UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Knowledge acquisition and reconstruction in complex networks

Lucas Guerreiro

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Lucas Guerreiro

Knowledge acquisition and reconstruction in complex networks

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
June 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G934k

Guerreiro, Lucas
Knowledge acquisition and reconstruction in complex networks / Lucas Guerreiro; orientador Diego Raphael Amancio. -- São Carlos, 2023.
106 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Network Science. 2. Knowledge Acquisition. 3. Sequences. 4. Network Topology. 5. Random Walks. I. Amancio, Diego Raphael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucas Guerreiro

Descoberta do conhecimento e reconstrução em redes complexas

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Junho de 2023

To my parents.

ACKNOWLEDGEMENTS

First, I want to thank God for the opportunity and strength given to pursue my dreams.

I would like to thank my parents, Itamar and Cecilia, that even though did not have the opportunity to study, they have always done everything they could for me. They are examples of simplicity and kindness.

I would like to thank my wife, Luana, for sharing a life of dedication, understanding, and love together.

I would like to thank my advisor, Prof. Diego Amancio, for all the support, friendship, patience, and knowledge shared with me. In the same way, I want to thank Filipi Nascimento for the partnership and guidance during my studies. I also want to express gratitude to all my professors, since the very beginning until this point, and specially for my professors at UNESP and USP for sharing their passion for doing science.

I want to thank Pecege, in the names of Prof. Daniel and Prof. Pedro, for the support and encouragement to study and believe in our projects. I want also to thank everybody in Skylar for being supportive and good friends.

Finally, I want to thank all my friends, family, and colleagues that have been a part of my journey up to this point.

“Not in knowledge is happiness,
but in the acquisition of knowledge.”
(Edgar Allan Poe)

RESUMO

GUERREIRO, L. **Descoberta do conhecimento e reconstrução em redes complexas**. 2023. 106 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Redes complexas vêm sendo empregadas nas mais diversas aplicações há algumas décadas. Sistemas complexos podem ser vistos em aplicações como transportes, redes de energia, internet, biologia e logística, dentre outras possíveis implementações. Em tais estruturas é possível que existam agentes percorrendo os nós e identificando novos conceitos e descobrindo a rede; este tipo de exploração é conhecido como descoberta do conhecimento e vem sendo pesquisado profundamente nas últimas décadas. Quando explorando uma rede, ou seja, descobrindo conhecimento nela, o caminho percorrido pode ser visto como uma sequência de nós visitados. Esta tese foca no estudo da relação entre topologias, dinâmicas e sequências em redes complexas. Com isso, no desenvolvimento desta tese pudemos observar o comportamento de diferentes dinâmicas em diferentes topologias quando adquirindo conhecimento. Além disso, propusemos um framework que com o auxílio de técnicas de aprendizado de máquina demonstrou a possibilidade de se recuperar qual a estrutura geradora da sequência sem conhecê-la. Por fim, avaliamos como as propriedades globais de uma rede são refletidas em estruturas geradas por sequências, ou seja, apresentamos uma análise se informações locais estão enviesadas ou se, de fato, podem representar uma visão real da rede como um todo; esta análise permitiu ainda identificar o impacto do tamanho das sequências na identificação das propriedades da rede. Com isso, os resultados apresentados nesta tese demonstraram o comportamento de diferentes estruturas no processo de descoberta do conhecimento. Destacamos ainda, a construção de um framework para classificação da topologia da rede e dinâmica utilizadas na geração de sequências. Tais resultados permitem a viabilização de diversas aplicações em ciência das redes, além de fundamentar conhecimentos para a área. Dentre os principais resultados atingidos, este trabalho permitiu identificar estruturas geradoras de sequências a partir de propriedades obtidas durante a reconstrução destas sequências como uma rede complexa e, ainda, foi possível observar que sequências pequenas permitem a identificação das estruturas com alta acurácia.

Palavras-chave: Ciência das Redes, Descoberta do Conhecimento, Sequências, Topologia de Redes, Caminhadas Aleatórias.

ABSTRACT

GUERREIRO, L. **Knowledge acquisition and reconstruction in complex networks**. 2023. 106 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Complex networks have been used in several applications in the past few decades. Complex systems can be observed in applications such as transportation, energy systems, internet, biology, and logistics, among other possible *implementations*. In such structures, there may be agents exploring nodes and identifying new concepts and discovering the network; this kind of exploration is known as Knowledge Acquisition and it has been deeply researched for the past decades. When exploring a network, i.e., acquiring knowledge in it, the path explored by an agent can be seen as a sequence of visited nodes. In this context, this thesis allowed us to observe the behavior of different network dynamics and topologies when acquiring knowledge. Moreover, using machine learning techniques, we have proposed a framework that showed it is possible to recover the generating structures that constructed an unknown sequence. Finally, we have evaluated how global properties of a network are reflected in structures generated by sequences. Thus, we have presented an analysis whether local information of a network are biased or it is indeed a real picture of the network as a whole; this analysis allowed us to measure the impact of the sequences size while identifying networks' properties. Hence, the results presented in this thesis have shown the behavior of different structures during the knowledge acquisition process. Lastly, we can highlight the framework built in this work, which allowed to classify which are the original topology and dynamics that generated a given sequence. Such results may enable several applications in network science, and pave the knowledge in this area. Among the main results, this work has allowed the proper identification of sequences' generating structures from the properties obtained during the reconstruction of such sequences as a complex network; and, moreover, it was possible to observe that small sequences allow the identification of the structures with high accuracy.

Keywords: Network Science, Knowledge Acquisition, Sequences, Network Topology, Random Walks.

LIST OF FIGURES

Figure 1 – Example of hypothetical undirected and directed networks 102

LIST OF ABBREVIATIONS AND ACRONYMS

BA	Barabási-Albert Model
ER	Erdös-Rényi Model
RW	Random walk
RWD	Random walk degree biased
RWID	Random walk biased by the inverse of the degree
SAW	Self-avoiding walk
TSAW	True self-avoiding walk
WAX	Waxman Model
WS	Watts-Strogatz Model

LIST OF SYMBOLS

G — Graph

V — Set of nodes

E — Set of edges

A — Adjacency matrix

a_{ij} — Position in A which refers to the adjacency between node i and node j

w_{ij} — Weight in the link that connects node i and node j

k — Network average degree

k_i — Degree of node i

k_{in} — In-degree

k_{out} — Out-degree

$d(u,v)$ — Distance between nodes u and v

u — node u

v — node v

CONTENTS

1 INTRODUCTION 23
1.1 Objectives 26
1.2 Thesis Organization 26

2 A COMPARATIVE ANALYSIS OF KNOWLEDGE ACQUISITION PERFORMANCE IN COMPLEX NETWORKS 29
2.1 Context 29
2.2 Contributions 30

3 RECOVERING NETWORK TOPOLOGY AND DYNAMICS VIA SEQUENCE CHARACTERIZATION 43
3.1 Context 43
3.2 Contributions 44

4 IDENTIFYING THE PERCEIVED LOCAL PROPERTIES OF NETWORKS RECONSTRUCTED FROM BIASED RANDOM WALKS . 69
4.1 Context 69
4.2 Contributions 70

5 CONCLUSION 95

BIBLIOGRAPHY 97

APPENDIX A COMPLEX NETWORKS AND KNOWLEDGE ACQUISITION 101
A.1 Concepts and Properties 101
A.2 Network Models 103
A.2.1 Erdös-Rényi model *103*
A.2.2 Watts-Strogatz model *103*
A.2.3 Barabási-Albert model *104*
A.2.4 Waxman model *104*
A.3 Network Dynamics 104
A.4 Knowledge Acquisition 106

CHAPTER

1

INTRODUCTION

The number of studies on complex networks has been increasing significantly over the past few decades. At the same time, new tools and computational capabilities are now available to researchers on many scientific fields. This has led to new methods for analysis in large data sets and allowed breakthroughs on the understanding and representation of many complex systems (NEWMAN, 2010). Furthermore, complex networks are becoming common when describing many real phenomena, such as the internet (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999; CHEN *et al.*, 2002), social networks (SCOTT, 2000; WASSERMAN; FAUST, 1994), airline routes (AMARAL *et al.*, 2000), biological systems (JEONG *et al.*, 2000; PODANI *et al.*, 2001; ITO *et al.*, 2001), economical trades (HOSSU *et al.*, 2009), among many others. This increase of interest on the field has brought attention to researches in the area, which are carrying breakthroughs in the theory and on these complex systems characteristics. Thus, the structures that compose complex networks have been studied due to the impact that optimizations and new perspectives on network science. Those advances can lead to relevant improvements on many interdisciplinary fields related to complex networks (NEWMAN, 2003).

The structure - or topology - of a network can be regarded as a representation of the space where every connection between elements in the modelled system can occur. The dynamics represent the way a certain agent walks over this structure (BOYER; SOLIS-SALAS, 2014; COSTA, 2006). For example, in airline routes (AMARAL *et al.*, 2000), the structure is composed by all the airports (i.e. nodes) as well as the routes (i.e. edges) between them. A possible dynamics in this system could be the set of rules that drive individuals travelling on this flights' network.

Moreover, real-life events and actions are usually taken into an unknown structure. For instance, a book is a result of words that can be seen as nodes, and consecutive words would have an edge linking them in a complex network structure. Therefore, the book has a *sequence* of symbols taken from this structure by a selected walking strategy over it; this walking dynamics is nothing less than the written work of the book author. Thus, when writing a book, the original complex network is unknown except for the vocabulary. Therefore, there is a relevant relationship

among these structures, and understanding how such roles are played can improve their use on real-life problems. This thesis focus on the analysis of the relationship between *topologies*, *dynamics*, and *sequences*, by exploring how these structures are interconnected and related.

As aforementioned, in real-life situations the original topology may be unknown, however they can be fit to some already studied model topology, such as Erdös-Rényi (ERDÖS; RÉNYI, 1959), Watts-Strogatz (WATTS; STROGATZ, 1998), or Barabási-Albert (BARABÁSI; ALBERT, 1999), for example. Similarly, the dynamics taken by an unknown agent or walking strategy can be approximated to an existing dynamics such as the traditional random walk (LOVÁSZ, 1996) or the true self-avoiding walk (KIM; PARK; YOOK, 2016; AMIT; PARISI; PELITI, 1983). Hence, we can approximate real-life events to a complex network composed by certain topology and dynamics. Thus, as opposite to the common-sense research and applications in complex networks, in which we may have some structure and dynamics and we find a sequence or series from the model, we may now, specially in real-life examples, have a sequence that is implicitly attached to a structure and dynamics that might be unknown. In this context, we will investigate the generating topology and dynamics of sequences, as those might be unknown in real-world problems.

There are some studies linking how at least two of our three concerning concepts (topology, dynamics, and sequences) relate. Arruda *et al.* (2017) have explored how knowledge acquisition operate on different settings, and they have also incorporated modifications to dynamics. From the structure point of view, this work demonstrates how variations applied to the network structure and dynamics may improve the set of sequences outcomes. Moreover, we can identify how these different sequences obtained in the exploration process have changed through new sets of parameters and models on the network basic structure.

Furthermore, Arruda *et al.* (2019) discuss the representation of visited nodes as a time series. In their work, the generated sequences - which are outcomes from network exploration by random walkers - were compressed and recovered, showing advances on the reconstruction of networks and gaining understanding on how generators and sequences work.

In terms of real-life applications based on the theory discussed by Arruda *et al.* (2019), the actions performed by random agents may be represented by complex networks. These events or sequences may be performed in a given topological structure, such as cars driving in a pre-determined city route or can be generated from an unknown source, such as a book, which is defined from a vocabulary but with no pre-determined topology. In either situation, the original topology and dynamics can be approximated to a network topology or even can be fit to an already known model. Therefore, these sequences or time series of occurrences will have a inherent structure and dynamics. To be more specific, whenever one person is walking to a given destination (which may even be unknown) we perform some kind of dynamics, and we have a structure (or network) that is the mapping of possible paths we may take; the path that we did take comprise a sequence of steps from an initial node to the target node.

As previously discussed, we can have other examples of dynamics performed on networks, such as knowledge discovery (ARRUDA *et al.*, 2017; COSTA, 2006). In this domain, we have a structure composed by its concept points, and when an agent is discovering information it performs a walk between the knowledge topics, producing the so-called knowledge discovery process.

In a purely scientific environment, we often describe a network and a dynamics and perform a task using them, such as an exploration to find nodes or, as stated previously, acquire knowledge. Hence, we perform a walk on a topology, and it will generate a sequence or time series of symbols. However, in real - or even artificial - situations, we may already have a sequence (e.g. a representation of steps in terms of a string of symbols) and finding the original network and the dynamics that originated such sequence would be interesting on the understanding of the given domain.

Some authors have tackled the problem of understanding the structure of topologies and dynamics, and their isolated relation to the sequences generated by walks on the networks (BOCCALETTI *et al.*, 2006). However, there are few studies on the recovery of the structures and dynamics based on the generated sequences, which could entail the understanding on how the network and dynamics behave; moreover, researches on this field have not given the necessary attention on the process of reconstructing the structures and dynamics by only the given sequence.

Going beyond on the benefits that such comprehension may bring, we also see the embeddings field as an interesting candidate to take advantage on our proposed work. In this context, several embedding techniques use a resulting random walk sequence as an input, which can be improved by our study. We can cite the *node2vec* algorihtm (GROVER; LESKOVEC, 2016), which learns representation for nodes in weighted or not weighted graphs, and also in directed or undirected networks; such algorithm could use our framework in order to improve their representation accuracy, which could also be exploited to other embedding algorithms such as *LINE* (TANG *et al.*, 2015) and *DeepWalk* (PEROZZI; AL-RFOU; SKIENA, 2014). Moreover, this understanding may also improve community detection algorithms such as *Infomap* (ROSVALL; BERGSTROM, 2008), by using the underlying comprehension of community structures, or based on reconstructed time-series.

Furthermore, the ideas presented on this thesis are not limited to a single field and can be explored in currently trending topics such as generative artificial intelligence and large language models (VASWANI *et al.*, 2017; DEVLIN *et al.*, 2019). In this context, generative models may predict the next word of a text (i.e., an ordered sequence of words) based on the probabilistic hypothesis given the context; therefore, there is a high correlation between these topics and knowledge acquisition, whereas the words can be modelled as nodes and the next word in the text is the most relevant node based on the concerning property.

1.1 Objectives

The general objectives of this thesis comprise analysis of knowledge acquisition behavior in complex networks. In this work, we define knowledge acquisition as the process of discovering new concepts, i.e. nodes and edges, in a complex network that represents a specific area. Thus, we investigate the behavior of random walkers when acquiring knowledge and analyze how the properties of reconstructed networks represent the original network - previous to the knowledge acquisition process.

Therefore, this thesis approaches a range of objectives that can be summarized as three goals as follows:

- **comprehend the efficiency of random walkers on the knowledge acquisition process**: we aim on comparing how efficient different random walkers are while acquiring knowledge. In this sense, we proposed a systematic comparison of several dynamics over different topologies, in order to infer the impact of the structures during the knowledge acquisition process.
- **develop a framework to recover topologies and dynamics from a sequence**: in the network science field, the detection of the originating topology and dynamics of a sequence is an interesting problem due to the possibility to understand unknown structures. Identifying the structure of a sequence may lead to further comprehension of the resulting time series obtained in the knowledge discovery process. Thus, we conduct a series of experiments that brought together a framework that identifies the original structures with high confidence using partial sequences.
- **identify the sequence size necessary to estimate the original complex network properties**: during the knowledge acquisition process, we can generate a sub-network from the original structure. In this context, we have conducted experiments to identify how different sequence sizes represent the original network properties, i.e. how big a reconstructed network should be in order to portrait relevant characteristics of the original network. These experiments can also shed light on the study of the bias created during networks' exploration.

In order to achieve such objectives, this work is concentrated in three articles which will be presented entirely in this thesis.

1.2 Thesis Organization

This thesis is organized in a structure of a collection of articles produced during the studies to achieve the goals of this work. Each of the three main chapters is composed by two

introductory sections that present the Context of the article and the Contributions of the paper. The first article has already been published, while the remaining two are still in the reviewing and production steps on the journals they were submitted. Therefore, as per the analysis of the referees these papers may be further reviewed, but they are already presented in this thesis due to time restrictions and are already considered a final version of the conducted studies. An important aspect of this thesis is that the three articles are self-contained, thus the methodologies and fundamentals are presented in the papers and can be understood during the reading of the articles. In order to present clarifications on possible remaining questions, in the end of the thesis we are presenting an appendix containing the basis content of complex networks necessary to the full comprehension of the concepts that were used.

