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 różnorakich 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 błądzenia 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 błądzeń losowych na deterministycznych grafach o dużej symetrii. Analitycznie i numerycznie wyznaczone zostały stany stacjonarne i czasy relaksacji zwykłego błądzenia przypadkowego odpowiadającego dyfuzji oraz błądzenia maksymalizującego entropię. Wyniki te pozwalają lepiej zrozumieć efekt lokalizacji i uwięzienia błądzenia przypadkowego w odseparowanych częściach sieci. Wreszcie w czwartym i piątym artykule analizowana jest użyteczność błądzeń 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 błądzenia losowe. Przeprowadzono również systematyczne porównanie opartych o błądzenie 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

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; nevertheless, 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** $k^{\mathrm{in}}(v)$ and **out-degree** $k^{\mathrm{out}}(v)$, which are the number of edges $(., v)$ pointing to v and the edges $(v, .)$ 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 $(v_1, e_1, v_2, e_2, \ldots, v_{t-1}, e_{t-1}, v_t)$, 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

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

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 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

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

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

$$\mathbf{L} = \mathbf{B}\mathbf{B}^T, \tag{1.3}$$

though the usual equivalent definition is

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

or in matrix notation

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{1.5}$$

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

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

$$\mathbf{L}_{\text{RW}} = \mathbf{D}^{-1}\mathbf{L} \tag{1.7}$$

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

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

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

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

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 $P(k) \propto \exp\left[-(\frac{k}{k_0})^{\beta}\right]$, or a power law with an exponential cut-off

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

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

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

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

$$l \propto \ln(N), \tag{1.12}$$

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

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

which leads to the mean degree of a nearest neighbour

$$\langle k \rangle_{\mathrm{NN}} = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{1.14}$$

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

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

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

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

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 $k(v)$

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

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

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

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 $\mathrm{Tr}(\mathbf{A}^3)$.

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]

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

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

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

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

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

and hence for large N it approaches the Poisson distribution

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

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

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

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 $\{N_k\}_{k=0,1,2,\ldots}$, 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 $v \in \{1, 2, \ldots, N\}$ 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

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

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]

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

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 $N \longrightarrow \infty$, 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])

$$P_\infty \propto (p - p_c)^\beta, \tag{2.1}$$

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:

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

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

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

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*

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

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 $N \longrightarrow \infty$ 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

$$\sum_k kQ(k) \geq 2, \tag{2.5}$$

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

$$\langle k \rangle_{\mathrm{NN}} = \frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2. \tag{2.6}$$

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

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

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]

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

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]

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

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 $N \longrightarrow \infty$. 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

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

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

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

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

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

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

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

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 $N \longrightarrow \infty$ one obtains

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

or equivalently

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

where $H_0(z) = \sum_n P_0(n) z^n$. Remembering that $H'(1) = \sum_n nP(n) = \langle n \rangle$ 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

$$2dp_c\phi = \frac{1}{\langle n \rangle_0}, \tag{2.16}$$

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

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

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 $V = \{v_1, v_2, \ldots, v_N\}$, 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 $P_{ij}\,\forall i,j = 1,\ldots,N$ 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 $\vec{\pi}(t) = (\pi_1(t),\ldots,\pi_N(t))^T$ 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

$$\vec{\pi}^T = \vec{\pi}^T\mathbf{P}, \tag{3.1}$$

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 $\vec{\pi}(t) \not\to \vec{\pi}$ when $t \to \infty$. 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 $\{\vec{\pi}(t)\}_{t=0,1,\ldots,\infty}$ 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

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

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

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

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

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

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

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

where $\vec{\Psi}_\alpha$ are eigenvectors of $\mathbf{S}$ associated with the eigenvalues $1 = |\Lambda_0| \geq |\Lambda_1| \geq \ldots \geq |\Lambda_{N-1}|$. 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*

$$\tau = 1/\ln \Lambda_1 \tag{3.5}$$

(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

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

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

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

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 $\gamma_{a_0 a_\tau} = (a_0, a_1, \ldots, a_\tau)$ of length τ. The probability of visiting this sequence of vertices starting at a_0 and the finishing at a_τ is

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

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]

$$\pi_i = \psi_{0i}^2, \tag{3.9}$$

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

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

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

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

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 $\mathbf{A}$:

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

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*

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

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 $\tilde{A}_{ij} \to A_{ij}\exp(U_{ij})$, 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 $\tilde{P}_{ij} \to P_{ij}\exp(U_{ij})$ 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

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

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]

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

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]

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

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

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

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

On expanding $(\mathbf{1} - \mathbf{P})^{-1}$ into Maclaurin series $\mathbf{1} + \mathbf{P} + \mathbf{P}^2 + \ldots$ 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 $\vec{e}\vec{\pi}^T$ 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

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

where $\vec{\mathbf{e}} = (1, 1, \ldots, 1)^T$ 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]

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

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 $\mathbb{R} \ni c(\alpha) \xrightarrow[\alpha\to 1]{} \infty$. The precise form of c is not important since it cancels out, leaving

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

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

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

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

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

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, $\vec{\mathbf{e}} \cdot \vec{M}_{*f} = \sum_i M_{if}$, 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 $\bar{M}_{\mathrm{MAX}} = \max_i\{\bar{M}_i\}$.

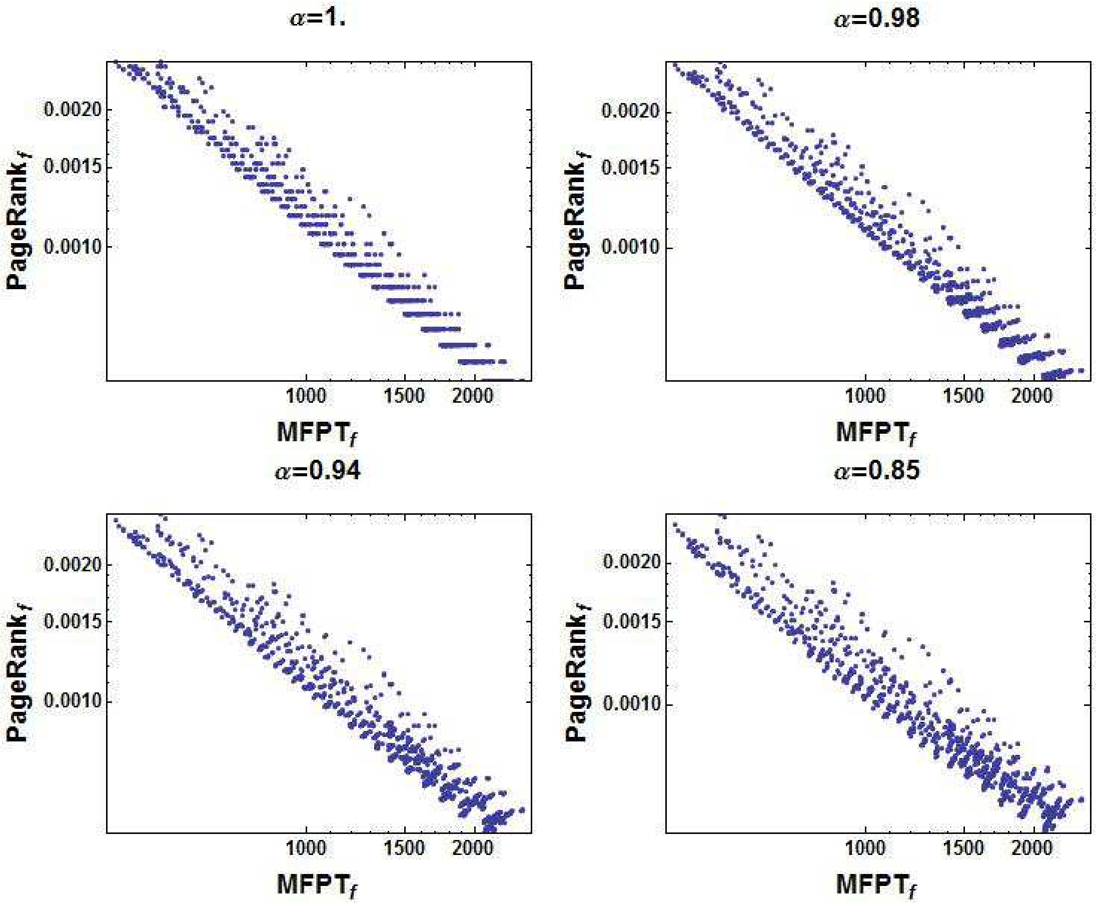

Figure 3.1: Log-log plot of the stationary probability of PageRank $\vec{\pi}(\alpha)$ versus total mean first-passage time $\sum_i M_{if}$ of reaching a vertex f for GRW on a sample network with power-law degree distribution. The case of $\alpha = 1$ is GRW, whose stationary states are proportional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

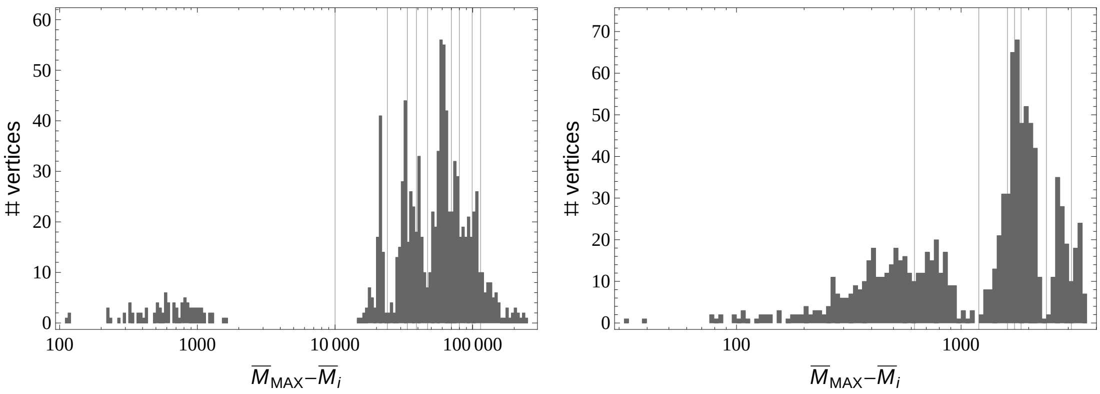

Figure 3.2: The histograms show the number of vertices i having a given MFPT value of going from the vertex to anywhere in the network; $\bar{M}_i = \sum_f M_{if}$ and $\bar{M}_{\text{MAX}} = \max_i\{\bar{M}_i\}$. MFPT is calculated for (left) MERW and (right) GRW. The vertical strokes are placed roughly 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.

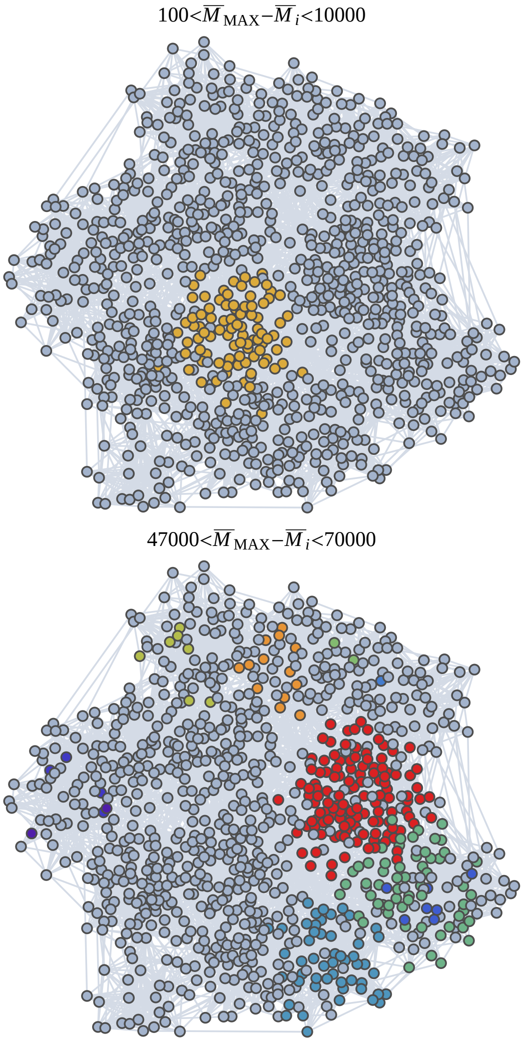
$100<\overline{M}_{\mathrm{MAX}}-\overline{M}_i<10000$
$10000<\overline{M}_{\mathrm{MAX}}-\overline{M}_i<24000$
$47000<\overline{M}_{\mathrm{MAX}}-\overline{M}_i<70000$
$114000<\overline{M}_{\mathrm{MAX}}-\overline{M}_i<300000$

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.

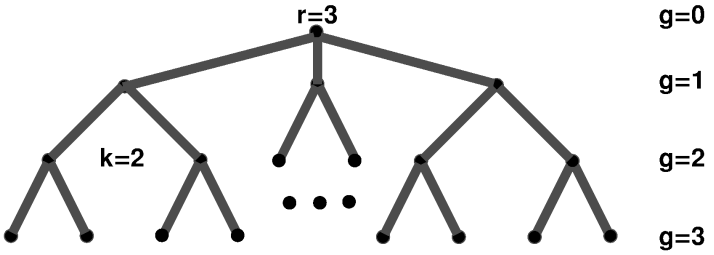

Figure 3.4: A Cayley tree with the branching number $k = 2$, except of the root $r = 3$. The last generation $g = 3$ of nodes is comprised of the leaves.

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.

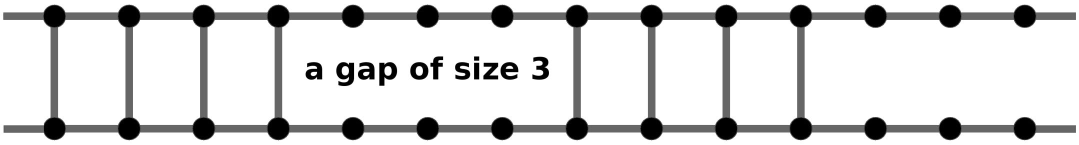

Figure 3.5: A ladder graph with 28 nodes and periodic boundary conditions marked by half-edges on both sides. The graph has two intact regions separated by two gaps of size $g = 3$.

This graph is comprised of two one-dimensional rings linked together, so that a vertex v_i is connected to its neighbours $v_{i\pm1}$, 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 $c : X \longrightarrow \mathbb{R}$ 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 $c(.)$ 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 $c(.)$ 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 $f : V(G) \longrightarrow V(G)$ 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 $(v, v_1, \ldots, v_{k-1}, v)$ of any length $k \in \mathbb{N}$, where $v_1, \ldots, v_{k-1} \in V$, 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

$$c_D(v) = k(v), v \in V. \tag{4.1}$$

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

$$\begin{aligned} c_{SS}(v) &= \pi_v, && \text{where } \vec{\pi} \text{ is a solution to} && (4.2)\\ \vec{\pi}^T\mathbf{P} &= \vec{\pi}^T, && && (4.3)\end{aligned}$$

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]

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

$$\mathbf{A}\vec{\psi}_0 = \lambda_0\vec{\psi}_0. \tag{4.5}$$

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]

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

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

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

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

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

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 $r = 0, \ldots, \infty$ 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

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

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

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

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

and after averaging over all source-target pairs

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

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 Kirchoff'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

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

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

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

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

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

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

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

$$\begin{aligned} c_I(v) &= \frac{e(G) - e(G\backslash\{v\})}{e(G)}, \text{where} \\ e(G) &= \frac{1}{N}\sum_{u\in G} c_S(u). \end{aligned} \tag{4.16}$$

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

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

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 $k(v) = k(v;H) + k(v;G\backslash H)$, 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 $k(v;G\backslash H) = \sum_{u \notin H} A_{uv}$.

Then, we call the subgraph H

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

$$\forall v \in V(H): \, k(v;H) > k(v;G\backslash H). \tag{5.1}$$

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

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

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

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

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

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

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]

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

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]

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

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 $v \in \{1, 2, \ldots, n\}$ 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 $(1, 1, \ldots, 1)$. 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)

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

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 $s_i = \mathrm{sgn}(v_{2i})$, 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 $\mathbf{L}_{\mathrm{RW}} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{P}$ [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

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

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

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

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 to close to the stationary state. In the same spirit [137–139] use the whole MFPT matrix

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

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]

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

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

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

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 multiplex[1] 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

[1] Briefly, 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 kindliness, 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 Wacław 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 hospitability 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. Hołyst, “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. Hołyst, “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 brownschen 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. Wacław, "Localization of the maximal entropy random walk," Phys. Rev. Lett. **102**, 160602 (2009).

[76] Z. Burda, J. Duda, J. M. Luck, and B. Wacław, "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. Hennessy (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 simmulated 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. MladeniÄŤ, 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-Ismail, 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. Petróczi, 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. KertĂ©sz, “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).

Eur. Phys. J. B **81**, 373–379 (2011)
DOI: 10.1140/epjb/e2011-10975-6

THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Shift of percolation thresholds for epidemic spread between static and dynamic small-world networks

J.K. Ochab[1,a] and P.F. Góra[1,2]

[1] M. Smoluchowski Institute of Physics, Jagiellonian University, Reymonta 4, 30-059 Kraków, Poland
[2] Mark Kac Complex Systems Research Centre, Jagiellonian University, Reymonta 4, 30-059 Kraków, Poland

Received 13 December 2010 / Received in final form 23 April 2011
Published online 7 June 2011 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2011

Abstract. The study compares the epidemic spread on static and dynamic small-world networks. They are constructed as a 2-dimensional Newman and Watts model (500 × 500 square lattice with additional shortcuts), where the dynamics involves rewiring shortcuts in every time step of the epidemic spread. We assume susceptible-infectious-removed (SIR) model of the disease. We study the behaviour of the epidemic over the range of shortcut probability per underlying bond $\phi = 0$–0.5. We calculate percolation thresholds for the epidemic outbreak, for which numerical results are checked against an approximate analytical model. We find a significant lowering of percolation thresholds on the dynamic network in the parameter range given. The result shows the behaviour of the epidemic on dynamic network is that of a static small world with the number of shortcuts increased by $20.7 \pm 1.4\%$, while the overall qualitative behaviour stays the same. We derive corrections to the analytical model which account for the effect. For both dynamic and static small worlds we observe suppression of the average epidemic size dependence on network size in comparison with the finite-size scaling known for regular lattice. We also study the effect of dynamics for several rewiring rates relative to infectious period of the disease.

1 Introduction

The epidemic modelling has become a significant and needed branch of complex systems research, as we have witnessed the recent epidemic threats and outbreaks of human diseases (H5N1 and H1N1 influenzas [1,2] or severe acute respiratory syndrome [3,4]) or animal (foot-and-mouth disease [5]) and plant diseases alike (e.g. Dutch elm disease [6] or rhizomania [7]). There are two crucial characteristics of the epidemic spread that make it complicated to be modelled on the one hand, and costly to be prevented in reality on the other: firstly, a number of infectious diseases exhibit long-range transmissions of varied nature, and secondly, the contact network of individuals affected by the disease may change in time as the epidemic spreads (which seems particularly relevant in the case of sexually transmitted diseases that are extensively discussed within the medical community [8,9]). These features make epidemiological models a part of larger studies of dynamics *on* complex networks, but also dynamics *of* complex networks.

Research findings of the epidemic spread on dynamic networks include its behaviour on adaptive networks, where the susceptible are able to avoid contact with the infected [10]), however a coupling between the epidemic and the network dynamics does not necessarily exist. For instance, in [11], spread of the aforementioned plant diseases is modelled by vectors performing random walk on the network, thus infecting individuals on their paths; Saramäki and Kaski [12] utilise SIR (susceptible-infectious-removed) mechanism on a dynamically changing small-world contact network, although mainly time development of the epidemic is of their interest. Likewise, in [13] (where focus is on the average epidemic size in time) nodes of the contact network can swap their edges at a given rate, preserving the degree distribution. It is also worth to note [14], where disease spread was simulated on a weighted contact network produced from *real* day-to-day encounters (as weights represent the frequency of encounters, the dynamics has been in a sense projected onto static weighted network).

While dynamic network models have been applied in the recent research, it seems that we lack comparative study on how the dynamics of the network influences the process that takes place on it. The aim of this paper is to find and quantify this effect for SIR epidemic spread on static and dynamic small-world networks. Based on known analytical calculations for static small-world network [15] we derive corrections accounting for the dynamics of the network, and check the results against numerical agent-based simulations.

[a] e-mail: jeremi.ochab@uj.edu.pl

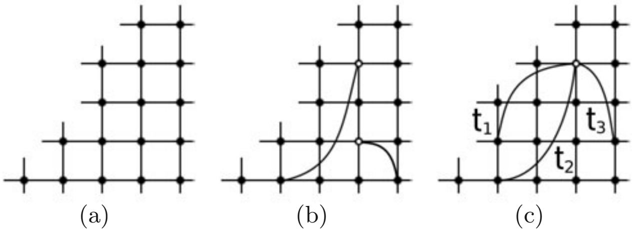

Fig. 1. (a) Regular 2D grid with periodic boundary conditions (torus). (b) Newman-Watts 2D small-world network: 2D grid with shortcuts added to it. (c) Dynamic small-world: all the long-range links connected to a set of source nodes randomly rewire in time.

2 Model

2.1 Network

We adopt Watts-Strogatz model of a small-world network [16] with the alteration by Newman and Watts [17]: first we take a 2-dimensional square lattice with $N = L^2$ nodes and $2N$ undirected edges (Fig. 1). To avoid some finite-size effects we impose periodic boundary conditions for the lattice (i.e. we get a torus). Then, we add a number of undirected edges between random pairs of nodes. The number of additional edges ('shortcuts') is set as $2\phi N$, hence ϕ is shortcut probability per underlying bond. Network with $\phi = 0$ is just a *regular lattice*. For nonzero ϕ we call the network a *static small-world*.

The third type of network is a *dynamic small-world*. One can construct it by randomly distributing shortcuts in every time step of simulation. Here, we choose $2\phi N$ nodes randomly, and keep them fixed for the whole run of the epidemic. In every time step we randomly launch shortcuts anchored in these nodes, which means the dynamics consists in rewiring one end of these shortcuts. For the sake of simplicity we allow for multiple shortcuts being incident with the same node, for shortcuts leading to nearest neighbours, and for loops being formed. The construction of the source nodes launching shortcuts allows for an easier interpretation of the network: the fixed nodes could correspond to centres of activity that can be identified as in the real world networks.

2.2 Epidemic

The SIR (susceptible-infectious-removed) model is adopted, where the disease is transmitted along the edges of the network in discrete time steps. The probability p of infecting a susceptible node by an infectious neighbour during one time step is set equal for short- and long-range links, both static and dynamic. The infectious period l of the disease is measured in discrete time units (we take $l = 3, 4$). Thus, an infectious node can transmit disease to susceptible nodes with probability p every turn for the period of l turns, and after that time it is removed, i.e. it cannot infect nor be reinfected. Every simulation starts with only one initially infecting node, all others being susceptible, and it ends when no node in the infectious state is left. Sample snapshots of the epidemic time development are presented in Figure 2.

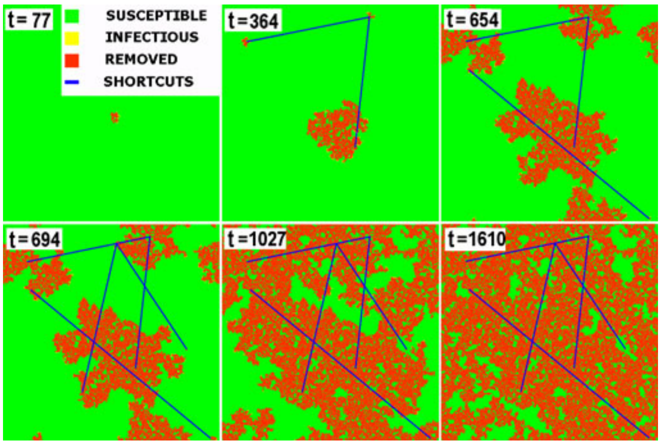

Fig. 2. (Color online) Snapshots of the epidemic spread slightly above percolation threshold. $L = 512$, the number of shortcuts is 10 (which gives $\phi = 2 \times 10^{-5}$). t gives the epidemic's time steps. The snapshots for $t = 364$, 694 show a dynamic infection (the two joined blue lines appear).

3 Numerical data

3.1 Parameters of simulations

The linear lattice size used for most calculations is $L = \sqrt{N} = 500$. In Section 5.2 we take sizes $L = 50$, 63, 79, 100, 126, 158, 199, 251, 315, 397, 500. The disease infectious period is set to $l = 3$ (for faster simulations reported in Sect. 5) or $l = 4$ (in Sect. 5.3 in order to get larger set of dynamic rates). The range of probability p scanned is $p = 0.05$–0.22 (depending on ϕ) with resolution of 1/1024, which translates to around $T = 0.15$–0.5. For every p and ϕ the epidemic is run 1024 times with random distributions of shortcuts each time. The fraction of shortcuts is $\phi = 0$–0.5, with steps of 0.025. The simulations are performed for both static and dynamic small-world network.

3.2 Calculating percolation threshold

In the study of the epidemic spread on networks, we stick to the percolation theory as a reference point. In the theory, a percolation threshold would be the value of p that generates an epidemic cluster spanning between the boundaries of the whole system. Otherwise, it is possible to define percolation as the point at which a cluster of macroscopic size forms (i.e. it occupies a finite fraction of the system for $N \to \infty$). We employ the latter to define percolation threshold (numerically) as the point at which the average epidemic's size divided by N rises above a certain value (here, set to 0.00115). The average is taken over a number of reruns for different shortcut drawings. As we can perform simulations only for finite sizes, we take the results for a relatively large network of $\sqrt{N} = 500$.

The choice of the threshold value is taken so as to calibrate the results for the static network to the previously confirmed analytical result. We take as the theoretical model [15], where the generating function and series expansion methods were used to find the approximate position of bond percolation transition in 2D small-world network, which corresponds to the epidemic spread on what we refer to as static small-world.

4 Theoretical analysis

We can account for the change between static and dynamic networks analytically using the model known for static small-world network [15]. In their paper, Newman et al. derive the expression for bond percolation thresholds for any d-dimensional hypercubic lattices with the addition of shortcuts:

$$T_c = \frac{1}{2d\phi\langle n_0\rangle}, \qquad (1)$$

where $\langle n_0\rangle$ is the average cluster size in bond percolation process. The authors then provide the approximation of $\langle n_0\rangle$ for the particular case of 2-dimensional square lattice, and thus for T_c on the small-world generated from the lattice. It is also known that Grassberger [18] related the probability of infection p to the probability T in bond percolation through

$$T = \sum_{t=1}^{l} p(1-p)^{t-1} = 1-(1-p)^l, \qquad (2)$$

where T is the so called *transmissibility* (it is the total probability of a node infecting one of its neighbours during the whole infectious period). In the case of 2-dimensional square lattice the bond percolation threshold is $T_c = 0.5$.

As the original theory has no time variable, it would be a hard task to introduce dynamics explicitly. The solution, however, is astonishingly simple. One can estimate the average number of nodes infected through shortcuts during infectious period l:

$$\langle N_{stat}\rangle = \phi_{stat}NT = \phi_{stat}N\sum_{t=1}^{l} p(1-p)^{t-1}, \qquad (3)$$

i.e. the number of shortcuts in the static network multiplied by the total probability of infecting a neighbouring node (this probability is the same for both regular links and shortcuts). The analogous expression for the dynamic network is found easily

$$\langle N_{dyn}\rangle = \phi_{dyn}N\sum_{i=1}^{l} i\binom{l}{i}p^i(1-p)^{l-i} = \phi_{dyn}Nlp, \qquad (4)$$

where the sum is an average number of infections transmitted by a single source of dynamic shortcuts for a given infectious period. It comes from the fact that a dynamic shortcut can pass infection several times (the factor p^i), while in the static case a node could infect only once (since nodes cannot be reinfected in the SIR model). This expression predicts lowering percolation thresholds, although numerical values of the shift are considerably smaller than the ones obtained from simulations.

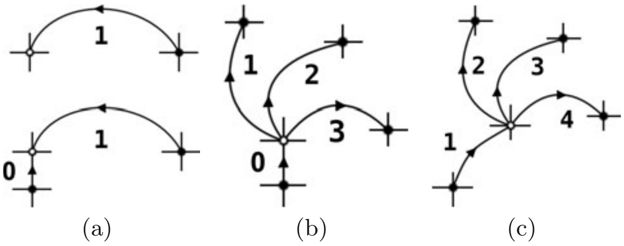

Fig. 3. (a) Infections through static shortcuts are symmetric. (b) Infection of the dynamic shortcuts' source through regular lattice. (c) Infection of the dynamic shortcuts' source through a shortcut.

