [150] B. H. Good, Y.-A. de Montjoye, and A. Clauset, “Performance of modularity maximization in practical contexts,” *Physical Review E* **81**, 046106 (2010).
 - [151] J. Reichardt and S. Bornholdt, “Statistical mechanics of community detection,” *Physical Review E* **74**, 016110 (2006).
 - [152] S. Fortunato and M. Barthelemy, “Resolution limit in community detection,” *Proceedings of the National Academy of Sciences of the USA* **104**, 36–41 (2007).
 - [153] J. Chen and B. Yuan, “Detecting functional modules in the yeast protein–protein interaction network,” *Bioinformatics* **22**, 2283–2290 (2006).

-
- [154] S. Gregory, “An algorithm to find overlapping community structure in networks,” in *Knowledge Discovery in Databases: PKDD 2007*, Lecture Notes in Computer Science, Vol. 4702, edited by J. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenić, and A. Skowron (Springer Berlin Heidelberg, 2007) pp. 91–102.
 - [155] S.-W. Son, H. Jeong, and J. D. Noh, “Random field ising model and community structure in complex networks,” *The European Physical Journal B-Condensed Matter and Complex Systems* **50**, 431–437 (2006).
 - [156] J. Baumes, M. K. Goldberg, M. S. Krishnamoorthy, M. Magdon-Ismael, and N. Preston, “Finding communities by clustering a graph into overlapping subgraphs,” in *IADIS International Conference on Applied Computing 2005*, Vol. 5, edited by N. Guimaraes and P. T. Isaias (IADIS, 2005) pp. 97–104.
 - [157] T. Nepusz, A. Petróczy, L. Négyessy, and F. Bazsó, “Fuzzy communities and the concept of bridgeness in complex networks,” *Physical Review E* **77**, 016107 (2008).
 - [158] I. Derényi, G. Palla, and T. Vicsek, “Clique percolation in random networks,” *Physical Review Letters* **94**, 160202 (2005).
 - [159] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature* **435**, 814–818 (2005).
 - [160] M. De Domenico, A. Sole, S. Gomez, and A. A., “Random walks on multiplex networks,” arXiv:1306.0519 [physics.soc-ph] (2013).
 - [161] A. Lancichinetti, S. Fortunato, and J. Kertész, “Detecting the overlapping and hierarchical community structure in complex networks,” *New Journal of Physics* **11**, 033015 (2009).
 - [162] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, “Inference and phase transitions in the detection of modules in sparse networks,” *Phys. Rev. Lett.* **107**, 065701 (2011).
 - [163] R. R. Nadakuditi and M. E. J. Newman, “Spectra of random graphs with arbitrary expected degrees,” *Phys. Rev. E* **87**, 012803 (2013).
 - [164] R. R. Nadakuditi and M. E. J. Newman, “Graph spectra and the detectability of community structure in networks,” *Phys. Rev. Lett.* **108**, 188701 (2012).
 - [165] F. Radicchi, “Detectability of communities in heterogeneous networks,” *Phys. Rev. E* **88**, 010801 (2013).