The remaining of this thesis is organized as follows. In Chapter 2, we explore an analysis on how knowledge is acquired for different topologies by different dynamics. In Chapter 3, we present our framework to recover network topologies and dynamics from sequences. In Chapter 4, we show our work on the identification of properties for partial sequences and biased random walkers. Finally, in Chapter 5, we present a general conclusion for this thesis.

CHAPTER

2

A COMPARATIVE ANALYSIS OF KNOWLEDGE ACQUISITION PERFORMANCE IN COMPLEX NETWORKS

A comparative analysis of knowledge acquisition performance in complex networks. Lucas Guerreiro, Filipi N. Silva, Diego R. Amancio. *Information Sciences 555*, May 2021, Pages 46-57.

2.1 Context

The theory and applications of exploring the knowledge acquisition process as a complex network problem has been studied thoroughly in the past decades (BARAT; CHAKRABARTI, 1995; MEERSCHAERT; SCALAS, 2006; COMIN *et al.*, 2020). In these processes, nodes may represent a concept and edges indicate interlinks between such concepts. The rules to determine how these edges and nodes are chosen are defined by how agents explore the structures.

The implication of random walkers in complex networks have been studied previously (LIMA *et al.*, 2018a; ARRUDA *et al.*, 2017; HERRERO, 2019), however there were no studies that approached the knowledge acquisition problem systematically. In Arruda *et al.* (2019), the authors have combined knowledge acquisition and information theory in order to understand how one may comprehend the compression task for such networks represented as sequences. The study has compared different topologies and dynamics - similarly to the study presented in this Chapter - but aiming on understanding how the compression impacts the network reconstruction. Interestingly, the task was successfully achieved and the results have shown that knitted networks performed better in such task.

Our study also connects knowledge acquisition and network science, by exploring systematically which network topologies and dynamics performs better in the knowledge acquisition

task. In this context, we have analyzed how many nodes are acquired for a distributed set of walk lengths. Therefore, our goal and motivation was to identify the behavior on different sets of topologies and dynamics in order to understand how such structures relate to each other. Thus, in this study we explore how efficient each random walker might be. Although one could just choose a brute force strategy (or bread-first search) in order to acquire knowledge in the structure, this would not be cost-effective due to the complexity of such task. Therefore, our study presents how to acquire knowledge efficiently and we also simulate the dynamics of real-life situations through theoretical walking strategies.

2.2 Contributions

The systematic analysis proposed in our paper has led to a few insights. First, we were able to identify the true self-avoiding walk as the best overall performer when acquiring knowledge. Second, our approach has demonstrated how the learning curves grow for different settings which has led to understanding the behavior of knowledge acquisition for the analyzed settings. Lastly, our proposed framework applied for a single property (node discovery) paved the way to our next studies that have analyzed other network properties in different contexts.

Information Sciences 555 (2021) 46–57

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A comparative analysis of knowledge acquisition performance in complex networks

Lucas Guerreiro [a], Filipi N. Silva [b], Diego R. Amancio [a,*]

[a] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil
[b] Indiana University Network Science Institute, Bloomington, Indiana 47408, USA

ARTICLE INFO

Article history:
Received 23 July 2020
Received in revised form 20 October 2020
Accepted 23 December 2020
Available online 29 December 2020

Keywords:
Knowledge acquistion
Network search
Network dynamics
Complex networks
Random walks

ABSTRACT

Discovery processes have been an important topic in the network science field. The exploration of nodes can be understood as the knowledge acquisition process taking place in the network, where nodes represent concepts and edges are the semantical relationships between concepts. While some studies have analyzed the performance of the knowledge acquisition process in particular network topologies, here we performed a systematic performance analysis in well-known dynamics and topologies. Several interesting results have been found. Overall, all learning curves displayed the same learning shape, with different speed rates. We also found ambiguities in the feature space describing the learning curves, meaning that the same knowledge acquisition curve can be generated in different combinations of network topology and dynamics. A surprising example of such patterns are the learning curves obtained from random and Waxman networks: despite the very distinct characteristics in terms of global structure, several curves from different models turned out to be similar. All in all, our results suggest that different learning strategies can lead to the same learning performance. From the network reconstruction point of view, however, this means that learning curves of observed sequences should be combined with other sequence features if one aims at inferring network topology from observed sequences.

1. Introduction

Many real-world systems can be naturally represented by sequences corresponding to chains of events or transitions between states, including human actions [10], machine workflow [43], scientists mobility [23] and language [15]. Communication can also be accomplished by encoding and decoding data into sequences of symbols or continuous signals. Indeed, a significant portion of datasets derived from real-world systems is available in this form. For a complex system, one can understand that sequences can be generated by a process driving the changes among states across a certain space of allowed transitions [6].

Network science has been employed to represent a great variety of complex systems [18,19,3,21,32,8]. In recent studies, complex networks have displayed the potential to represent the space of transitions between states for many types of systems [15,28,29]. In this context, the driving processes generating sequences are represented by stochastic walks of a variety of heuristics. An example of this case is the knowledge acquisition process [7], in which nodes represent knowledge that is

* Corresponding author.
E-mail address: diego@icmc.usp.br (D.R. Amancio).

https://doi.org/10.1016/j.ins.2020.12.060
0020-0255/

connected according to how related they are. One or multiple agents (such as researchers) navigate in this knowledge space, which is unknown from the start, and discoveries are made when the agents visit new nodes. In such a system, sequences are derived by the paths taken by the agents.

While Markov chains [39] are a simple way to model and recover the inherent network of transition probabilities, it relies on considering that the studied phenomenon is driven by a simple stochastic process with no *a priori knowledge of its space. Many real-systems, however, may present more intricate driving stochastic dynamics (which may depend on long term memory or properties of the nodes, for instance). An example of that system is urban transportation, where agents navigate across a system of roads with possibly predefined origin and destinations. The paths taken by connecting these endpoints cannot be driven solely based on local probabilities. Also, the inherent space of state transitions can display a variety of different topologies [18] in contrast to more well-defined structures, such as regular graphs, as a consequence, even simple stochastic dynamics can lead to intricate sequences [6].*

In many real-world problems, only the sequences generated by the system are observed. Thus, having a way to discriminate characteristics that are either consequence of the dynamics or from the network can lead to a better understanding of the studied phenomenon. A simple property derived from sequences that can be differently impacted by both of these aspects is the rate of appearance of new symbols. This corresponds to the exploration coverage of a network under the action of a walking dynamics, which is also related to the learning curve in a knowledge acquisition process. This property is also related to how well an agent performs in discovering knowledge.

To our knowledge, no previous study focused on a systematic analysis among the dynamics, networks, and the sequences generated by them. Here we analyzed the learning curves for sequences obtained from four random walk dynamics and four network models with different topological structures. At first, we are interested in knowing if the learning curves are already good criteria for determining both the model and the dynamics used to generate a sequence.

Our analysis revealed that, among the considered stochastic walk dynamics using only local network information, the *true self-avoiding dynamics* (TSAW) was found to present the best performance in coverage rate for the considered network models. In addition to that, different patterns for the performances of coverage rate were observed. Aside from TSAW, the ranking based on performance of exploration for different sets of walk dynamics tends to depend on the network structure. For instance, when the stochastic walk is biased according to the node degree, better performance is attained when the network is sparse and the walks are biased towards preferring highly connected nodes. On the other hand, if the network is denser, better performance is reached when the walk avoids highly connected nodes. We also encountered situations in which there exists ambiguity in the coverage property for certain combinations of dynamics and network models. This indicates that it would be possible to swap the dynamics and the inherent structure and even so, attain similar learning curves. These developments could shed a light on the analysis of the mechanisms leading to text generation, for instance, to better understand how the vocabulary grows along with the text.

The following section explores the related literature to the problem of modeling real-world phenomena in terms of networks, dynamics, and sequences. Next, the methodology is presented alongside the description of the considered network models and dynamics. Results are presented together with discussions, which is followed by conclusions.

2. Related works

Random walks (RW) have been studied in many networked applications [9,38,14,16]. In the early studies of the emergent network science field, the properties of RW was investigated in power-law distributed networks. In [1], the authors compared the efficiency of random and self-avoiding walks in transferring messages through the network. Hubs were found to play the role of centralizing and distributing information to other nodes. Most importantly, this finding revealed that the efficiency of discovering new nodes depends on the topology of the underlying network.

The process of network discovering has been approached by several recent studies [6,33,7,46,25]. In [7], the authors compared the learning speed of several dynamics for particular network topologies. Specifically, they analyzed how effective different dynamics are when discovering new nodes in the network. In addition to traditional random walks, this study considered also random walks with Lévy flights [45]. Thus, the agents were allowed to visit any node in the network in the next step with a certain probability. The authors found that more frequent jumps favors the discovery rate, specially in Barabási-Albert networks. In particular topologies, though, jumps were found not to be as effective. This is the case of geographic networks. Another interesting finding is that the discovery of new nodes occurs with different speed in different network regions. The core – as identified via accessibility (entropy diversity) [48,5] – tends to be covered faster than the network borders.

In [33] the authors studied the efficiency of agents walking over the network to learn the structure of the network. Differently from other works, the authors considered a model where knowledge discovered by different agents is integrated in a specific entity of the system. This system is referred to as *network brain*. This type of dynamics was intended to represent e.g. the knowledge acquisition when mapping communities of similar interests in the Web. The most surprising result arising from this study is the fact that the learning behavior, considering variations of the self-avoiding walk, has a very weak dependence on the considered dynamics and network topologies.

The problem of knowledge acquisition in networks has also been studied in the context of information theory applications [6]. In [6], distinct random walks are performed over different topologies. The sequence of visited nodes generates a

sequence of symbols, which is further analyzed in function of the observed compression ratio – computed via Huffman coding. Finally, such a sequence is used to reconstruct the original network, and the error is analyzed for distinct topologies and agent dynamics. Several interesting results have been found using the framework combining knowledge acquisition and information theory. Interestingly, the best performance in the framework constructed for representing the phenomena of compression (during transmission) and reconstruction of networks revealed that a simple knitted network model [17] yielded the best performance. This finding is compatible with the idea that language is optimized for transmission [12], since knitted networks are representations of co-occurrence language networks [13,44,47,37].

The study reported in [29] aimed at identifying key Physics concepts from students' representations of perceived similarity between distinct topics. The representation used in this work was a concept network, where nodes represent the concepts (in the sense of quantities, laws, models, or experiments), and edges represent similarities between these concepts, such as actions for determining a model or the realization of a experiment using some law [46]. The paper studies these concept networks using subgraph and communicability betweenness centrality. The most relevant concept networks were identified using an importance ranking coefficient, which is a normalized geometric mean of the considered centrality measurements. While this study does not relies on random walks to represent the acquired network, the concepts networks are used as examples of networks representing the knowledge acquired by students, according to unknown knowledge acquisition dynamics.

The study conducted in [25] analyzed the properties of self-avoiding walks (SAW) in clustered scale-free networks. The study investigated how the number of SAWs changes as the desired walk length increases. The main result of the paper shows that, for scale-free networks with same average degree, there are more SAWs in clustered networks when compared to unclustered networks. This result suggests that the modular organization in the same topological family of networks may impact the discovery process in the network. Differently from most of the works in the literature, here we analyze the knowledge acquisition problem in terms of a generalist point of view. We analyze whether different network topologies and dynamics can lead to the same behavior in the observed learning curves. In other works, we analyze the behavior of learning curves by comparing, *at the same time*, different configurations of network topology and agents dynamics.

3. Methodology

The main objective of this paper is to compare the efficiency of different walking strategy to discover new nodes in the network. In this knowledge acquisition approach, we consider that a single agent is walking through the network, and every new node discovered is regarded as a new piece of knowledge. This process of discovery of nodes is relevant in practice because it is useful in network search applications where topology is not known on a global scale [49,30]. Here we refer to knowledge acquisition process in a wider context. While a random walk on a network is conceptually similar to the process of reading and writing a text, the notion of knowledge acquisition goes beyond the cognitive process of learning. We refer to knowledge in a broader way, where any complex system represented by a network can be discovered via exploration. In our definition, a new knowledge is acquired when a new site (node) is visited. For this reason, in this work, we do not restrict our analysis to dynamics or topologies that are related to cognitive processes. We focus our study on the behavior of nodes exploration in any complex system. For this reason, we studied the dynamics performance for well-known random walks and network topologies. Most importantly, we analyze the behavior of "learning curves" for each pair topology/dynamics in order to analyze whether different combinations of topology and random walks can lead to the same learning curve (and vice versa).

The adopted methodology is illustrated in Fig. 1 and summarized in the following steps:

1. *Network topology*: we selected different network topologies. We have selected well-known network models reproducing the characteristics of real-world networks. A brief description of the adopted models is provided in Section 3.1.

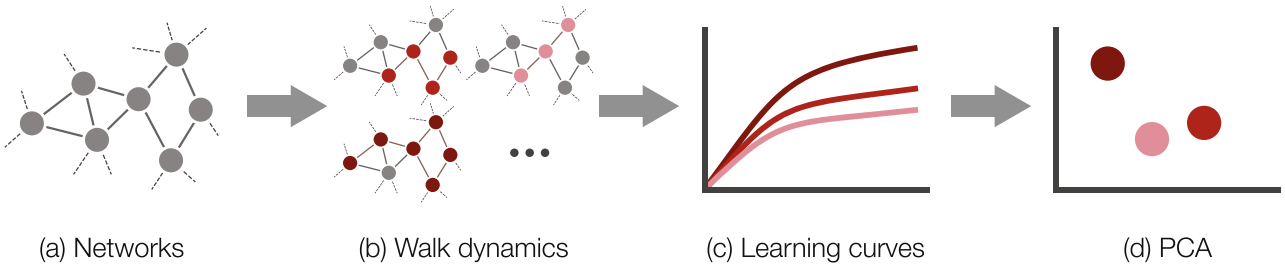


Fig. 1. Methodology employed to analysis the behavior of learning curves. In (a), we selected different network topologies. In (b), dynamics based on variations of random walks were considered to explore the networks. In (c), we obtain the learning curves describing how many nodes are discovered as the network is explored. Finally, in (d) each curve is mapped into a 2-dimensional space and similarities in the behavior of learning curves for different topologies and dynamics are analyzed.

2. *Network dynamics*: different ways to walk over the networks were considered, including dynamics based on traditional random walks and dynamics biased towards particular neighbor properties. A brief description of the adopted network dynamics is provided in Section 3.2.
3. *Learning curves*: For each pair of topology and dynamics, we obtain the learning curves. This learning curve describes how fast new nodes are discovered as the dynamics unfolds (see Section 3.3).
4. *Cluster analysis*: in this phase, each learning curves are mapped into a vector. This is used to measure the similarity between two curves. Similar curves are then identified via cluster analysis. This step is important to show that the behavior curve A brief description of this process is provided in Section 3.4.

3.1. Network topology

Artificial networks were built for each set of network models. The following parameters were used to create the networks: number of nodes $(N) = \{500, 1000, 5000\}$ and average degree $(\langle k \rangle) = \{4, 6, 8, 10\}$. We have worked with four well-known undirected network topology models:

- *Erdős-Rényi (ER)*: this model generates random networks. In this fashion, nodes have similar degrees. The probability of creating an edge is equally distributed among the nodes.
- *Barabási-Albert (BA)*: this topology implements the scale-free model, inherent to many real networks. BA networks are characterized by a few hubs with a very high degree, while most nodes have small degrees.
- *Waxman (Wax)*: this a traditional geographic model, which comprehends a set of nodes in a two dimensional space that incorporates new edges through an algorithm in which the probability decays exponentially as the distance between each pair of nodes grows. More specifically, the probability of two nodes to be linked is given by:

$$\pi_{ij} = a \exp(d_{ij}/\beta), \tag{1}$$

  where a is a normalization factor, d_{ij} is the geographic distance between nodes v_i and v_j and β is a parameter that defines the connectivity of the network.
- *Modular Networks (LFR)*: networks with community structure were implemented using the methodology described in [31]. In this model, each community is represented as a scale-free network. In addition to the number of nodes and average degree, additional parameters can be considered to generate the networks. The main parameters describing this model are the number of communities (n_C), the minus exponent for the degree sequence (t_1), the minus exponent for the community size distribution (t_2), the maximum degree $(\max_k)$, and the the mixing parameter (μ), which determines the fraction of edges linking distinct network communities. Here we used $n_C = 5, t_1 = 3, t_2 = 0, \mu = 0.20$. The maximum degree $\max_k$ were chosen so as to obtain networks with the desired average degree $\langle k \rangle$.