Figures 3a–3c explain why the above expression is not yet correct: it is derived only for the source nodes passing the disease on, while it disregards the fact that the node may itself become infected via long-range link. Since on the static network there is no difference between shortcuts' source and target nodes, we can attach the factor $\phi N/2$ to both infection graphs presented in Figure 3a. For dynamic network, the graphs in Figures 3b, 3c for infecting a source node through a regular link and through a dynamic link give different counts of how many shortcuts were used. The former was given in equation (4) as lp, and the latter actually utilises the same formula, but with the substitution $l \to l+1$. In total, we get

$$\langle N_{dyn}\rangle = \phi_{dyn}N/2lp + \phi_{dyn}N/2\,(l+1)\,p. \qquad (5)$$

We assume that $\langle N_{dyn}\rangle = \langle N_{stat}\rangle$ if the epidemic on both networks has the same percolation threshold. Thus, we can obtain the ratio of the two shortcut densities

$$\begin{aligned} r(T,l) = \phi_{stat}/\phi_{dyn} &= \frac{p\,(l+1/2)}{T} \\ &= \frac{\left[1-(1-T)^{1/l}\right](l+1/2)}{T}, \end{aligned} \qquad (6)$$

where p is the probability of infection in one time step and l infectious period of a disease. Now, we can calculate $T_c(r\phi)$ numerically, just as we do it with the fitted $T_c[(1+v)\phi)]$ in Figure 4. The ratio in equation (6) was used to plot the lower solid line in Figure 4 and the predicted thresholds in Figure 5.

5 Results

5.1 Shift of percolation thresholds

In Figure 4 we plot numerical and theoretical values of percolation thresholds T_c for both static and dynamic

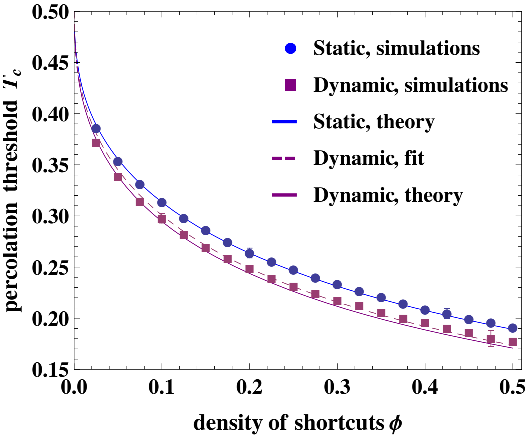

Fig. 4. (Color online) Circles: dataset for the static small-world. Squares: dataset for the dynamic network. The solid blue (upper) line is the analytic approximation [15] for $T_c(\phi)$ and the dashed line gives $T_c[(1+v)\phi)]$, with the fit parameter $v = 0.207 \pm 0.014$. The solid purple (lower) line represents theoretical approximation from Section 5.3. Error bars are of the size of the plot markers, unless visible. Infectious period $l = 3$.

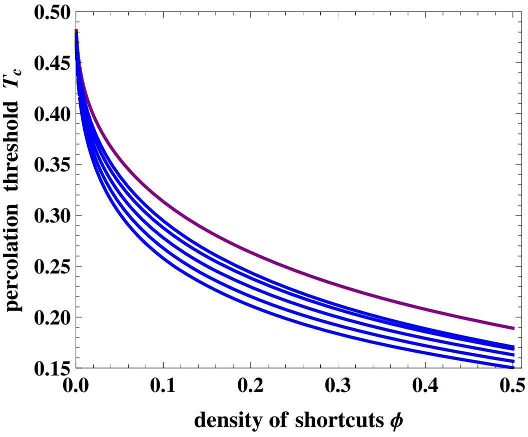

Fig. 5. (Color online) The shift of epidemic thresholds for different infectious periods. Uppermost curve – static network, $T_c(\phi)$. Lower curves – dynamic networks, $T_c(r\phi)$ with infectious periods l = 3, 6, 9, 12, 15, and $r(T,l)$ given in equation (6).

small-worlds. The resulting $T_c(\phi)$ data points for static small-world network agree with the analytical approximation [15], which confirms the validity of calibration procedure. As the lower dataset marks the effect of network dynamics, the difference between the two networks proves to be systematic and significant. The dashed line is a fit $T_c[(1 + v)\phi]$ of the analytical model for the static network, where the fitted parameter v may be interpreted as a virtual percentage of additional shortcuts needed to obtain the dynamic network percolation thresholds. It follows from the fit that percolation thresholds for dynamic network are lower as if the shortcut density were $(1+v)\phi$ (where $v = 0.207 \pm 0.014$ is the fitted parameter). Nonetheless, qualitatively the epidemic on dynamic small world behaves in the same way as on the static one for the given range of parameters ($\phi = 0.5$ corresponds to every node in the network having on average two additional links).

The analytical correction slightly exceeds the values of simulation data points, but the overall agreement is satisfactory. The difference between the analytical solution and the observed behaviour does not exceed the shift between static and dynamic networks obtained from simulations. The discrepancy might be due to the method of calculating percolation thresholds from numerical data or due to the approximate nature of the correction.

In Figure 5 we give theoretical predictions for epidemic thresholds for longer infectious periods, using the result of Section 4. As can be seen, the lowering of thresholds with respect to the static case may be much larger. Unfortunately, producing numerical results for the longer infectious periods becomes more costly.

The decrease in percolation threshold T_c may be understood in terms of increasing the average node degrees as we add shortcuts to the network (i.e. we increase ϕ). Following May and Anderson [8] who described HIV infection dynamics, we might estimate the invasion threshold using the equation for the reproductive rate of infection R_0 ("which is the average number of secondary infections produced by one infected individual in the early stages of an epidemic"):

$$R_0 = \beta c D, \tag{7}$$

in the original notation (β here corresponds to transmissibility T, D to infectious period l, and c is "the average rate at which new sexual partners are acquired" in the context of the original paper, and the average number of dynamic links here). Thus, one expects epidemic outbreak for $R_0 > 1$. As c is given by the ratio of node degree moments $c = \langle k^2\rangle/\langle k\rangle = 1+4\phi$ (the second equality assumes Poissonian distribution of the number of dynamic links in our model), we get $\beta_c \sim 1/l(1 + 4\phi)$. This relation indeed explains the general decrease of the thresholds with higher ϕ and with longer infectious periods l, although it does not take into account the underlying regular lattice, and it does not predict correct numerical values within the discussed model.

5.2 Suppression of finite-size scaling

The primary motivation of checking finite-size scaling for the system was to utilise it to determine the percolation thresholds very accurately (as the shift of thresholds observed in Fig. 4 is relatively small), and to arrive at threshold value for infinite system size. Yet, it is worth noting at this point that the knowledge of thresholds for infinite system sizes would not usually be appropriate for evaluation of risks in the real epidemic, given the sizes of some real networks. To study the size of finite-size effects is thus vital on its own right.

In Figure 6b the convergence of the average epidemic size to the threshold behaviour can be observed, and the significant dependence on system size ranges up to the epidemic size of around $0.5N$ and interval of transmissibility of length around 0.08 (the numbers are very rough

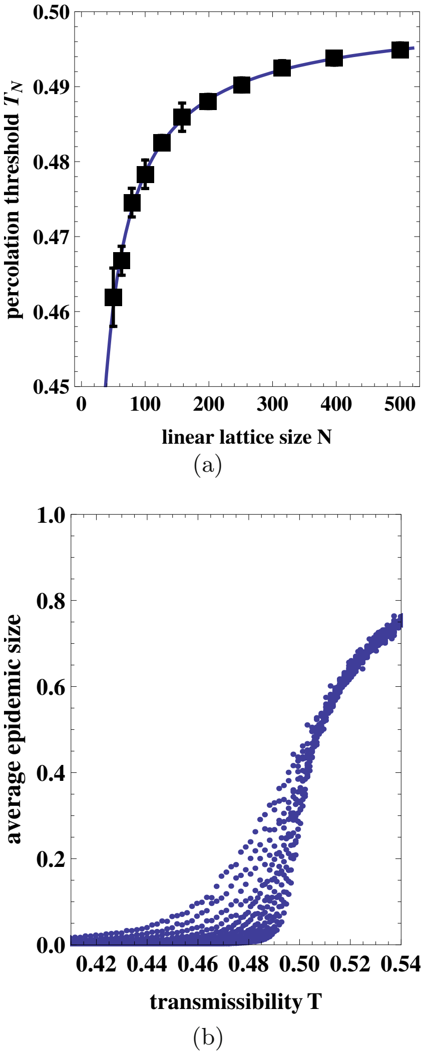

Fig. 6. (Color online) Behaviour of the epidemic outbreak magnitude for various system sizes (linear size vary between L = 50–500, left- and rightmost data points in (b), respectively). (a) Finite-size scaling $T_N = T_\infty - L^{-2/\nu}$ on regular lattice. The points correspond to values of T at the level of the epidemic size 0.1. (b) The extent of size dependence for regular lattice.

estimates). As presented in Figure 6a, one may check that sections of the plot for a given average epidemic size obey scaling of the form

$$T_N = T_\infty - N^{-1/\nu} = T_\infty - L^{-2/\nu}, \tag{8}$$

where T_N are the values of transmissibility for a given system size N and a set section position, and T_∞ is the percolation threshold for infinite system size. For regular lattice T_∞ is fitted correctly for various section positions as 0.500 ± 0.005 (the error may vary for different sections, but does not exceed the given value).

It appears that the dependence on system size for small-world networks (both static and dynamic) is dissimilar to the one of regular lattices, as can be seen in Figure 7 (ϕ = 0.05). It is suppressed to smaller values of the average epidemic size. For the shortcuts density $\phi = 0.5$ the dependence on system size is already visible only below the epidemic size of 0.03. Because the dependence of the epidemic size on size of the system becomes of the order of magnitude of statistical fluctuations (the quality of the data can already be seen in Fig. 7), any attempts to utilise finite-size scaling for determining percolation threshold are not viable. Indeed, the errors do not allow us to check if the same form of finite-size dependence as in equation (8) holds.

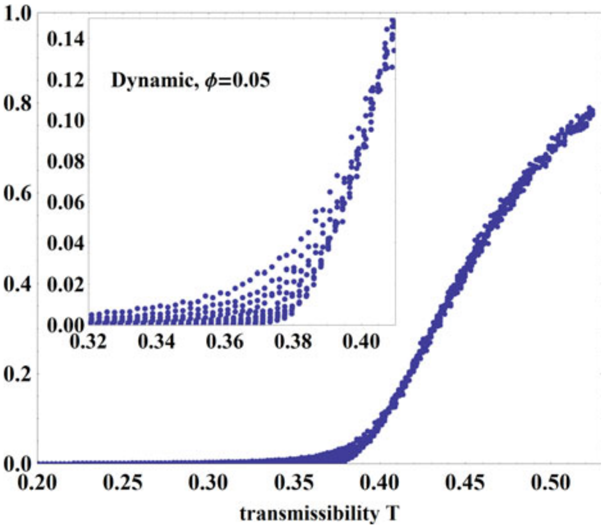

Fig. 7. (Color online) For dynamic small world size dependence of the epidemic outbreak magnitude is suppressed. Inset shows enlarged region around percolation threshold.

5.3 Dependence on the rate of dynamics

One can generalise the theoretical analysis for various rates of dynamics, given the formula in equation (4). To explain this, let us notice that there are two time scales in the model: the infectious period l of the disease and the duration $1/d$ between consecutive rewirings of dynamic links (both measured in discrete time steps of the epidemic spread). As the choice of infectious period l only rescales the total probability of infection $T = T(p, l)$, we can dispose of it, and the crucial parameter ld that accounts for the shift of percolation thresholds is defined as the number of shortcut movements during infectious period.

Obviously, for a static network we get $d = 0$, while for all the above analysis of dynamic network we have $ld = 3$ ($l = 3$ and the rewiring was performed every turn, so $d = 1$). Depending on the interpretation of the model, we could also consider $d > 1$. However, if p is to be the probability of infection *during one time step* it is reasonable that shortcuts rewiring faster than one time step would infect with appropriately smaller probability, and there would be no further shift of percolation thresholds.

Since the epidemic spreads with discrete time, which results in sums as in equation (4), we are interested in rational numbers $d \in [0, 1] \cap \mathbb{Q}$, particularly of the form

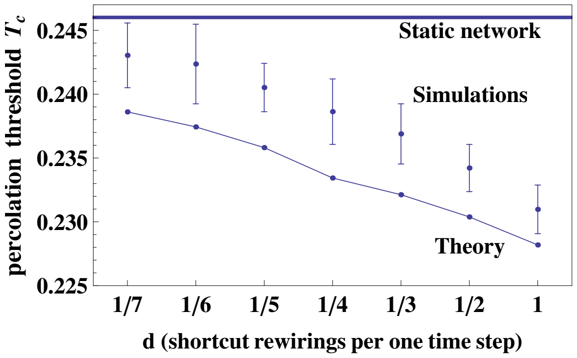

Fig. 8. (Color online) Dependence on dynamics for $\phi = 0.25$, infectious period $l = 4$.

$1/i, i \in \mathbb{Z}$. What we need is N_{dyn} calculated in a similar way to that in equation (4). Here, we take $l = 4$, $d = 1, 1/2, \ldots, 1/7$, and we plot both the numerical and theoretical results for $\phi = 0.25$ in Figure 8. Theoretical derivation is to be found in the appendix. The theoretical approach gives slightly exceeding values (the scale should be noted), which is the same effect as discussed in the Section 5.1.

6 Discussion

We have shown that introducing dynamics of the long-range links in a small-world network significantly lowers an epidemic threshold in terms of probability of disease transmission, although the overall dependence on number of shortcuts stays the same. Consequently, the risk of an epidemic outbreak is higher than in any calculations involving static models. The effect remains secondary to the influence that the introduction of additional of shortcuts has on the spread of the disease. It should be noted that the shift of percolation thresholds depends on the relative measure of dynamics of the network with respect to the process on the network (rewiring rate and infectious period, respectively). Any accurate analytical calculation or simulation should take this quantity as a significant parameter, to be estimated for a particular disease and type of the network.

As in reality we consider only finite-size networks, and real epidemic sizes do not usually reach values of the order of even 10% of the system size, the information on finite-size effects seemed very much needed. That the epidemic outbreak magnitude does not depend on the system size for small-world networks as much as it does for regular lattices means that we should not expect the epidemic outbreaks below transmissibility threshold value. Thus, finite-size effects seem to become secondary, as well.

The usefulness of such a model for risk prediction still depends on our knowledge of the probability of transmission (p or T) of a given disease, which is not easy to obtain for diseases spreading outside of well controlled environments like hospitals. Relatively good estimates, thanks to the nature of transmission, exist for syphilis. Transmissibility of the disease is reviewed in [9], where authors give values ranging from 9.2% to 63% per partner, and decide on 60% as the lower boundary for *untreated* disease. This seems to be well above the epidemic threshold, irrespective of very different network topology for such diseases. However, this also shows that errors on estimates of transmission probabilities exceed the effect of threshold shifting studied here.

Though the 2-dimensional network structure used here may be said to correspond mainly to that of plantations, it is worth noting its generality: nodes may be interpreted as plants, animals or humans, but also on a larger scale as farms, households, or cities and airports; in turn, long-range links could mean wind (on farms), disease vectors, occasional human contacts, or airline connections. Still, it has some other fairly realistic characteristics: according to [14], who analysed the structure of human social interactions, 'the majority of encounters (76.70%; 75.26–78.07) occur with individuals never again encountered by the participant during the 14 days of the survey'. This may mean that about 24% of the repeated contacts corresponds roughly to our regular underlying lattice with $z = 4$ neighbours for each node, while the 76% correspond to around $3z$ dynamic contacts distributed over 14 days. This gives on average $\phi \approx 0.20$ for simulation with daily time steps, which lies within the parameter range studied in this paper.

This work is supported by the International Ph.D. Projects Programme of the Foundation for Polish Science within the European Regional Development Fund of the European Union, agreement No. MPD/2009/6.

Appendix A: Dependence on the rate of dynamics

Below we present the way to calculate N_{dyn} for infectious periods $l = 4, 5$ (in the simulation we set $l = 4$, but we need to take into account also the process from Fig. 3c, which in a sense increases infectious period by 1). Let us define

$$\begin{aligned}
A_0(p,l) &= 1 - (1-p)^l \equiv T(l), \quad l \geq 1\\
A_1(p,l) &= T(1)\,[1 - T(l-1)] + [1 - T(1)]\,T(l-1)\\
&\quad + 2T(p,1)T(l-1)\\
A_2(p,l) &= T(2)\,[1 - T(l-2)] + [1 - T(2)]\,T(l-2)\\
&\quad + 2T(2)T(l-2)\\
A_{11}(p,l) &= 3T(1)^2T(l-2) + 2T(1)^2\,[1 - T(l-2)]\\
&\quad + 2T(1)T(l-2)\,[1 - T(1)] +\\
&\quad + T(l-2)\,[1 - T(1)]^2 \qquad\qquad (A.1)\\
&\quad + 2T(1)\,[1 - T(1)]\,[1 - T(l-2)]\\
A_{12}(p,l) &= 3T(1)T(2)T(l-3) + 2T(1)T(2)\,[1 - T(l-3)]\\
&\quad + T(1)\,[1 - T(2)]\,T(l-3)\\
&\quad + [1 - T(1)]\,T(2)T(l-3)\\
&\quad + T(1)\,[1 - T(2)]\,[1 - T(l-3)]\\
&\quad + [1 - T(1)]\,T(2)\,[1 - T(l-3)]\\
&\quad + [1 - T(1)]\,[1 - T(2)]\,T(l-3),
\end{aligned}$$

where we substituted $T(1)$ for p on the right-hand sides, and we leave out the argument p in $T(p,l)$ to simplify the notation. Those quantities correspond to the average number of infections during one infectious period depending on when the rewiring takes place. One can present those diagrammatically (here for $l = 5$) as

$$\begin{aligned} A\ _0(p,5)\ &= \cdot\cdot\cdot\cdot\cdot \\ A\ _1(p,5)\ &= \cdot|\cdot\cdot\cdot\cdot + \cdot\cdot\cdot\cdot|\cdot = 2\cdot|\cdot\cdot\cdot\cdot \\ A\ _2(p,5)\ &= \cdot\cdot|\cdot\cdot\cdot + \cdot\cdot\cdot|\cdot\cdot = 2\cdot\cdot|\cdot\cdot\cdot \\ A\ _{11}(p,5) &= \cdot|\cdot\cdot\cdot|\cdot \\ A\ _{12}(p,5) &= \cdot|\cdot\cdot|\cdot\cdot + \cdot\cdot|\cdot\cdot|\cdot = 2\cdot|\cdot\cdot|\cdot\cdot \end{aligned} \tag{A.2}$$

where the symbol '$|$' marks rewiring, and '$\cdot$' one epidemic time step during infectious period. For instance $\cdot|\cdot\cdot$ would correspond to three turns with one rewiring, during which either 0, 1 or 2 infections are possible. The derivation involves only very easy combinatorics, but for longer infectious periods one would need to repeat these calculations to obtain more terms and different prefactors. Now, one can easily obtain expressions for N_{dyn} for any $1/d \in \mathbb{Z}$. Below we give only the general expression for $1/d \geq l$:

$$\begin{aligned} N_{dyn} = \frac{\phi_{dyn}N}{2} d\{&[2A_1(l) + A_2(l) + (1/d + 1 - l)A_0(l)] \\ &+ [2A_1(l{+}1){+}2A_2(l{+}1){+}(1/d{-}l)A_0(l)]\} \end{aligned} \tag{A.3}$$

where $l = 4$. The first term in the brackets corresponds to Figure 3b and the second to Figure 3c. For greater numbers of rewiring per turn d, we need to consider the terms A_{11}, A_{12}. The result is plotted against simulated data in Figure 8.

References

1. World Health Organization, Avian influenza (H5N1), http://www.who.int/csr/disease/avian_influenza/en/ (2010)
2. World Health Organization, Swine influenza (H1N1), http://www.who.int/csr/disease/swineflu/en/ (2010)
3. World Health Organization, Severe acute respiratory syndrome (SARS), http://www.who.int/csr/sars/en/ (2010)
4. C. Dye, N. Gay, Science **300**, 1884 (2003)
5. M.J. Keeling, M.E.J. Woolhouse, D.J. Shaw, L. Matthews, M. Chase-Topping, D.T. Haydon, S.J. Cornell, J. Kappey, J. Wilesmith, B.T. Grenfell, Science **294**, 813 (2001)
6. J. Swinton, C.A. Gilligan, Proc. Trans. R. Soc. B **351**, 605 (1996)
7. A.J. Stacey, J.E. Truscott, M.J.C. Asher, C.A. Gilligan, Phytopathology **94**, 209 (2004)
8. R.M. May, R.M. Anderson, Nature **326**, 137 (1987)
9. G.P. Garnett, S.O. Aral, D.V. Hoyle, W. Cates Jr, R.M. Anderson, Sex. Transm. Dis. **24**, 185 (1997)
10. T. Gross, C.J.D. D'Lima, B. Blasius, Phys. Rev. Lett. **96**, 208701 (2006)
11. B. Dybiec, A. Kleczkowski, C.A. Gilligan, J. R. Soc. Interface **6**, 941 (2009)
12. J. Saramäki, K. Kaski, J. Theor. Biol. **234**, 413 (2005)
13. E. Volz, L.A. Meyers, Proc. R. Soc. B **274**, 2925 (2007)
14. J.M. Read, K.T.D. Eames, W.J. Edmunds, J. R. Soc. Interface **5**, 1001 (2008)
15. M.E.J. Newman, I. Jensen, R.M. Ziff, Phys. Rev. E **65**, 021904 (2002)
16. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
17. M.E.J. Newman, D.J. Watts, Phys. Lett. A **263**, 341 (1999)
18. P. Grassberger, Math. Biosci. **63**, 157 (1983)

Exact solution for statics and dynamics of maximal-entropy random walks on Cayley trees

J. K. Ochab* and Z. Burda†

Marian Smoluchowski Institute of Physics and Mark Kac Complex Systems Research Center, Jagiellonian University, Reymonta 4, PL-30-059 Kraków, Poland

(Received 12 January 2012; published 24 February 2012)

We provide analytical solutions for two types of random walk: generic random walk (GRW) and maximal-entropy random walk (MERW) on a Cayley tree with arbitrary branching number, root degree, and number of generations. For MERW, we obtain the stationary state given by the squared elements of the eigenvector associated with the largest eigenvalue λ_0 of the adjacency matrix. We discuss the dynamics, depending on the second largest eigenvalue λ_1, of the probability distribution approaching to the stationary state. We find different scaling of the relaxation time with the system size, which is generically shorter for MERW than for GRW. We also signal that depending on the initial conditions, there are relaxations associated with lower eigenvalues which are induced by symmetries of the tree. In general, we find that there are three regimes of a tree structure resulting in different statics and dynamics of MERW; these correspond to strongly, critically, and weakly branched roots.

DOI: 10.1103/PhysRevE.85.021145 PACS number(s): 05.40.Fb, 02.50.Ga, 89.75.Hc, 89.70.Cf

I. INTRODUCTION

After the theory of Brownian motion and diffusive processes was formulated in the seminal works by Einstein [1] and Smoluchowski [2], random walk (RW) models, which stem from time or space discretization of these processes, have continuously attracted attention. The most celebrated ones include the Polya random walk on a lattice [3] and its generalizations to arbitrary graphs. RW has been discussed in thousands of papers and textbooks in statistical physics, economics, biophysics, engineering, particle physics, etc., and still is an active area of research.

Mathematically speaking, RW is a Markov chain which describes the trajectory of a particle taking successive random steps. For instance, in the case of Polya random walk, at each time step the particle jumps onto one of the neighboring nodes with equal probability. The generalization of this process to any graph is what we call the ordinary or generic random walk (GRW).

Another kind of RW, i.e., one that maximizes the entropy of paths and hence is named maximal-entropy random walk (MERW), has been investigated recently [4,5]. The same principle of entropy maximization earlier led to the biological concept of evolutionary entropy [6,7]. It was also used in the problem of importance sampling where it served as an optimal sampling algorithm [8]. Now, MERW enters also the realm of complex networks [9–13]. Its defining feature results in equiprobability of paths of given length and end points, which means that if information is sent between two places, MERW makes it impossible to resolve which route the information has traveled. Such an ensemble of equiprobable paths is a natural choice for a measure used in Feynman path integrals in discrete quantum gravity models with curved spacetime [5]. Another unprecedented feature of this RW is the localization phenomenon on diluted or defective lattices, where most of the stationary probability is localized in the largest nearly spherical region free of defects [4,5]. It has been illustrated with an interactive online demonstration [14]. In this paper, we show not only how stationary states of GRW and MERW differ, but also how their dynamics differs. More precisely, we have analytically determined stationary probability distributions and relaxation times of GRW and MERW on Cayley trees. In particular, we have found that there are three regimes of a tree structure, which depend on its root's degree, resulting in different stationary states and relaxation times of MERW. This type of random walk in comparison to GRW has stationary probability centered around the root of the tree and its relaxation is generically faster, with the time scaling as a logarithm of the system size.

The paper is organized as follows: we begin with Sec. II defining GRW and MERW in general. In Sec. III, we restrict our considerations to Cayley trees, for whose adjacency matrix we solve the eigenvalue problem by generalizing the method given in [15]. The scheme presented there is utilized in Sec. IV, where we determine the eigenvector to the largest eigenvalue of the adjacency matrix, and then in Sec. V, we generalize part of this result to eigenvectors associated with next-to-leading eigenvalues. Section VI presents the solution for the eigenvalue problem of GRW transition matrix, repeating the order of arguments from Sec. III. Based on results from previous sections, Sec. VII describes stationary distributions of GRW and MERW on Cayley trees. Sections VIII and IX concern relaxation times of those two random walks, with general remarks in the former and particular results in the latter. Details concerning the solution of eigenproblems are to be found in Appendices A and B.

II. GENERALITIES

Let us consider a discrete time random walk on a finite connected undirected graph. We are interested in a class of random walks with a stochastic matrix $\mathbf{P}$ that is constant in time. An element $P_{ij} \geqslant 0$ of this matrix encodes the probability that a particle being on a node i at time t hops to a node j at time $t+1$. These matrix elements fulfill the condition $\sum_j P_{ij} = 1$ for all i, which means that the number of particles is conserved. Additionally, let us assume that particles are allowed to hop only to a neighboring node. This can be

*jeremi.ochab@uj.edu.pl
†zdzislaw.burda@uj.edu.pl

formulated as $P_{ij} \leqslant A_{ij}$, where A_{ij} is the corresponding element of the adjacency matrix $\mathbf{A}$ of the graph: $A_{ij} = 1$ if i and j are neighbors, and $A_{ij} = 0$ otherwise. The generic random walk (GRW) is realized by the following stochastic matrix:

$$P_{ij} = \frac{A_{ij}}{k_i}, \qquad (1)$$

where $k_i = \sum_j A_{ij}$ denotes the node degree. The factor $1/k_i$ in the above formula produces uniform probability of selecting one of k_i neighbors of the node i. Clearly this choice maximizes entropy of neighbor selection and corresponds to the standard Einstein-Smoluchowski-Polya random walk. The stationary state[1] is given by $\pi_i = k_i / \sum_j k_j$. The other important type of random walk, maximal-entropy random walk (MERW), maximizes the entropy of random trajectories. In other words, one looks for a stochastic matrix that maximizes entropy for trajectories of given length and given end points. This is a global principle similar to the least action principle. It leads to the following stochastic matrix:

$$P_{ij} = \frac{A_{ij}}{\lambda_0} \frac{\psi_{0j}}{\psi_{0i}}, \qquad (2)$$

where λ_0 is the largest eigenvalue of the adjacency matrix $\mathbf{A}$ and ψ_{0i} is the ith element of the corresponding eigenvector $\vec{\psi}_0$. By virtue of the Frobenius-Perron theorem, all elements of this vector are strictly positive, because the adjacency matrix $\mathbf{A}$ is irreducible. For a stochastic matrix to maximize the entropy of an ensemble of paths, the choice (2) is unique. The stationary state of the stochastic matrix $\mathbf{P}$ is given by Shannon-Parry measure [16],

$$\pi_i = \psi_{0i}^2. \qquad (3)$$

The last formula intriguingly relates MERW to quantum mechanics. Namely, ψ_{0i} can be interpreted as the wave function of the ground state of the operator $-\mathbf{A}$ and ψ_{0i}^2 as the probability of finding a particle in this state [4,5]. The two types, (1) and (2), of a random walk have in general completely different properties, although on a k-regular graph exceptionally they are identical.