A visualization of the considered models for selected parameters is illustrated in Fig. 2. The visualizations were generated using the *Networks3d* software [42]. It is clear that for different models the nodes with highest degrees (orangish nodes) are distributed in different ways.

We also included two real-world networks for comparison: citation network of complex networks papers (CN) [42] and a Wikipedia network comprising only articles related to the subcategories of Physics and Mathematics [7].

3.2. Network dynamics

The network dynamics aims at generating a sequence of visited nodes, which represents the set of acquired knowledge. The dynamics observed by visiting sequentially network nodes has an analogy with the process of generating written texts [6]. If we consider that, at each step, a symbol is generated to represent that the current node has been visited, after N_S steps we have a sequence of symbols (i.e. a text) comprising N_S words (or tokens). The learning curve can thus be seen as the vocabulary size observed for a given text length. While in written texts the relationship between vocabulary size and text length is well described by the Heaps' Law [35], the learning curve observed in network discovery processes tends to follow a different pattern [7].

In order to recover the symbols from these models we have worked with the following walk dynamics: traditional random walk (RW) [34], random walk biased by degree (RWD) [11], random walked biased by the inverse of the degree (RWID) [11], and true self-avoiding walk (TSAW) [28,4]. These walks have been widely employed to study the dynamics of learning curves in the last few years [6,7,33]. The main differences among these walk dynamics are detailed below:

- *Traditional random walks*: the random walk dynamics is one of the most used in literature, and a very simple one. If the walker is at node v_i and Γ_i is the set of neighbors of v_i, all nodes in Γ_i have the same probability to be chosen as next node in the walk. In other words, the probability of transition from v_i to $v_j \in \Gamma_i$ is $p_{ij} = k_i^{-1}$.
- *Degree-biased random walk*: in this walking dynamics, a higher probability of transition p_{ij} is given to those neighbors with higher degrees. Mathematically, p_{ij} is proportional to the degree k_j of $v_j \in \Gamma_i$:

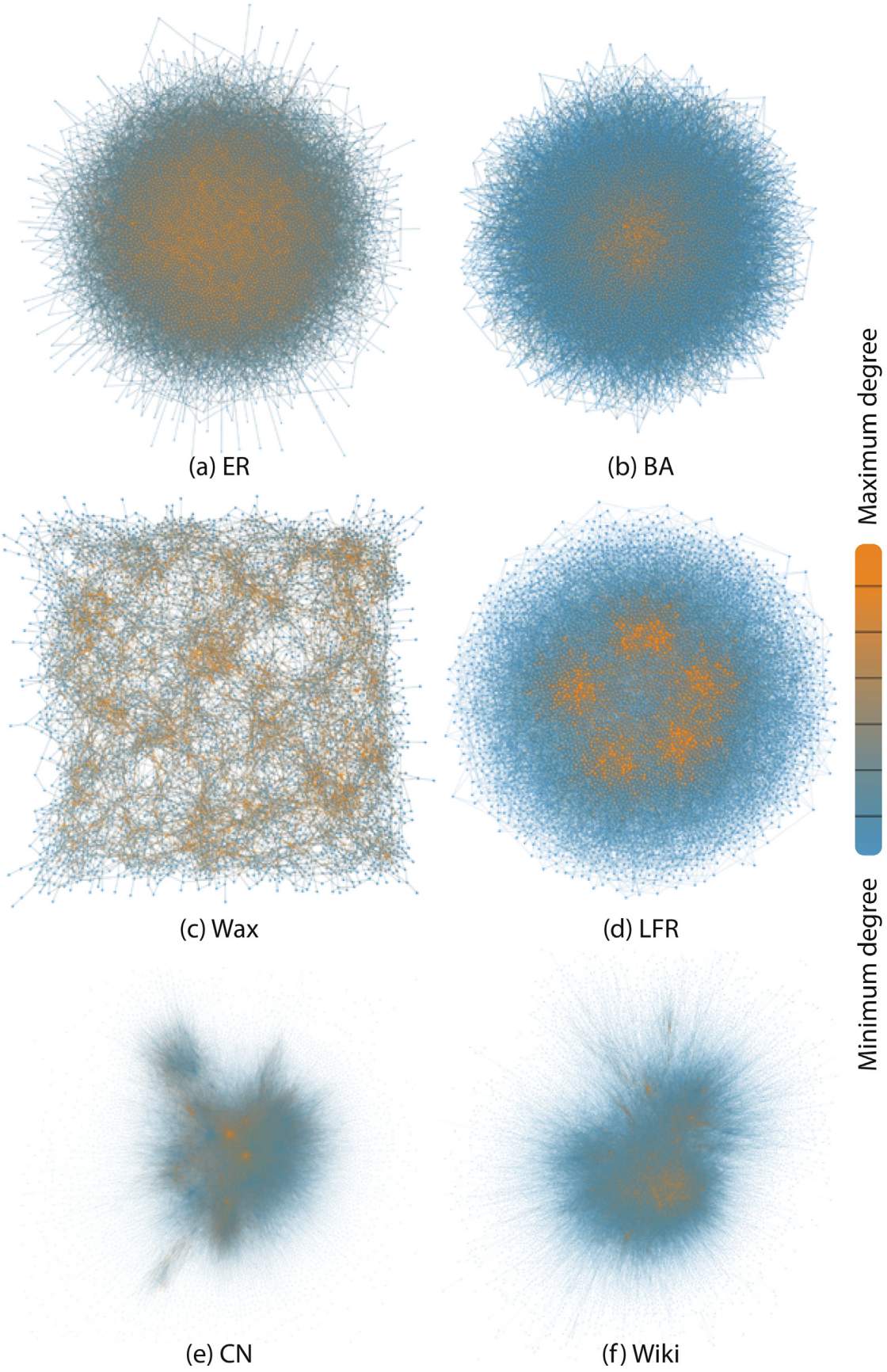


Fig. 2. Force-directed visualizations of the considered network models and real networks. Different colors correspond to node degree in respective to the maximum of each network. The visualizations were generated using the *Networks3d* software [42]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$p_{ij} = \frac{k_j}{\sum_{l\in\Gamma_i} k_l}. \tag{2}$$

In other words, the RWD dynamics always tries to explore the network by prioritizing visits to nodes with the highest number of neighbors.

- *Low degree-biased random walk*: a different variation of the traditional random walk is the walk biased towards the inverse of the degree. In this case, the probability of transition from v_i to $v_j \in \Gamma_i$ is:

$$p_{ij} = \frac{k_j^{-1}}{\sum_{l\in\Gamma_i} k_l^{-1}}. \tag{3}$$

Therefore, in this case, the walker tends to select nodes with low-degree in the next step of the random walk.

- *True self-avoiding walk*: in a true self-avoiding walk dynamics, already visited nodes are avoided. This is achieved this by memorizing edges that have already been visited. The transition probability is computed as

$$p_{ij} = \frac{e^{-\lambda f_{ij}}}{\sum_{l\in\Gamma_i} e^{-\lambda f_{il}}}, \tag{4}$$

where f_{ij} is the frequency of visits to the edge linking nodes v_i and v_j. The parameter $\lambda > 0$ corresponds to the exponential decay factor for which the probabilities decrease with the number visits. In this study, we use $\lambda = \ln 2$.
The main advantage of this dynamics is that it tends to present a higher learning rate when many nodes have already been visited. When the walker is visiting a region with no visited nodes, this random walk behaves similarly to the RW dynamics. In some studies, it has been shown that TSAWs tend to display the best learning performance [28]. Here we use TSAWs as a reference to probe how close is the performance of other dynamics (in different topologies) when compared to TSAW. Most importantly, we aim at verifying if there are other combinations of topology/dynamics yielding learning curves similar to the ones observed with TSAWs.

3.3. Learning curves

The measure used to characterize each dynamics is the so-called learning rate. This is an important property in network science and is related to many processes on complex networks, including knowledge acquisition, discovery processes, diffusion and spreading [20]. For each pair of network and random walk dynamics, we considered 5000 iterations (steps). Learning curves are then obtained as the fraction of the total number of *different* nodes visited after a given number of steps.

3.4. Principal component analysis

Principal Component Analysis (PCA) is a statistical method for dimensionality-reduction widely used to simplify large data sets described by a large number of features. This is achieved by reducing the number of variables from the original data set of D dimensions into a set of $d < D$ features. As a result, most of the information from the original data is kept because redundancy (measured via co-variance) is removed [40,26,27]. A data projection via PCA involves the following sequence: (i) data standardization; (ii) co-variance matrix computation; (iii) identification of the eigenvectors and their corresponding eigenvalues from the co-variance matrix; (iv) identification of the d principal components as the eigenvectors associated to the d largest eigenvalues; and (v) projection of the original data points into the principal components [24]. More specifically, the removal of correlations is accomplished via a n-dimensional rotation applied to the original feature spaces. A detailed proof of how such a rotation minimizes the correlations among the axes of the rotated space can be found in [24].

Here different learning curves are compared and similar learning curves is observed. To quantify the similarity between curves we represent each curve as n-dimensional vector, where the ith position of the vector represents the fraction of nodes visited after the ith step. Because such a representation of curves yields several strongly correlated features, we use PCA [24] to remove possible correlations. In fact, as we shall show, two dimensions of the PCA (i.e. $d = 2$) accounts for more than 95% of the data variation.

After the learning curves are represented in a two-dimensional space, clusters can be identified. Because our objective is to analyze whether similar learning curves can be obtained with different topology/dynamics choices, the identification of clusters was performed via visual inspection. However, a scenario with several instances could also be analyzed by using traditional clustering algorithms [41].

4. Results and discussion

Our analyses take into account the exploration coverage over time for agents discovering knowledge in network models as they explore nodes through edges. The first step is obtaining the learning curves for the considered pairs of dynamics (RW, TSAW, RWD, and RWID) and network models (ER, BA, Wax, and LFR models). For each network model setup, we generated 5

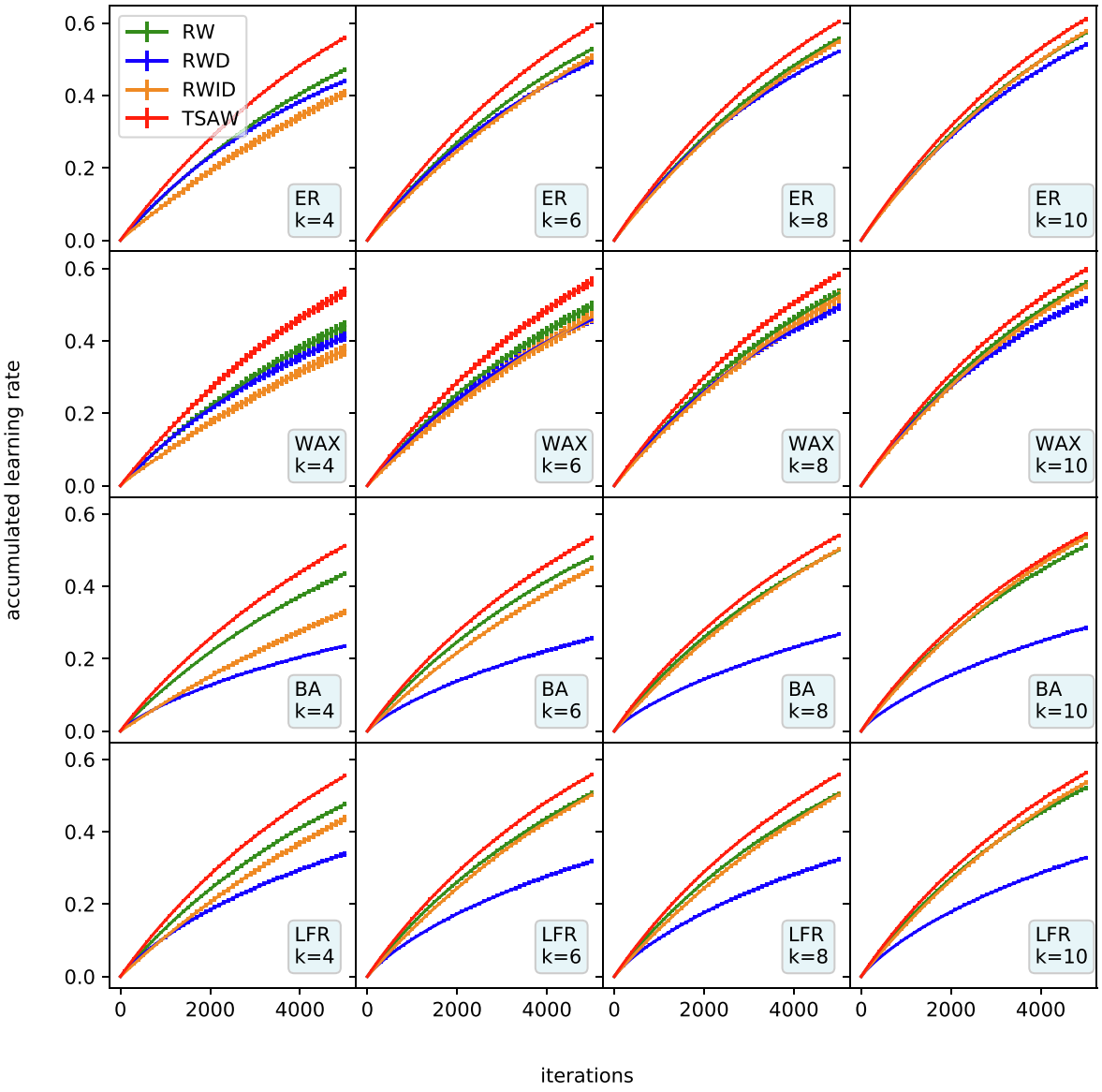


Fig. 3. Learning curves for $N = 5000$ nodes and the models ER, BA, Wax and LFR. Each row and column correspond to different network topologies and average degrees, respectively.

networks and recorded the learning curves for 50 realizations of each dynamics. The starting position of each realization was drawn uniformly from the network nodes and for each configuration we computed the average and standard deviation of the coverage (learning) curves. The resulting curves are shown in Fig. 3. Each row and column corresponds to different network models and average degree, respectively. The panels contain curves colored according to the considered dynamics.

An initial observation shows that the TSAW dynamics outperformed the other dynamics in all the experiments, corroborating previous studies in which TSAW was found to be among the most optimal stochastic walks [7]. On the other hand, the RWD and RWID dynamic resulted in the worst performance among the considered configurations.

All curves seem to present similar shapes but different growing speeds, with faster coverage as $\langle k \rangle$ increases, a behavior that is stronger for the RWD and RWID dynamics. In particular, for ER, the performance among the dynamics becomes substantially similar as the average degree increases. This indicates that the considered dynamics performs very similarly for denser networks. An exception to this rule is the RWD for the BA and LFR. In these cases, the performance of RWD gets slightly worse as network connectivity increases. This is probably related to the fact that a scale-free network (such as BA or LFR) allows the existence of extremely connected nodes in which a walker could get stuck given its preference to move to nodes with high degrees.

Another important aspect of the analysis is how the ranking of dynamics performance change amongs the experiments. In general, TSAW is followed by RW, except for the LFR and BA networks with high connectivity. In this case, RWID attains a second place. This reveals that, in these networks, avoiding hubs can be a good strategy to explore them more quickly. When

the degree is lower, however, RWD performs better than RWID, indicating that, in this case, it is preferable to reach the hubs than avoiding them to attain better performance.

In addition to the previous analyses, we observe two distinct patterns for the behavior of the curves among the network models, one for ER and Wax, and another for BA and LFR. While these pairs do not necessarily display exactly the same behavior, the performance rankings of the dynamics within these pairs of models do not change much.

We also analyzed the differences (rates of growth or learning rate) of the cumulative discovery curves. Fig. 4 shows the obtained rate curves for all the considered configurations. Both the ranks and other overall observations drawn from the cumulative curves can also be drawn for the rate curves. In sum, TSAW also seems to present the highest rates of growth among the considered random walks. In addition, all curves seems to have the same shape, with the exception of RWD in BA and LFR networks. In both cases, the decrease in the observed rate is more abrupt during the first steps. This might be related to the fact that both models have hubs, and RWD tends to explore already visited hubs, as mentioned before. The ranks of learning rate also confirms that TSAW tends to be followed by RW, especially for BA and LFR networks.