The stochastic matrix is not symmetric in general, so it may have different right and left eigenvectors:

$$\mathbf{P}\vec{\Psi}_\alpha = \Lambda_\alpha \vec{\Psi}_\alpha, \quad \vec{\Phi}_\alpha \mathbf{P} = \Lambda_\alpha \vec{\Phi}_\alpha. \qquad (4)$$

Throughout the paper, we consider left eigenvectors to be rows and right eigenvectors to be columns. It can be easily seen that all the eigenvalues and eigenvectors of the stochastic matrix $\mathbf{P}$ can be expressed in terms of eigenvalues λ_α and eigenvectors of $\vec{\psi}_\alpha$ of the adjacency matrix $\mathbf{A}$:

$$\Lambda_\alpha = \frac{\lambda_\alpha}{\lambda_0}, \quad \Psi_{\alpha i} = \frac{\psi_{\alpha i}}{\psi_{0i}}, \quad \Phi_{\alpha i} = \psi_{\alpha i}\psi_{0i}. \qquad (5)$$

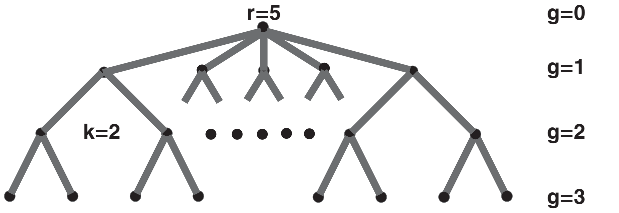

FIG. 1. Cayley tree with root degree $r = 5$, branching number $k = 2$, and $G = 3$ generations.

In particular, $\Lambda_0 = 1, \Psi_{0i} = 1$, and $\Phi_{0i} = \psi_{0i}^2 = \pi_{0i}$ for all i. The spectral decomposition of $\mathbf{P}$ reads

$$P_{ij} = \sum_\alpha \Lambda_\alpha \Psi_{\alpha i} \Phi_{\alpha j} = \sum_\alpha \frac{\lambda_\alpha \psi_{\alpha i} \psi_{\alpha j}}{\lambda_0} \frac{\psi_{0j}}{\psi_{0i}}. \qquad (6)$$

Thus, clearly all properties of MERW are encoded in the spectral decomposition of the adjacency matrix of a given graph. In what follows, we analyze the spectral properties of adjacency matrices for Cayley trees, derive the stationary state and dynamical characteristics of MERW on these trees, and compare them to GRW.

III. CAYLEY TREE

Let us consider a Cayley tree with G generations of nodes and a branching number k defined as the number of edges that connect a given node to nodes belonging to the next generation. We assume that the root of the tree has r edges, which in general may be different from k (see Fig. 1), and by convention, it belongs to the zeroth generation. Consequently, the zeroth generation contains one node, $n_0 = 1$, the first one $n_1 = r$ nodes, the second one $n_2 = rk$, the third one $n_3 = rk^2$, and so forth. The total number of nodes in the tree is $n = \sum_{g=0}^{G} n_g = 1 + r(k^G - 1)/(k - 1)$.

The adjacency matrix of the underlying graph reads

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{B_0} & & & & \\ \mathbf{B_0^T} & \mathbf{0} & \mathbf{B_1} & & & \\ & \mathbf{B_1^T} & \mathbf{0} & \mathbf{B_2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{0} & \mathbf{B_{G-1}} \\ & & & & \mathbf{B_{G-1}^T} & \mathbf{0} \end{pmatrix}, \qquad (7)$$

where the next-to-diagonal blocks $\mathbf{B_g}$ are rectangular matrices of dimensions $n_g \times n_{g+1}$:

$$\mathbf{B_g} = \begin{pmatrix} 1\ldots 1 & & & \\ & 1\ldots 1 & & \\ & & \ddots & \ddots & \\ & & & & 1\ldots 1 \end{pmatrix}. \qquad (8)$$

Each line of $\mathbf{B_g}$ contains k unities corresponding to branches leading to the descendent generation. The block $\mathbf{B_0}$ reduces to a single-row matrix with r unities. The matrices $\mathbf{B_g^T}$ are the transposes of $\mathbf{B_g}$'s.

[1] A stationary state exists if a graph is not bipartite, but, even for bipartite graphs, a semistationary state can be defined by averaging the probability distribution over two consecutive time steps.

A. Eigenvalues of the adjacency matrix

In this section, we calculate eigenvalues of the adjacency matrix of the Cayley tree using the method described in [15]. The eigenvalues are given by solutions of the equation

$$0 = \det(\mathbf{A} - \lambda\mathbf{1}) = \begin{vmatrix} \mathbf{D_0} & \mathbf{B_0} & & & & \\ \mathbf{B_0^T} & \mathbf{D_1} & \mathbf{B_1} & & & \\ & \mathbf{B_1^T} & \mathbf{D_2} & \mathbf{B_2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{D_{G-1}} & \mathbf{B_{G-1}} \\ & & & & \mathbf{B_{G-1}^T} & \mathbf{D_G} \end{vmatrix}, \tag{9}$$

where the diagonal blocks $\mathbf{D_g} = -\lambda\mathbf{1}$ are of size $n_g \times n_g$, with $n_0 = 1$, $n_1 = r$, and $n_g = rk^{g-1}$ for $g > 1$. In order to calculate the determinant, we use a sequence of elementary transformations such as additions of multiple of a row to another row, leaving the determinant invariant. This way the matrix is reduced to a triangular form with zeros above the diagonal. First, we annihilate nonzero elements of the block $\mathbf{B_{G-1}}$ by multiplying rows that contain $-\lambda$ in the diagonal block $\mathbf{D_G}$ by $1/\lambda$ and adding them to the corresponding rows in $\mathbf{B_{G-1}}$ that contain unities. This way all elements of $\mathbf{B_{G-1}}$ are turned to zero but at the same time the diagonal block $\mathbf{D_{G-1}}$ is modified to $\mathbf{D'_{G-1}} = -a_{G-1}\mathbf{1}$, where $a_{G-1} = -\lambda + k/\lambda$. Now, this procedure can be repeated to set the block $\mathbf{B_{G-2}}$ to zero by multiplying rows that contain diagonal elements of $\mathbf{D'_{G-1}}$ by $1/a_{G-1}$ and adding them to rows that contain unities in $\mathbf{B_{G-2}}$. While doing so, we see that the diagonal block $\mathbf{D_{G-2}}$ has been modified to $\mathbf{D'_{G-2}} = -a_{G-2}\mathbf{1}$, where $a_{G-1} = -\lambda - k/a_{G-2}$. Proceeding with this scheme recursively for the whole matrix, we eventually obtain a triangular matrix determinant

$$\det(\mathbf{A} - \lambda\mathbf{1}) = \begin{vmatrix} \mathbf{D'_0} & & & & & \\ \mathbf{B_0^T} & \mathbf{D'_1} & & & \mathbf{0} & \\ & \mathbf{B_1^T} & \mathbf{D'_2} & & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{D'_{G-1}} & \\ & & & & \mathbf{B_{G-1}^T} & \mathbf{D'_G} \end{vmatrix}, \tag{10}$$

with diagonal blocks $\mathbf{D'_g} = a_g\mathbf{1}$ of size $n_g \times n_g$ whose coefficients are given by

$$\begin{aligned} a_G &= -\lambda, \\ a_g &= -\lambda - k/a_{g+1} \quad \text{for} \quad g = G-1,\ldots,1, \\ a_0 &= -\lambda - r/a_1. \end{aligned} \tag{11}$$

The diagonal coefficients $a_G(\lambda) = -\lambda$, $a_{G-1}(\lambda) = -\lambda - k/\lambda$, $a_{G-2} = -\lambda - k/(-\lambda - k/\lambda)$, etc., are nested fractions in the argument λ. Hence, Eq. (9) for eigenvalues λ takes the following form:

$$\prod_{g=0}^{G}[a_g(\lambda)]^{n_g} = 0. \tag{12}$$

It is convenient to rewrite the left-hand side of the above equation as a product of polynomials instead of fractions. There is a natural set of polynomials which can be constructed from a_g's to this end:

$$\begin{aligned} A_0(\lambda) &= a_G = -\lambda, \\ A_1(\lambda) &= a_G a_{G-1} = \lambda^2 - k, \\ A_2(\lambda) &= a_G a_{G-1} a_{G-2} = -\lambda(\lambda^2 - 2k), \\ &\ldots \\ A_g(\lambda) &= -\lambda A_{g-1}(\lambda) - kA_{g-2}(\lambda) \quad \text{for} \quad g < G, \\ A_G(\lambda) &= -\lambda A_{G-1}(\lambda) - rA_{G-2}(\lambda). \end{aligned} \tag{13}$$

The recursive formula given above is derived by noticing that $A_g = A_{g-1}a_{G-g} = A_{g-1}(-\lambda - k/a_{G-g+1}) = -\lambda A_{g-1} - kA_{g-2}$. The exception is $g = G$, since then, in the last step, the coefficient k has to be replaced by r. Expressed in terms of polynomials A_g, the equation (12) reads

$$\prod_{g=0}^{G}[A_g(\lambda)]^{m_g} = 0, \tag{14}$$

where $m_G = 1$ and $m_{G-g} = n_g - n_{g-1}$, for $g = 1,2,\ldots,G$, or, equivalently, $m_{G-1} = r - 1$, $m_{G-g} = r(k-1)k^{g-2}$ for $g = 2,3,\ldots,G$. A simple analysis of the last equation shows that $A_g(\lambda)$ are polynomials of order $g+1$. Moreover, all odd order polynomials have a root equal to zero. Later, we shall see that the equation $A_g(\lambda) = 0$ has $g+1$ real roots, and that if λ is a root, then $-\lambda$ also is. The total number of real roots of Eq. (14) counted with degeneracy m_g is $\sum_g (g+1)m_g = \sum_g n_g = n$, so Eq. (14) gives all n eigenvalues of the adjacency matrix. The equation $A_0(\lambda) = 0$ gives eigenvalues $\lambda = 0$ with the degeneracy $m_G = r(k-1)k^{G-2}$, the equation $A_1(\lambda) = 0$ gives eigenvalues $\pm\sqrt{k}$ with the degeneracy $m_G = r(k-1)k^{G-3}$, etc. It should be noticed that some eigenvalues may be solutions of $A_g(\lambda) = 0$ for different g. For instance, $\lambda = 0$ is a root of $A_g(\lambda) = 0$ for all even g, so the total degeneracy of the eigenvalue $\lambda = 0$ is $\sum_g (2g+1)m_{2g}$.

It turns out that the solutions of equations $A_g(\lambda) = 0$ can be found systematically. The polynomials $A_g(\lambda)$ for $g < G$ (13) can be written in a concise form using an auxiliary parameter θ (see Appendix A):

$$A_g = k^{(g+1)/2}\frac{\sin[(g+2)\theta]}{\sin\theta}, \tag{15}$$

where

$$\cos\theta = -\frac{\lambda}{2\sqrt{k}}. \tag{16}$$

It can be checked by inspection that these equations indeed reproduce the polynomials (13). For example, for $g = 0$, one retrieves $A_0 = \sqrt{k}\sin(2\theta)/\sin\theta = 2\sqrt{k}\cos\theta = -\lambda$; for $g = 1$, $A_1 = k\sin(3\theta)/\sin\theta = k[4(\cos\theta)^2 - 1] = \lambda^2 - k$, etc., in agreement with (13). The equation for A_G can be obtained by combining the last equation in (13), $A_G = -\lambda A_{G-1} - kA_{G-2}$, with the explicit form of A_{G-1} and A_{G-2} (15), which yields

$$A_G = k^{(G-1)/2}\frac{k\sin[(G+2)\theta] + (k-r)\sin(G\theta)}{\sin\theta}, \tag{17}$$

where θ is given by (16). When the root of the tree has $r = k$ neighbors (equal to the branching number of the tree), the last equation reduces to the one for remaining generations (15).

The eigenvalues of the adjacency matrix can be determined by finding values of the auxiliary parameter θ for which

A_g (15) and A_G (17) are zero, and inserting these values to the formula $\lambda = -2\sqrt{k}\cos\theta$ (16). As can be seen, A_g (15) for $g < G$ is equal to zero for $\theta \neq 0$, fulfilling the equation

$$\sin[(g+2)\theta] = 0 \tag{18}$$

that has $g+1$ solutions,

$$\lambda_{g,j} = 2\sqrt{k}\cos\left(\frac{\pi j}{g+2}\right) \quad \text{for} \quad j = 1,\ldots,g+1. \tag{19}$$

Each eigenvalue in this series is m_g times degenerated, as follows from (14). The situation is slightly more complicated for $g = G$, since the equation $A_G = 0$ amounts to an equation for θ,

$$k\sin[(G+2)\theta] + (k-r)\sin(G\theta) = 0, \tag{20}$$

that can be solved analytically only for $r = k$ or $r = 2k$. In the first case, exactly the same formula as for $g < G$ (19) is obtained,

$$\lambda_{G,j} = 2\sqrt{k}\cos\left(\frac{\pi j}{G+2}\right) \quad \text{for} \quad j = 1,\ldots,G+1, \tag{21}$$

while in the second one,

$$\lambda_{G,j} = 2\sqrt{k}\cos\left[\frac{\pi(j-1/2)}{G+1}\right] \quad \text{for} \quad j = 1,\ldots,G+1. \tag{22}$$

For other values of r, one has to solve (20) numerically. The largest eigenvalue of the adjacency matrix is $\lambda_0 = \lambda_{G,1}$. For $r = k$, it is equal to

$$\lambda_0 = \lambda_{G,1} = 2\sqrt{k}\cos\left(\frac{\pi}{G+2}\right), \tag{23}$$

while for $r = 2k$, it is

$$\lambda_0 = \lambda_{G,1} = 2\sqrt{k}\cos\left(\frac{\pi}{2G+2}\right). \tag{24}$$

For other values of r, the eigenvalue λ_0 can be determined approximately, as discussed in Appendix B. The solutions can be divided into three classes with respect to values of r: the first class for $r \in (0,2k - 2k/G)$, the second one for $r \in (2k - 2k/G, 2k + 2k/G)$, and the third one for $r \in (2k + 2k/G, +\infty)$. In the large G limit, i.e., for $G \gg 2k$, the second class reduces to a single integer value of $r = 2k$ for which the solution is known (24). The first class corresponds to the values $r < 2k$ for which the approximate solution reads

$$\lambda_0 = 2\sqrt{k}\cos\frac{\pi}{G+\delta}, \tag{25}$$

where

$$\delta \approx \frac{2k}{2k-r}, \tag{26}$$

as explained in Appendix B. For the third class, $r > 2k$, the equation (20) has no real solutions in the range $[0,\pi/(G+1)]$, and the largest eigenvalue λ_0 is obtained from a purely imaginary solution for θ. The corresponding equations change from trigonometric to hyperbolic. For large G, the solution can be approximated by

$$\lambda_0 = \frac{2\sqrt{k}}{\sqrt{1-x^2}}, \tag{27}$$

where

$$x = z\left[1 - 2\left(\frac{1-z}{1+z}\right)^{G+1}\right] \tag{28}$$

and

$$z = 1 - \frac{2k}{r}. \tag{29}$$

Again, we refer the reader to Appendix B for details. One sees that x approaches z exponentially as G grows, so for large G, one can substitute x by z in (27) to eventually obtain

$$\lambda_0 \approx \frac{r}{\sqrt{r-k}}. \tag{30}$$

As can be seen, the largest eigenvalue for trees with a strongly branched root, $r > 2k$, behaves differently as compared to trees with a weakly branched root, $r < 2k$. This eigenvalue is now larger than $2\sqrt{k}$, while it was smaller in the previous case; it grows with r, and it is weakly dependent on G.

IV. THE EIGENVECTOR TO THE LEADING EIGENVALUE

In order to obtain the stationary state of MERW, the largest eigenvalue λ_0 and the squared elements of the eigenvector $\vec{\psi}_0$ associated with this eigenvalue are needed:

$$(\mathbf{A} - \lambda_0\mathbf{1})\vec{\psi}_0 = 0. \tag{31}$$

The ground state $\vec{\psi}_0$ has a helpful symmetry in the sense that all elements ψ_{0i} for nodes in a given generation g are identical. So the problem can be simplified by ascribing the same value ψ_g to all nodes in the generation (henceforth, when we write out the elements of the eigenvector, we omit the index corresponding to the eigenvalue):

$$\vec{\psi}_0 = (\psi_0, \underbrace{\psi_1,\ldots,\psi_1}_{n_1},\ldots,\underbrace{\psi_G,\ldots,\psi_G}_{n_G}). \tag{32}$$

Effectively, instead of n equations for ψ_{0i}, $i = 1,\ldots,n$, (31), there are just $(G+1)$ independent equations for ψ_g, $g = 0,\ldots,G$, remaining:

$$\begin{aligned} -\lambda_0\psi_0 + r\psi_1 &= 0, \\ \psi_{g-1} - \lambda_0\psi_g + k\psi_{g+1} &= 0 \quad \text{for} \quad g = 1,\ldots,G-2, \\ \psi_{G-1} - \lambda_0\psi_G &= 0. \end{aligned} \tag{33}$$

This recurrence can be solved starting from the end, $g = G$, and decreasing g to 0. For convenience, we introduce coefficients

$$C_g = \frac{\psi_{G-g}}{\psi_G} \tag{34}$$

that invert the order of the recurrence. They correspond to the original values normalized to ψ_G, in particular $C_0 = 1$. The recurrence relations (33) are equivalent to

$$C_g = \lambda_0 C_{g-1} - kC_{g-2} \quad \text{for} \quad g = 2,\ldots,G, \tag{35}$$

with the initial condition $C_0 = 1$, $C_1 = \lambda_0$. Let us note that the recurrence relation is identical as for A_g (13) when λ_0 is replaced by $-\lambda_0$. The initial condition is also identical, except that the counter of the recurrence is shifted by one, so the solution can be copied: $C_g(\lambda_0) = A_{g-1}(-\lambda_0)$ to obtain

$$C_g = k^{g/2}\frac{\sin[(g+1)\theta]}{\sin\theta} \quad \text{for} \quad g = 0,\ldots,G, \tag{36}$$

where $\cos\theta = \lambda_0/2\sqrt{k}$. The first equation in (33) $-\lambda_0\psi_0 + r\psi_1 = 0$, which corresponds to an equation $-\lambda_0 C_G + rC_{G-1} = 0$, that is automatically fulfilled for C_G and C_{G-1} given by (36) under substitution of $\lambda_0 = 2\sqrt{k}\cos\theta$ and $r\sin(G\theta) = k\sin[(G+2)\theta] + k\sin(G\theta)$ according to Eq. (20).

This concludes our calculations of the eigenvector to the leading eigenvalue of the adjacency matrix. Using (34), we have

$$\psi_g = C_{G-g}\psi_G = \frac{C_{G-g}}{\sum_h C_h^2} \tag{37}$$

for all nodes in the gth generation. The value ψ_G is chosen to ensure the proper normalization, $\sum_g \psi_{0,g}^2 = 1$.

V. THE EIGENVECTOR TO NEXT-TO-LEADING EIGENVALUES

In the case of the eigenvector $\vec{\psi}_1$ to the eigenvalue λ_1, we exploit the fact that it is symmetric within each of r principal branches of the tree (which means that for given generation g within the branch, all the elements ψ_g are the same; once again, when writing out the elements of the vector, we omit the index corresponding to the number of the eigenvalue). In appropriate coordinates, the elements belonging to these principal branches can be separated,

$$\vec{\psi}_1 = (\psi_0, \alpha_1\vec{\phi}, \dots, \alpha_r\vec{\phi}), \tag{38}$$

where the branches may have different multiplicative factors $\alpha_1, \dots, \alpha_r$ and the vector

$$\vec{\phi} = (\psi_1, \underbrace{\psi_2, \dots, \psi_2}_{n_2/r}, \dots, \underbrace{\psi_G, \dots, \psi_G}_{n_G/r}) \tag{39}$$

represents the relative value of the eigenvector elements in each branch. The multiplicities n_g are evenly distributed among the r branches, hence the factor $1/r$.

We obtain $(G+1)$ independent equations for ψ_g, $g = 0, \dots, G$, in analogy to Eq. (33):

$$-\lambda_1\psi_0 + (\alpha_1 + \dots + \alpha_r)\psi_1 = 0, \tag{40a}$$
$$\psi_0/\alpha_i - \lambda_1\psi_1 + k\psi_2 = 0 \quad \text{for} \quad i = 1, \dots, r, \tag{40b}$$
$$\psi_{g-1} - \lambda_1\psi_g + k\psi_{g+1} = 0 \quad \text{for} \quad g = 2, \dots, G-1,$$
$$\psi_{G-1} - \lambda_1\psi_G = 0. \tag{40c}$$

The only difference is the first two equalities above, which show how the r branches couple together at the root of the tree. The rest of the equalities stay the same, as the recurrence progresses only within a given branch and the factor α_i is eliminated.

For each of the branches, the system is solved starting from $g = G$ and decreasing g to 1. Until this point, the solution is the same as before (36).

Now, we check if (40a) and (40b) are consistent with this solution. Clearly, in Eq. (40b), the terms $-\lambda_0\psi_1 + k\psi_2 = k^{G/2}\frac{\sin[(G+1)\theta]}{\sin\theta}\psi_G = 0$, because λ_1 corresponds to the value $\theta = \frac{\pi}{G+1}$. Thus, after rewriting, Eqs. (40a) and (40b) take the form

$$\alpha_1 + \dots + \alpha_r = 0, \tag{41a}$$
$$\psi_0 = 0. \tag{41b}$$

In fact, $\psi_0 = 0$ is consistent with the explicit solution $C_G \propto \sin[(G+1)\theta] = 0$. If we recall the form of eigenvalues given in (19), of which one special case was $\lambda_1 = \lambda_{G-1,1}$, it is noticeable that for each $g = 0, \dots, G-1$, the eigenvalue $\lambda_{g,1}$ corresponds to the angle $\frac{\pi}{g+2}$ and so the solution of the recurrence equation vanishes for generation $G - g$. This is the point at which the symmetry of the corresponding eigenvector is broken. Such a vector to the eigenvalue $\lambda_{g,1}$ has the elements $\psi_{g'} = 0$ for $g' < G - g$, and the symmetric values of $\psi_{g'}$ for $g' \geqslant G - g$. We do not discuss here the eigenvectors to the other eigenvalues.

The last point concerns the multiplication factors of branches $\alpha_1, \dots, \alpha_r$. Noticeably, one of them can incorporate the normalization factor ψ_G, which leaves r free parameters. There are, however, $r - 1$ eigenvectors in the eigenspace of λ_1, so there are in fact $r(r-1)$ parameters [for lower eigenvalues, one needs to include the degeneration according to (14)]. Now, there are also constraints: $r - 1$ normalization conditions, $r - 1$ constraints in (41b), and $\binom{r-1}{2} = \frac{(r-1)(r-2)}{2}$ pairwise orthogonalizations. This leads to the number

$$r(r-1) - (r-1) - (r-1) - \frac{(r-1)(r-2)}{2}$$
$$= \frac{(r-1)(r-2)}{2} \tag{42}$$

of free parameters, which are the allowed rotations $O(r-1)$ of the eigenspace.

To illustrate this with a simple example, let us take $r = 3$, which gives $r - 1 = 2$ eigenvectors to λ_1:

$$\vec{\psi} = (0, \alpha_1\psi_1, \dots, \alpha_3\psi_1, \dots, \alpha_1\psi_G, \dots, \alpha_3\psi_G), \tag{43}$$
$$\vec{\phi} = (0, \beta_1\phi_1, \dots, \beta_3\phi_1, \dots, \beta_1\phi_G, \dots, \beta_3\phi_G). \tag{44}$$

Two normalization conditions, for $\vec{\phi}$ and $\vec{\psi}$, rid Eqs. (41a) of two parameters,

$$\alpha_1 + \alpha_2 + 1 = 0, \tag{45a}$$
$$\beta_1 + \beta_2 + 1 = 0, \tag{45b}$$

while the orthogonalization $\vec{\phi} \cdot \vec{\psi} = 0$ gives

$$\alpha_1\beta_1 + \alpha_2\beta_2 + 1 = 0, \tag{46}$$

and finally the symmetric relation between the two vectors is obtained, leaving one free parameter that rotates them,

$$2 + \alpha_1 + \beta_1 + 2\alpha_1\beta_1 = 0. \tag{47}$$

VI. THE EIGENVALUES OF THE GRW TRANSITION MATRIX

Under the same procedure of transforming the determinant to the triangular form, as explained in Sec. III, the transition matrix of generic random walk defined in (1) leads to similar recurrence equations as in (11):

$$\begin{aligned} a_G &= -\lambda, \\ a_{G-1} &= -\lambda - \frac{k}{k+1}\frac{1}{a_G}, \\ a_l &= -\lambda - \frac{k}{(k+1)^2}\frac{1}{a_{l+1}} \quad \text{for} \quad g = G-2, \dots, 1, \\ a_0 &= -\lambda - \frac{1}{k+1}\frac{1}{a_1}. \end{aligned} \tag{48}$$

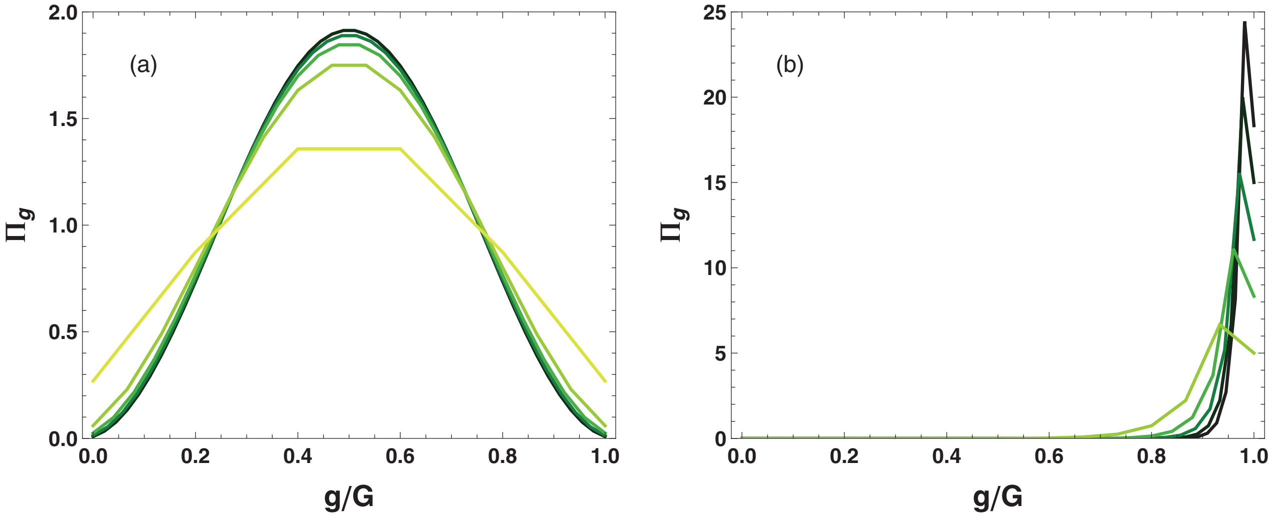

FIG. 2. (Color online) Finite-size effect: the broken lines correspond to the distribution Π_g for $G = 5,\dots,45$ in steps of 10, for a tree with $k = r = 3$. (a) The distributions for MERW. For $r < k$, the corresponding curves would be skewed and would approach the limiting distribution from the right, while for $r > k$, they would approach from the left. (b) The distributions for GRW. The larger the number of generations, the more peaked the distribution.

In the last equality, the factor r appears in the numerator and denominator, so it cancels out, and the equations remain r independent.

We proceed as before and define

$$A_g(\lambda) = \prod_{j=0}^{g} a_{G-j}(\lambda) \quad \text{for} \quad g = 0,\dots,G, \tag{49}$$

and hence we get the recurrence relations

$$\begin{aligned} A_g(\lambda) &= -\lambda A_{g-1}(\lambda) - \frac{k}{(k+1)^2}A_{g-2}(\lambda) \quad \text{for} \quad g < G, \\ A_G(\lambda) &= -\lambda A_{G-1}(\lambda) - \frac{1}{(k+1)^2}A_{G-2}(\lambda). \end{aligned} \tag{50}$$

The general solution for $g < G$ is (see Appendix A for details)

$$A_g(\lambda) = \left[\frac{k}{(k+1)^2}\right]^{(g+1)/2} \frac{\sin[(g+2)\theta] - k\sin(g\theta)}{\sin\theta} \quad \text{for} \quad g = 0,\dots,G-1, \tag{51}$$

where $\cos\theta = \frac{\lambda}{2}\sqrt{\frac{(k+1)^2}{k}}$. $A_G(\lambda)$ can be found by inserting the above solution to Eq. (50),

$$A_G(\lambda) = \left[\frac{k}{(k+1)^2}\right]^{(G+1)/2} \frac{(2k\cos\theta - 1 - k^2)\sin(G\theta)}{k\sin\theta}. \tag{52}$$

Now, the eigenvalues of GRW transition matrix are determined by finding values of the auxiliary parameter θ for which A_g (51) and A_G (52) are zero. We first solve the equation for $g = G$, which factorizes into two parts,

$$2k\cos(2\theta) = 1 + k^2, \tag{53}$$

whose solution is the largest eigenvalue of the transition matrix,

$$\lambda_0 = 1, \tag{54}$$

the second part being

$$\sin(G\theta) = 0, \tag{55}$$

which gives

$$\lambda_{G,j} = 2\sqrt{\frac{k}{(k+1)^2}}\cos\left(\frac{\pi j}{G}\right) \quad \text{for} \quad j = 1,\dots,G. \tag{56}$$

For $g < G$, we obtain

$$\sin[(g+2)\theta] = k\sin(g\theta), \tag{57}$$

which has the same form as Eq. (20), but with different coefficients. For $k > 1$ in (57), we enter the same range of parameters as for $r \in (2k + 2k/G, +\infty)$ in Eq. (20), which means that the solution leading to the largest eigenvalue in a given series is imaginary. The corresponding equations change from trigonometric to hyperbolic. Under substitution $k = \frac{z+1}{1-z}$, $z = \frac{k-1}{k+1}$, where z was given in (29), definition (28) reads

$$x = \frac{k-1}{k+1}[1 - 2k^{-(g+1)}]. \tag{58}$$

We are particularly interested in the second largest eigenvalue (corresponding to the series $g = G - 1$). For large G, the solution is approximated by

$$\lambda_1 = 2\sqrt{\frac{k}{(k+1)^2}}\frac{1}{\sqrt{1-x^2}}, \tag{59}$$

and it can be easily seen that λ_1 exponentially approaches $\lambda_0 = 1$ for large G.

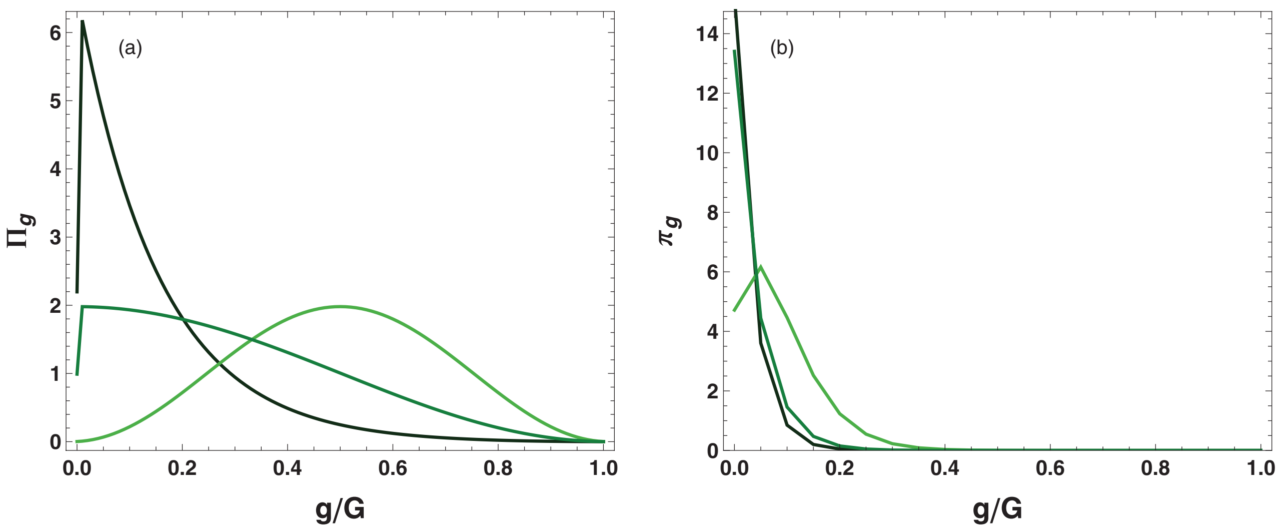

FIG. 3. (Color online) (a) Plots for Π_g, which is a total probability for a generation. Curves for large G: a sine square for $r < 2k$, cosine square for $r = 2k$, and hyperbolic sine for $r > 2k$ ($k = 3$ and $r = 3,6,9$). (b) Probability per one node π_g for $r = 3,6,9$, and $G = 20$.

VII. STATIONARY STATES OF GRW AND MERW ON CAYLEY TREES

As mentioned earlier, a stationary state for a random walk on a graph exists if the graph is not bipartite. In the case of bipartite graphs, a semistationary state can be defined by averaging probability distributions over two consecutive steps (because even and odd times are independent) or by averaging the state over initial configurations.

The stationary state of GRW is given by the linear dependence on the degree of the vertices,

$$\pi_i = \frac{k_i}{\sum_j k_j} \quad \text{for} \quad i = 1, \ldots, n, \tag{60}$$

so the distribution is flat (degree $k_i = k + 1$), but for the root (degree r) and leaves (degree 1). If we sum the probabilities over whole generations, the exponential factor appears as

$$\Pi_g = n_g \pi_i = k^{g-1}\frac{k^2 - 1}{k^G - 1} \quad \text{for} \quad g = 1, \ldots, G, \quad \text{and} \quad i \in g. \tag{61}$$

The stationary state of maximal-entropy random walk is given by squared elements of $\vec{\psi}_0$, the eigenvector to the largest eigenvalue of the adjacency matrix:

$$\pi_i = \psi_{0i}^2. \tag{62}$$

Remembering the solution (37),

$$\pi_i \propto k^{G-g} \sin[(G - g + 1)\theta]^2 \quad \text{for} \quad g = 0, \ldots, G, \quad \text{and} \quad i \in g, \tag{63}$$

where we omitted the normalization factor. As we sum the stationary probability over $i \in g$, we get

$$\Pi_g = n_g \pi_i \propto k^{G-1} \sin[(G - g + 1)\theta]^2 \quad \text{for} \quad g = 1, \ldots, G, \quad \text{and} \quad i \in g, \tag{64}$$

where the only exception is $g = 0$ with its $n_0 = 1$. Exemplary probability distributions Π_g for MERW and GRW are shown in Fig. 2.

Now, as this result depends on θ, and the solutions for λ depend on whether $r < 2k$ [Eq. (21)], $r = 2k$ [Eq. (22)], or $r > 2k$ [Eq. (27)], this means that we can get different distributions for different choices of r. For $r < 2k$, parameter $\theta \approx \frac{\pi}{G+\delta}$ and the distribution remains a sine square; for $r = 2k$, $\theta = \frac{\pi/2}{G+1}$ and the distribution becomes a cosine square; for $r > 2k$, $\theta = i \;\operatorname{arctanh} x$ [where x is given in (28) and i is the imaginary unit], thus we obtain a hyperbolic sine. Figure 3 illustrates these cases. An interactive demonstration showing these results as well as finite-size effects is available online [17].

VIII. RELAXATION TIMES

A. General considerations

Let us denote the probability of finding a particle at a node i at time t of random walk by $\pi_i(t)$, and the probability distribution on the whole graph $\{\pi_i(t)\}_{i=1,\ldots,n}$ by $\vec{\pi}(t)$. Given the initial probability distribution $\vec{\pi}(0)$ and the stochastic matrix $\mathbf{P}$, one can determine the distribution at any time t,

$$\vec{\pi}(t) = \vec{\pi}(0)\mathbf{P}^t. \tag{65}$$

Using the spectral decomposition of the stochastic matrix (6), one can rewrite the last equation as

$$\vec{\pi}(t) = \sum_\alpha c_\alpha \Lambda_\alpha^t \vec{\Phi}_\alpha, \tag{66}$$

where c_α is a spectral coefficient: $c_\alpha = \vec{\pi}(0) \cdot \vec{\Psi}_\alpha = \sum_i \pi_i(0)\Psi_{\alpha i}$. In particular, $c_0 = \sum_i \pi_i(0) = 1$. In general, all eigenvalues Λ_α of the stochastic matrix are known to be located inside or on the unit circle in the complex plane $|\Lambda_\alpha| \leqslant 1$. In the limit $t \to \infty$, all terms in the sum on the right-hand side of the last equation for $|\Lambda_\alpha| < 1$ are suppressed exponentially, and only those for $|\Lambda_\alpha| = 1$ survive. The stochastic matrices for GRW or MERW on trees have only two eigenvalues on the unit circle:[2] $\Lambda_0 = 1$ and $\Lambda_n = -1$, so for large t, one has

$$\vec{\pi}(t) \approx c_0\vec{\Phi}_0 + (-1)^t c_n\vec{\Phi}_n. \tag{67}$$

The eigenvectors associated with the eigenvalue $\Lambda_0 = 1$ are $\Psi_{0i} = 1$ for all i, $\Phi_{0i} = \psi_{0i}^2 = \pi_i$. In order to write down the eigenvectors associated with the eigenvalue $\Lambda_n = -1$, it is convenient to bipartition the graph into nodes belonging to generations numbered by odd and even g. Naturally, the "odd" nodes are neighbors of "even" ones only, and vice versa. Elements of the eigenvectors are $\Psi_{nj_o} = 1$, $\Psi_{nj_e} = -1$, $\Phi_{nj_o} = \pi_{j_o}$, and $\Phi_{nj_e} = -\pi_{j_e}$, where the index j_o runs over odd nodes and j_e runs over even nodes. This gives, for large t,

$$\begin{aligned}\pi_{j_o}(2t) = 2\sigma\pi_{j_o}, \quad \pi_{j_e}(2t) = 2(1-\sigma)\pi_{j_e},\\ \pi_{j_o}(2t+1) = 2(1-\sigma)\pi_{j_o}, \quad \pi_{j_e}(2t+1) = 2\sigma\pi_{j_e},\end{aligned} \tag{68}$$

where σ is the probability that a particle is in the odd part. Clearly, $c_n = 2\sigma - 1$, and for $\sigma = 1/2$, the stationary state is recovered. The equations above tell us that the probability distribution oscillates between odd and even nodes. In a single step of a random walk, particles disappear from odd nodes to appear on even ones, and vice versa. If one traces the state of the random walk process every second step, one sees that the distributions of particles on odd and even nodes approach the stationary state in each partition. The relaxation to the stationary state is generically governed by the next-to-leading eigenvalue Λ_1 and its negative partner $\Lambda_{n-1} = -\Lambda_1$. The corresponding term in the spectral decomposition (66) reads $\sum[c_1\vec{\Phi}_1 + (-1)^t c_{n-1}\vec{\Phi}_{n-1}]\Lambda_1^t$, and its contribution to the sum vanishes exponentially as $\exp(-t/\tau_1)$, where $\tau_1 = [-\ln(\Lambda_1)]^{-1} = [\ln(\lambda_0/\lambda_1)]^{-1}$. The symbolic sum $\sum$ indicates that all eigenvectors in the eigenspaces of Λ_1 and Λ_{n-1} are taken into account. The exception is the case when the corresponding spectral coefficients c_1 and c_{n-1} vanish, since then also the corresponding term vanishes. In that case, the next-to-leading contribution in the large t limit comes from a lower eigenvalue Λ_k, the largest with a nonvanishing spectral coefficient.

Thus, by τ_1, we denote the generic relaxation time, the largest one, and by τ_2, the one associated with $\lambda_{G,2}$ (in the Sec. VIII H, we explain what symmetries lead to this relaxation). As there are several tree parameter regimes which yield different results for adjacency matrix eigenvalues, the relaxation times for MERW in those cases are also different. As explained in Appendix B, for eigenvalues of the adjacency matrix, the relation $\lambda_{G-1,1} > \lambda_{G,2}$ always holds, so the second largest eigenvalue is $\lambda_1 = \lambda_{G-1,1}$, unless some special parameters k,r are chosen. Thus, we discuss below the strongly, critically, and weakly branched root, and then some special cases. The discussion of relaxation for GRW and remarks on numerical measurements conclude this section. An interactive demonstration illustrating the results concerning relaxation is available online [18].

B. Strongly branched root

The strongest root branching that yields qualitatively distinct behavior of MERW is $r > 2k$, where $k > 1$ is assumed. The largest eigenvalue λ_0 is given by (27), while the second largest eigenvalue, with multiplicity $r - 1$, belongs to the second level of hierarchy of eigenvalues,

$$\lambda_1 = \lambda_{G-1,1} = 2\sqrt{k}\cos\left(\frac{\pi}{G+1}\right). \tag{69}$$

Thus, the generic relaxation time reads

$$\tau_1 = -\left\{\ln\left[\sqrt{1-x^2}\cos\left(\frac{\pi}{G+1}\right)\right]\right\}^{-1}, \tag{70}$$

where x, defined in (28), approaches $\frac{r-2k}{r}$ exponentially fast when $G \to \infty$. Hence, asymptotically,

$$\tau_1 \cong c + \frac{c^2\pi^2}{2}\frac{1}{G^2} + \cdots \longrightarrow c = \text{const}, \tag{71}$$

where

$$c = \left[\ln\frac{r}{2\sqrt{(r-k)k}}\right]^{-1}, \tag{72}$$

which gives an extremely fast relaxation, with the relaxation time converging to a constant for large G. A faster relaxation resulting from symmetry and associated with the eigenvalue $\lambda_{G,2}$ can be found as well; however, the relaxation time might only be improved by a multiplicative constant.

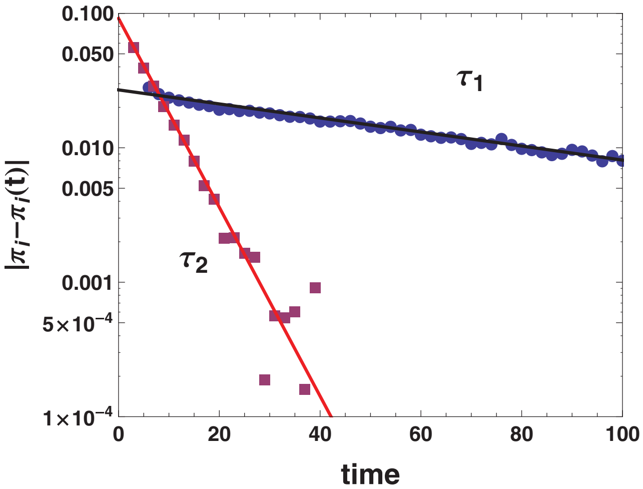

FIG. 4. (Color online) Cayley tree with $k = r = 2$ and $G = 8$: the mild slope (circles) corresponds to the generic relaxation and the steep slope (squares) corresponds to a symmetry-induced one. The data points were generated in a Monte Carlo simulation with 5×10^4 random walkers, all starting either from a single node in the third generation, which leads to the typical behavior (generic relaxation), or from the root (the most symmetric initial condition), which always leads to a faster relaxation. The distance between the stationary state π_i and the probability $\pi_i(t)$, measured at a single node i belonging to the second generation, is shown on a logarithmic scale. The lines represent theoretical slopes corresponding to $\tau_1 = [\ln(\lambda_0/\lambda_1)]^{-1} \approx 83$ and $\tau_2 = [\ln(\lambda_0/\lambda_{G,2})]^{-1} \approx 6$.

[2] More generally, since trees are bipartite, one can show that if Λ is an eigenvalue, then also $-\Lambda$ is.

C. Critically branched root

The behavior of MERW changes for the special case of $r = 2k$, $k > 1$, as could be observed in the stationary states. The largest eigenvalue is given by (22) and the second largest eigenvalue, $\lambda_1 = \lambda_{G-1,1}$, as before, hence the asymptotic relaxation time is

$$\tau_1 \cong \frac{8G^2}{3\pi^2} + \frac{16G}{3\pi^2} + \cdots . \tag{73}$$

The symmetry-induced relaxation corresponding to $\lambda_{G,2} = 2\sqrt{k}\cos(\frac{3\pi/2}{G+1})$ produces asymptotic behavior with the same scaling with respect to the number of generations,

$$\tau_2 \cong \frac{1}{\pi^2}G^2 + \frac{2}{\pi^2}G + \cdots . \tag{74}$$

It is worth noting that while the number of vertices $n \sim k^G$, the probability distribution relaxes as a logarithm of the system size $\tau_1, \tau_2 \sim \ln n$, which still is rather fast.

D. Weakly branched root

After passing the critical value of $r = 2k$, the tree enters the regime of a weakly branched root, where $1 < r < 2k$, $k > 1$. The only exact solution for λ_0 in this range of parameters is $r = k$ in (21), otherwise there is the approximation (25) at our disposal. The second largest eigenvalue is the same as above, $\lambda_1 = \lambda_{G-1,1}$. Hence, the generic relaxation follows

$$\tau_1 \cong \frac{2k - r}{r\pi^2}G^3 + \frac{3(4k - r)}{2r\pi^2}G^2 + \cdots , \tag{75}$$

and the faster relaxation relying on $\lambda_{G,2}$ gives

$$\tau_2 \cong \frac{2}{3\pi^2}G^2 - \frac{8k}{3\pi^2(r - 2k)}G + \cdots . \tag{76}$$

Noticeably, the generic relaxation time τ_1 is G times longer than τ_2 and than both relaxation times for the tree with a critically branched root.

E. Planted tree

Until now, we have considered only the root of degree $r > 1$, where all the levels in the hierarchy of the eigenvalues have a nonzero degeneracy. Trees with a root of degree $r = 1$ (known as *planted trees*) are a special case, because the level $\lambda_{G-1,j}$ of the hierarchy has degeneracy $m_{G-1} = r - 1 = 0$. Thus, the second largest eigenvalue is $\lambda_1 = \lambda_{G-2,1}$, while λ_0 is approximated by (25), and the generic relaxation time is given by

$$\tau_1 \cong \frac{2k - 1}{2k\pi^2}G^3 + \frac{3}{2\pi^2}G^2 + \cdots . \tag{77}$$

The faster relaxation remains associated with the eigenvalue $\lambda_{G,2}$, so the asymptote (76) is still valid for τ_2 after inserting $r = 1$.

F. Linear chain

Parameters $k = 1$, $r = 1$ produce a particularly degenerate case of a Cayley tree, namely, a linear chain. While $m_{G-1} = r - 1 = 0$ and $m_{G-g} = r(k - 1)k^{g-2} = 0$, there remains only one level in the hierarchy of the eigenvalues of the adjacency matrix,

$$\lambda_{G,j} = 2\sqrt{k}\cos\left(\frac{j\pi}{G + 2}\right), \quad j = 1, \ldots , G + 1. \tag{78}$$

Naturally, $\lambda_{G,i} > \lambda_{G,j}$ for $i < j$, so $\lambda_0 = \lambda_{G,1}$ and $\lambda_1 = \lambda_{G,2}$, hence

$$\tau_1 \cong \frac{2G^2}{3\pi^2} + \frac{8G}{3\pi^2} + \cdots , \tag{79}$$

and the relaxation connected with the third eigenvalue,

$$\tau_2 \cong \frac{G^2}{4\pi^2} + \frac{G}{\pi^2} + \cdots . \tag{80}$$

However, if the number of generations G is odd (n even), then there does not exist a central vertex where this relaxation could be measured. If G is even ($n = G + 1$ is odd; it actually might be translated to $r' = 2, k' = 1, G' = G/2$ Cayley tree, although the solution differs from the previous ones), then one central node exists and the faster relaxation can be measured there or if some symmetric initial conditions are taken.

Finally, let us notice that the system size is $n = G + 1$ and the scaling is $\tau_1, \tau_2 \sim n^2$. This is the same result as for a simple diffusion, which is modeled by GRW.

G. GRW relaxation times

For GRW, $\lambda_0 = 1$ and the second largest eigenvalue is given by (59) for all $k > 1$. It follows that the relaxation time is given by

$$\tau_1 \cong 2\left\{\ln\left[1 + \frac{(k - 1)^2}{4k}\frac{4k^G - 1}{4k^{G2}}\right]\right\}^{-1} . \tag{81}$$

After using Taylor expansion, in the limit of large G,

$$\tau_1 \cong \frac{8k}{(k - 1)^2}k^G , \tag{82}$$

which means that $\tau_1 \sim n$. Just as in the case of MERW, for $r = 1$, one level of the hierarchy of eigenvalues vanishes and λ_1 has to be taken as the solution of Eq. (57), with $g = G - 2$ instead of $g = G - 1$.

The eigenvalue associated with the faster relaxation is $\lambda_{G,1}$ and it leads to the characteristic time

$$\tau_2 \cong c - \frac{c^2\pi^2}{2}\frac{1}{G^2} + \cdots \longrightarrow c = \text{const}, \tag{83}$$

where

$$c = -\left\{\ln\left[2\sqrt{\frac{k}{(k + 1)^2}}\right]\right\}^{-1} . \tag{84}$$

H. Numerical measurements

It is possible to measure the relaxation process in two ways: either by explicitly taking powers of the transition matrix or by Monte Carlo simulation with N walkers traversing the graph.

In the former case, compute the transition matrix $\mathbf{P}$, choose the initial conditions (initial probabilities for any vertex of the graph), obtain the power of the transition matrix $\mathbf{P}^t$ (one might use spectral decomposition for that, although for large t

better precision is needed) corresponding to probabilities after t steps, and measure the difference between the stationary state that we have found theoretically. One might need to take the average of two consecutive steps to avoid the odd-even blinking.

In the case of Monte Carlo, let N walkers start from node a (or a set of nodes), let every sweep for each of those walkers draw a random number, and check it against the transition matrix to know in which direction the walker should go. At a node b, measure the number of random walkers at the sweep t, normalize it to the total number of walkers, and subtract the stationary state probability.

We have confirmed the theoretical relaxation times in both ways.

The difference between the stationary state and the probability at time t might be averaged over all nodes of the tree. However, to observe both the generic and the faster relaxation, one might do one of the following:

(1) take one initial vertex with probability 1, and one measuring vertex, or

(2) take r initial vertices with probabilities $p_1, p_2, \ldots, p_r$, and one measuring vertex.

In the first case, if the initial vertex *or* the vertex at which one measures probabilities is the root, then the observed relaxation time is τ_2 and τ_1 otherwise. In the second case, if the vertices and probabilities are chosen symmetrically (e.g., for $r = 2$, the two neighbors of the root with probabilities $1/2$ each), then one also sees τ_2 if measuring the relaxation in the generation $g = 1$. An interactive demonstration allowing the study of this behavior is available online [18].

In general, one might spot other relaxations upon specific choices of initial conditions. This may be seen as eliminating contributions from given eigenvalues in the spectral decomposition of **P** (6), as explained in Sec. VIII A. Intuitively, this is the same phenomenon as interference of waves, although we deal with probability waves here.

IX. CONCLUSIONS

In this paper, we have analytically derived the form of the stationary state for GRW and MERW on Cayley trees, which shows that the stationary probability of the latter is centered around the root of a tree, in contrast to the flat distribution of the former. The dynamics of the probability approaching to the stationary state have proven to be generically faster for MERW (logarithmic with respect to the system size) than for GRW (linear with respect to the system size).

While maximal-entropy random walk is defined so as to keep all paths of a given length between two given points equiprobable, it might be considered a process capable of hiding the route that the information has traveled, e.g., on the Internet. The properties of stationary probability distribution of MERW have already been used to enhance centrality measures in complex networks [10]. Considering the faster dynamics of MERW and the connection of eigenvalues of the adjacency matrix to the paths' statistics (which are a basis for a number of community detection algorithms [19]), this type of random walk may prove useful in finding community structures on complex networks.

ACKNOWLEDGMENTS

The authors are grateful to B. Waclaw for valuable discussions. Project operated within the Foundation for Polish Science International Ph.D. Projects Programme co-financed by the European Regional Development Fund covering, under the Agreement No. MPD/2009/6, the Jagiellonian University International Ph.D. Studies in Physics of Complex Systems. The research of Z.B. was partially supported by the Polish Ministry of Science Grant No. N N202 229137 (2009–2012).

APPENDIX A: DIFFERENCE EQUATIONS

In this Appendix, we provide the reader with a detailed solution of the recurrence equations (13) resulting in

$$\begin{aligned} A_g &= -\lambda A_{g-1} - kA_{g-2} \quad \text{for} \quad g < G, \\ A_G &= -\lambda A_{G-1} - rA_{G-2}. \end{aligned} \tag{A1}$$

These difference equations can be solved with two initial conditions,

$$A_0 = -\lambda, \quad A_{-1} = 1, \tag{A2}$$

where the first condition is found in Eq. (13) and the second condition is chosen so as to stay in agreement with the recurrence relation (indeed, $A_1 = -\lambda A_0 - kA_{-1} = \lambda^2 - k$).

The characteristic polynomial of this difference equation yields $\alpha^2 + \lambda\alpha + k = 0$, resulting in $\alpha = \frac{1}{2}(-\lambda \pm i\sqrt{4k - \lambda^2})$, and using the notation

$$\begin{aligned} \cos\theta &= -\lambda/2\sqrt{k}, \\ \sin\theta &= \sqrt{1 - (\lambda/2\sqrt{k})^2}, \end{aligned} \tag{A3}$$

the general solution is obtained as

$$\begin{aligned} A_g &= k^{(g+1)/2}[\alpha_1 \cos(g\theta) + \alpha_2 \sin(g\theta)] \\ &\quad \text{for} \quad g = 0, \ldots, G-1. \end{aligned} \tag{A4}$$

The first and second initial condition, respectively, lead to

$$\begin{aligned} \alpha_1 &= 2\cos\theta, \\ \alpha_2 &= \frac{\cos(2\theta)}{\sin\theta}, \end{aligned} \tag{A5}$$

after insertion of which the solution takes the form

$$\begin{aligned} A_g &= k^{(g+1)/2} \frac{\sin(2\theta)\cos(\theta g) + \cos(2\theta)\sin(\theta g)}{\sin\theta} \\ &= k^{(g+1)/2} \frac{\sin[\theta(G+2)]}{\sin\theta} \quad \text{for } g < G. \end{aligned} \tag{A6}$$

The last value, A_G, is calculated separately due to the root having degree r that may be different from k:

$$A_G = k^{(G-1)/2} \frac{k\sin[\theta(G+2)] + (k-r)\sin(\theta G)}{\sin\theta}. \tag{A7}$$

In the case of GRW, the recurrence equations are given by (50). The solution proceeds analogously; however, due to different coefficients, the initial conditions need to be adjusted accordingly:

$$A_0 = -\lambda, \quad A_{-1} = k + 1. \tag{A8}$$

The general form of the solution remains the same as given above, but for the prefactor $k^{(g+1)/2}$ substituted with $[k/(k+1)^2]^{(g+1)/2}$. The first and second initial conditions give

$$\alpha_1 = 2\cos\theta, \quad \alpha_2 = \frac{\cos(2\theta) - k}{\sin\theta}, \tag{A9}$$

which eventually lead to the solutions [(51) and (52)].

APPENDIX B: TRIGONOMETRIC EQUATIONS

In this Appendix, we derive in more detail the approximate solutions to the trigonometric equations that appeared earlier in the paper. Equation (20) can be illustrated with Fig. 5. For $r = k$ and $r = 2k$, the analytical solutions (21) and (22) are found immediately. As mentioned in Sec. III, for other values of r, the solutions can be divided into three classes with respect to values of r: the first class for $r \in (0,2k - 2k/G)$, the second one for $r \in (2k - 2k/G, 2k + 2k/G)$, and the third one for $r \in (2k + 2k/G, +\infty)$. In the large G limit, that is, for $G \gg 2k$, the second class reduces to a single integer value of $r = 2k$ (although for small G, one can find several values, e.g., for $G = 3, k = 3, r = 7$, the solution is still real).