To summarize the main characteristics of the obtained learning curves, we applied PCA as a way to reduce their dimension. For each experiment, we derive a set of 50 features corresponding to the values of the learning rate curves (i.e., the derivatives shown in Fig. 4) at epochs 100 iterations apart (see Section 3.4). In other words, each curve was described by the following set of features $(f_1, f_2, f_3 \ldots f_{50})$, where f_i corresponds to the observed fraction of discovered nodes after $100 \times i$ steps were taken by the walker. While we could have sampled the obtained learning curve in a higher frequency rate,

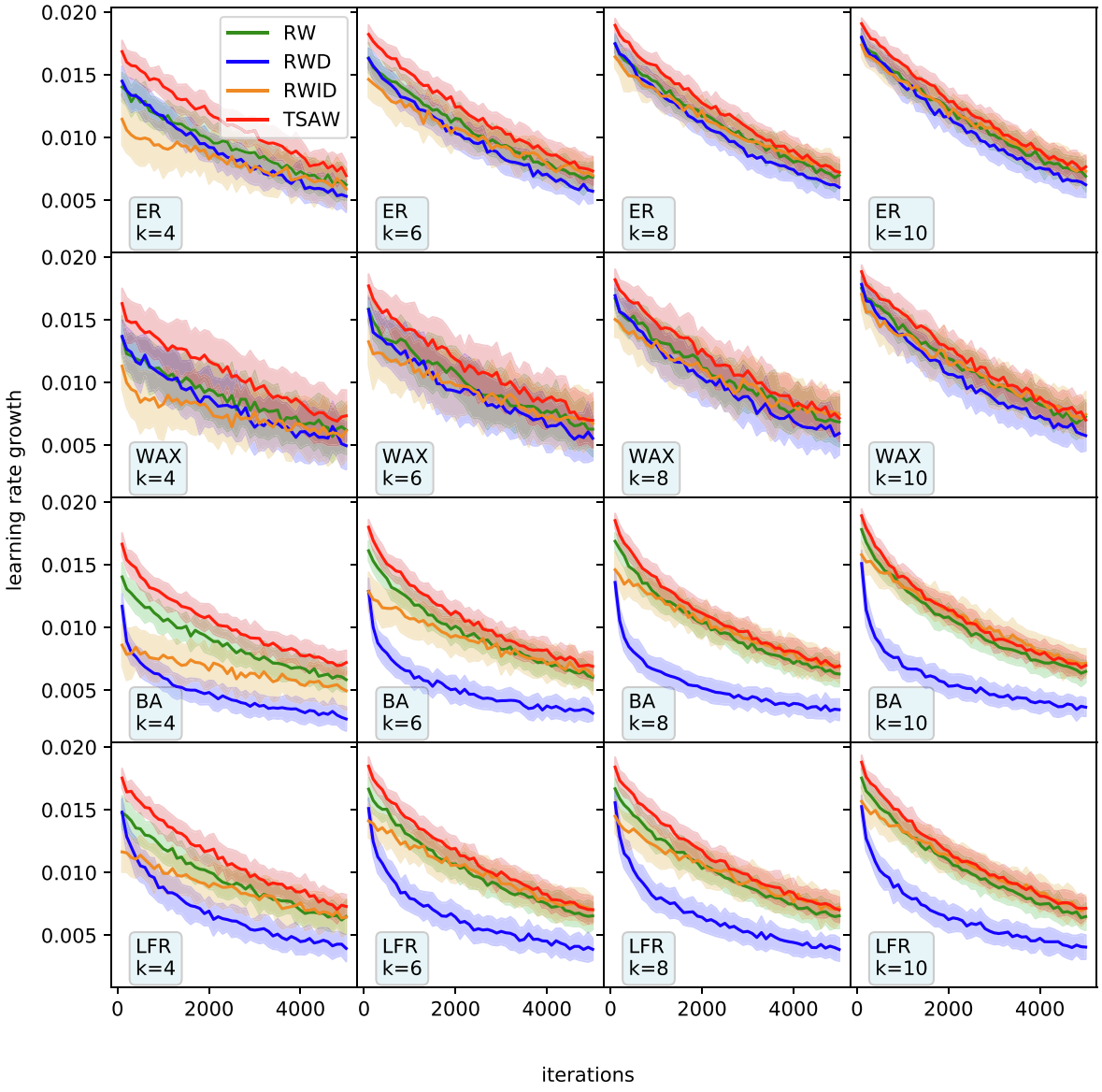


Fig. 4. Learning rates for the considered models and $N = 5000$. Each curve indicates the growth of the number of discovered nodes across the simulation epochs. Error bounds represent the standard deviations of the data points among their realizations. Such variations naturally arise given the probabilistic nature of random walks. Error bounds also include variations in different realizations of network topologies.

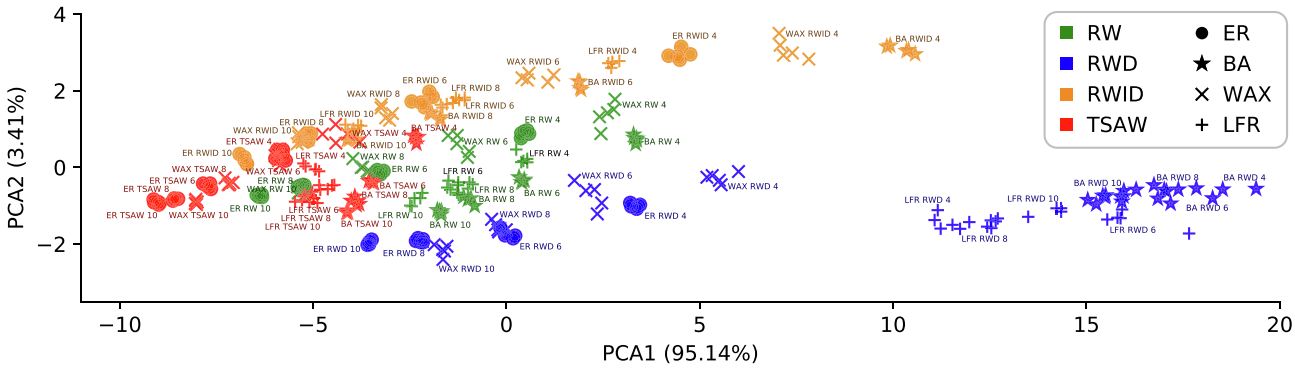


Fig. 5. PCA results for ER, BA, Wax, and LFR for $N = 5000$ nodes. Each instance represents a learning curve obtained for a specific pair of network topology and agent dynamics. Interestingly, in some cases, different combinations of topology/dynamics can lead to similar learning curves.

preliminary results have shown that 50 or more features are equally discriminative, i.e. they lead to a similar PCA distribution. As we shall show, this happens because the derived features are strongly correlated to each other.

The obtained data projection, shown in Fig. 5, reveals that almost 100% of the variance in the curves can be explained by only two components. In particular, the first component covers about 95.1% of the variance. This outcome indicates a high correlation among the curves. At the positive extreme of the first principal component, we find a separated group corresponding to the curves obtained for RWD dynamics simulated on the BA and LFR networks. These correspond to the curves with worst performance among the considered experiments. The RWID curves spread across the PCA1 axis, revealing its diversified behavior with each curve depending on the network model and connectivity.

Along the negative segment of the first principal component, we observe a substantial overlap among the curves for different experiment configurations. This region corresponds to configurations of high node degree or simulated through the TSAW dynamics. Among the notable overlapping configurations are ER and Wax. This is a surprising result, since they present very distinct characteristics in terms of global structure. At least three other regions are shared by different combinations of networks and dynamics. This includes those obtained from ER, Wax and LFR models when the dynamics are TSAW for LFR, and RW for the others. Another example are the RW curves for the BA, Wax, and ER. These results indicate that just by looking at the coverage performance curves it is not trivial to distinguish between network models and dynamics. In addition to that analysis, we also probed whether the results can be explained by the degree assortativity of the networks. Fig. S1 of the supplementary material shows the positions of the curves considering TSAW dynamics according to the obtained PCA colored by degree assortativity. We could not find any specific pattern as the adopted network models present low degree assortativity.

The profile of the PCA axes in the original space, shown in Fig. 6, reveals that the first principal component (PCA1) is almost flat along the iterations. This indicates that all epochs are equally important for the principal component. Conversely, PCA2 seems to capture the difference of rates at the beginning and end of the curves. To further explore these aspects we plotted together all the averaged cumulative learning curves of the considered configurations colored by PCA1 and PCA2. This result is shown in Fig. 7. We note that PCA1 (a) indeed correspond to the inverse of total learning coverage, which is somewhat independent from the shape of the curves. A second order effect seems to be captured by PCA2 (b), corresponding to how fast the rates of the learning curves are increasing across the epochs. This becomes more clear when all the curves are aligned so that the starts and ends match, as shown in (c). Curves with low values of PCA2 tends to be more concave (pre-

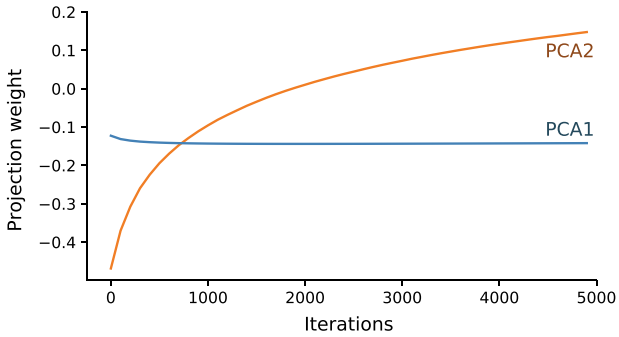


Fig. 6. Projection profiles of PCA1 and PCA2 axes along the original space.

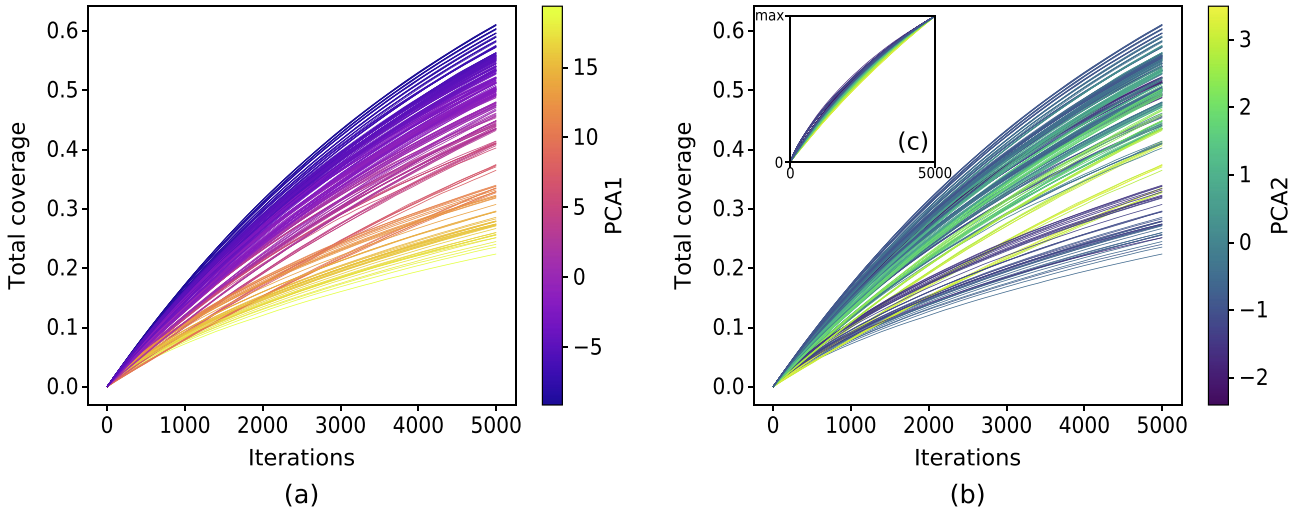


Fig. 7. Averaged learning curves for all the considered configurations. The color of each curve indicates the PCA1 (a) or PCA2 (b). The insight (c) shows all the curves normalized by their respective maximum value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

senting high curvature) and vice versa. All in all, PCA1 corresponds to the average learning speed, while PCA2 seems to be related to the acceleration of the curves.

We also analyzed the PCA of the learning curves obtained for real networks. Since these networks have different sizes and density compared to the considered model realizations previously employed in this section, we generated a unique PCA for each real network. This was accomplished by using the models to generate networks of similar number of nodes and node degree as in the real networks. The results are shown in Figs. S2 and S3 of the supplementary information. The PCA of both real networks resulted in positions very far from any of the existing models, indicating that even sharing topological properties with the adopted models, such as scale-free degree distribution and community structure, the real networks still have different characteristic learning curves. Also, in both networks, the curves obtained for the RW dynamics behaves similarly to those obtained for the TSAW dynamics. Finally, the walk dynamics biased by the degree (RWD) and inverse of the degree (RWID) are located far from each other and from the other dynamics.

5. Conclusion

With many real-world phenomena being modeled and represented as sequences, one way to characterize their respective complex system is by separating the dynamics encoding the sequences from their underlying state space. In this context, a certain stochastic walk dynamics acts as the encoder while a complex network can be used to represent the state space. While this framework has been used to model several real-world problems, no systematic analysis of the relationships among these three aspects of the systems exists in the literature.

In this paper, we performed a systematic analysis of the behavior of different dynamics in well-known network topologies. Whenever a dynamics (or exploration strategy) is performed on a network, one obtains a sequence of visited nodes. We aimed at studying how both topology and network dynamics affects the observed sequence of visited nodes. Here we focused in one property of the sequences, the total number of different visited nodes. This property has many applications in network science, and is oftentimes related to the process of knowledge acquisition [7,6]. In a semantic network, for example, each visited node can be considered as a new learned concept.

We adopted a framework to study the behavior of learning curves. For each combination of network topology and dynamics, we obtained the corresponding learning curves. Then, each learning curve was mapped into a two-dimensional space via PCA. This allowed us to compare curves in a more systematic way, with the advantage of removing correlations while keeping the variability of the original learning curves space.

Several interesting results have been found with our approach. Overall we found that true self avoiding walks outperformed all other dynamics, while the variations of random walks biased towards high or low degree displayed the worst learning curve performances. Despite such differences in performance, we found that all learning curves presented similar shapes. A further investigation of growth rates (i.e. the derivatives) of learning curves revealed that no additional information can be obtained from such an analysis. This means that the learning curves are sufficient to discriminate different network topologies and dynamics.

The Principal Component Analysis confirmed that, despite distinct performances, all curves shapes are similar. This could be confirmed by the fact that curves could be mapped into a two-dimensional space virtually without any lost in the original data variation. Surprisingly, the first component accounted for 95% of the original variation. The visualization provided by

PCA allowed us to observe some interesting patterns. Some regions were found to share different combinations of topologies and dynamics. For example, similar learning curves were found in ER and Wax, showing that the same behavior can be obtained even in very distinct network topologies. The PCA visualization also revealed the variability of learning curves with different topologies. While RWD and RWID were found to be very dependent upon topology, learning curves obtained with TSAW dynamics were found to be much less sensitive to distinct network topologies.

The ambiguity of the behavior of learning curves observed in the PCA space can be useful in practical scenarios. For example, in a knowledge acquisition scenario, the network topology can represent how concepts are linked to each other, while the chosen dynamics can be interpreted as the methodology used to cover the concepts being taught. In such educational scenario, our results suggest that one can be able to deliver the same learning experience by adopting completely different knowledge organization (i.e. network topology) and teaching sequence (i.e. network dynamics).

The obtained results have some implications for text analysis. If the network that should be discovered is a semantic (or cognitive) network, one could analyze whether different authors experience different types of accelerations in learning curves. In the same way, one could derive complexity metrics, based on how easy it is to discover a specific concept. Thus, one could guess that texts with poorly connected concepts tend to be more complex than other texts with words that can be more easily explored. In a similar fashion, one could identify how easy a hapax legonema can be found in such semantic networks [2]. We expect that particular classes of words in texts may have different learning behaviors. Finally, one could also apply similar ideas to analyze how learning processes are affected in the presence of cognitive impairment [36,22].

Our results show that when one uses learning curves to describe sequences of visited nodes ambiguous behaviors may arise. In other words, sequences with similar behavior can be observed from distinct pairs of topology/dynamics. This result suggests that the reconstruction of the processes underlying network construction and topology cannot rely only on learning curves as descriptive features of sequences. For this reason, in future works, we intend to study additional sequence features to identify a minimum set of sequence descriptors that are able to discriminate both the topology and dynamics generating the observed sequence. Because sequences are used to construct embeddings, further studies can analyze if similar embeddings can be obtained from distinct topologies and walks.

CRediT authorship contribution statement

Lucas Guerreiro: Software, Validation, Formal analysis, Investigation, Data curation, Writing - review & editing, Visualization. **Filipi N. Silva:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Diego R. Amancio:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

D.R.A. acknowledges financial support from CNPq-Brazil (Grant No. 304026/2018–2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.ins.2020.12.060.

References

[1] L.A. Adamic, R.M. Lukose, A.R. Puniyani, B.A. Huberman, Search in power-law networks, Phys. Rev. E 64 (Sep 2001) 046135.
[2] A.E. Allahverdyan, W. Deng, Q.A. Wang, Explaining zipf's law via a mental lexicon, Phys. Rev. E 88 (6) (2013) 062804.
[3] D.R. Amancio, O.N. Oliveira Jr, L.d.F. Costa, Topological-collaborative approach for disambiguating authors' names in collaborative networks, Scientometrics 102 (1) (2015) 465–485.
[4] D.J. Amit, G. Parisi, L. Peliti, Asymptotic behavior of the true self-avoiding walk, Phys. Rev. B 27 (Feb 1983) 1635–1645.
[5] H.F. Arruda, L.F. Costa, D.R. Amancio, Using complex networks for text classification: discriminating informative and imaginative documents, EPL (Europhys. Lett.) 113 (2) (2016) 28007.
[6] H.F. Arruda, F.N. Silva, C.H. Comin, D.R. Amancio, L.F. Costa, Connecting network science and information theory, Phys. A Stat. Mech. Appl. 515 (2019) 641–648.
[7] H.F. Arruda, F.N. Silva, L.F. Costa, D.R. Amancio, Knowledge acquisition: a complex networks approach, Inf. Sci. 421 (2017) 154–166.
[8] K. Ban, M. Perc, Z. Levnajić, Robust clustering of languages across wikipedia growth, Roy. Soc. Open Sci. 4 (10) (2017) 171217.
[9] K. Barat, B.K. Chakrabarti, Statistics of self-avoiding walks on random lattices, Phys. Rep. 258 (6) (1995) 377–411.
[10] H. Barbosa, M. Barthelemy, G. Ghoshal, C.R. James, M. Lenormand, T. Louail, R. Menezes, J.J. Ramasco, F. Simini, M. Tomasini, Human mobility: models and applications, Phys. Rep. 734 (2018) 1–74.
[11] M. Bonaventura, V. Nicosia, V. Latora, Characteristic times of biased random walks on complex networks, Phys. Rev. E 89 (Jan 2014) 012803.

[12] R.F. Cancho, R.V. Solé, Least effort and the origins of scaling in human language, Proc. Nat. Acad. Sci. 100 (3) (2003) 788–791.
[13] N. Castro, M. Stella, The multiplex structure of the mental lexicon influences picture naming in people with aphasia, J. Complex Networks 7 (6) (2019) 913–931.
[14] C.H. Comin, T. Peron, F.N. Silva, D.R. Amancio, F.A. Rodrigues, L.F. Costa, Complex systems: features, similarity and connectivity, Phys. Rep. 861 (2020) 1–41.
[15] E.A. Corrêa Jr., V.Q. Marinho, D.R. Amancio, Semantic flow in language networks discriminates texts by genre and publication date, Phys. A: Stat. Mech. Appl. (2020) 124895..
[16] E.A. Corrêa Jr, F.N. Silva, L.F. Costa, D.R. Amancio, Patterns of authors contribution in scientific manuscripts, J. Inf. 11 (2) (2017) 498–510.
[17] L.F. Costa, Knitted complex networks. arXiv: 0711.2736, 2007.
[18] L.F. Costa, O.N. Oliveira Jr, G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, L.E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Adv. Phys. 60 (3) (2011) 329–412.
[19] L.F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: asurvey of measurements, Adv. Phys. 56 (1) (2007) 167–242.
[20] L.F. Costa, G. Travieso, Exploring complex networks through random walks, Phys. Rev. E 75 (1) (2007) 016102.
[21] A.S. da Mata, Complex networks: a mini-review, Braz. J. Phys. (2020)..
[22] L.B. dos Santos, E.A.C. Júnior, O.N. Oliveira Jr, D.R. Amancio, L.L. Mansur, S.M. Aluísio, Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts, ACL 1 (2017) 1284–1296.
[23] C. Franzoni, G. Scellato, P. Stephan, Foreign-born scientists: mobility patterns for 16 countries, Nat. Biotechnol. 30 (12) (2012) 1250–1253.
[24] F.L. Gewers, G.R. Ferreira, H.F. Arruda, F.N. Silva, C.H. Comin, D.R. Amancio, L.F. Costa, Principal component analysis: a natural approach to data exploration. arXiv:1804.02502, 2018..
[25] C. Herrero, Self-avoiding walks and connective constants in clustered scale-free networks, Phys. Rev. E 99 (2019) 01.
[26] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[27] I.T. Jolliffe, Principal Component Analysis, second Ed., Springer Series in Statistics, Springer-Verlag New York, New York, NY, 2002.
[28] Y. Kim, S. Park, S.-H. Yook, Network exploration using true self-avoiding walks, Phys. Rev. E 94 (2016) 042309.
[29] I.T. Koponen, M. Nousiainen, Concept networks in learning: finding key concepts in learners' representations of the interlinked structure of scientific knowledge, J. Complex Networks 2 (2) (2014) 187–202..
[30] A. Kumar, Y. Goswami, M. Santhanam, Distinct nodes visited by random walkers on scale-free networks, Phys. A Stat. Mech. Appl. 532 (2019) 121875.
[31] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (2008) 046110.
[32] Z. Levnajić, Derivative-variable correlation reveals the structure of dynamical networks, Eur. Phys. J. B 86 (7) (2013) 298.
[33] T.S. Lima, H.F. Arruda, F.N. Silva, C.H. Comin, D.R. Amancio, L.F. Costa, The dynamics of knowledge acquisition via self-learning in complex networks, Chaos Interdisc. J. Nonlinear Sci. 28 (8) (2018) 083106..
[34] L. Lovász, Random walks on graphs: a survey, in: D. Miklós, V.T. Sós, T. Szőnyi (Eds.), Combinatorics, Paul Erdős is Eighty, vol. 2, János Bolyai Mathematical Society, Budapest, 1996, pp. 353–398..
[35] L. Lü, Z.-K. Zhang, T. Zhou, Zipf's law leads to heaps' law: analyzing their relation in finite-size systems, PLoS One 5 (12) (2010).
[36] N.B. Lundin, P.M. Todd, M.N. Jones, J.E. Avery, B.F. O'Donnell, W.P. Hetrick, Semantic search in psychosis: modeling local exploitation and global exploration, Schizophrenia Bull. Open 1 (1) (2020), sgaa011.
[37] V.Q. Marinho, G. Hirst, D.R. Amancio, Authorship attribution via network motifs identification, in: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), IEEE, 2016, pp. 355–360.
[38] M.M. Meerschaert, E. Scalas, Coupled continuous time random walks in finance, Phys. A Stat. Mech. Appl. 370 (1) (2006) 114–118.
[39] J.R. Norris, Markov Chains, 2, Cambridge University Press, 1998.
[40] K. Pearson, LIII. on lines and planes of closest fit to systems of points in space, London Edinburgh Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572.
[41] M.Z. Rodriguez, C.H. Comin, D. Casanova, O.M. Bruno, D.R. Amancio, L.F. Costa, F.A. Rodrigues, Clustering algorithms: a comparative approach, PLoS One 14 (1) (2019).
[42] F.N. Silva, D.R. Amancio, M. Bardosova, L.F. Costa, O.N. Oliveira, Using network science and text analytics to produce surveys in a scientific topic, J. Inf. 10 (2) (2016) 487–502.
[43] M. Sipser, Introduction to the theory of computation, ACM Sigact News 27 (1) (1996) 27–29.
[44] M. Stella, A. Zaytseva, Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth, PeerJ Comput. Sci. 6 (2020) e255.
[45] D.K. Sutantyo, S. Kernbach, P. Levi, V.A. Nepomnyashchikh, Multi-robot searching algorithm using lévy flight and artificial potential field, in: 2010 IEEE Safety Security and Rescue Robotics, IEEE, 2010, pp. 1–6.
[46] P. Thagard, Conceptual Revolutions, Princeton University Press, Princeton, NJ, US, 1992.
[47] J.V. Tohalino, D.R. Amancio, Extractive multi-document summarization using multilayer networks, Phys. A Stat. Mech. Appl. 503 (2018) 526–539.
[48] B.A.N. Travençolo, L.F. Costa, Accessibility in complex networks, Phys. Lett. A 373 (1) (2008) 89–95.
[49] S.-J. Yang, Exploring complex networks by walking on them, Phys. Rev. E 71 (1) (2005) 016107.

CHAPTER

3

RECOVERING NETWORK TOPOLOGY AND DYNAMICS VIA SEQUENCE CHARACTERIZATION

Recovering network topology and dynamics via sequence characterization. Lucas Guerreiro, Filipi N. Silva, Diego R. Amancio. *Knowledge-Based Systems.* Current status: Under Review.

3.1 Context

The analysis of complex networks exploration as time series has been previously studied, especially in the knowledge acquisition process (ARRUDA *et al.*, 2017; LIMA *et al.*, 2018a; ARRUDA *et al.*, 2019). In Arruda *et al.* (2019), for example, the authors have discussed symbols as an outcoming of the combination of a network (generator) and a dynamics (agent), in which the symbols are resulted from an exploration within the network. Therefore, a sequence can be understood as the time series of symbols obtained during the exploration of a network by a random walker. In terms of real-world situations, this can be seen, for instance, whenever an agent is walking over a city. Such agent will explore streets and corners and will recognize some places, therefore the time series of the visited places will generate a sequence of symbols.

Although such topic had been studied deeply over the past few decades, the previous works focused on the relationship arisen from known topologies and dynamics to the sequences. Thus, our proposed work enlightens the reverse path, i.e., from the sequences to the topologies and dynamics. Therefore, our main question comes about: is it possible to find out the original network topology and the dynamics used to generate a sequence? We aimed on recovering the structures that generated the sequences by analyzing only the sequences and their underlying properties.

Hence, the motivation of this work was to prove whether it is possible to identify the generator structures of a sequence, and, consequently, observe the most distinguishing properties of reconstructed networks, which can lead to a deeper understanding of the knowledge acquisition process in complex networks.

3.2 Contributions

The main contribution of this paper was to present a novel approach to determine the original network topology and the used dynamics based solely on the generated sequence by such structures. Our proposed methodology abstracts properties from a reconstructed network from the sequence, such properties are further input into different machine learning classifiers in order to identify their respective original structures.

The results of this paper allowed us to proof that it is indeed possible to recover the aforementioned structures having little information on the topology and dynamics used. Our study has shown that some properties obtained from the network are more relevant on the recovery, such as the average degree of the reconstructed network. Moreover, some rather simple classifiers such as Random Forest and Decision Tree had high performance on the classification task.

Finally, although limited to a certain range of topologies and dynamics, this paper could imply the possibility to recover the original structures, which can lead to further developments in the area considering that no previous studies had approached this problem. Real-world situations may also be explored, where one may identify the network mapping from the resulting walk over the inherent structures by fitting it to some known model.

Recovering network topology and dynamics via sequence characterization

Lucas Guerreiro[1], Filipi N. Silva[2] and Diego R. Amancio[1]

[1]*Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil*

[2]*Indiana University Network Science Institute, Bloomington, Indiana 47408, USA*

Abstract

Sequences arise in many real-world scenarios; thus, identifying the mechanisms behind symbol generation is essential to understanding many complex systems. This paper analyzes sequences generated by agents walking on a networked topology. Given that in many real scenarios, the underlying processes generating the sequence is hidden, we investigate whether the reconstruction of the network via the co-occurrence method is useful to recover both the network topology and agent dynamics generating sequences. We found that the characterization of reconstructed networks provides valuable information regarding the process and topology used to create the sequences. In a machine learning approach considering 16 combinations of network topology and agent dynamics as classes, we obtained an accuracy of 87% with sequences generated with less than 40% of nodes visited. More extensive sequences turned out to generate improved machine learning models. Our findings suggest that the proposed methodology could be extended to classify sequences and understand the mechanisms behind sequence generation.

I. INTRODUCTION

Many human behavior phenomena are linked to the generation of sequences [15]. Examples include: the sequence of places people visit in a touristic city [41]; users visited websites such as social media profiles and posts [29]; videos viewed by individuals in a streaming platform [1]; everyday decisions, such as where to eat, where to work; topics of papers produced by scientists, pieces of narratives (in text or movies), or even music [20, 21]. In recent years, complex networks [18] have been used to represent and model the structure of these systems. In most cases, however, the available information is not in the form of networks but sequences.

Some techniques can be used to infer the underlying network from a sequence or set of sequences. Examples are the reconstruction of language models based on text data [2, 4], where nodes represent words or pieces of text, and human mobility networks [40], with nodes representing places. In such models, an edge exist for every adjacent nodes appearing in a set of sequences (e.g., of words or of places). Sequences can then be understood as trajectories of nodes that are performed in a network according to a certain dynamics, such as a walk [6]. Such a type of reconstruction process is taken place gradually as a discrete knowledge acquisition process [7]. To simulate such a process, agents are initially scattered in a network. Next, they perform walks (e.g., random walks) that generate sequences of nodes. For each iteration, each agent partially reconstructs the network from its own history of visited nodes.

Exploring the properties of the simulated knowledge acquisition process can help understanding real-world problems related to the agents' perception or effectiveness in reconstructing networks from sequences. For instance, in a social media platform, depending on how and how long a user (i.e., an agent) navigates across related posts, they may find a certain post to be central, which may not correspond to the view of other users navigating using different heuristics. In general, the networks reconstructed by users using different walk dynamics can be substantially different. In this context, an important question is if we can recover both the generating dynamics (walks) and the generating network independently. In this work, we explore this question from the perspective of network measurements. In particular, we check if the topological properties (e.g., average degree, transitivity, etc) can be used to recover the network structure and identify the walk dynamics. This differs from

previous works [24], in which only the learning curve profiles were used to characterize the generated networks.

Our analysis starts with the realizations of well-known random network models, which cover several characteristics present in real-world networks, such as scale-free distribution [8], small-world [44], presence of community structure [31], and locality [45]. Then, different walk dynamics are performed in such networks, resulting in sequences of nodes, which are used to create partial reconstructions of each network. Next, network measurements are computed for each network. Finally, we applied classification algorithms to identify the original network model and walk dynamics by looking solely on the generated sequences.

Our results indicate that it is indeed possible to recover the generating model and dynamics from network measurements obtained from sequences. As expected, longer sequences lead to better accuracy. However, we found different combinations of network characteristics and walk dynamics result in different performances for shorter sequences. For instance, the classifiers have difficulty distinguishing from the uniform random models (Erdős-Rényi model [16]) and geographic models (Waxman model [45]).

The following section discusses related works in the literature as we can see that the concepts of sequences originating from topologies and dynamics are being widely explored and with interesting findings. Later, in the methodology section we present the steps to produce our experiment as well as the configurations of our work. Finally, we present the results and discussions of our study, and the conclusions we have found.

II. RELATED WORKS

The problem of exploring complex networks via walk dynamics has been addressed in several contexts [9, 13, 14, 36]. One particular issue is finding the best dynamics strategy to discover networks nodes (and/or edges) in a optimized way [7, 32]. In [7] the authors probed the performance of knowledge acquisition regarding the true self-avoiding dynamics and a Lévy flight-based dynamics on distinct topologies. The influence of variations in the dynamics parameters was analyzed. The authors found that the global impact of parameters variations on knowledge acquisition is surprisingly low. Conversely, the parameter selection is more effective when evaluating the knowledge acquisition problem locally. All in all, this study observed that, when performing collective discovery in a structure, the dynamics

parameters do not affect significantly the performance.

In a related study, [24] investigated how knowledge is acquired on different configurations of models and dynamics. The learning time regarding nodes discovery rate in a given period was the main focus of the study. The authors reported that the efficiency in acquiring knowledge depends on how the network is explored and on the topology of knowledge representation. The true self-avoiding dynamics was found to be effective in most scenarios. Most importantly, the same learning behavior could be achieved with different pairs of network structures and agent dynamics. This means that it is possible to generate the same efficiency in learning by changing both knowledge organization and the way knowledge is acquired.

The problem of analyzing how agents learn the structure of networks has also been explored with variations in the way knowledge is transmitted and stored. In [32] the concept of a *network brain* was proposed. The idea behind this model comprehends a centralizing structure that receives the knowledge discovered by multiple agents walking over the network. Interestingly, the authors report that neither the topology nor the dynamics strongly impacts the learning efficiency.