As regards the first class, $r < 2k$, an approximation of the smallest θ (the largest λ) for large G can be derived in the following way: Let us transform Eq. (21) into

$$\tan[(G+1)\theta] = \frac{r}{r-2k}\tan\theta. \tag{B1}$$

In the limit $G \to \infty$, we expect $\theta \to 0$ (as we do observe such behavior for $r = k$ and $r = 2k$), and upon Taylor expansion, we obtain

$$\tan[(G+1)\theta] \cong \frac{r}{r-2k}(\theta + \theta^3/3), \tag{B2a}$$

$$(G+1)\left(\theta - \frac{\pi}{G+1}\right) \cong \arctan\left[\frac{r}{r-2k}(\theta+\theta^3/3)\right], \tag{B2b}$$

$$(G+1)\left(\theta - \frac{\pi}{G+1}\right) \cong \frac{r}{r-2k}\theta + O(\theta^3), \tag{B2c}$$

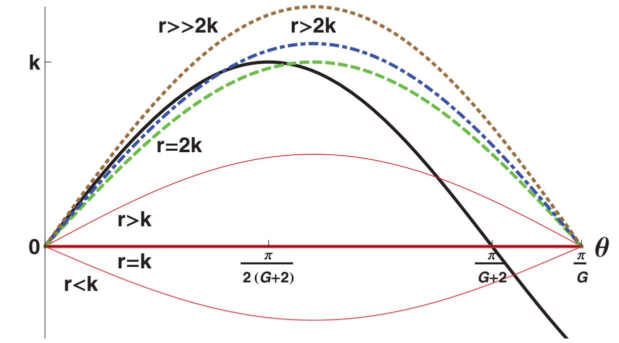

FIG. 5. (Color online) The intersection of the black curve with the other ones marks the solution of Eq. (20). The uppermost brown dotted curve corresponding to a strongly branched root shows no real solutions. The blue dot-dashed sine is an example of the rare case of a strongly branched root with a real solution. The green dashed line is the critically branched root and the red continuous lines correspond to weakly branched roots.

which, when having denoted by $\delta \approx \frac{2k}{2k-r}$, finally leads to

$$\theta \cong \frac{\pi}{G+\delta}, \tag{B3}$$

and produces the asymptotic solution (25) for the first level of eigenvalues in the limit $G \to \infty$ for any branching parameters $k, r < 2k$.

For the third class, $r > 2k$, Eq. (20) has no real solutions in the range $(0, \frac{\pi}{G+1})$ and the largest eigenvalue λ_0 is obtained from a purely imaginary solution for θ. The corresponding equations change from trigonometric to hyperbolic, so after transformation of (21), one gets

$$\tanh[(G+1)\theta] = \frac{r}{r-2k}\tanh\theta. \tag{B4}$$

For $G \to \infty$, this equation approaches

$$1 = \frac{r}{r-2k}\tanh\theta^*, \tag{B5}$$

which gives

$$\theta^* = \operatorname{arctanh}\left(\frac{r-2k}{r}\right). \tag{B6}$$

With the notation $z = 1 - \frac{2k}{r}$, and after utilizing the identity $\operatorname{arctanh}(z) = \frac{1}{2}\ln(\frac{1+z}{1-z})$,

$$(G+1)\theta = \frac{1}{2}\ln\left(\frac{1}{z}\tanh\theta + 1\right) - \frac{1}{2}\ln\left(1 - \frac{1}{z}\tanh\theta\right). \tag{B7}$$

For large G, the first term on the right-hand side approaches $\frac{1}{2}\ln 2$, while the left-hand side approaches $(G+1)\theta^*$. After rearranging this equation,

$$\theta \cong \operatorname{arctanh}(z\{1 - \exp[\ln 2 - 2(G+1)\theta^*]\}), \tag{B8}$$

and finally under substitution of θ^*,

$$\theta \cong \operatorname{arctanh}\left\{z\left[1 - 2\left(\frac{1+z}{1-z}\right)^{-(G+1)}\right]\right\}. \tag{B9}$$

The final solution (27) for λ_0 is due to the identity $\cos(i\,\operatorname{arctanh}x) = \frac{1}{\sqrt{1-x^2}}$.

The last remark concerns the problem of which eigenvalue $\lambda_{g,j}$ is the second largest one. If $r > 1$, the level $G-1$ of the eigenvalue hierarchy exists. The eigenvalue $\lambda_{G-1,1}$ is defined by the angle $\theta_{G-1,1} = \frac{\pi}{G+1}$, whereas the second eigenvalue in the first level $\lambda_{G,2}$ is defined by an angle $\theta_{G,2} > \frac{\pi}{G}$. The latter information can be easily deduced from Fig. 5, where the intersections below the angle $\frac{\pi}{G}$ correspond to the largest eigenvalue. Thus, $\theta_{G-1,1} < \theta_{G,2}$, and, consequently, $\lambda_{G-1,1} > \lambda_{G,2}$. As this argument holds in general, $\lambda_1 = \lambda_{G-1,1}$.

[1] A. Einstein, Ann. Phys. (Leipzig) **322**(8), 549 (1905); **324**(2), 371 (1906).

[2] M. Smoluchowski, Ann. Phys. (Leipzig) **326**(14), 756 (1906).

[3] G. Polya, Math. Ann. **84**, 149 (1921).

[4] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Phys. Rev. Lett. **102**, 160602 (2009).

[5] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Acta Phys. Pol. B **41**, 949 (2010).

[6] L. Demetrius, V. M. Gundlach, and G. Ochs, Theor. Popul. Biol. **65**, 211 (2004).

[7] L. Demetrius and T. Manke, Physica A **346**, 682 (2005).

[8] J. H. Hetherington, Phys. Rev. A **30**, 2713 (1984).

[9] V. Zlatic, A. Gabrielli, and G. Caldarelli, Phys. Rev. E **82**, 066109 (2010).

[10] J.-C. Delvenne and A.-S. Libert, Phys. Rev. E **83**, 046117 (2011).

[11] R. Sinatra, J. Gomez-Gardenes, R. Lambiotte, Vincenzo Nicosia, and Vito Latora, Phys. Rev. E **83**, 030103 (2011).

[12] C. Monthus and T. Garel, J. Phys. A **44**, 085001 (2011).

[13] K. Anand, G. Bianconi, and S. Severini, Phys. Rev. E **83**, 036109 (2011).

[14] B. Waclaw, *Generic Random Walk and Maximal-Entropy Random Walk*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/GenericRandomWalkAndMaximalEntropyRandomWalk/.

[15] W. M. X. Zimmer and G. M. Obermair, J. Phys. A **11**, 1119 (1978).

[16] W. Parry, Trans. Amer. Math. Soc. **112**, 55 (1964).

[17] J. K. Ochab, *Stationary States of Maximal-Entropy Random Walk and Generic Random Walk on Cayley Trees*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/StationaryStatesOfMaximalEntropyRandomWalkAndGenericRandomWa/.

[18] J. K. Ochab, *Dynamics of Maximal-Entropy Random Walk and Generic Random Walk on Cayley Trees*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/DynamicsOfMaximalEntropyRandomWalkAndGenericRandomWalkOnCayl/.

[19] S. Fortunato, Phys. Rep. **486**, 75 (2010).

MAXIMAL ENTROPY RANDOM WALK: SOLVABLE CASES OF DYNAMICS*

J.K. Ochab

The Marian Smoluchowski Institute of Physics, Jagiellonian University
Reymonta 4, 30-059 Kraków, Poland
jeremi.ochab@uj.edu.pl

(Received April 2, 2012)

We focus on the study of dynamics of two kinds of random walk: generic random walk (GRW) and maximal entropy random walk (MERW) on two model networks: Cayley trees and ladder graphs. The stationary probability distribution for MERW is given by the squared components of the eigenvector associated with the largest eigenvalue λ_0 of the adjacency matrix of a graph, while the dynamics of the probability distribution approaching to the stationary state depends on the second largest eigenvalue λ_1. Firstly, we give analytic solutions for Cayley trees with arbitrary branching number, root degree, and number of generations. We determine three regimes of a tree structure corresponding to strongly, critically, and weakly branched roots. Each of them results in different statics and dynamics of MERW. We show how the relaxation times, generically shorter for MERW than for GRW, scale with the graph size. Secondly, we give numerical results for ladder graphs with symmetric defects. MERW shows a clear exponential growth of the relaxation time with the size of defective regions, which indicates trapping of a particle within highly entropic intact region and its escaping that resembles quantum tunneling through a potential barrier. GRW shows standard diffusive dependence irrespective of the defects.

DOI:10.5506/APhysPolB.43.1143
PACS numbers: 05.40.Fb, 02.50.Ga, 89.75.Hc, 89.70.Cf

1. Introduction

After Einstein [1] and Smoluchowski [2] gave explanations of Brownian motion and originated the theory of diffusive processes, there has been an unceasing research on models of random walk (RW), which may be regarded as time or space discretization of these processes. Thousands of papers and textbooks in statistical physics, particle physics, engineering, economics, biophysics, *etc.*, have been published.

[*] Presented at the XXIV Marian Smoluchowski Symposium on Statistical Physics, “Insights into Stochastic Nonequilibrium”, Zakopane, Poland, September 17–22, 2011.

From the mathematical perspective, RW is a Markov chain describing the random consecutive steps of a particle. As an example, the well-known Polya random walk on a lattice [3] at each time performs equiprobable steps to any of the neighboring nodes. This process, generalized to any graph, is known as the ordinary or generic random walk (GRW).

RW can also maximize the entropy of paths, and hence we call it the maximal entropy random walk (MERW); lately, this type has been studied in [4,5]. This principle of entropy maximization, which is a global one alike the least action principle, earlier brought about the biological concept of evolutionary entropy [6,7]. It also served as an optimal sampling algorithm in the problem of importance sampling [8]. MERW has also begun to be used in the study of complex networks [9,10,11,12,13].

The defining feature of MERW makes the paths of given length and end-points equiprobable. This leads to an unprecedented feature that the stationary probability on diluted lattices localizes in the biggest spherical region [4,5]. An interactive online demonstration [14] illustrates this feature. In this paper, we focus on how the dynamics of GRW and MERW differs. More precisely, we show analytic expressions for stationary probability distributions and relaxation times of GRW and MERW on Cayley trees; we also give numerical results for ladder graphs, showing that the relaxation time for MERW grows exponentially with the size of defective regions as opposed to diffusion behavior for GRW.

In this paper, in Sec. 2 we provide definitions and notes on the two types of random walk. In Sec. 3, we give several analytical results concerning Cayley trees (involving eigenproblem solution for the adjacency matrix, discussion of stationary state and relaxation). Lastly, in Sec. 4, we show numerical results concerning relaxation process on a class of ladder graphs.

2. General considerations

Let us consider a discrete time random walk defined by a constant stochastic matrix $\boldsymbol{P}$, on a finite connected undirected graph. The probability that a random walker which can be found on a node i at time t hops to a node j at time $t+1$ is encoded by the element $P_{ij} \geq 0$ of this matrix. Another condition fulfilled by this matrix element is $\sum_j P_{ij} = 1$ for all i. If we denote by $\boldsymbol{A}$ the adjacency matrix of the graph ($A_{ij} = 1$ if i and j are neighbors, and $A_{ij} = 0$ otherwise), we can formulate an additional condition: $P_{ij} \leq A_{ij}$, which means that particles are allowed to jump between neighboring nodes only. The stochastic matrix corresponding to the generic random walk (GRW) is given by

$$P_{ij} = \frac{A_{ij}}{k_i}\,, \tag{1}$$

where $k_i = \sum_j A_{ij}$ is the node degree, and the probability of selecting one of k_i neighbors of the node i is uniform. This means that the entropy of neighbor selection is maximized and shows that this is the standard Einstein–Smoluchowski–Polya random walk. Lastly, the stationary state of GRW is given by $\pi_i = k_i / \sum_j k_j$.

On the other hand, maximal entropy random walk (MERW) maximizes the entropy of choosing a trajectory of given length and end-points. This principle leads to

$$P_{ij} = \frac{A_{ij}}{\lambda_0} \frac{\psi_{0j}}{\psi_{0i}} , \tag{2}$$

where λ_0 is the largest eigenvalue of the adjacency matrix $\boldsymbol{A}$ and ψ_{0i} is the ith component of the corresponding eigenvector $\vec{\psi}_0$. From the Frobenius–Perron theorem and from the fact that the adjacency matrix $\boldsymbol{A}$ is irreducible it follows that all elements of $\vec{\psi}_0$ are strictly positive. Shannon–Parry measure [15] then describes the stationary state of $\boldsymbol{P}$

$$\pi_i = \psi_{0i}^2 . \tag{3}$$

Intriguingly, this equation forms a connection between MERW and quantum mechanics, as one may interpret $\vec{\psi}_0$ as the wave function of the ground state of the operator $-\boldsymbol{A}$ and consequently ψ_{0i}^2 becomes the probability of finding a particle in this state [4, 5]. The two random walks, (1) and (2), in general exhibit altogether different behaviors except for the case of k-regular graphs, where they coincide.

3. Cayley tree

We define a Cayley tree with a branching number k, which is the number of edges leading from a given node to the next generation of nodes, and the number of generations G. The root of the tree is assumed to have a degree r and it belongs to the zeroth generation (see Fig. 1). The number of nodes in the zeroth generation is therefore $n_0 = 1$, in the first $n_1 = r$ nodes, in the second $n_2 = rk$, in the third one $n_3 = rk^2$, *etc.* The tree has n nodes in total: $n = \sum_{g=0}^{G} n_g = 1 + r(k^G - 1)/(k - 1)$.

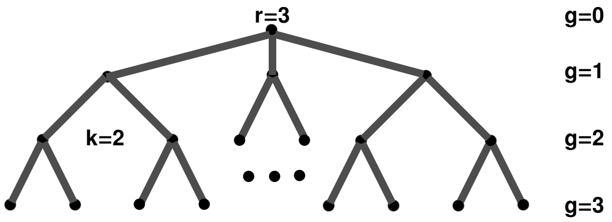

Fig. 1. A Cayley tree with root degree $r = 3$, branching number $k = 2$, and $G = 3$ generations.

3.1. Eigenvalues of the adjacency matrix

This section is devoted to calculation of eigenvalues of the adjacency matrix of Cayley tree, which can be determined by solving the equation

$$\det(\boldsymbol{A} - \lambda\mathbf{1}) = 0\,. \tag{4}$$

The determinant can be calculated with the use of a sequence of elementary transformations that leave it invariant, *e.g.*, additions of multiple of a row or column to another row or column. Thus, the determinant can be reduced to a triangular form with zeros above the diagonal, as first presented in [16]. Details of this procedure can be found in [17]. The triangular form of the determinant allows to rewrite (4) as a product of the diagonal coefficients

$$\prod_{g=0}^{G} [A_g(\lambda)]^{m_g} = 0\,, \tag{5}$$

where $m_G = 1$ and $m_{G-g} = n_g - n_{g-1}$, for $g = 1, 2, \dots, G$, and $A_g(\lambda)$ are polynomials w.r.t. λ given by the recursive equations

$$\begin{aligned} A_0(\lambda) &= -\lambda\,, \\ A_g(\lambda) &= -\lambda A_{g-1}(\lambda) - kA_{g-2}(\lambda)\,, \qquad \text{for } g < G\,, \\ A_G(\lambda) &= -\lambda A_{G-1}(\lambda) - rA_{G-2}(\lambda)\,. \end{aligned} \tag{6}$$

Notice that for $g = G$ the coefficient k is replaced by r, which is a consequence of the tree structure allowing arbitrary root degree. To complete the set of equations we have to take initial condition $A_{-1} = 1$. The real roots of equation (5) counted with the degeneracy m_g give the total number of $\sum_g (g+1)m_g = \sum_g n_g = n$, which means all n eigenvalues of the adjacency matrix are retrieved.

The recurrence (6) can be solved

$$A_g = k^{(g+1)/2}\frac{\sin[(g+2)\theta]}{\sin\theta}\,, \qquad \text{for} g < G\,, \tag{7}$$

where $\cos\theta = -\lambda/(2\sqrt{k})$ and θ is an auxiliary parameter. To obtain the polynomial A_G one needs to combine the last equation in (6) $A_G = -\lambda A_{G-1} - kA_{G-2}$ with the solutions for A_{G-1} and A_{G-2} (7), which yields

$$A_G = k^{(G-1)/2}\frac{k\sin[(G+2)\theta] + (k-r)\sin(G\theta)}{\sin\theta}\,. \tag{8}$$

Now, instead of (5) we can consider equations $A_g = 0$ and $A_G = 0$ to find the solutions for θ and then determine the eigenvalues of the adjacency matrix using the formula $\lambda = -2\sqrt{k}\cos\theta$. The first equation, leading to

$$\sin[(g+2)\theta] = 0\,, \tag{9}$$

has $g+1$ solutions

$$\lambda_{g,j} = 2\sqrt{k}\cos\left(\frac{\pi j}{g+2}\right), \qquad \text{for } j = 1,\ldots,g+1\,. \tag{10}$$

As follows from (5), each eigenvalue $\lambda_{g,j}$ in this series has multiplicity m_g. The equation $A_G = 0$ produces

$$k\sin[(G+2)\theta] + (k-r)\sin(G\theta) = 0\,, \tag{11}$$

for which analytical solutions exist in the case $r = k$

$$\lambda_{G,j} = 2\sqrt{k}\cos\left(\frac{\pi j}{G+2}\right), \qquad \text{for } j = 1,\ldots,G+1\,, \tag{12}$$

and in the case $r = 2k$

$$\lambda_{G,j} = 2\sqrt{k}\cos\left[\frac{\pi(j-1/2)}{G+1}\right], \qquad \text{for } j = 1,\ldots,G+1\,. \tag{13}$$

Other choices of r involve numerical solving of (11).

It can be shown that the largest eigenvalue of the adjacency matrix is $\lambda_0 = \lambda_{G,1}$. It belongs to one of three classes of solutions depending on r, which takes values $r \in (0, 2k - 2k/G)$ in the first class, $r \in (2k - 2k/G, 2k + 2k/G)$ in the second, and $r \in (2k + 2k/G, +\infty)$ in the third.

For large G (*i.e.*, $G \gg 2k$) the second interval becomes just a single integer value $r = 2k$. The first class allows values $r < 2k$ for which an approximate solution exists

$$\lambda_0 = 2\sqrt{k}\cos\frac{\pi}{G+\delta}\,, \tag{14}$$

where $\delta \approx 2k/(2k-r)$, or the exact solution for $r = k$ (12). In the third class, $r > 2k$, there are no real solutions of (11) for $\theta \in (0, \pi/(G+1))$ and λ_0 corresponds to a purely imaginary θ. The trigonometric equation (11) is thus replaced by a hyperbolic one. In the limit of large G the approximate solution is

$$\lambda_0 = \frac{2\sqrt{k}}{\sqrt{1-x^2}}\,, \tag{15}$$

where

$$x = z\left[1 - 2\left(\frac{1-z}{1+z}\right)^{G+1}\right] \quad \text{and} \quad z = 1 - \frac{2k}{r}\,. \tag{16}$$

3.2. The eigenvalues of the GRW transition matrix

The stochastic matrix of generic random walk (1) can be subjected to the same procedure as explained in Sec. 3.1. Transforming its determinant to the triangular form generates analogous recursion as in (6)

$$\begin{aligned} A_0 &= -\lambda\,, \\ A_g &= -\lambda A_{g-1} - \frac{k}{(k+1)^2} A_{g-2}\,, \qquad \text{for } g = 2,\dots,G-1\,, \\ A_G &= -\lambda A_{G-1} - \frac{1}{k+1} A_{G-2}\,. \end{aligned} \tag{17}$$

We take $A_{-1} = k+1$ as an initial condition that agrees with the rest of equations and proceed as before, solving this recurrence to obtain eigenvalues from the equations $A_g = 0$ and $A_G = 0$. From $A_G = 0$ one gets

$$2k\cos(2\theta) = 1 + k^2 \quad \text{and} \quad \sin(G\theta) = 0\,, \tag{18}$$

whose solutions lead to, respectively,

$$\lambda_0 = 1\,, \quad \text{and} \quad \lambda_{G,j} = 2\sqrt{\frac{k}{(k+1)^2}}\cos\left(\frac{\pi j}{G}\right)\,, \quad \text{for } j = 1,\dots,G\,, \tag{19}$$

and from $A_g = 0$

$$\sin[(g+2)\theta] = k\sin(g\theta)\,. \tag{20}$$

The last equation has the identical form as (11) except for different coefficients. The class of solutions of (11) with $r \in (2k + 2k/G, +\infty)$ corresponds to value $k > 1$ in the above equation. Hence, the value of θ that leads to the largest eigenvalue in a given series is imaginary. Once again, the trigonometric equations (20) change into hyperbolic ones. Upon replacements $k = \frac{z+1}{1-z}$, $z = \frac{k-1}{k+1}$, we end up with (16) rewritten as $x = \frac{k-1}{k+1}\left[1 - 2k^{-(g+1)}\right]$. In the large G limit, the second largest eigenvalue is thus approximated by

$$\lambda_1 = 2\sqrt{\frac{k}{(k+1)^2}}\frac{1}{\sqrt{1-x^2}} \tag{21}$$

and clearly the second largest eigenvalue λ_1 approaches $\lambda_0 = 1$ exponentially in G.

3.3. Stationary states of GRW and MERW on Cayley trees

A random walk on a graph has a stationary probability distribution if the graph is not bipartite. If a graph is bipartite, one can define a semi-stationary state: it involves either averaging probability distributions over two consecutive time steps t and $t+1$ (because the distributions for even and odd times are independent) or averaging the distribution over initial conditions.

GRW leads to the stationary occupation probabilities

$$\pi_i = \frac{k_i}{\sum_j k_j}\,, \quad \text{for} \quad i = 1, \ldots, n\,, \tag{22}$$

which comprise a flat distribution for nodes of degree $k_i = k+1$ and the exception of the root having r neighbors and leaves neighboring with just one node. As nodes in each generation have equal stationary probabilities we can sum over them $\Pi_g = n_g \pi_i \propto k^{g-1}$, which produces the exponential factor.

The stationary probabilities of MERW are equal to the squared components of $\vec{\psi}_0$. All elements ψ_{0i} of this vector have the same values for i belonging to a given generation g. This simplifies the description of the stationary state so that we may write ψ_g for all nodes in the generation g (we omit the first index, which numbers the corresponding eigenvalue). Exact solution for ψ_g can be obtained by solving a recurrence equation analogous to (6)

$$\pi_i = \psi_{0i}^2 \propto k^{G-g} \sin[(G-g+1)\theta]^2\,, \quad \text{for} \quad g = 0, \ldots, G \ \text{ and } i \in g\,, \tag{23}$$

where the normalization constant has been omitted. After summing over whole generation $i \in g$, the probabilities become

$$\Pi_g = n_g \pi_i \propto k^{G-1} \sin[(G-g+1)\theta]^2\,, \quad \text{for} \quad g = 1, \ldots, G \ \text{ and } i \in g\,, \tag{24}$$

where the case $g = 0$ with its $n_0 = 1$ needs a separate treatment.

This result depends on the choice of r, k through θ and λ. For $r < 2k$, parameter $\theta \approx \frac{\pi}{G+\delta}$ and the limiting distribution is a sine square; for $r = 2k$, $\theta = \frac{\pi/2}{G+1}$ and the distribution is a cosine square; for $r > 2k$, $\theta = i\ \text{arctanh}\, x$ (where x is defined in (16), while i is the imaginary unit), which yields a hyperbolic sine. These limiting results as well as finite-size effects are showed in an online interactive demonstration [18].

3.4. Relaxation times

A stochastic matrix does not have to be symmetric, thus its right and left eigenvectors may differ: $\boldsymbol{P}\vec{\Psi}_\alpha = \Lambda_\alpha \vec{\Psi}_\alpha, \vec{\Phi}_\alpha \boldsymbol{P} = \Lambda_\alpha \vec{\Phi}_\alpha$. Hence, there exists a spectral decomposition of $\boldsymbol{P}$

$$P_{ij} = \sum_\alpha \Lambda_\alpha \Psi_{\alpha i} \Phi_{\alpha j}\,, \tag{25}$$

where for MERW one can make replacements: $\varLambda_\alpha = \lambda_\alpha/\lambda_0, \varPsi_{\alpha i} = \psi_{\alpha i}/\psi_{0i}$, $\varPhi_{\alpha i} = \psi_{\alpha i}\psi_{0i}$. The spectral decomposition of the adjacency matrix of a given graph thus contains information about all properties of MERW.

From the knowledge of the initial probability distribution $\vec{\pi}(0)$ and the transition matrix $\boldsymbol{P}$ the distribution can be determined at any time t

$$\vec{\pi}(t) = \vec{\pi}(0)\boldsymbol{P}^t\,, \tag{26}$$

where the elements $\pi_i(t),\ i = 1,\ldots,n$ of $\vec{\pi}(t)$ denote the probability of finding a particle performing a random walk at a node i at time t.

The last equation can be reformulated utilizing the spectral decomposition of the stochastic matrix (25)

$$\vec{\pi}(t) = \sum_\alpha c_\alpha \varLambda_\alpha^t \vec{\varPhi}_\alpha\,, \tag{27}$$

where c_α denotes a spectral coefficient: $c_\alpha = \vec{\pi}(0)\cdot\vec{\varPsi}_\alpha = \sum_i \pi_i(0)\varPsi_{\alpha i}$.

Generally, all eigenvalues $\varLambda_\alpha$ of $\boldsymbol{P}$ are located inside or on the unit circle in the complex plane $|\varLambda_\alpha| \leq 1$ and in the limit of infinite t on the right-hand side of (27) only $|\varLambda_\alpha| = 1$ survive, while all the other terms vanish exponentially.

TABLE I

Relaxation times τ_1 for large G. All rows except for the last one refer to MERW. The symbols $\lambda_{g,j}$ correspond to one of the equations (10), (12), or (13), whichever is appropriate for the choice of parameters k, r. In the first row: $c = \left(\ln\frac{r}{2\sqrt{(r-k)k}}\right)^{-1}$. While the number of vertices $n \sim k^G$, the probability distribution relaxes a logarithm of the system size $\tau_1 \sim \ln n$.

Regime	λ_0	λ_1	τ_1
Strongly branched: $r > 2k,\ k > 1$	Eq. (15)	$\lambda_{G-1,1}$	$c + \frac{c^2\pi^2}{2}\frac{1}{G^2} + \ldots$
Critically branched: $r = 2k,\ k > 1$	$\lambda_{G,1}$	$\lambda_{G-1,1}$	$\frac{8G^2}{3\pi^2} + \frac{16G}{3\pi^2} + \ldots$
Weakly branched: $1 < r < 2k,\ k > 1$	$\lambda_{G,1}$	$\lambda_{G-1,1}$	$\frac{2k-r}{r\pi^2}G^3 + \frac{3(4k-r)}{2r\pi^2}G^2 + \ldots$
Planted tree: $r = 1$	$\approx$ Eq. (14)	$\lambda_{G-2,1}$	$\frac{2k-1}{2k\pi^2}G^3 + \frac{3}{2\pi^2}G^2 + \ldots$
Linear chain: $k = 1,\ r = 1$	$\lambda_{G,1}$	$\lambda_{G,2}$	$\frac{2G^2}{3\pi^2} + \frac{8G}{3\pi^2} + \ldots$
GRW: $r > 1,\ k > 1$	1	Eq. (21)	$\frac{8k}{(k-1)^2}k^G$

For both GRW and MERW on a tree only two eigenvalues on the unit circle are left $\Lambda_0 = 1$ and $\Lambda_n = -1$ due to bipartiteness of the graph. For $t \to \infty$ the relaxation to the stationary state is generically governed by the second largest eigenvalue Λ_1 and its negative counterpart $\Lambda_{n-1} = -\Lambda_1$. The corresponding term in the spectral decomposition (27) decreases exponentially as $\exp(-t/\tau_1)$, where $\tau_1 = [-\ln(\Lambda_1)]^{-1} = [\ln(\lambda_0/\lambda_1)]^{-1}$.

Thus, τ_1 is what we call the generic relaxation time, which is the largest one. We note, however, that there are symmetries that lead also to other relaxation times. As the eigenvalues of the adjacency matrix depend on the tree parameters, also the relaxation times for MERW fall into several classes. The relaxation times for large G are given in Table I. It is noteworthy that whereas the probability distribution for GRW relaxes linearly with the system size $\tau_1 \sim n \sim k^G$, for MERW it is as fast as a logarithm of the system size $\tau_1 \sim \ln n$. Derivations and further details expanding the note on symmetries can be found in [17]. An online interactive demonstration [19] can also facilitate understanding of these results.

4. Ladder graph

In this section, we discuss a particular class of ladder graphs (exemplary ladder graph can be seen in Fig. 2). A ladder graph consists of two chains of integer length $n/2$ which are connected by rungs, *i.e.* node i of one chain is connected to node i' of the second one, then $i+1$ to $i'+1$ and so forth. We also impose periodic boundary conditions producing a ring, where node $i + n/2$ is connected to node $i+1$, and node $i' + n/2$ to node $i'+1$. This structure is symmetric with respect to reflection $i \to i'$, and so the graph is a quasi one-dimensional system. It is a 3-regular graph, although we remove some rungs from the ladder to introduce defects, so that MERW and GRW are not equivalent on this graph anymore.