Some works also have investigated random walks as generator of symbols data [6, 24]. In [6] the authors argue that a sequence of symbols can be seen as being generated from walks on a networked topology. The authors analyzed how the performance of different combinations of topologies and dynamics may impact the performance of knowledge acquisition. The study also analyzed how well the network is reconstructed when it is transmitted as a sequence of visited nodes that are compressed before transmission. The so-called knitted networks – similar to word adjacency networks – were found to display the best performance when considering both compression efficiency and recovery of sequential data.

The reconstruction of networks is another relevant topic related to the current study [19, 23, 30, 47]. One of the most well-known methods is the visibility graph [30]. The method is based on the idea that each node in a time series of values can see other nodes, and the spatial visibility criteria are used to generate edges. Such a rather simple idea has been proven to be effective on the reconstruction of networks, being able to translate, for example, from a periodic time series to a regular graph, while using a random series as a random graph. While the work proposes a reconstruction method, it does not consider that the original time series stems from a particular walk dynamics.

Differently from previous works, here we aim to identify the main properties of the recon-

structed networks. More specifically, we investigate whether the reconstruction of networks via a co-occurrence strategy is able to identify the underlying topology and dynamics generating the observed sequence of symbols.

Even though sequences have already been analyzed as a result of dynamics occurring on a networked structure [6], no previous studies have inferred the underlying structure and mechanisms using machine learning techniques. This work proposes a framework that may uncover the original network and dynamics generating a sequence. We also explore the sequence lengths that can lead to high accuracy in identifying the topology and dynamics. While very large sequences are more likely to provide improved prediction accuracy, here we also analyzed how short a sequence can be so that one is able to identify the topology and dynamics generating the observed sequence.

III. METHODOLOGY

In this paper, we probe whether the characteristics of time series can be useful to infer the topology and walk dynamics associated to the observed sequence, i.e. the sequence generated by walking over the network. We adopted the following methodology to test our hypothesis. First, we generated artificial network topologies. Well-known walks dynamics were also used to generate the sequence of symbols. The obtained sequences are used to create a network, in the network reconstruction step. Then, the structural properties of the reconstructed are extracted, and they are used as input to the machine learning method aiming at identifying the associated topological and dynamical properties used to generate the sequence. The proposed framework is illustrated in Figure 1 and the main steps are summarized below.

1. *Network topology*: we selected four well-known network models to generate the topology of the networks. The chosen topologies includes random networks and models with more realistic features, including heterogeneity in connectivity.

2. *Network dynamics*: given a network, an agent walks over the network in order to generate a sequence of symbols. We also have used four well-known network dynamics to explore network. As a result, a sequence of visited nodes is generated.

3. *Network reconstruction*: the sequence of generated symbols by the agent walking

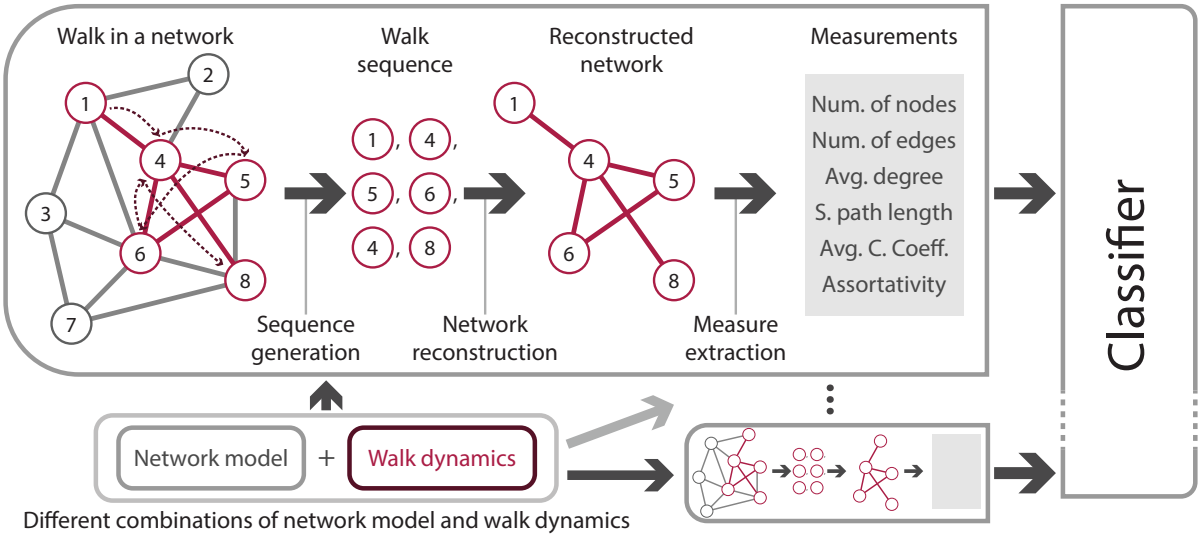


FIG. 1. Schematics of the methodology. First, we generate multiple iterations of the considered network models. Then, for each considered walk dynamics, we sample a number of walk sequences. Each walk sequence is then used to generate a reconstructed network. This is accomplished by connecting nodes according to adjacency of nodes in the sequence. For instance, for a sequence of nodes $\{1, 4, 5, 6, 4, 8\}$, the following edges are reconstructed: $\{(1, 4), (4, 5), (5, 6), (6, 4), (4, 8)\}$. Next, we compute a set of network measurements for each walk sequence. Finally, the features of each network are used to train a classifier that try to recover the corresponding model, walk dynamics or both.

through the network is used to reconstruct the network. This process uses a co-occurrence approach to create a representation of the original network.

4. *Properties extraction*: features are extracted from the reconstructed networks. The extracted features are then used in the machine learning algorithms. We used 6 features to characterize the structure of the reconstructed networks.

5. *Structures classification*: features are fed into different machine learning methods in order to recover both topology and the dynamics that generated the observed sequence. In summary, given an unknown sequence, we aim at identifying the network model and dynamics that generated it.

A. Network topologies

We considered four well-known network topologies. The adopted network parameters are similar to those employed in similar papers studying the relationship between network

topology and dynamics [24]. We adopted the following undirected and unweighted network models [16, 17, 31]:

- *Erdős-Rényi (ER)*: this model generates random graphs. To create this network, each candidate edge is established based on a global probability p. Random networks are typically characterized by low shortest path lengths and low values of local connectivity (clustering coefficient).

- *Barabási-Albert (BA)*: the BA model [8] reproduces the scale-free distribution of node degree. The model adds, at each step, a new node the has probability of linking to other nodes of the network. Let k_i be the degree of node i. The probability of i to receive a link from the new node (p_i) is proportional to k_i, i.e. $p = k_i / \sum_j k_j$. Because a new node has a preference to connect with more connected nodes, a few hubs arise.

- *Waxman (Wax)*: this model implements a geographic network. In order to construct such a network, all nodes are randomly distributed into a two dimensional space. The probability of a link existing between two nodes takes into consideration the distance between nodes, so that nearby nodes have a higher probability of being connected to each other [45].

- *Modular Networks (LFR)*: we have also used a topology that reproduces the modularity of real-world networks. We adopted the implementation proposed in [31]. This model also reproduces the scale-free behavior within network community. The parameters employed to generate the communities is similar to those used in related studies [24]. The following parameters were considered here: number of communities (n_C), exponent for the degree sequence (t_1), community size distribution (t_2) and mixing parameter (μ) [31]. As in related works, we adopted the following values: $n_C = 5$, $t_1 = 3$, $t_2 = 0$ and $\mu = 0.2$. The maximum node degree is chosen in order to obtain the desired average degree of the network.

We considered networks comprising $N = 5000$ nodes with the following values of average degree $\langle k \rangle = \{4, 6, 8, 10\}$. A wider range values for N was not considered because, in preliminary experiments, we have not observed significant variations in the obtained results.

B. Network dynamics

The networks are explored with four different agent dynamics. We have selected the traditional random walk dynamics (RW) [34] and three variations of this random walk: degree-biased random walk (RWD) [10], random walk biased towards the inverse of the degree (RWID) [10], and the true self-avoiding walk (TSAW) [5, 28]. As illustrated in Figure 1, the agent dynamics are the rules to walk over the network. The adopted dynamics are described below:

- *Random Walk (RW)*: the traditional random walk randomly chooses one of the neighbors of the current node based on a uniform distribution. The probability of transition from node i to j is $p_{ij} = k_i^{-1}$, where k_i is the degree of i.

- *Random Walk biased towards the Degree (RWD)*: this dynamics is a variation of the traditional RW. Here, the agent has a higher probability of visiting nodes with larger degree. The transition probability p_{ij} is computed as:

$$p_{ij} = \frac{k_j}{\sum_{l \in \Gamma_i} k_l}, \tag{1}$$

  where Γ_i is the set comprising the neighbors of i.

- *Random Walk biased towards the Inverse of the Degree (RWID)*: here the agent tends to visit nodes with smaller degrees, according to the following equation for the transition probability:

$$p_{ij} = \frac{k_j^{-1}}{\sum_{l \in \Gamma_i} k_l^{-1}}. \tag{2}$$

- *True Self-Avoiding Walk (TSAW)*: this dynamics is a variation of the self-avoiding random walk [27]. In this dynamics, the agent avoids visiting nodes previously visited. As observed in related works, this walks tends to explore more efficiently the network [6, 24]. The probability transition for the TSAW is computed as:

$$p_{ij} = \frac{e^{-\lambda f_{ij}}}{\sum_{l \in \Gamma_i} e^{-\lambda f_{il}}}, \tag{3}$$

  where f_{ij} denotes the frequency that the edge linking nodes i and j has been visited. The parameter λ is a positive constant that controls how likely an agent will visit

an edge that has already been visited. Similar to related studies, we are using $\lambda = \ln 2$ [6, 24].

In order to analyze the dependence of the results with the length of the random walk (S), we considered different values of S. This includes very short walks and also a number of steps compatible with the network size considered. We used $(S) = \{10, 50, 100, 500, 1000, 2000, 5000\}$.

C. Network reconstruction

The next step in the framework presented in Figure 1 is to reconstruct the network. The network is reconstructed based on the resulting time series of visited nodes. Each pair of adjacent nodes observed in the sequence of visited nodes is linked with an edge. This is similar to the co-occurrence strategy used to create networks from sequential data. This co-occurrence model is particularly used when modeling texts as networks [2, 4, 33, 35, 42]. For each network model and random walk, we considered 1000 realizations.

D. Properties extraction

We have selected the following properties to characterize the reconstructed networks:

1. *Number of nodes*: the size of the reconstructed network will also represent the number of nodes discovered after the random walk. We intend to analyze whether the total of discovered nodes is a relevant information to identify the topology and dynamics used to generate the sequence. Related works have shown that the number of discovered nodes can vary according to the adopted topology and dynamics [24], with the highest efficiency observed for the TSAW random walk. This measure alone, however, is not enough to discriminate all combination of network topology and agent dynamics [24].

2. *Number of edges*: similarly to previous property, we extracted the number of discovered edges, i.e. the total number of edges of the reconstructed network.

3. *Average degree*: we also explore the influence of the reconstructed networks' average degree on the identification of the originating structures. Although this information can be recovered from the both number of nodes and edges, we included this information in order to discuss the results in terms of this well-known quantity as well.

4. *Shortest path length*: this property quantifies the typical shortest path length between all nodes of the reconstructed network. We also measure – in the reconstructed network – the shortest path length between the first and last nodes of the sequence used to reconstruct the network.

5. *Average clustering coefficient*: this property quantifies the local density of edges of a node, which is directly related to the number of triangles. The clustering coefficient C_i of a given node i is:

$$C_i = \frac{2T(i)}{k_i(k_i - 1)}, \tag{4}$$

where $T(i)$ denotes the number of triangles including node i. Equivalently, $T(i)$ is the number of links between neighbors of i.

6. *Assortativity*: this coefficient measures whether nodes with high degree tends to be linked with other highly connected nodes. The assortativity can be measured in the of the degree correlation of linked nodes [46].

The adopted measures are meant to characterize the network locally and globally.

Because the network features can be extracted along the evolution of the network reconstruction, two strategies for the network characterization were considered. The strategy referred to as "last value" considered only the network generated at the end of the random walk. We also considered the approach referred to as "all values". Here, the above measurements are extracted considering the evolution of the network as symbols are generated and incorporated in the network. For each random walk length, we extracted the features along the evolution by considering 10 intermediary values. For example, for the random walk considering 1,000 steps, we extracted the network features when the agent completes 100, 200, 300 ... steps.

E. Topology and dynamics classification

The properties extracted from the reconstructed network are features that are used to characterize samples in multi-class classifiers [3], where classes can be: (i) the topology of the network; (ii) the dynamics used by the agent; and (iii) both (i) and (ii). This experiment aims to understand how accurate one can estimate the topology and dynamics generating the observed sequence of symbols.

The goal of the classification, as discussed previously, is to obtain a classifier model that can successfully classify the samples, this case the properties of the generated network, into known structures. Therefore, we may have a generalist model that will be able to predict unknown sequences and infer which topology and/or dynamics generated such sequence, based only on the resulting properties of the sequence's reconstructed network.

In order to evaluate whether the extracted properties can be used to detect the mechanisms and structure generating a sequence, we used well-known supervised classifiers [3]. The following classifiers were chosen to detect patterns in the data: Decision Trees (DT) [38], Random Forest (RF) [12], Stochastic Gradient Descent (SGD) [11], Multilayer Perceptron (MLP) [26], k-Nearest Neighbors (KNN) [22], Linear Discriminant Analysis (LDA) [25], and Gaussian Naive Bayes (GNB) [39]. The parameters were optimized according to the approach suggested in related works [3].

IV. RESULTS AND DISCUSSION

A. Evolution of the reconstructed network

To analyze the evolution of the considered metrics as network are reconstructed, we show, in Figure 2 the values of the considered metrics for different walk lengths, models and dynamics. Here we considered $\langle k \rangle = 4$, however similar results were obtained for different values of average degree (result not shown).

The results show a similar behavior for LFR and BA models with the RWD dynamics in most properties curves. Also, we observe that the curves for both LFR and BA when considering the RWD dynamics correspond to the smallest growth along the walk with a large gap between them and the other dynamics that increases as more steps are considered (i.e. as the network grows). The results also revealed that is that the TSAW dynamics displayed the highest efficient to discover nodes and edges. Concerning the discovery of edges, we can notice that the gap between TSAW and other dynamics is even larger. These results achieved by the TSAW dynamics are consistent with similar findings [24].

As for the average degree, we found that the RWD dynamics on both LFR and BA models discovers new edges much faster than nodes, leading thus to a higher estimated degree compared to the other combination of models an dynamics. This effect might stem

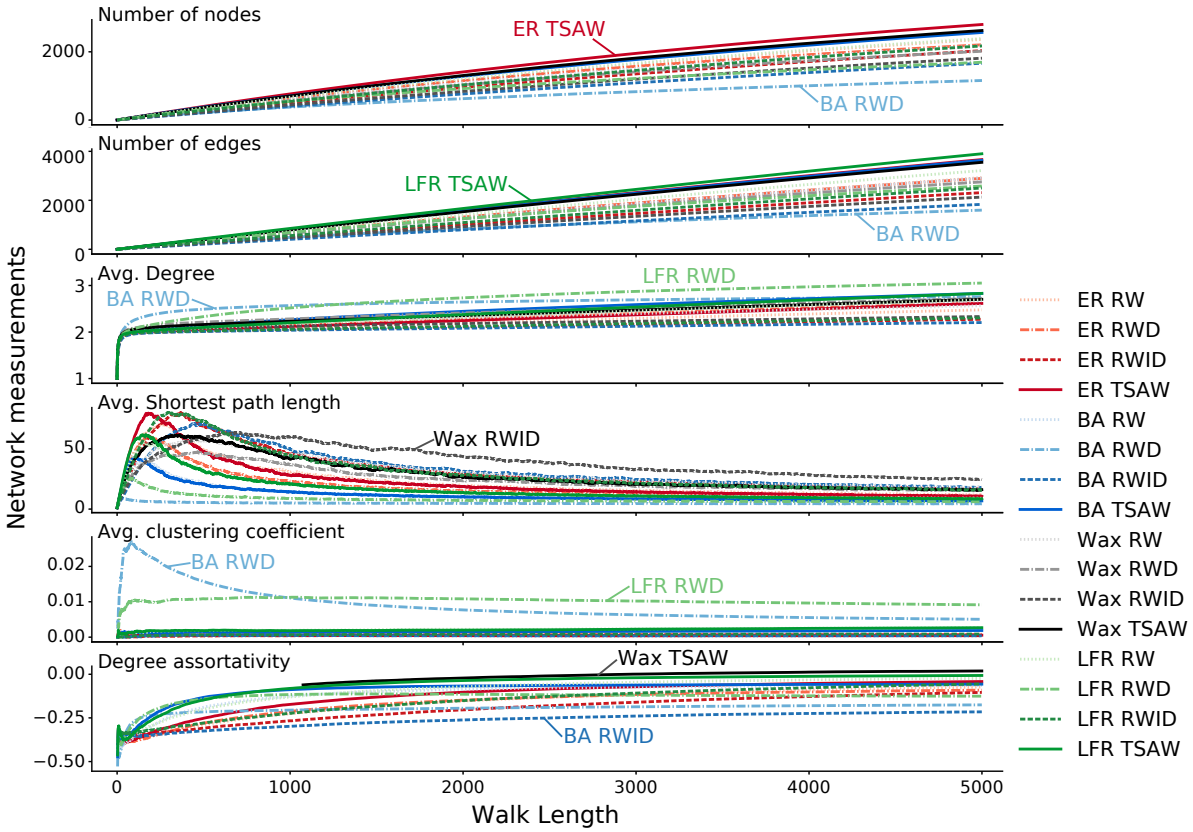


FIG. 2. Evolution of the network metrics obtained from the reconstructed networks when considering different values of walk length. The original network was created considering $\langle k \rangle = 4$ in all network models.

.

from the fact that the TSAW takes into account the visited edges, so already visited edges tend to be less visited in future interactions.

Interestingly, the shortest path length initially takes high values in almost all cases. This happens because when new nodes are discovered and not revisited, the network is similar to a line graph. For larger values of walk length, the shortest path length seems to converge. Once again the RWD dynamics for both LFR and BA networks displayed a different behavior: the estimated value along the first steps are not as high as the ones observed for the other combinations of network and dynamics.

For the clustering coefficient property, while most curves remain close to each other, LFR and BA for the RWD dynamics, again, seem to estimate much higher values of cluster coefficient compared to the other configurations. This might be a consequence of this dynamics being sensitive to hubs, which is present in both LFR and BA topologies. Finally, for the

degree assortativity, after larger variation for short sequences, the assortativity seems to increase (and converge) for all combinations of network models and dynamics as more nodes are visited along the walk.

In general, simpler features, such as the number of nodes, edges and the average degree and degree associativity resulted in monotonically increasing curves, with no crosses or substantial changes between. In contrast, different patterns where observed for the clustering coefficient and shortest path length.

B. Detecting the network topology and walk dynamics

While in Figure 2 one can observe in fact some differences in the behavior of curves, the discrimination between network models, dynamics or even both is not visually evident. Because we are interested in recovering the original network model and dynamics that generated the sequence (i.e. the first two steps in Figure 1), we conducted a machine learning experiment. Our machine learning experiment was carried out in a threefold way, depending on the information being recovered: network model, agent dynamics, or both. For each configuration of network model, we considered 100 random realizations. Concerning the different dynamics, each realization started from a random seed, and 1,000 realizations were considered. We used half of the data for training.

In Table I we show the accuracy rate obtained when identifying the model generating the sequence, considering the cross-validation analysis adopted. Two approaches to compose the features are considered. The approach referred to as "all" extract the features obtained as the reconstructed network evolves, while "last" only considers the last configuration of the reconstructed network. We show the results obtained for different walk lengths $s = \{10, 50, 100, 500, 1000, 2000, 5000\}$. The results show that – as expected – the discriminability is not significant for short walks. However, it is interesting to note that an accuracy higher than 50% can be obtained with only 100 steps (for the RF classifier). The best accuracy also significant is improved when considering 500 steps, reaching an accuracy higher than 86% when less than 20% of network is discovered. An almost perfect discrimination of network models is achieved with a sequence of 5000 symbols (walk length). It is also worth noting also that, in almost all cases, the best classification is obtained with the Random Forest classifier. In almost all scenarios, we also noted that the performance considering

the "last" and "all" approaches are similar, meaning that the evolution of the reconstructed network does not provide a significant amount of information for this classification task.

TABLE I. Accuracy rate (%) for the classification considering the four network models: ER, BA, Wax and LFR (see Section III A). Columns represent the number of steps taken by the agents. When considering 5,000 steps in the walk, the best result is obtained with the Random Forest classifier. The network model generating the sequence can be identified with an accuracy of 98.59%. The best results for each walk length are highlighted.

Method		10	50	100	500	1,000	2,000	5,000
DT	last	28.7	42.0	47.1	69.9	81.6	90.5	97.6
	all	28.9	35.4	43.3	70.6	81.6	90.7	97.8
RF	last	28.7	42.9	49.0	75.1	85.6	93.0	98.3
	all	**29.0**	42.0	**53.0**	**78.1**	**86.6**	**93.5**	**98.6**
SGD	last	25.8	29.7	36.7	49.0	48.4	47.6	49.4
	all	25.7	28.2	40.6	45.0	53.5	43.7	52.8
MLP	last	28.2	43.3	52.1	72.5	80.0	86.2	76.3
	all	28.3	**43.6**	52.3	74.5	80.8	80.5	39.1
KNN	last	26.5	37.3	46.4	65.5	72.4	80.4	91.0
	all	26.1	34.0	40.1	63.8	74.2	82.2	93.6
LDA	last	28.0	37.6	41.9	58.4	64.4	72.5	74.0
	all	28.0	38.2	42.8	60.6	67.7	73.8	81.5
GNB	last	27.2	37.7	45.1	53.7	55.7	57.9	53.8
	all	26.5	35.5	42.0	53.5	55.3	56.3	58.7

In Table II we show the accuracy rates obtained when detecting the walk dynamics used to generate the sequence. The results show that the best accuracy rate is similar to the one obtained when identifying the network models (98.76%). The dynamics are retrieved with higher accuracy than models for short sequences (10, 50 and 100), while model recovery accuracy is higher for longer sequences (500, 1000, and 2000). Surprisingly, when less than 2% of the network is recovered (100 steps) one is able to identify the walk dynamics with an accuracy higher than 50%. Likewise, one can reach almost 80% of accuracy when less than 10% of the network is discovered (500 nodes). In order to recover the dynamics with an accuracy higher than 85%, 1000 steps are required. Concerning the methods, the Random Forest classifier once again achieved most of the highest accuracies, along with the "all" strategy. The MLP algorithm displayed competitive results, specially for the sequence sizes of 50 and 100.

Table III shows the results obtained when identifying both the network topology and walk dynamics generating the observed sequence. While this is a much difficult task since we are discriminating 16 classes (i.e. four different topologies and four different walks),

TABLE II. Accuracy rate (%) for the classification considering the four network dynamics: RW, RWD, RWID and TSAW (see Section III A). Columns represent the number of steps taken by the agents. When considering 5,000 steps in the walk, the best result is obtained with the Random Forest classifier. The network dynamics generating the sequence can be identified with an accuracy of 98.76%. The best results for each walk length are highlighted.

Method		10	50	100	500	1,000	2,000	5,000
DT	last	40.0	54.3	55.3	65.9	77.3	89.1	97.9
	all	**40.4**	46.9	51.3	65.8	76.6	88.7	97.9
RF	last	40.0	54.7	59.9	71.0	81.8	92.0	98.8
	all	**40.4**	54.2	61.1	**72.9**	**82.2**	**92.1**	**98.8**
SGD	last	37.8	51.7	56.4	62.8	63.7	57.3	66.0
	all	38.2	51.8	57.3	63.8	63.3	69.7	70.6
MLP	last	39.7	55.8	60.7	67.0	71.3	78.7	83.7
	all	40.3	**56.0**	**61.4**	69.0	72.3	75.8	87.6
KNN	last	27.7	50.9	56.1	66.8	73.5	82.4	92.5
	all	29.3	49.7	55.5	65.0	72.4	83.4	95.1
LDA	last	39.7	53.8	57.9	62.0	64.4	66.7	69.3
	all	39.6	54.2	58.2	63.2	65.2	68.1	79.2
GNB	last	36.0	51.8	55.8	61.6	63.1	63.1	60.6
	all	35.6	49.4	54.2	60.6	62.4	63.4	62.6

high accuracy rates can be obtained, specially when for walks comprising more than 1000 steps. At the best scenario, the accuracy rate reaches 97.57%. When very short walks are considered, the generated sequence does not provide much discriminative information, which leads to typical low accuracy rates. The results suggest that it is possible to retrieve the originating model and dynamics by looking only to the reconstructed network properties, given that the sequence length is long enough.

In order to better understand the errors in the adopted classifiers, we analyzed confusion matrices. We selected the matrix corresponding to the results obtained with 500 steps because a higher number of steps usually leads to a very high accuracy rates. Conversely, very short walks usually leads to higher error rates mainly because almost all of the network can not be discovered with only a few steps. Similar results were also obtained when considering 1000 steps. The confusion matrix obtained from the classification of network models and dynamics are displayed in Figures 3 and 4, respectively.

The confusion matrix obtained for the network models reveals that the Wax model can be predicted with the highest accuracy. If a sequence is generated from a Wax network, it can be predicted with an accuracy of roughly 89%. Conversely, network with community structure are the ones generating the highest error rates. 22% of all sequences generated

TABLE III. Accuracy rate (%) for the classification considering both network models and walk dynamics: {ER, BA, Wax, LFR} × {RW, RWD, RWID, TSAW} (see Section III A). Columns represent the number of steps taken by the agents. When considering 5,000 steps in the walk, the best result is obtained with decision trees from the Random Forest classifier. Both the network topology and walk dynamics generating the sequence can be identified with an accuracy of 97.57%. The best results for each walk length are highlighted.

Method		10	50	100	500	1,000	2,000	5,000
DT	last	12.1	22.3	27.0	49.3	65.6	82.1	96.0
	all	**12.4**	18.2	24.5	49.9	65.4	81.9	96.3
RF	last	12.1	23.0	28.8	55.3	72.1	86.7	97.2
	all	**12.4**	23.2	32.2	**59.2**	**73.4**	**87.3**	**97.6**
SGD	last	7.9	16.4	19.6	32.3	30.7	35.8	29.2
	all	7.1	12.8	23.5	28.8	39.2	40.3	42.8
MLP	last	11.7	24.3	31.7	54.4	65.1	75.6	86.0
	all	12.1	**25.0**	**33.0**	55.1	63.4	74.3	88.4
KNN	last	8.3	19.6	26.7	46.0	57.4	71.7	87.8
	all	10.0	17.9	23.5	42.6	56.5	72.8	91.0
LDA	last	11.7	21.4	26.4	41.9	49.1	55.4	66.2
	all	11.9	21.9	27.2	43.0	50.3	58.6	74.8
GNB	last	10.0	21.1	27.1	43.6	47.8	48.8	49.2
	all	9.7	18.9	25.0	41.4	46.7	49.1	50.8

on LFR networks are classified as a BA network. While this can be explained by the fact that both BA and LFR networks present long tail distributions, they still have different mesoscale organization. A high error rate also associates LFR sequences as if they were generated on ER networks.

The confusion matrix for the agent dynamics classification (Figure 4) shows that RWID is the one with the highest accuracy: if the sequence is generated via a RWID walk, it can be recovered with an accuracy of 73%. In a similar fashion, RWD and TSAW had similar results in terms of general accuracy. Most errors in the RWID classification occurred as RW predictions, and the same behavior can be observed for RWD and TSAW, i.e., most of the mispredictions in this classifier occurred as RW guesses. Interestingly, random walks has a low accuracy. Its behavior is classified as RWD, RWID and TSAW dynamics with similar probability. In sum, the lack of accuracy in the classification of dynamics occur mostly when detecting that walk is the traditional unbiased random walk. This lack of correspondence might happen when the agent explores homogeneous region, causing thus no distinctive effect for degree-biased random walk. In a similar fashion, when the network is not fully explored, the true self-avoiding walk is seldomly applied and the TASW behaves like a RW

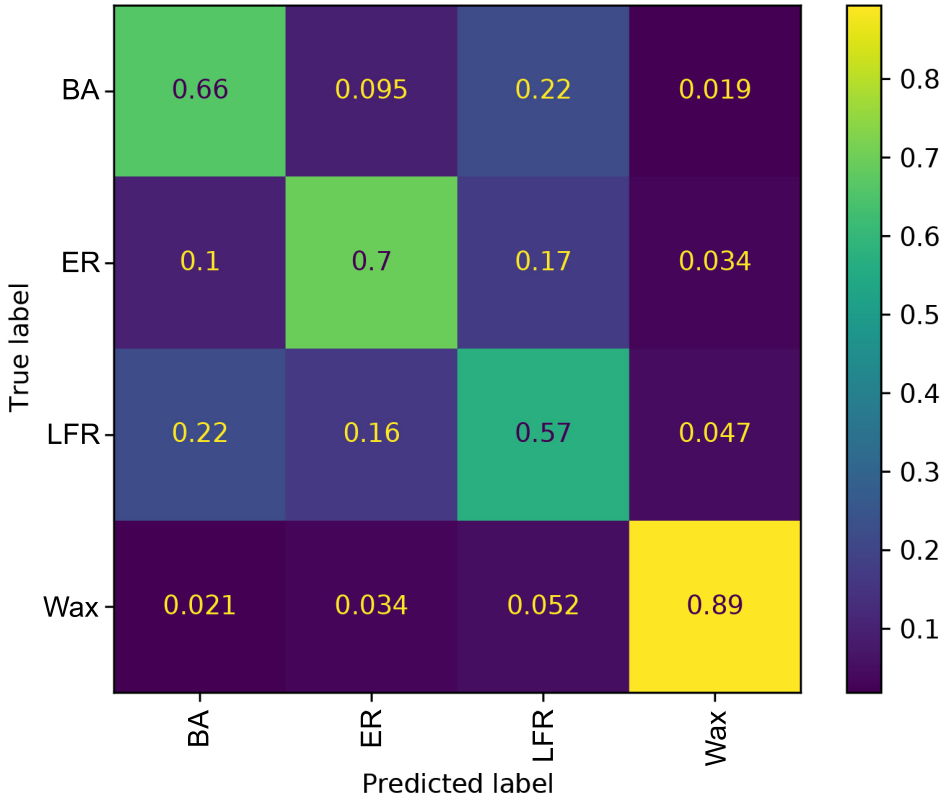


FIG. 3. Confusion matrix considering decision trees in the classification of sequences. The classification considered four network models: Erdős-Rényi (ER), Barabási-Albert (BA), Waxman (Wax) and Modular Networks (LFR).

walk.

In order to better understand the contribution of the adopted features (see Section III D) to recover the structure and dynamics generating the observed sequences, we probed features are the most relevant in the classification task. We used the Gini relevance index to compute the relevance of the features. This index is used in many different contexts [37, 43]. The most relevant features for all the sequences sizes studied for model and dynamics are shown in Figures 5 and 6, respectively. For each subplot, the features are sorted in decreasing order of importance. We only show the results for the "all" approach.