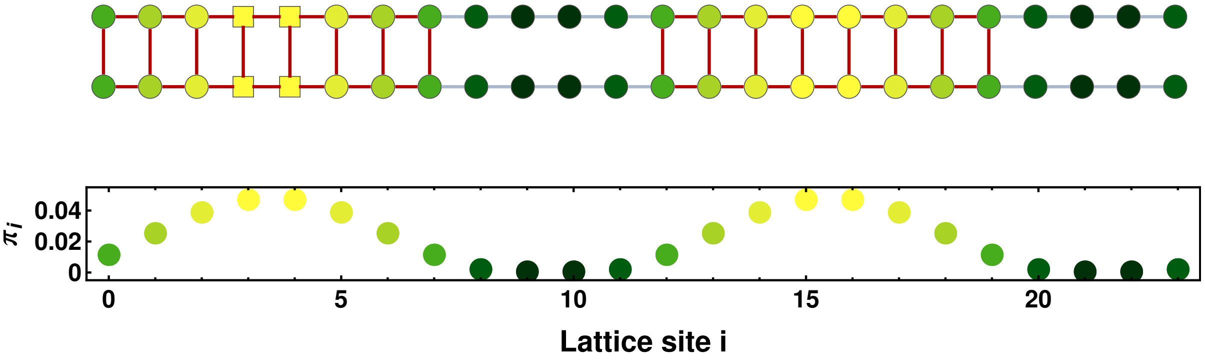

Fig. 2. (Color online) Stationary probability for a ladder graph with periodic boundary conditions: the probability localizes in the intact regions. The square nodes represent initial condition that would be chosen for this graph.

For the adjacency matrix of the graph $\boldsymbol{A}$, its largest eigenvalue λ_0 and the eigenvector $\vec{\psi}_0$ associated with it, the stationary solution for MERW is given as the ground state of the tight-binding equation

$$\left(\boldsymbol{H}\vec{\psi}_0\right)_a = \left(-\Delta\vec{\psi}_0\right)_a + V_a\psi_{0,a} = E_0\psi_{0,a}\,, \tag{28}$$

where the Hamiltonian is defined as $H_{ab} = k_{\max}\delta_{ab} - A_{ab}$, with the Kronecker delta δ_{ab}, maximum degree of the graph $k_{\max}$, and $V_a = k_{\max} - k_a$, $E_0 = k_{\max} - \lambda_0$. For a ladder graph with defects this equation yields

$$2\psi_{0,a} - \psi_{0,a-1} - \psi_{0,a+1} + V_a\psi_{0,a} = E_0\psi_{0,a}\,, \tag{29}$$

where $E_0 = 3 - \lambda_0$ and $V_a = 0$ or 1 (rung present or absent). Stationary states of a number of ladder graphs (with one, two, or a number of random defects) were discussed in Section 6 of [5].

Additionally, we impose a symmetry on those defects: there can only be two equal regions intact and two equal regions with rungs removed (gaps). We take the initial probability 1 at the center of one of the intact regions (this may be 2 or 4 nodes, depending on whether the length of the region is odd or even, see Fig. 2). The systems under study have $n = 48 - 512$ total number of nodes and the number of deleted rungs separating two regions (the gap size) varies between $g = 1 - 10$.

We measure the probability $P(t)$ summed over one whole region (as the regions are equal in size, $P_\infty = 1/2$ is its stationary value). It might be understood as a macroscopic measure of the process taking place in this region. As expected, the probability flows from one the initial intact region to the other one until equilibrium ($P(t) = 1/2$ in both regions) is attained.

We fit the numerical results to exponential dependence on time t: $P(t) \sim \exp[-a(t-b)]$, where a and b are fitting parameters from which we extract the relaxation time τ, which is the characteristic time scale of an exponential approach to the stationary state. The results for the behavior of relaxations times for GRW and MERW are given in Table II. It turns out that for MERW there is a clear dependence of the relaxation on the gap size for a given lattice size (example in Fig. 3 (a) for $n = 96$): $a(g) = \exp(-c\,g - d)$,

TABLE II

Relaxation times τ as functions of the system size n and gap size g, where $c(n) = c_\infty - fn^{-1/\nu}$ and d, c_∞, f, ν are fitted constant.

GRW	$\tau(n,g) = c\,n^d$, $\qquad d = \text{const.} = 2$
MERW	$\tau(n,g) = \exp\left[c(n)\cdot g\right]$

$b(g) = \exp(c\,g + d)$, where c, d are constants with respect to the gap size g. After extracting this dependence, only the dependence on the system size remains in the function $c = c(n)$, which is very well fitted with a power law (Fig. 3 (b): $c(n) = c_\infty - fn^{-1/\nu}$ (best-fit value parameters are $c_\infty = 0.9643 \pm 0.0078$, $f = 58 \pm 42$, $\nu = 0.773 \pm 0.098$). Thus, the macroscopic probability depends on time, system size, and gap size

$$|P(t;g,n) - P_\infty| \propto \exp\left\{-t\exp\left[-c(n)\,g\right]\right\}. \tag{30}$$

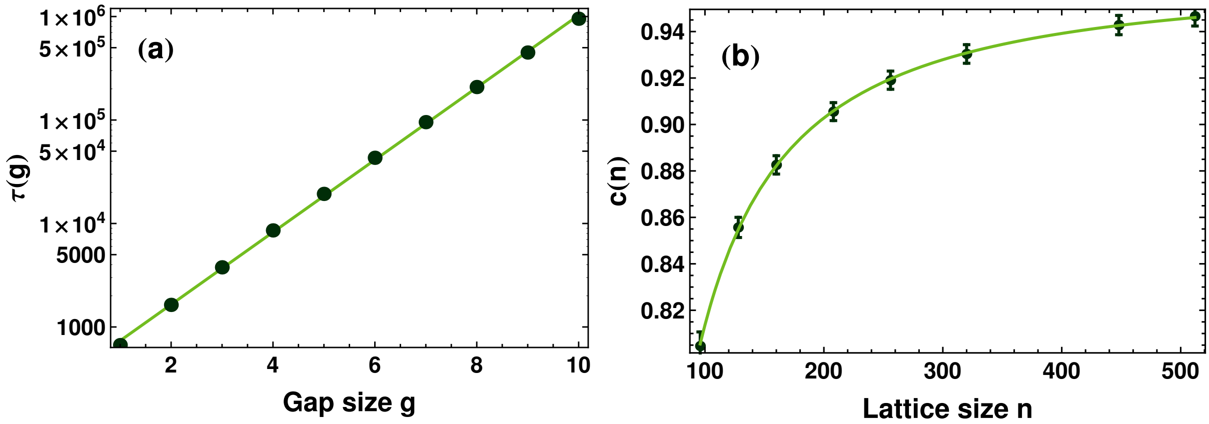

Fig. 3. Maximal Entropy Random Walk: (a) Logarithmic plot shows an exponential dependence of the relaxation time on the gap size (an exemplary system size, $n = 96$) reminding of quantum tunneling, (b) the dependence of the relaxation time on the system size, $c(n) = c_\infty - fn^{-1/\nu}$ [see (30)]. Continuous lines are the best fits of an exponential function and power law, respectively.

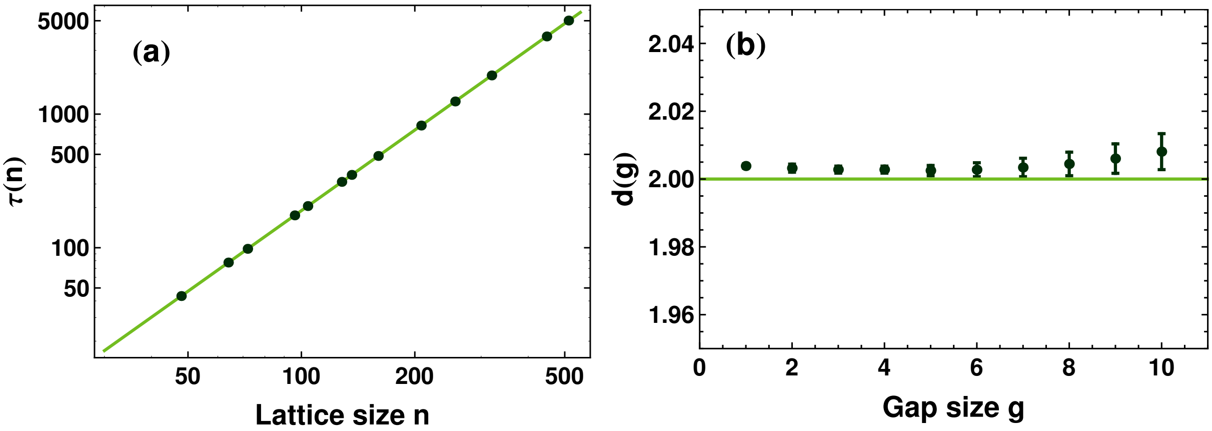

Fig. 4. Generic Random Walk: (a) log–log plot shows power law dependence of relaxation time, expected for a diffusion process (continuous line is the best fit; gap size $g = 1$), (b) the best-fit exponents d of the power law show independence from the gap size. The errors result from finite-size effects.

For GRW, both fitted parameters a and b have shown no dependence from the gap size g, although they do depend on the system size: $a(n) = cn^{-d}$, $b(n) = c'n^{d'}$, where $d', d \approx 2$ (see Fig. 4). This produces the familiar behavior $\tau \sim n^2$ which is expected for a one-dimensional random walk.

5. Conclusions

In this paper, we have discussed the dynamics of generic random walk and maximal entropy random walk on two classes of graphs. For Cayley trees, we have shown the analytic form of generic relaxation times governing how fast probability distributions of those random walks approach their stationary states. MERW has proven to be generically faster (logarithmic with respect to the system size) than GRW (linear w.r.t. the system size). However, on defective ladder graphs the relaxation of probability seems to show opposite behavior: while GRW relaxes diffusively, the relaxation times for MERW are much longer, growing exponentially with the size of the defective region.

These results indicate that MERW might exhibit comparatively fast relaxation within intact or homogeneous regions (like a Cayley tree) but inhibits the relaxation process between regions separated by defects, bottlenecks or bridges. While qualities of MERW's stationary states have already been utilized to improve centrality measures in complex networks [10], its dynamic properties and a close relation between eigenvalues of the adjacency matrix and the statistics of paths may be of use in community search algorithms on complex networks (a number of algorithms based on random walks, path enumeration and spectral properties of the adjacency matrix are reviewed in [20]). As a more speculative idea, it is also worth remembering that MERW keeps all paths of a given length between any two endpoints equiprobable, which makes it capable of hiding the route information travels, *e.g.* over the Internet.

The author would like to thank Z. Burda and B. Waclaw for fruitful discussions. Project operated within the Foundation for Polish Science International Ph.D. Projects Programme co-financed by the European Regional Development Fund covering, under the agreement no. MPD/2009/6, the Jagiellonian University International Ph.D. Studies in Physics of Complex Systems.

REFERENCES

[1] A. Einstein, *Ann. Phys. (Leipzig)* **17**, 549 (1905); **19**, 371 (1906).

[2] M. Smoluchowski, *Ann. Phys. (Leipzig)* **21**, 756 (1906).

[3] G. Polya, *Math. Ann.* **84**, 149 (1921).

[4] Z. Burda, J. Duda, J.M. Luck, B. Waclaw, *Phys. Rev. Lett.* **102**, 160602 (2009).

[5] Z. Burda, J. Duda, J.M. Luck, B. Waclaw, *Acta Phys. Pol. B* **41**, 949 (2010).

[6] L. Demetrius, V.M. Gundlach, G. Ochs, *Theor. Popul. Biol.* **65**, 211 (2004).

[7] L. Demetrius, T. Manke, *Physica A* **346**, 682 (2005).

[8] J.H. Hetherington, *Phys. Rev.* **A30**, 2713 (1984).

[9] V. Zlatic, A. Gabrielli, G. Caldarelli, *Phys. Rev.* **E82**, 066109 (2010).

[10] J.-C. Delvenne, A.-S. Libert, *Phys. Rev.* **E83**, 046117 (2011).

[11] R. Sinatra *et al.*, *Phys. Rev.* **E83**, 030103 (2011).

[12] C. Monthus, T. Garel, *J. Phys. A: Math. Theor.* **44**, 085001 (2011).

[13] K. Anand, G. Bianconi, S. Severini, *Phys. Rev.* **E83**, 036109 (2011).

[14] B. Waclaw, Generic Random Walk and Maximal Entropy Random Walk, Wolfram Demonstration Project, `http://demonstrations.wolfram.com/GenericRandomWalkAndMaximalEntropyRandomWalk/`

[15] W. Parry, *Trans. Amer. Math. Soc.* **112**, 55 (1964).

[16] W.M.X. Zimmer, G.M. Obermair, *J. Phys. A: Math. Gen.* **11**, 1119 (1978).

[17] J.K. Ochab, Z. Burda, *Phys. Rev.* **E85**, 021145 (2012) `arXiv:1201.1420v1 [cond-mat.stat-mech]`.

[18] J.K. Ochab, Stationary states of Maximal Entropy Random Walk and Generic Random Walk on Cayley Trees, Wolfram Demonstration Project, `http://demonstrations.wolfram.com/StationaryStatesOfMaximalEntropyRandomWalkAndGenericRandomWa/`

[19] J.K. Ochab, Dynamics of Maximal Entropy Random Walk and Generic Random Walk on Cayley Trees, Wolfram Demonstration Project, `http://demonstrations.wolfram.com/DynamicsOfMaximalEntropyRandomWalkAndGenericRandomWalkOnayl/`

[20] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).

Maximal-entropy random walk unifies centrality measures

J. K. Ochab[*]

Marian Smoluchowski Institute of Physics, Jagiellonian University, Reymonta 4, PL-30-059 Kraków, Poland

(Received 25 June 2012; published 12 December 2012)

This paper compares a number of centrality measures and several (dis-)similarity matrices with which they can be defined. These matrices, which are used among others in community detection methods, represent quantities connected to enumeration of paths on a graph and to random walks. Relationships between some of these matrices are derived in the paper. These relationships are inherited by the centrality measures. They include measures based on the principal eigenvector of the adjacency matrix, path enumeration, as well as on the stationary state, stochastic matrix, or mean first-passage times of a random walk. As the random walk defining the centrality measure can be arbitrarily chosen, we pay particular attention to the maximal-entropy random walk, which serves as a very distinct alternative to the ordinary (diffusive) random walk used in network analysis. The various importance measures, defined both with the use of ordinary random walk and the maximal-entropy random walk, are compared numerically on a set of benchmark graphs with varying mixing parameter and are grouped with the use of the agglomerative clustering technique. It is shown that centrality measures defined with the two different random walks cluster into two separate groups. In particular, the group of centrality measures defined by the maximal-entropy random walk does not cluster with any other measures on change of graphs' parameters, and members of this group produce mutually closer results than members of the group defined by the ordinary random walk.

DOI: 10.1103/PhysRevE.86.066109 PACS number(s): 89.75.Hc, 02.50.Ga, 05.40.Fb, 89.70.Cf

I. INTRODUCTION

Graphs represent abstracted relationships between entities. They form a structure on which a process may take place, which is often formalized into the mathematical concept called random walk. Together, graphs and random walks can constitute a model for citations in scientific collaboration networks, dissemination of information on social networks, or data transmission on the Internet. Instead of these kinds of information transfer, more tangible subjects may be considered, such as molecule movement on physical or biological networks. Whatever the exact nature of the phenomenon, the natural question arises: Which entity in the network is the most influential, be it a gene or a transcription factor, an overloaded hub, a frequented web site, or a renowned researcher.

A number of importance (or centrality) measures answering that question have been invented to study social (e.g., Ref. [1]; Ref. [2] is an extensive resource) or telecommunication networks (e.g., HITS [3] and PageRank [4]). A significant portion of ideas defining the measures originate from graph theory (e.g., the degree of a vertex, enumeration of paths, or the principal eigenvector of the adjacency matrix) and the theory of Markov chains (e.g., stationary states of random walks, their stochastic matrices, and mean first-passage times). Likewise, most of these approaches have been widely utilized in algorithms of community detection [5].

In this paper, we show that a number of these ideas can be formulated in a common framework. A group of centrality measures defined with the use of a given random walk produce nearly equivalent results, and the results for such a group are more distinct from other measures if they make use of the maximal-entropy random walk (MERW; also called Ruelle-Bowens random walk).

The random walks (RWs) discussed here exist in discrete time and space, in general on any graph, where a step from a vertex (a node) to one of its nearest neighbors takes a unit of time. The stochastic nature of the process expresses itself in the probabilities of taking a step from one node to another. A set of all such probabilities for the whole graph can be stored in a stochastic matrix, which may serve as an object uniquely defining a given RW. For instance, equal probabilities of going from a node to any of its nearest neighbors are enough to define what is called here the generic random walk (GRW), one that is well known and commonly used. However, a sequence of nodes, a path, traversed by a random walker, say a particle, can also be attributed a probability, which is given by the product of the one-step probabilities. A well-chosen set of those path probabilities can also be taken as a RW's definition. This is the case of MERW, for which all paths of a given length and end points are equally probable. These two approaches result in RWs that significantly differ in their stationary and dynamic behavior. It is often the case that either GRW or some biased RWs are better suited for particular problems. However, the author believes that MERW should serve alongside GRW as a null model of random processes on networks. The reason is of fundamental nature: It is GRW that maximizes entropy locally (entropy of the nearest neighbor selection, which results in equal one-step probabilities outgoing from a vertex), and it is MERW that maximizes entropy globally (entropy of the path selection, which results in the equiprobable paths). Since a random walk can be seen as an ensemble of all paths it can generate, it is MERW that yields the largest entropy for that ensemble [6,7] (to be precise, the quantity maximized is the entropy production rate). Thus, in a sense, it is the most random of random walks.

We supplement the above statement with a couple more practical reasons. MERW exhibits behaviors that may be of general interest: Its stationary distribution localizes on diluted lattices [6–8], its relaxation to stationary state is extremely fast on Cayley trees [9,10], and it is very slow between two

[*] jeremi.ochab@uj.edu.pl

identical connected k-regular regions [11]. The equiprobable paths that MERW produces are the natural candidates for an ensemble used in Feynman path integrals (in models of discrete quantum gravity with curved space-time) [7] or in the optimal sampling algorithm in the path-integral Monte Carlo methods [12]. Since entropy maximization is a global principle, conceptually analogical to the least action principle, it was also studied in biology and has led to the concept of evolutionary entropy [13]. The same authors have found the value of entropy for a given graph useful in selection of robust networks [14]. Last, MERW has begun to be used as a tool for analysis of complex networks [15–19].

The rest of this paper is organized as follows. We begin with short definitions of the two random walks under consideration in Sec. II, which allows us to derive relationships between several (dis-)similarity matrices in Sec. III; this sets a framework for revision of a number of known centrality measures in Sec. IV and, finally, for their numerical comparison in Sec. V. We conclude with Sec. VI.

II. GENERIC AND MAXIMAL-ENTROPY RANDOM WALKS

Let us consider a finite connected undirected graph. We define a discrete-time random walk on this graph by a *stochastic matrix* (also transition or Markov matrix) $\mathbf{P}$. Its entry $P_{ij} \geqslant 0$ is the probability that if a random walker stays on a node i at a time t, it will step to a node j at time $t+1$. Any row P_{i*} contains the probabilities of moving to all neighbors of i, and since the walker cannot disappear from the graph nor be created, they all sum up to unity, $\sum_j P_{ij} = 1$. As we assume that the walker can only move to neighboring nodes, the stochastic matrix can have a nonzero entry only if the adjacency matrix of the graph has a nonzero entry at the same place. Shortly, $\forall i,j : P_{ij} \leqslant A_{ij}$, where $\mathbf{A}$ is the *adjacency matrix* of the graph. Elements of this matrix can take two values: $A_{ij} = 1$ if i and j are neighbors and $A_{ij} = 0$ otherwise. The binary values express the fact that the edges of the graph are unweighted. The assumption of undirected edges results in symmetry of the adjacency matrix $\mathbf{A} = \mathbf{A}^T$, which means that the edges can pass information both ways; this does *not* imply symmetry of the stochastic matrix. Both $\mathbf{P}$ and $\mathbf{A}$ are assumed to be time independent.

The probability that the random walker stays at a given vertex i of the graph at a given time t is encoded in the i-th element of the vector $\vec{\pi}(t)^T = [\pi_1(t), \ldots, \pi_N(t)]$. Thus, the initial distribution of probabilities is $\vec{\pi}(0)^T$, and the distribution after t steps $\vec{\pi}(t)^T = \vec{\pi}(t-1)^T\mathbf{P} = \vec{\pi}(0)^T\mathbf{P}^t$, where the stochastic matrix has been multiplied t times.

A quantity of interest, given by a solution of

$$\vec{\pi}^T = \vec{\pi}^T\mathbf{P}, \tag{1}$$

is the *stationary probability distribution* (or stationary state), which may be understood as the probability distribution after infinite time. We assume it exists.[1]

The ordinary or, as we call it, generic random walk corresponds to the standard random walks used by Einstein, Smoluchowski, or Polya. It is realized by the following stochastic matrix:

$$P_{ij} = \frac{A_{ij}}{k_i}, \tag{2}$$

where $k_i = \sum_j A_{ij}$ denotes a *degree* of i-th node (i.e., the number of its nearest neighbors). Its stationary state is proportional to the degrees and is given by $\pi_i = k_i / \sum_j k_j$. An i-th row of the matrix contains uniform probabilities, each equal to $1/k_i$, of selecting any of the k_i neighbors of the node i. Thus, the entropy of neighbor selection is maximal.

The other type of RW, introduced earlier, maximizes the entropy of random trajectories and, hence, is called here the maximal-entropy random walk. This maximization condition leads to a unique stochastic matrix,

$$P_{ij} = \frac{A_{ij}}{\lambda_0}\frac{\psi_{0j}}{\psi_{0i}}, \tag{3}$$

where λ_0 is the largest eigenvalue of the adjacency matrix $\mathbf{A}$ and ψ_{0i} is the i-th element of the principal eigenvector $\vec{\psi}_0$. Since the adjacency matrix is irreducible, the Frobenius-Perron theorem guarantees that all elements of this vector are strictly positive, thus the condition $P_{ij} \leqslant A_{ij}$ is fulfilled. It can be checked that matrix multiplication $\sum_k P_{ik}P_{kj}$ makes the k-labeled eigenvector elements in the numerator and denominator cancel out, which leaves the path probabilities independent of the intermediate nodes. That is how the aforementioned equiprobability of paths is expressed.

MERW has the stationary probability distribution given by Shannon-Parry measure [20]

$$\pi_i = \psi_{0i}^2. \tag{4}$$

Let us note that this formula allows us to interpret ψ_{0i} as the wave function of the ground state of the operator $-\mathbf{A}$ and ψ_{0i}^2 as the probability of finding a particle in this state [6,7], thus relating MERW to quantum mechanics.

It is easily seen that the two RWs, (2) and (3), are identical on k-regular graphs (i.e., graphs whose all nodes have uniform degree of k, e.g., grids or complete graphs). This should be considered an exception, as, in general, their properties are entirely distinct.

III. RELATIONS BETWEEN THE STOCHASTIC MATRIX, ITS DISTANCE MATRIX, MEAN FIRST-PASSAGE TIME MATRIX, AND THE RESOLVENT OF ADJACENCY MATRIX

A. Properties of the stochastic matrix

In general, a stochastic matrix may be not symmetric, and so it may have different right, $\vec{\Psi}_\alpha$, and left, $\vec{\Phi}_\alpha^T$, eigenvectors,

$$\mathbf{P}\vec{\Psi}_\alpha = \Lambda_\alpha\vec{\Psi}_\alpha, \quad \vec{\Phi}_\alpha^T\mathbf{P} = \Lambda_\alpha\vec{\Phi}_\alpha^T, \tag{5}$$

which results in a spectral decomposition,

$$\mathbf{P} = \sum_\alpha \Lambda_\alpha\vec{\Psi}_\alpha\vec{\Phi}_\alpha^T. \tag{6}$$

Clearly, the first left eigenvector is exactly the stationary state vector from (1), $\vec{\Phi}_0 = \vec{\pi}$. We define a diagonal matrix

[1] A stationary state exists if an undirected graph is not bipartite, but even for bipartite graphs a semistationary state can be defined by averaging probability distribution over two consecutive time steps.

$\mathbf{D} \equiv \mathrm{diag}(\vec{\pi})^{-1}$, whose diagonal entries equal to the inverses of the stationary state vector's elements.

With the use of it we narrow down the set of stochastic matrices to a more manageable one. Thus, let us consider a class of random walks whose stochastic matrix can be transformed into a symmetric matrix as given,

$$\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{P}\mathbf{D}^{1/2}. \tag{7}$$

It follows that

$$\vec{\Phi}_\alpha = \mathbf{D}^{-1}\vec{\Psi}_\alpha. \tag{8}$$

This relation does not hold for any random walk but is clearly obtained for GRW and MERW, which satisfy (7), as shown below.

In matrix notation the stochastic matrix (2) for GRW can be written as

$$\mathbf{P} = \mathrm{diag}(k_1,k_2,\dots,k_N)^{-1}\mathbf{A}. \tag{9}$$

Since the stationary state of GRW is proportional to the node degrees $(k_1,k_2,\dots,k_N)$, the diagonal matrix yields $\mathbf{D} = \mathrm{diag}(k_1,k_2,\dots,k_N)^{-1}\sum_j k_j$. At the same time, the symmetric adjacency matrix can be decomposed into $\mathbf{A} = \sum_\alpha \lambda_\alpha \vec{\psi}_\alpha \vec{\psi}_\alpha^{\,T}$, with notation defined in Sec. II. Substitution of these two relations into the equation above yields

$$\mathbf{P} = \frac{1}{\sum_j k_j}\sum_\alpha \lambda_\alpha \mathbf{D}\vec{\psi}_\alpha\vec{\psi}_\alpha^{\,T}. \tag{10}$$

Comparing this formula with (6) one can easily see that the eigenvectors are given by

$$\vec{\Psi}_\alpha = \mathbf{D}\vec{\psi}_\alpha, \quad \vec{\Phi}_\alpha = \vec{\psi}_\alpha, \tag{11}$$

which are related as given in (8).

Similarly, MERW allows for expression of all the eigenvalues and eigenvectors of the stochastic matrix $\mathbf{P}$ (3) in terms of eigenvalues λ_α and eigenvectors of $\vec{\psi}_\alpha$ of the adjacency matrix $\mathbf{A}$,

$$\Lambda_\alpha = \frac{\lambda_\alpha}{\lambda_0}, \quad \vec{\Psi}_\alpha = \mathbf{D}^{1/2}\vec{\psi}_\alpha, \quad \vec{\Phi}_\alpha = \mathbf{D}^{-1/2}\vec{\psi}_\alpha, \tag{12}$$

where $\mathbf{D} = \mathrm{diag}(\psi_{01}^2,\psi_{02}^2,\dots,\psi_{0N}^2)^{-1}$. In particular, $\Lambda_0 = 1, \Psi_{0i} = 1$, and $\Phi_{0i} = \psi_{0i}^2 = \pi_{0i}$ for all i. The spectral decomposition of $\mathbf{P}$ then reads

$$P_{ij} = \sum_\alpha \Lambda_\alpha \Psi_{\alpha i}\Phi_{\alpha j} = \sum_\alpha \frac{\lambda_\alpha}{\lambda_0}\psi_{\alpha i}\psi_{\alpha j}\frac{\psi_{0j}}{\psi_{0i}}. \tag{13}$$

Thus, clearly all properties of MERW are encoded in the spectral decomposition of the adjacency matrix of a given graph; it allows for an easier derivation of, for example, the stationary state and dynamical characteristics of MERW for Cayley trees [9,10,21].

B. (Dis-)similarity matrices

Taking the powers of the stochastic matrix has been utilized in methods of both assessing centrality [22] and finding communities [23,24]. The distance matrix used by Latapy and Pons [24] was given by

$$r(t)_{ij} = \sqrt{\sum_k \frac{[(\mathbf{P}^t)_{ik} - (\mathbf{P}^t)_{jk}]^2}{\pi_k}}, \tag{14}$$

where $\mathbf{P}$ and $\vec{\pi}$ were meant to correspond to GRW (we extend it to a class of random walks). Intuitively, it is assumed that the probability distribution of a random walk outgoing from a node i after t steps [represented by the row $(\mathbf{P}^t)_{i*}$] is a quantity that characterizes the node i; in fact, it tells you how the node i sees the graph after t steps of passing information according to the RW. The difference of these viewpoints between nodes i and j defines the distance between them. The division by π_k is a way of normalizing the contribution of a vertex's centrality to that distance.

These authors mention that $\mathbf{r}^2$, the *entrywise* square of this distance matrix, is equivalent to

$$r^2(t)_{ij} = \sum_{\alpha=1}^{N-1} \Lambda_\alpha^{2t}(\Psi_{\alpha i} - \Psi_{\alpha j})^2, \tag{15}$$

based on spectral decomposition of $\mathbf{P}$ (6).

The remarks from the previous subsection allow us to make further observations. For any RW for which $\mathbf{S}$ defined in (7) is symmetric, and specifically in the case of MERW and GRW, the spectral decomposition (6) leads to the compact form

$$\mathbf{r}^2(t) = \mathbf{D}[(\mathbf{P}^{2t})_{\mathrm{dg}}\mathbf{E} - (\mathbf{P}^{2t})^T] + [\mathbf{E}(\mathbf{P}^{2t})_{\mathrm{dg}} - \mathbf{P}^{2t}]\mathbf{D}, \tag{16}$$

where $(\mathbf{P}^{2t})_{\mathrm{dg}}$ is a matrix with entries $(\mathbf{P}^{2t})_{ii}$ on the diagonal and zeros otherwise. This is a new formula, which, however, very much resembles a symmetrized version of a quantity known as *mean first-passage time* matrix.

The mean first-passage time (MFPT) matrix $\mathbf{M}$ is a useful concept for studying RWs. Its elements M_{if} encode the average time to reach the final vertex f from the initial vertex i for the first time. We invoke a neat construction of the matrix given by Kemeny and Snell [25,26]: first, let us define the *fundamental matrix*

$$\mathbf{Z} = (\mathbf{1} - \mathbf{P} + \vec{e}\vec{\pi}^T)^{-1}, \tag{17}$$

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

$$\mathbf{M} = (\mathbf{E}\mathbf{Z}_{\mathrm{dg}} - \mathbf{Z})\mathbf{D}, \tag{18}$$

where $\mathbf{E}$ is a matrix of all ones, $\mathbf{Z}_{\mathrm{dg}}$ is a diagonal matrix with elements $(\mathbf{Z}_{\mathrm{dg}})_{ii} = Z_{ii}$, and $\mathbf{D}$ was introduced in (7).

The cited authors defined the fundamental matrix so as to contain all the powers of the stochastic matrix $\mathbf{P}$, which follows from expansion of $(\mathbf{1} - \mathbf{P})^{-1}$ in a series $\mathbf{1} + \mathbf{P} + \mathbf{P}^2 + \cdots$. However, as they remark, matrix $\mathbf{1} - \mathbf{P}$ is noninvertible and consequently the expansion does not exist. The correction $\vec{e}\vec{\pi}^T$ allows for a well-defined inversion. In fact, instead of the fundamental matrix one may use other so-called *generalized inverses* (the formalism is summarized in Ref. [27]), although we use (17) for its conceptual and computational simplicity.

Drawing on the analogy between $\mathbf{r}^2(t)$ and $\mathbf{M}$ that we have spotted, we may redefine (14), and take $\sum_{t=0}^{\infty}\mathbf{P}^t$ instead of $\mathbf{P}^t$ to account for all the powers of the stochastic matrix. This infinite sum

$$\pi_f \sum_{t=0}^{\infty}(\mathbf{P}^t)_{fi} = \sqrt{\pi_f}G_{fi}\sqrt{\pi_i} \tag{19}$$

reproduces the path-integral (MERW) and field-theoretical (GRW) propagator $\mathbf{G}$ of a free relativistic particle, as has been shown in Ref. [7], which supports the view that the stationary probability (4) is reminiscent of the square of a wave function.

Nevertheless, the matrix $\mathbf{G}$ needs further elaboration. An elementary and more general definition than in (19) makes an entry $\mathbf{G}_{fi}$ represent the number of trajectories of all lengths between nodes i and f. We recall that the count of paths of length t is conveniently given by the powers of the adjacency matrix $(\mathbf{A}^t)_{fi}$. As the number of paths dramatically grows with their length, however, a normalizing parameter $e^{\mu} > e^{\mu_0} \equiv \lambda_0$ has to be introduced for the sum of paths to converge,

$$\mathbf{G}(\mu) = \sum_{t=0}^{\infty} e^{-\mu t}\mathbf{A}^t. \tag{20}$$

From the point of view of paths' statistics, $\mathbf{G}(\mu)$ defines the grand-canonical ensemble of paths. An element $G_{fi}(\mu)$ corresponds to the grand-canonical partition function, μ to the chemical potential, and the average path length is $\langle t\rangle_{fi} = -(\ln G)'_{fi}(\mu)$.

The role of μ is equivalent to a cutoff of a path length of the order $T = 1/\Delta\mu$, where $\Delta\mu \equiv \mu - \mu_0$. In Eqs. (19) above and (22) below, $\mathbf{G}$ stands for $\mathbf{G}(\mu_0)$, where the special choice of $\mu_0 = -\ln\lambda_0$ explicitly relates the propagator to the graph structure by the largest eigenvalue of the adjacency matrix. In this limit, $\mu \longrightarrow \mu_0$, infinite paths begin to dominate the average, and $\mathbf{G}(\mu)$ has a singularity.

However, $\mu = -\ln\lambda$ very close to μ_0 can be taken, yielding

$$\mathbf{G}(\mu) = \sum_{t=0}^{\infty}\frac{\mathbf{A}^t}{\lambda^t} = \frac{1}{\lambda}(\lambda\mathbf{1} - \mathbf{A})^{-1}, \tag{21}$$

which is the resolvent of the adjacency matrix. Clearly, at $\mu = \mu_0$ the right-hand side is ill defined. To define $G(\mu)$ at the singularity, the matrix $\lambda_0\mathbf{1} - \mathbf{A}$ has to be projected to the subspace perpendicular to $\vec{\psi}_0$ before inversion. It can be done similarly as in (17) by taking $(\lambda_0\mathbf{1} - \mathbf{A} + \lambda_0\vec{\psi}_0\vec{\psi}_0^{\,T})^{-1}$, which eliminates the zeroth eigenmode. This is expected and advantageous in community finding methods, as discussed in Ref. [28].

Finally, on substitution of the infinite sum over t in place of $\mathbf{P}^t$ in (16) we obtain

$$\mathbf{r}^2 = \mathbf{D}(\mathbf{G}^2)_{\mathrm{dg}}\mathbf{E} - 2\sqrt{\mathbf{D}}\mathbf{G}^2\sqrt{\mathbf{D}} + \mathbf{E}(\mathbf{G}^2)_{\mathrm{dg}}\mathbf{D}, \tag{22}$$

with $\mathbf{G}$ functioning as an analog of the fundamental matrix $\mathbf{Z}$ and where the time dependence has been eliminated. Thanks to the symmetry of the matrix $\mathbf{r}^2$, however, the singular mode of $\mathbf{G}$ cancels out, even without the projection discussed in the paragraph above. This stands in contrast to the definition of MFPT with the use of the fundamental matrix, which is nonsymmetric, and where the singularity was elimated by hand.

IV. CENTRALITY MEASURES

The above considerations constitute a common framework for a number of centrality measures. Below, the connections between them are reviewed and established.

A. Centrality based on paths

The original concept of counting paths to assess centrality was introduced in 1953 [29]. The idea is to count all the paths that lead to a vertex whose importance we measure. For a given path length t, the number of such paths between vertices i and f is given by the element $(A^t)_{fi}$ of the t-th power of the adjacency matrix. This corresponds exactly to the definition shown in (20). Below, we rewrite the original definition from Ref. [30] in terms of the propagator $\mathbf{G}(\mu)$.

The importance of the final vertex f is then given by the element I_f of the vector $\vec{\mathbf{I}} = (\mathbf{G}(\mu) - \mathbf{1})\vec{\mathbf{e}} \approx c_0(\mu)\vec{\psi}_0$, where the uniform vector $\vec{\mathbf{e}}$ was chosen as a set of initial conditions (the importance is measured uniformly with respect to all initial vertices), and the proportionality to the principal eigenvector holds near μ_0, with some constant of proportionality c_0 which depends on μ. Squared elements of the principal eigenvector of $\mathbf{A}$ are the stationary probabilities of MERW, which means they correspond to the contribution from infinite paths. In the limit $\mu \longrightarrow \mu_0$ the constant $c_0(\mu)$ diverges. As a result, the contribution to the centrality of other eigenvectors, corresponding to paths of shorter lengths, is negligible. We explain the nuances of this divergence at the end of the previous section.

In Sec. V, where we perform numerical analysis, we do not restrict the values of μ to be strictly greater than μ_0. Instead, we do effectively the same thing by setting $\mu = \mu_0$ and limiting the maximal length of the enumerated paths with finite sums taken in (20). The maximal length of the paths is set equal to the diameter of a given graph. We also remark that in the numerical analysis the elements of $\vec{\mathbf{I}}$ are squared so they correspond to the stationary probability of MERW.

The path weights exponential with respect to the path's length t [i.e., $e^{-\mu t}$ in the notation we use in (20)] were also employed in Ref. [22]. These authors, however, proposed additional restrictions. The idea was to reduce the number of paths one takes into account, e.g., by taking only the shortest paths or k-short paths (i.e., paths of length smaller than k). The heuristic explanation is that the path between given two nodes that transmits the information the fastest is the crucial one, and perhaps, in the real world, the longer paths would not have been used. Similar idea motivates taking only k-short vertex-disjoint paths (i.e., paths that additionally do not have any nodes in common apart from the initial and terminal one). Unfortunately, these ideas are harder to trace analytically, so we just use them for the sake of comparison. In Sec. V we take only the shortest paths without any constraints on their length.

The exponential weighting of paths, as $e^{-\mu t}$ in (20), is not the only possible choice. Alternatively, factorial weights $\beta^t/t!$ might be introduced [31,32]. As in the previous case, these weights guarantee convergence of the infinite sum in (20). Mathematically speaking, instead of producing a resolvent operator, as in (21), the similarity matrix takes the form of another well-known operator, the *heat kernel* $K(-\beta) = (e^{\beta\mathbf{A}})$. We stress that what usually is called the heat kernel of a graph has a Laplacian matrix in place of $\mathbf{A}$ and time in place of $-\beta$. It is then the solution of the heat equation and is thought to represent the flow of information on the graph in time. Here, the analogy lies in the form of the operator, where the adjacency matrix seems to play the role of the graph Laplacian. The interpretation the authors of the cited paper give, however, differs: They think of the graph as a network of balls (nodes) and springs (edges). The adjacency matrix becomes then the Hamiltonian of the system, and $K(-\beta)$ becomes a Green's function. The parameter β can then represent an inverse temperature of a heat bath the system is immersed in. Effectively, $\beta < 1$ suppresses and $\beta > 1$ allows for longer paths to be taken into account. The resulting similarity matrix has been used in a method of community detection.

We provide the reader with this type of path weighting as an alternative. It is, however, the former weight choice (21) that generates the unique maximally entropic random walk. Those weights make MERW directly reflect the structure of the graph, which is explicit in the transition matrix definition (3) or, conversely, appropriately weighted paths gain the interpretation of a random walk.

B. Centrality based on powers of the transition matrix

Equation (19) shows that path enumeration is equivalent to the propagation of MERW. Let us note, however, that the walks are also weighted by the ratio $\psi_{0f}/\psi_{0i} = \sqrt{\pi_f/\pi_i}$ of stationary probabilities of the two vertices. It is a reasonable intuition that the importance of a random walk trajectory depends on the importance of the initial and final vertices. It seems that the problem of calculating centrality by employing the transition matrix becomes self-consistent (importance calculated from paths, whose weights depend on the importance) and, thus, eliminates arbitrariness.

The method of assessing centrality by summing consecutive powers of the transition matrix is stated in Ref. [22],

$$\vec{\mathbf{I}}^T = \sum_{t=1}^{T} \vec{\pi}(0)^T \mathbf{P}^t, \tag{23}$$

where for simplicity we choose uniform initial probability distribution $\vec{\pi}(0)^T$. Intuitively, the influence of the initial vertex on its surroundings is estimated with T steps of a random walk, which corresponds to the appropriate choice of μ in path enumeration approach, as explained above in Eq. (21). This parameter controls whether local effects or the stationary state is favored, with $\vec{\mathbf{I}}$ approaching the stationary probability distribution for large T. The number of steps is usually kept rather small, due to computation costs.

As noted at the beginning of this subsection, what results from our study is that for MERW the definition (23) is very similar to counting paths. Even for relatively small T, it produces results very close to the stationary probability distribution. Obviously, one expects it for large T, but the nontrivial fact we have checked is that, on average, MERW reaches the stationary state faster than GRW on the benchmark graphs used. To be precise, the probability distribution of a random walk comes closer to the stationary state as an exponent in time, $|\pi_i(t) - \pi_i| \propto \exp(-t/\tau)$, with the characteristic relaxation time τ, which is shorter for MERW than for GRW (even twice as short for graphs with very strong modular structure; however, this is less visible for nearly random graphs). To a large extent we have accounted for that behavior, as MERW seems to relax very fast within connected regions (proven for Cayley trees [9,10]), although it takes a long time to relax between two identical connected regions [11].

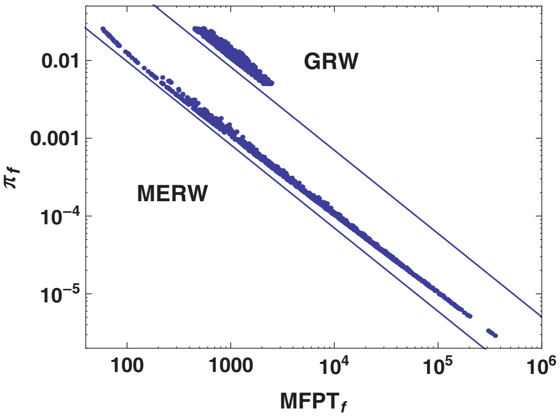

FIG. 1. (Color online) The stationary probability distribution of MERW and GRW as a function of the averaged rows of MFPT matrix (average MFPT of reaching a vertex f from any initial vertex) for a sample graph with $N = 1000$ vertices. Solid lines have best fit slopes -1.027 ± 0.001 (MERW) and -1.070 ± 0.005 (GRW). The correlation for GRW is weaker, since the degrees of the graph take values 10–50 and accordingly the stationary state is quantized. π_f for GRW is multiplied by 10 for clarity.

C. Centrality based on mean first-passage times, stationary distributions, and the principal eigenvector of the adjacency matrix

As shown in Sec. III there is a close analogy between $\mathbf{r}^2$, which uses powers of the transition matrix discussed above, and mean first-passage times matrix $\mathbf{M}$. The centrality based on MFPT matrix is given by the inverse of $\sum_i M_{if}$, where the sum represents the average time the information needs to reach the final vertex f from anywhere in the graph. This definition is called *Markov centrality* in Ref. [22].

Clearly, the multiplication by $\mathbf{D}$ in the definition of MFPT (18) causes the approximate trend $M_{if} \sim \pi_f^{-1}$. However, since previous studies usually assumed the Markov process to be GRW, we complement them with comparison of the dependence between MERW and GRW, shown in Fig. 1. Since the stationary state of MERW is typically distributed over a wide range of values (even on almost regular graphs [7]), MFPTs correlate with it very strongly. Hence, for MERW the information extracted from MFPT matrix and the stationary distribution is largely equivalent. Especially on bounded-degree graphs, their values extend much further than for GRW, whose stationary distribution is proportional to vertex degrees and, thus, also bounded.

This observation begs the question: Which random walk should be chosen to define a centrality measure in terms of the stationary distribution? Thus far, the one used most widely is GRW, whose stationary state is produced by the simplest version of the prominent PageRank [4]. The two random walk centralities, however, have already been compared in Ref. [16] and the conclusion was, among others, that MERW has "a larger discriminating power between the best and worst pages" and is sensitive to link farms.

Nonetheless, in these studies the connection to other methods has been missing. We indicate that both paths' statistics and random walks are linked to the idea of calculating centrality as an eigenvector associated with the largest eigenvalue of the adjacency matrix, which is a concept as old as the economic and sociological papers from 1965 [33] and 1972 [34]. In the latter, this centrality was derived from the assumption that $\vec{\mathbf{I}}_t = \mathbf{A}^t\vec{\mathbf{e}}/\lambda_0^t$ is the t-th order importance measure and that an objective measure should be taken in $t = \infty$, convergence thus requiring the factor λ_0^{-t}. Clearly, this formula is simply the canonical ensemble version of the one based on paths (20), and the proposed eigenvector centrality is the square root of the stationary state of MERW (4). This is also reminiscent of the HITS algorithm [3] for directed graphs, which nevertheless uses eigenvectors of $\mathbf{A}^T\mathbf{A}$.

We note that just as centrality may be defined with the use of the principal eigenvector of the adjacency matrix or the stochastic matrix (the stationary state vector), there is a family of community detection methods analyzing the rest of the eigenvectors (often it is the spectrum of Laplacian that is analyzed). In fact, each of the methods of assessing centrality mentioned above has a number of counterparts that in a similar manner try to find the community structure of a network. In Ref. [28], we present a comparison between GRW and MERW in performance of some community finding methods based on the concepts presented above.

V. COMPARISON

We check the affinity of different centrality measures described above (together with the closeness and betweenness centrality given for reference; see Ref. [35]) by comparing the result they produce for a sample of graphs. For a given graph, each centrality measure produces a vector $\vec{\mathbf{I}}$, whose consecutive elements are centrality values of the corresponding nodes. To compare results of a pair of methods on that graph, we take the corresponding pair of vectors and measure the rms distance between them (cosine or Pearson correlation distance have been checked as well and have generated similar results). After repeating this computation for all pairs of centrality measures we obtain a square matrix. Since each graph from the sample produces one such matrix, we take the average (entrywise) over the whole sample. The entries of the resultant matrix represent the average distance between a pair of centrality measures. Finally, this distance matrix is used as input for an agglomerative clustering algorithm with average weights, which generates the dendrograms in Fig. 2. The heights of their branches correspond to the distances between pairs of clusters. The maximum standard deviation of the distance matrix entries is smaller than 0.61%, hence, the results of the clustering algorithm should be correct for most graphs in the sample.

For example, in the dendrogram on the left in Fig. 2, the nearest centralities are *1* and *2* (which denote the stationary state of MERW and its MFPT centrality), and they were the first to be clustered together. Next, methods *5* and *7* were clustered (the shortest paths' centrality and the MFPT centrality of GRW), and so on. It can be seen that the closeness and betweenness (*9* and *10*) are always clustered at the very end, which means they are very distinct from the other methods. This is expected, as they are based on a different concept of importance and could, for instance, assign a high centrality score to a node near a bottleneck (i.e., very narrow, local bridge between two communities), even though it is poorly connected. The methods *4* and *5* depend on the maximum path length T taken into account (it is set to the diameter of the graph, which varies between 4 and 10), so their assignment might differ for parameter choices other than shown here.

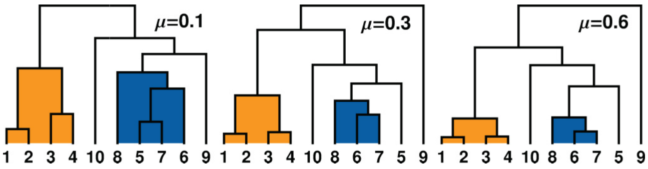

FIG. 2. (Color online) The dendrograms correspond to benchmark graphs with the mixing parameter $\mu = 0.1,0.3,0.6$. The labels *1*, *2*, and *3* represent MERW's stationary state, MFPT, and $\mathbf{P}^t$ (orange cluster). *6*, *7*, and *8* are GRW's analogs (blue cluster). *4* denotes weighted paths' and *5* shortest paths' centrality. *9*, *10* are closeness and betweenness. In *3* and *8*, the maximal power of $\mathbf{P}$ is $T = 5$. In *4* and *5* $\mu = \ln \lambda_0$ and T equals the diameter of the graph.

More importantly, another look at the dendrograms reveals that when the parameters of the benchmark graphs change there are two clusters of methods that do not mix with each other. One includes centralities derived from MERW (*1*, *2*, and *3*: centralities based on the stationary state, MFPT, and $\mathbf{P}^t$, respectively) and *4*, which is based on weighted paths (the orange cluster in Fig. 2), while the other includes centralities derived from GRW (*6*, *7*, and *8*, again the stationary state, MFPT, and $\mathbf{P}^t$) and *5*, which is based on shortest paths (the blue cluster). Thus, methods utilizing GRW are close to each other; however, for graphs with easily distinguishable communities they can cluster together with other centrality measures. The methods utilizing MERW are all connected to path enumeration (20), as predicted in Sec. IV A, and they never intermingle with the other centrality measures. The average distance of this whole group from other methods analyzed is greater than the analogous distance for the group of GRW methods, whereas the average distance between the members of this group is smaller than the corresponding value for GRW methods. This means that, indeed, the centralities defined by MERW comprise a distinct, close-knit family, and produce equivalent results.

A. Benchmarks

In the analysis in the previous section Lancichinetti-Fortunato-Radicchi benchmark graphs [36] were utilized. Since they were designed to benchmark community finding algorithms, they contain communities with preset size distribution, constructed with the use of the *planted partition model*. In short, the model is based on fixed probabilities, p_{in} and p_{out}, which determine if two given nodes should be linked (in this case they have been assigned to the same communities or to different ones, respectively). Although the graphs constructed that way are locally random, they model a range of possible real-world structures, and so they serve our purpose in testing centrality measures.

We follow the notation used by the authors of the benchmarks. Thus, by μ we denote the mixing parameter [this should not be confused with the usage of μ in (20), where chemical potential is meant; the context should make which meaning is intended unambiguous], which is the fraction of links a given node shares with the nodes outside its community. The parameter is approximately equal for all nodes in a graph. For chosen values of μ, we take 100 graphs with $N = 200$ vertices; their exponents for the degree and community size distributions are, respectively, $\tau_1 = -2$ and $\tau_2 = -1$. The average and maximum degrees are 10 and 30, and the community sizes range from 5 to 35.

VI. CONCLUSIONS

In this paper it has been shown that the random walk distance matrix $\mathbf{r}^2(t)$ defined in (14), when modified to account for walks of all lengths, is equivalent to a symmetric version of the mean-first passage matrix $\mathbf{M}$ (18), where the fundamental matrix $\mathbf{Z}$ is substituted with the propagator $\mathbf{G}$.

This observation also leads to the conclusion that a number of known centrality measures are nearly identical if the random walk under consideration is the maximal-entropy random walk. This common perspective includes measures related to the properties of graphs (the eigenvector centrality and centrality based on enumeration of weighted paths) and those related to random walks (their stationary state, powers of their transition matrix, and, finally, their MFPT matrix), as reviewed in Sec. IV.

A numerical investigation on a set of benchmark graphs confirms this thesis, showing that there is a group of centrality measures related to GRW that tend to produce similar results and an even more homogeneous and distinct group of centralities related to MERW. To quote Bonacich [34]: “Three different approaches to calculating popularity scores have almost the same solution [. . .]. This is an economy; three approaches are reduced to just one. This is the main point of the paper.”

ACKNOWLEDGMENTS

Special thanks for numerous discussions are due to Zdzisław Burda. 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.

[1] L. Freeman, Social Networks **1**, 215 (1979).

[2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, UK, 1994).

[3] J. Kleinberg, J. Assoc. Comput. Mach. **46**, 604 (1999).

[4] S. Brin and L. Page, Comput. Networks ISDN Syst. **30**, 107 (1998).

[5] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[6] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Phys. Rev. Lett. **102**, 160602 (2009).

[7] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Acta Phys. Pol. B **41**, 949 (2010).

[8] B. Waclaw, *Generic Random Walk and Maximal Entropy Random Walk*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/GenericRandomWalkAndMaximalEntropyRandomWalk/.

[9] J. K. Ochab and Z. Burda, Phys. Rev. E **85**, 021145 (2012).

[10] J. K. Ochab, *Dynamics of Maximal Entropy Random Walk and Generic Random Walk on Cayley Trees*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/DynamicsOfMaximalEntropyRandomWalkAndGenericRandomWalkOnCayl/.

[11] J. K. Ochab, Acta Phys. Pol. B **43**, 1143 (2012).

[12] J. H. Hetherington, Phys. Rev. A **30**, 2713 (1984).

[13] L. Demetrius, V. M. Gundlach, and G. Ochs, Theor. Popul. Biol. **65**, 211 (2004).

[14] L. Demetrius and T. Manke, Physica A **346**, 682 (2005).

[15] V. Zlatic, A. Gabrielli, and G. Caldarelli, Phys. Rev. E **82**, 066109 (2010).

[16] J.-C. Delvenne and A.-S. Libert, Phys. Rev. E **83**, 046117 (2011).

[17] R. Sinatra, J. Gomez-Gardenes, R. Lambiotte, V. Nicosia, and V. Latora, Phys. Rev. E **83**, 030103 (2011).

[18] C. Monthus and T. Garel, J. Phys. A **44**, 085001 (2011).

[19] K. Anand, G. Bianconi, and S. Severini, Phys. Rev. E **83**, 036109 (2011).

[20] W. Parry, Trans. Amer. Math. Soc. **112**, 55 (1964).

[21] J. K. Ochab, *Stationary States of Maximal Entropy Random Walk and Generic Random Walk on Cayley Trees*, Wolfram Demonstration Project, http://demonstrations.wolfram.com/StationaryStatesOfMaximalEntropyRandomWalkAndGenericRandomWa/.

[22] S. White and P. Smyth, in *KDD ’03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., 2003* (ACM, New York, 2003), pp. 266–275.

[23] D. Harel and Y. Koren, in *FST TCS ’01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science* (Springer-Verlag, London, 2001), pp. 18–41.

[24] M. Latapy and P. Pons, Lect. Notes Comput. Sci. **3733**, 284 (2005).

[25] J. G. Kemeny and J. L. Snell, *Finite Markov Chains* (Springer-Verlag, New York, 1976).

[26] C. M. Grinstead and J. L. Snell, *Introduction to Probability* (American Mathematical Society, Providence, RI, 1997).

[27] J. J. Hunter, Res. Lett. Inf. Math. Sci. **3**, 99 (2002).

[28] J. K. Ochab and Z. Burda, arXiv:1208.3688v1.

[29] L. Katz, Psychometrika **18**, 39 (1953).

[30] P. Bonacich, Amer. J. Sociol. **92**, 1170 (1987).

[31] E. Estrada and N. Hatano, Phys. Rev. E **77**, 036111 (2008).

[32] E. Estrada and N. Hatano, Appl. Math. Comput. **214**, 500 (2009).

[33] C. H. Hubbell, Sociometry **28**, 377 (1965).

[34] P. Bonacich, J. Math. Sociol. **2**, 113 (1972).

[35] Dirk Koschützki *et al.*, Lect. Notes Comput. Sci. **3418**, 16 (2005).

[36] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).

Eur. Phys. J. Special Topics **216**, 73–81 (2013)
© EDP Sciences, Springer-Verlag 2013
DOI: 10.1140/epjst/e2013-01730-6

THE EUROPEAN
PHYSICAL JOURNAL
SPECIAL TOPICS

Regular Article

Maximal entropy random walk in community finding

J.K. Ochab[a] and Z. Burda[b]

Marian Smoluchowski Institute of Physics and Mark Kac Complex Systems Research Center
Jagiellonian University, Reymonta 4, 30-059 Kraków, Poland

Received 30 November 2012 / Received in final form 6 December 2012
Published online XX January 2013

Abstract. The aim of this paper is to check feasibility of using the maximal-entropy random walk in algorithms finding communities in complex networks. A number of such algorithms exploit an ordinary or a biased random walk for this purpose. Their key part is a (dis)similarity matrix, according to which nodes are grouped. This study encompasses the use of a stochastic matrix of a random walk, its mean first-passage time matrix, and a matrix of weighted paths count. We briefly indicate the connection between those quantities and propose substituting the maximal-entropy random walk for the previously chosen models. This unique random walk maximises the entropy of ensembles of paths of given length and endpoints, which results in equiprobability of those paths. We compare the performance of the selected algorithms on LFR benchmark graphs. The results show that the change in performance depends very strongly on the particular algorithm, and can lead to slight improvements as well as to significant deterioration.

1 Introduction

Relationships between entities can be represented as a graph structure upon which some process takes place, be it information or opinion spread on social networks, including citation and collaboration networks, WWW or the Internet, or perhaps a physical process (molecular motion) on physical or biological networks. One of the natural questions to be asked is whether there are groups of entities which are connected stronger to each other than to the rest of the network. Due to the sociological legacy, these are called *communities*, but they can comprise researchers, websites, genes or transcription factors as well.

A plenitude of methods have been devised to find such communities, and a plenitude of definitions have been conceived to tell what it is that we really look for. These definitions and methods have been thoroughly reviewed in [1]. A particular subgroup of algorithms is based on random walks (RWs), since intuitively a random walker is

[a] e-mail: jeremi.ochab@uj.edu.pl
[b] e-mail: zdzislaw.burda@uj.edu.pl

expected to spend a longer time inside the well-connected graph regions, and there should be only a slim chance that it crosses from one to another.

The most common choice for such algorithms has been the well-known random walk defined by equal probabilities of going from a node to any of its nearest neighbours, which we call the generic random walk (GRW). On the contrary, maximal-entropy random walk (MERW) ensures equiprobability of all paths of a given length and endpoints. Although for many problems, GRW and biased RWs are often more suitable, MERW deserves particular interest: while the former maximises the entropy locally (entropy of the nearest neighbour selection), the latter maximises the entropy globally (entropy of the path selection) [2,3]. Among its curious behaviours, MERW exhibits localization of its stationary distribution on diluted lattices [2–4] and Cayley trees [5,6], it also relaxes extremely fast on these trees [5,7], while it does very slowly between two identical connected regions [8]. Thus, we believe MERW can serve alongside GRW as a null model of random processes on networks.

It is noteworthy that equiprobable paths (as generated by MERW) are the natural choice for an ensemble used in Feynman path integrals (e.g., discrete quantum gravity models with curved space-time) [3] or in the optimal sampling algorithm in the path-integral Monte Carlo methods [9]. Entropy maximization is a global principle much like the least action principle. It has earlier led to the biological concept of evolutionary entropy [10]. Interestingly, the value of entropy for a given graph, as defined by MERW, has been found useful for selection of robust networks [11]. Finally, it has begun to be used in the study of complex networks [12–16].

2 Generic and maximal-entropy random walks

Let us consider a discrete time random walk on a finite connected undirected graph, with its stochastic matrix P being constant in time. An element $P_{ij} \geq 0$ of this matrix encodes the probability that a walker that stands on a node i at time t hops to a node j at time $t+1$. These matrix elements fulfil the condition $\sum_j P_{ij} = 1$ for all i, which means that the number of walkers is conserved. An additional assumption allows the walkers to hop only to a neighbouring node. This can be formulated as $P_{ij} \leq A_{ij}$, where A_{ij} is the corresponding element of the adjacency matrix A of the graph: $A_{ij} = 1$ if i and j are neighbours, and $A_{ij} = 0$ otherwise.

For any time t, the probability of a walker staying on a given vertex of the graph is encoded in the vector $\pi(t) = (\pi_1(t), \ldots, \pi_N(t))^T$. The initial distribution of particles is $\pi(0)$, and the distribution after t steps $\pi(t)^T = \pi(0)^T \mathsf{P}^t$. A quantity of interest is the stationary probability distribution, which we assume to exist. Then, it is given by a solution of

$$\pi^T = \pi^T \mathsf{P}, \tag{1}$$

and may be regarded as the probability distribution after infinite time.

GRW is realised by the following stochastic matrix:

$$P_{ij} = \frac{A_{ij}}{k_i}, \tag{2}$$

where $k_i = \sum_j A_{ij}$ denotes the node degree. The factor $1/k_i$ in the above formula produces a uniform probability of selecting one of k_i neighbours of the node i. This choice maximises the entropy of neighbour selection and corresponds to the standard Einstein-Smoluchowski-Polya random walk. The stationary probability distribution of GRW is given by $\pi_i = k_i / \sum_j k_j$.

The other type of random walk, MERW, is defined by a stochastic matrix that maximises entropy of a set of trajectories with a given length and end-points.

This is a global principle similar to the least action principle. It leads to the following stochastic matrix:

$$P_{ij} = \frac{A_{ij}}{\lambda_0} \frac{\psi_{0j}}{\psi_{0i}}, \tag{3}$$

where λ_0 is the largest eigenvalue of the adjacency matrix A, and ψ_{0i} is the i-th element of the corresponding eigenvector $\boldsymbol{\psi}_0$. By virtue of the Frobenius-Perron theorem, all elements of this vector are of the same sign, because the adjacency matrix A is irreducible. For a stochastic matrix to maximise the entropy of an ensemble of paths, the choice (3) is unique.

The defining condition of entropy maximization leads to equiprobability of paths. More precisely, let us take a sequence of nodes $\gamma_{a_0 a_\tau} = (a_0, a_1, \ldots, a_\tau)$, which is a path of τ steps with the initial node a_0 and the final node a_τ. The probability of visiting this sequence of nodes is

$$P(\gamma_{a_0 a_\tau}) = P_{a_0 a_1} P_{a_1 a_2} \cdots P_{a_{\tau-1} a_\tau}, \tag{4}$$

which results from the Markov property of the random walk. Upon substitution of MERW's stochastic matrix, one obtains

$$P(\gamma_{a_0 a_\tau}) = \frac{1}{\lambda_0^\tau} \frac{\psi_{0a_0}}{\psi_{0a_\tau}}, \tag{5}$$

which depends only on the number of steps and on the two ending points, but is independent of the intermediate nodes. This is what we mean by equal probability of paths of a given length and end-points. Consequently, the probability measure on this ensemble of paths is uniform, and its entropy is maximal.

The stationary state of MERW is given by Shannon-Parry measure [17]:

$$\pi_i = \psi_{0i}^2. \tag{6}$$

The last formula forms a connection between MERW and quantum mechanics, since ψ_{0i} can be understood as the wave function of the ground state of the operator $-\mathsf{A}$ and ψ_{0i}^2 as the probability of finding a particle in this state [2,3]. The two types of random walk, (2) and (3), behave identically on k-regular graphs. In general, however, they have completely disparate properties.

3 (Dis)similarity matrices for community finding algorithms

Methods of both assessing centrality [18] and finding communities [19,20] have widely utilised calculating powers of the stochastic matrix. The one by Latapy and Pons [20] uses the dissimilarity matrix

$$r(t)_{ij} = \sqrt{\frac{\sum_k [(\mathsf{P}^t)_{ik} - (\mathsf{P}^t)_{jk}]^2}{\pi_k}}, \tag{7}$$

where the division by π_k is supposed to reduce the effect of centrality of a vertex. Originally, P and π corresponding to GRW were chosen.

Another approach is an explicit use of the mean first-passage times (MFPT) [21–23]. MFPT matrix M is a useful and well-studied quantity characterising RWs. Its construction with the use of the fundamental matrix Z is given in [24,25]

$$\mathsf{Z} = (1 - \mathsf{P} + \boldsymbol{e}\pi^T)^{-1}, \tag{8}$$

$$\mathsf{M} = (\mathsf{E}\mathsf{Z}_d - \mathsf{Z})\mathsf{D}, \tag{9}$$

where 1 is the identity matrix, $\mathrm{e} = (1, 1, ..., 1)^T$, E is a matrix of all ones, Z_d is a diagonal matrix with elements $(\mathsf{Z}_d)_{ii} = Z_{ii}$, and D is a diagonal matrix with elements $(\mathsf{D})_{ii} = 1/\pi_i$. The elements M_{ij} encode the average time to reach the vertex j from i for the first time (in general $M_{ij} \neq M_{ji}$).

The last approach we discuss is a similarity matrix containing the average number of paths between two given nodes (which is just A^t) with weights that depend on the length of the path.

$$\mathsf{G}(\mu) = \sum_{t=0}^{\infty} e^{-\mu t}\mathsf{A}^t. \tag{10}$$

For $e^\mu \equiv \lambda > \lambda_0$, the sum is convergent and can be carried out with the use of spectral decomposition of A. From the point of view of statistics of the paths $\mathsf{G}(\mu)$ defines the grand-canonical ensemble of paths. An element $G_{fi}(\mu)$ corresponds to the grand canonical partition function, μ corresponds to the chemical potential, and the average path length is $\langle t \rangle_{fi} = -(\ln G)'_{fi}(\mu)$. To avoid a conflicting notation, henceforth we use $\lambda \equiv e^\mu$, whereas the symbol μ will be exclusively reserved for the mixing parameter of benchmark graphs (see Sect. 4.2).

In the case of MERW and GRW (generally, for any RW for which $\mathsf{D}^{-1/2}\mathsf{P}\mathsf{D}^{1/2}$ is symmetric) it can be shown that these three quantities are intimately related constituting a common framework for a number of centrality measures [3,26].

4 Community finding algorithms

4.1 Comparison

Each of the above quantities has an analogic centrality measure: r has the stationary state centrality and centralities defined by summation of powers of the stochastic matrix, G has the eigenvector centrality and centralities defined by path enumeration, and M has a centrality defined by the inverse of its average rows [26]. These are natural counterparts to some community finding methods.

Just as centrality may be defined with the use of the principal eigenvector of the adjacency matrix or the stochastic matrix (then, the eigenvector is the stationary state), there is a family of community finding methods analysing the rest of the eigenvectors (often it is the spectrum of Laplacian that is analysed) [27–32]. However, having the two random walks at hand, we are more interested in methods that utilize their characteristics. Particularly, we try to assess what difference it makes, when we switch between those two random walks.

Below, we present several available methods that originally use GRW as the random walk of choice. These algorithms have not been previously systematically compared on benchmark graphs (described in detail in Sect. 4.2). We measure their performance on a set of such graphs, and compare it with the performance of the same methods, in which we substituted MERW for GRW.

There are a number of methods using powers of the transition matrix. For instance, [19] use the matrix

$$\mathsf{P}^{\leq T} \equiv \sum_{t=1}^{T} \mathsf{P}^t, \tag{11}$$

where P corresponded to GRW, and T was taken around 2–3. The assumption is that two nodes are close to each other if the corresponding rows of $\mathsf{P}^{\leq t}$ matrix are similar.

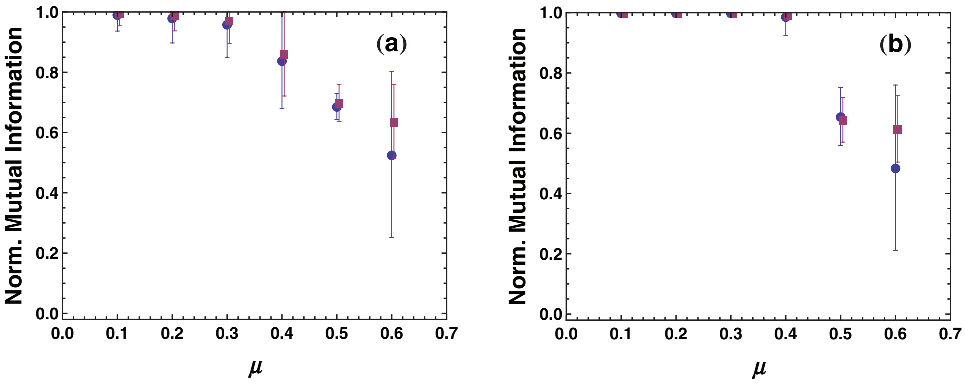

Fig. 1. Comparison of community detection efficiency between MERW (squares; $T = 2$, 4 iterations) and GRW (circles; $T = 3$, 3 iterations) transition matrix used in the first iteration of algorithm [19] on benchmark graphs. Graph size: (a) $N = 200$, (b) $N = 1000$. Normalized mutual information (NMI) equal to 1 means a perfect match between the communities found and the preset community structure; NMI equal to 0 means no information on the real community structure. The two random walks provide comparable performance of the algorithm.

One of the proposed similarity functions between two vectors is

$$\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \exp\left(2T - \sum_{i=1}^{N} |x_i - y_i|\right) - 1. \tag{12}$$

In this formula, if $T = 1$, the vectors $\mathbf{x}, \mathbf{y}$ are rows of the stochastic matrix. Hence, the elements of each of them sum up to 1. There are T stochastic matrices summed in (11), hence in general the elements of each vector sum up to T. If the two vectors are maximally different, the sum in (12) becomes $2T$, and the similarity reaches the lower boundary value of 0.

The algorithm consists in replacing edge weights of the original graph with the elements of the similarity matrix, so that external (intercommunity) links get smaller weights, and the internal ones get larger weights. The procedure is iterated until the differences between weights become large enough, and the weights below a given threshold can be disposed of. What remains is the communities. It is viable to use the transition matrix of MERW only in the first iteration step. As illustrated in Fig. 1, MERW produces slightly better results, especially for considerable μ. The normalized mutual information is equal to 1 when the algorithm finds the same community structure as planted in the graph, and it is equal to 0 if the two partitions are statistically independent. (Details of the benchmark graphs parameters are described in Sect. 4.2.)

Next, Pons and Latapy [20] introduced an algorithm using the quantity given in (7) as a distance matrix between nodes of the graph. Their algorithm is an agglomerative one: it starts with each node being a community, and then, based on the distance matrix, it merges the two closest adjacent communities. The condition of maximal modularity chooses the partition from the resulting dendrogram. We refer to the original paper for details.

Figure 2 shows the performance of the algorithm. For large networks, the algorithm is very good independently of the random walk chosen. For small networks, MERW considerably decreases the efficiency of the algorithm for small μ; the precise reasons for that are not established. The general tendency of the algorithms to

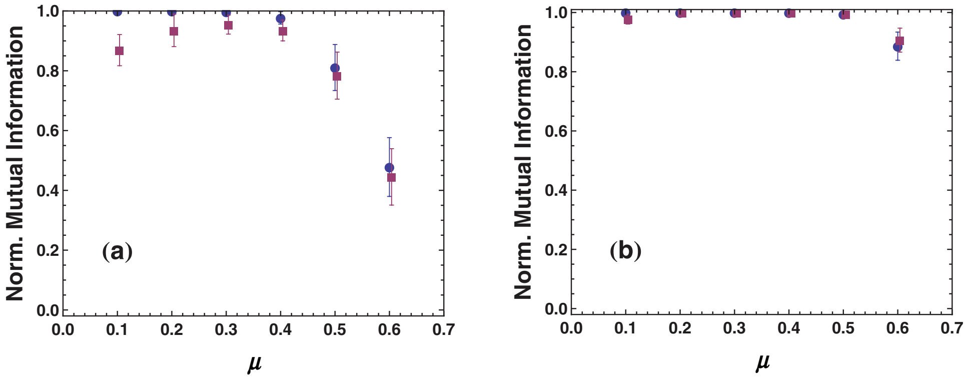

Fig. 2. Comparison of community detection efficiency between MERW (squares; summed powers of P^t, $t = 1 - 3$) and GRW (circles; $t = 3$) for the algorithm of Pons and Latapy [20]. Graph size:(a) $N = 200$, (b) $N = 1000$. It is the best among the algorithms discussed. MERW slightly decreases its performance for small μ.

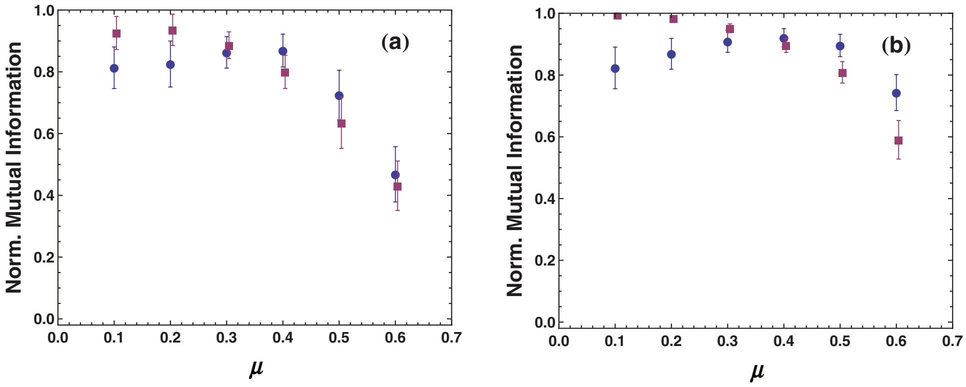

Fig. 3. Comparison between λ_0^t (MERW, squares) and $t!$ (circles) path weights. Graph size:(a) $N = 200$, (b) $N = 1000$. MERW gives better performance for smaller μ, while factorial weighting for larger. The overall performance is satisfactory for a method based on agglomerative clustering.

perform worse for smaller networks is probably due to small average node degree, which may result in single nodes detaching easier from their communities.

In (10), the weights $e^{-\mu t}$ produce the resolvent operator of A, but also factorial weights $\beta^t/t!$ might be introduced [33,34], yielding the heat kernel. To analyse the resulting matrix one needs to remove the zeroth eigenmode of A, so that G is well-defined. The choice $e^{\mu} = \lambda_0$ is directly related to MERW.

The procedure [33,34] goes on, producing a matrix with 0s and 1s in place of negative and positive entries of G. The original idea involved finding all maximal cliques (maximal complete subgraphs) of the graph represented by this matrix. Since this is computationally strenuous, we use a much simpler approach and carry out hierarchical clustering on that matrix. To obtain communities, we take the dendrogram section which maximises the modularity [35]. This algorithm, however, should be considered as only a very rough approach, just for the sake of preliminary comparison. It can be seen in Fig. 3 that exponential weights works better for small μ, while factorial weights give a reasonable performance for larger values of mixing parameter.

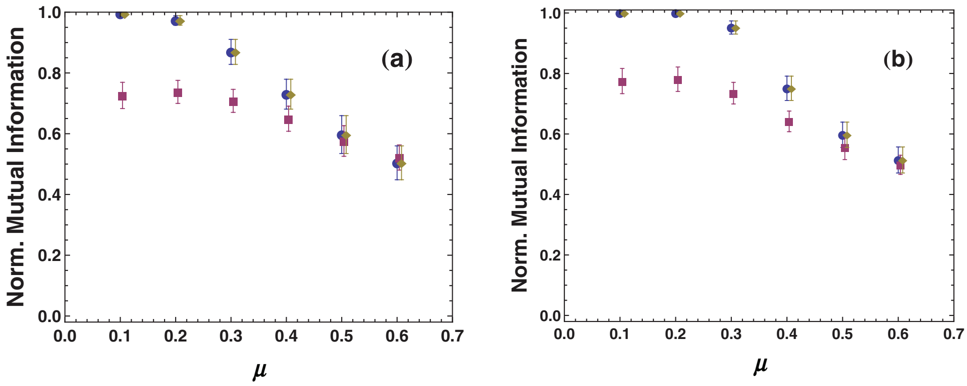

Fig. 4. Comparison between *Netwalk* [23] using MERW (squares), GRW (circles) and biased RW (diamonds)on benchmark graphs. Graph size:(a) $N = 200$, (b) $N = 1000$. The algorithm becomes unreliable for relatively small μ. MERW considerably reduces its performance for the whole parameter range.

Lastly, one may look at the methods grouping the nodes according to their MFPT values. In [21,23], a similarity matrix is introduced that computes the total of differences between MFPTs of random walkers incoming to particular nodes a and b from any initial node

$$\Lambda_{ab} = \frac{\sqrt{\sum_{c\neq a,b}^{N} |M_{ac} - M_{bc}|^2}}{N-2}. \tag{13}$$

On this basis, the authors developed an algorithm called *Netwalk*. We skip the details of the algorithm and refer the reader to the original papers. In this case, the outcome of the comparison between MFPTs of different random walks (we also implement a biased random walk used originally by *Netwalk*), in Fig. 4, shows that MERW should not be used in this algorithm. The original algorithm, however, works well only for very small μ, and in general its performance is unexpectedly unreliable even for large network size.

4.2 Benchmark graphs

The algorithms in Sect. 4.1 are compared to the use of unweighted undirected benchmark graphs introduced in [36] by Lancichinetti, Fortunato, and Radicchi (LFR) in a manner analogous to the authors' later work [37]. These graphs were designed specifically to benchmark community detection methods, and they are characterized with a preset power-law distribution of node degrees, and more importantly, also with a power-law distribution of community sizes. They are constructed based on the *planted partition model*, in which two nodes that are *a priori* assigned to the same community are linked with probability p_{in}, and with probability p_{out} if they are assigned to different communities. This means that each community is a random subgraph. The LFR benchmark graphs are parametrized in a similar manner with the mixing parameter μ, which is the fraction of links that a given node shares with the nodes outside its community, and may be thought of as a fixed ratio p_{out}/p_{in}. The parameter μ is approximately the same for all nodes in a graph.

We take 100 benchmark graphs with $N = 200, 1000$ nodes; their exponents for the degree distribution and for the community size distribution are respectively $\tau_1 = -2$ and $\tau_2 = -1$. For $N = 200$ the parameters are: the average degree of 10, maximum degree of 30, and the minimum and maximum community sizes are taken to be 5 and 35. For $N = 1000$: the average degree of 20, maximum degree of 50, and the minimum and maximum community sizes are 20 and 100, respectively. The mixing parameter μ is set to $\mu = 0.1 - 0.6$. For the upper bound, most of the algorithms start to have severe problem with detecting communities.

To check how good partition has been found, we use the normalised mutual information (NMI) [38]. NMI treats node assignments to communities as probabilities. As a result, it measures the statistical independence of two assignments (probability distributions) yielding 1 if they are equivalent, and 0 if they are statistically independent. We always measure NMI of the partition obtained from a given algorithm with respect to the partition planted in the benchmark. Let us note that the definition of a community here relies on the planted partition model, which means that the performance of algorithms is checked in accordance with this particular definition.

5 Conclusions

We have briefly introduced the concept of maximal-entropy random walk and reviewed some of its features, while in the main body of this paper we compared the performance of several community finding algorithm, in which MERW-based (dis)similarity matrices substituted the original ones.

The results obtained by the most reliable method checked here, made by Latapy and Pons, are comparable for GRW and MERW, although we note a significant worsening for small networks when using the latter.

The other methods have not been previously compared on LFR benchmark graphs. The one by Harel and Koren is generally unreliable for $\mu > 0.4$. However, its performance is slightly improved by MERW for both small and large networks. By contrast, MERW does not suit for *Netwalk*. Even for GRW, which was used originally, this algorithm produces a markedly unsatisfactory results for the medium range of the mixing parameter in comparison with the available state-of-the-art methods. The method based on factorial path weighting has considerable problems for small μ. Surprisingly, switching to exponential weighting, which corresponds to MERW, produces better results than *Netwalk*. In general, it performs reasonably well, even though the algorithm used a simple hierarchical clustering as a temporary means for the sake of comparison.

Meanwhile, MERW exhibits a surprising localisation and relaxation properties on some defective regular graphs. This case study shows that on the LFR benchmark graphs, which are locally random, this random walk can offer a performance of community finding methods comparable to that of GRW. It remains to be investigated if the behaviour of MERW on other types of graphs, including real-world networks, is more distinctive. Further effort is also needed to determine whether the development of a dedicated algorithm which makes better use of the information contained in this type of random walk is possible.

Project operated within the Foundation for Polish Science International Ph.D. Projects Programme co-financed by the European Regional Development Fund covering, under the agreement no. MPD/2009/6, the Jagiellonian University International Ph.D. Studies in Physics of Complex Systems.

References

1. S. Fortunato, Phys. Rep. **486**, 75 (2010)
2. Z. Burda, J. Duda, J.M. Luck, B. Waclaw, Phys. Rev. Lett. **102**, 160602 (2009)
3. Z. Burda, J. Duda, J.M. Luck, B. Waclaw, Acta Phys. Pol. B **41**, 949 (2010)

4. B. Waclaw, *Generic Random Walk and Maximal Entropy Random Walk*, Wolfram Demonstration Project
5. J.K. Ochab, Z. Burda, Phys. Rev. E **85**, 021145 (2012)
6. J.K. Ochab, *Stationary States of Maximal Entropy Random Walk and Generic Random Walk on Cayley trees*, Wolfram Demonstration Project
7. J.K. Ochab, *Dynamics of Maximal Entropy Random Walk and Generic Random Walk on Cayley trees*, Wolfram Demonstration Project
8. J.K. Ochab, Acta Phys. Pol. B **43**, 1143 (2012)
9. J.H. Hetherington, Phys. Rev. A **30**, 2713 (1984)
10. L. Demetrius, V.M. Gundlach, G. Ochs, Theor. Popul. Biol. **65**, 211 (2004)
11. L. Demetrius, T. Manke, Phys. A **346**, 682 (2005)
12. V. Zlatic, A. Gabrielli, G. Caldarelli, Phys. Rev. E **82**, 066109 (2010)
13. J.-C. Delvenne, A.-S. Libert, Phys. Rev. E **83**, 046117 (2011)
14. R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, V. Latora, Phys. Rev. E **83**, 030103 (2011)
15. C. Monthus, T. Garel, J. Phys. A: Math. Theor. **44**, 085001 (2011)
16. K. Anand, G. Bianconi, S. Severini, Phys. Rev. E **83**, 036109 (2011)
17. W. Parry, Trans. Amer. Math. Soc. **112**, 55 (1964)
18. S. White, P. Smyth, *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, 2003* (ACM, New York, USA, 2003), pp. 266–275
19. D. Harel, Y. Koren, *FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science* (Springer-Verlag, London, UK, 2001), pp. 18–41
20. M. Latapy, P. Pons, Lect. Notes Comput. Sci. **3733**, 284 (2005)
21. H. Zhou, Phys. Rev. E **67**, 061901 (2003)
22. H. Zhou, Phys. Rev. E **67** (4), 041908 (2003)
23. H. Zhou, R. Lipowsky, Lect. Notes Comput. Sci. **3038**, 1062 (2004)
24. J.G. Kemeny, J.L. Snell, *Finite Markov Chains* (Springer Verlag, New York, 1976)
25. C.M. Grinstead, J.L. Snell, *Introduction to Probability* (American Mathematical Society, Providence, RI, 1997)
26. J. Ochab, arxiv: 1206.4094 (2012)
27. L. Donetti, M.A. Munoz, J. Stat. Mech. P10012 (2004)
28. K.A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003)
29. I. Simonsen, Phys. A **357**, 317 (2005)
30. J. Shi, J. Malik, IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888 (2000)
31. M. Meila, J. Shi, in: *AI and STATISTICS (AISTATS)* (2001)
32. A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Phys. A **352**, 669 (2005)
33. E. Estrada, N. Hatano, Phys. Rev. E **77**, 036111 (2008)
34. E. Estrada, N. Hatano, Appl. Math. Comput. **214**, 500 (2009)
35. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
36. A. Lancichinetti, S. Fortunato, F. Radicchi, Phys. Rev. E **78**, 046110 (2008)
37. A. Lancichinetti, S. Fortunato, Phys. Rev. E **80**, 056117 (2009)
38. L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, J. Stat. Mech.: Theory Exp., P09008 (2005)

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Wydział
Fizyki
Astronomii
i Informatyki
Stosowanej

dr hab. P. F. Góra
Wydział Fizyki, Astronomii i Informatyki Stosowanej UJ
30-059 Kraków, Reymonta 4
Tel. 12 663 55 66
e-mail pawel.gora@uj.edu.pl

Kraków, 5 lipca 2013

Oświadczenie o współautorstwie

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)

Mój wkład w powstanie niniejszej pracy polegał na zaproponowaniu wstępnej koncepcji badań, a następnie na krytycznym śledzeniu i dyskutowaniu kolejnych wyników, sugerowaniu ewentualnych poprawek i kolejnych rzeczy, które warto sprawdzić. Uczestniczyłem także w przygotowywaniu manuskryptu i w przygotowywaniu odpowiedzi dla recenzentów. Mój wkład w tę pracę oceniam na 15%.

Instytut Fizyki
im. Mariana Smoluchowskiego
UNIWERSYTET JAGIELLOŃSKI
Reymonta 4, 30-059 Kraków
POLAND

Prof. Zdzisław Burda

tel. +48 12 663 57 24
fax. +48 12 633 40 79
zdzislaw.burda@uj.edu.pl

Kraków, 4 lipca 2013

Oświadczenie

Poniższe oświadczenie składam w związku z przewodem doktorskim pana mgra Jeremiego Ochaba.

Pan Ochab jest głównym autorem artykułów:

- 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)
- J.K. Ochab, Z. Burda, "Maximal entropy random walk in community finding", Eur. Phys. J-Spec. Top. 216, 73-81 (2013)

W trakcie pracy nad nimi pan Ochab sformułował problem badawczy, przeprowadził większość obliczeń i napisał tekst. Mój udział polegał na bieżącej dyskusji oraz na sprawdzaniu wzorów, tekstu i wniosków. Szacuję go na 20%.

Zdzisław Burda

Zdzisław Burda