We found a similar behavior for all three classification scenarios. The degree property is a relevant feature for most of the sequence sizes, meaning that the relationship between vocabulary size (i.e. the number of different nodes discovered [24]) and the sequence length is a discriminative feature. Conversely, the shortest path property is typically the least relevant feature in larger sequences, while being among the most relevant features for smaller

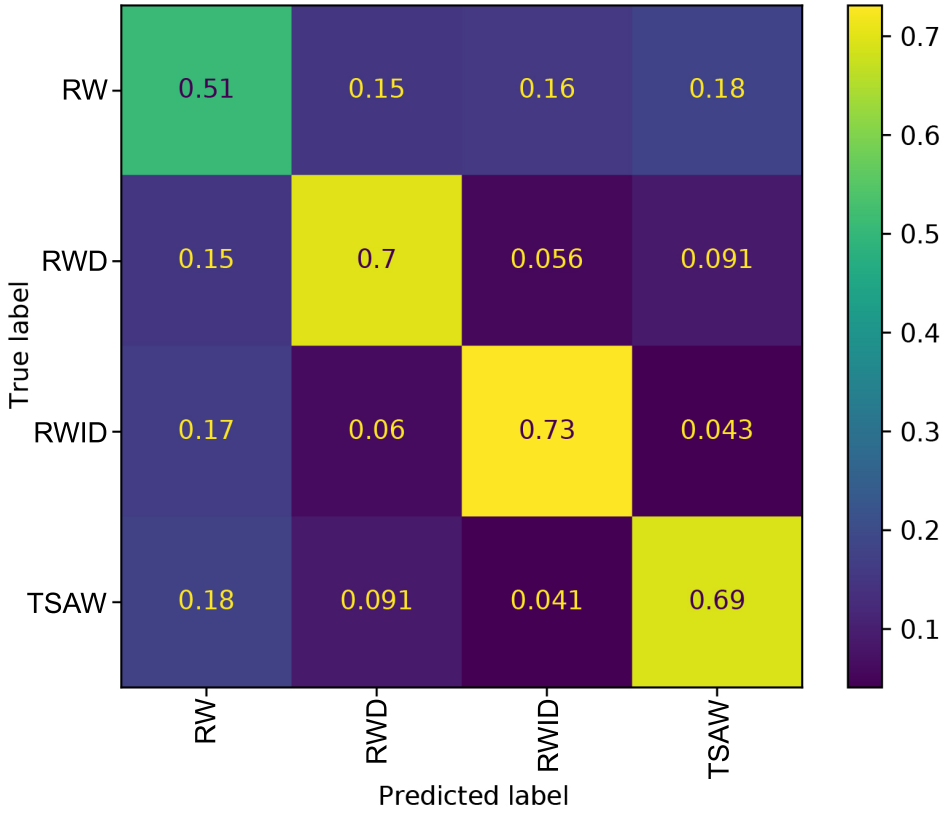


FIG. 4. Confusion matrix considering decision trees in the classification of sequences. The classification considered four network dynamics: random walk (RW), degree-biased random walk (RWD), inverse degree biased random walk (RWID) and true self-avoiding random walk (TSAW).

sequences. The total number of edges is not a relevant feature for smaller sequences, but it is one of the most important for larger sequences; therefore, as the network structure has more nodes, the number of edges becomes a significant feature. All in all, the results show that the best accuracy is obtained with larger sequences, and in those scenarios, local features such as the number of nodes, edges and clustering coefficient are among the most relevant features to identify the network model and agent dynamics generating the observed sequences.

V. CONCLUSION

In the current paper, we analyzed whether the observation of sequences as a result of agent walking over a network topology can provide accurate information regarding the process generating the sequence. We used the observed sequence to reconstruct a network via co-

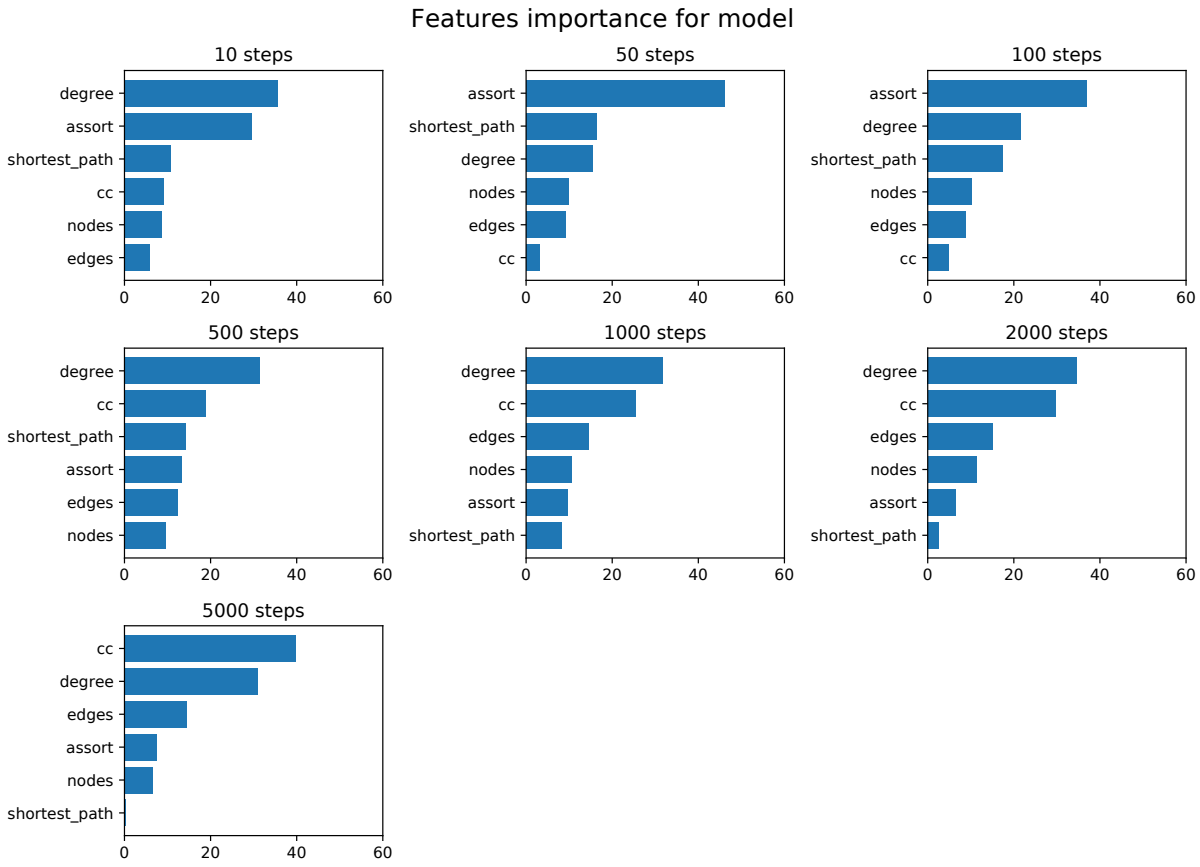


FIG. 5. Features importance for the classification of *network models*. The values in the x-axis are proportional to the features relevance in the classification task.

occurrence links, and then the observed network properties were used to infer both the original network models and the rule used by the agent to walk on the network. The properties of the reconstructed network were then used to characterize the reconstructed network in machine learning models. Our experiments were performed using 4 topological models and 4 random walk rules.

Our results revealed that one can predict both the network topology and agent dynamics with high accuracy provided that the observed sequence has a minimum length. When predicting only the network topology, the correct topology could be found with accuracy higher than 86% when less than 20% of nodes are visited. This accuracy increases to 93% when visiting less than 40% of nodes. When predicting the walk rule used by the agent, we found a slight lower accuracy. When 20% and 40% of all nodes were visited, we recovered the agent dynamics with an accuracy of 82% and 92%, respectively. Our models were are trained to infer both the topology and agent dynamics in the same model. In this case, the

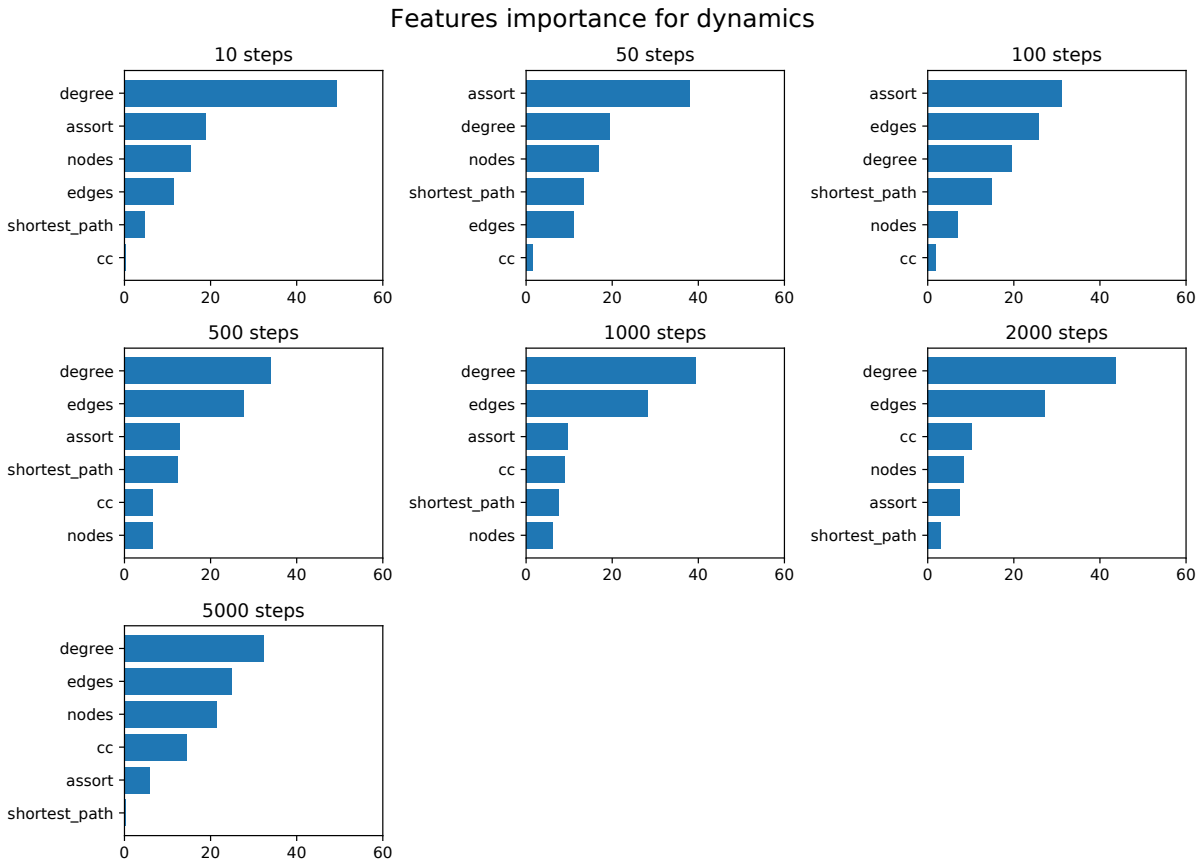


FIG. 6. Features importance for the classification of *network dynamics*. The values in the x-axis are proportional to the features relevance in the classification task.

model also displayed excellent performance, especially for walk length larger than 40% of the network. When considering shorter sequences, we found that distinct combinations of network models and walk dynamics result in different performances.

We showed, as a proof of principle, that it is possible to recover the generating model and dynamics from features extracted from reconstructed networks. Future works could address the problem of finding the best method to reliably recover the most suitable network structure and dynamics from sequences. This work has focused on the systematical analysis on artificial networks, and another question for future works is checking how robust the measures are across different strategies and sizes to reconstruct networks from sequences. In terms of applications, the developments of this work could be applied to identify patterns in real-world sequences, such as clickstreams [1] across the web, or from users exploring social media and video content. In such case, the dynamics performed sequences and those suggested by recommendation systems could be compared.

ACKNOWLEDGMENTS

D.R.A. acknowledges financial support from CNPq-Brazil (grant no. 311074/2021-9) and São Paulo Research Foundation (FAPESP grant no. 20/06271-0).

---

[1] L. Aguiar and B. Martens. Digital music consumption on the internet: Evidence from clickstream data. *Information Economics and Policy*, 34:27–43, 2016.

[2] C. Akimushkin, D. R. Amancio, and O. N. Oliveira Jr. On the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications*, 495:49–58, 2018.

[3] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. da Fontoura Costa. A systematic comparison of supervised classifiers. *PloS one*, 9(4):e94137, 2014.

[4] D. R. Amancio, O. N. Oliveira Jr, and L. da F Costa. Using complex networks to quantify consistency in the use of words. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(01):P01004, 2012.

[5] D. J. Amit, G. Parisi, and L. Peliti. Asymptotic behavior of the "true" self-avoiding walk. *Phys. Rev. B*, 27:1635–1645, Feb 1983.

[6] H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. Connecting network science and information theory. *Physica A: Statistical Mechanics and its Applications*, 515:641 – 648, 2019.

[7] H. F. Arruda, F. N. Silva, L. F. Costa, and D. R. Amancio. Knowledge acquisition: A complex networks approach. *Information Sciences*, 421:154 – 166, 2017.

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[9] K. Barat and B. K. Chakrabarti. Statistics of self-avoiding walks on random lattices. *Physics Reports*, 258(6):377–411, 1995.

[10] M. Bonaventura, V. Nicosia, and V. Latora. Characteristic times of biased random walks on complex networks. *Phys. Rev. E*, 89:012803, Jan 2014.

[11] L. Bottou. *On-Line Learning and Stochastic Approximations*, page 9–42. Cambridge University Press, USA, 1999.

[12] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[13] C. H. Comin, T. Peron, F. N. Silva, D. R. Amancio, F. A. Rodrigues, and L. F. Costa. Complex systems: Features, similarity and connectivity. *Physics Reports*, 861:1–41, 2020.

[14] E. A. Corrêa Jr, F. N. Silva, L. F. Costa, and D. R. Amancio. Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, 11(2):498–510, 2017.

[15] A. Dorle, F. Li, W. Song, and S. Li. Learning discriminative virtual sequences for time series classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2001–2004, 2020.

[16] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[17] P. Erdös and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[18] E. Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

[19] W. Fang, X. Gao, S. Huang, M. Jiang, and S. Liu. Reconstructing time series into a complex network to assess the evolution dynamics of the correlations among energy prices. *Open Physics*, 16(1):346–354, 2018.

[20] M. Fell and C. Sporleder. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 620–631, 2014.

[21] H. Ferraz de Arruda, F. Nascimento Silva, V. Queiroz Marinho, D. Raphael Amancio, and L. da Fontoura Costa. Representation of texts as complex networks: a mesoscopic approach. *Journal of Complex Networks*, 6(1):125–144, 2018.

[22] E. Fix and J. L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238, 1989.

[23] Z. Gao and N. Jin. Complex network from time series based on phase space reconstruction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(3):033137, 2009.

[24] L. Guerreiro, F. N. Silva, and D. R. Amancio. A comparative analysis of knowledge acquisition performance in complex networks. *Information Sciences*, 555:46 – 57, 2021.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[26] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, USA, 2nd edition, 1998.

[27] C. P. Herrero. Self-avoiding walks on scale-free networks. *Phys. Rev. E*, 71:016103, Jan 2005.

[28] Y. Kim, S. Park, and S.-H. Yook. Network exploration using true self-avoiding walks. *Phys. Rev. E*, 94:042309, Oct 2016.

[29] D. Koehn, S. Lessmann, and M. Schaal. Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, 150:113342, 2020.

[30] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975, 2008.

[31] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78:046110, Oct 2008.

[32] T. S. Lima, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. The dynamics of knowledge acquisition via self-learning in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(8):083106, 2018.

[33] H. Liu and J. Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.

[34] L. Lovász. Random walks on graphs: A survey. In D. Miklós, V. T. Sós, and T. Szőnyi, editors, *Combinatorics, Paul Erdős is Eighty*, volume 2, pages 353–398. János Bolyai Mathematical Society, Budapest, 1996.

[35] J. Machicao, E. A. Corrêa Jr, G. H. Miranda, D. R. Amancio, and O. M. Bruno. Authorship attribution based on life-like network automata. *PloS one*, 13(3):e0193703, 2018.

[36] M. M. Meerschaert and E. Scalas. Coupled continuous time random walks in finance. *Physica A: Statistical Mechanics and its Applications*, 370(1):114–118, 2006.

[37] S. Nembrini, I. R. König, and M. N. Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

[38] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986.

[39] R. D. S. Raizada and Y. Lee. Smoothness without smoothing: Why gaussian naive bayes is not naive for multi-subject searchlight studies. *PLoS ONE*, 8, 2013.

[40] T. Ramiadantsoa, C. J. E. Metcalf, A. H. Raherinandrasana, S. Randrianarisoa, B. L. Rice, A. Wesolowski, F. M. Randriatsarafara, and F. Rasambainarivo. Existing human mobility data sources poorly predicted the spatial spread of sars-cov-2 in madagascar. *Epidemics*, 38:100534, 2022.

[41] R. D. Rodrigues, L. Zhao, Q. Zheng, and J. Zhang. A tourist walk approach for internal and external outlier detection. *Neurocomputing*, 393:203–213, 2020.

[42] M. Stella, A. Kapuza, C. Cramer, and S. Uzzo. Mapping computational thinking mindsets between educational levels with cognitive network science. *Journal of Complex Networks*, 9(6):cnab020, 2021.

[43] J. A. Tohalino and D. R. Amancio. On predicting research grants productivity via machine learning. *Journal of Informetrics*, 16(2):101260, 2022.

[44] D. Watts. Strogatz-small world network nature. *Nature*, 393:440–442, 1998.

[45] B. Waxman. Routing of multipoint connections. *IEEE J. Sel. Area Comm*, 1:286–292, 01 1988.

[46] Y. Yuan, J. Yan, and P. Zhang. Assortativity measures for weighted and directed networks. *Journal of Complex Networks*, 9(2):cnab017, 2021.

[47] J. Zhang and M. Small. Complex network from pseudoperiodic time series: Topology versus dynamics. *Phys. Rev. Lett.*, 96:238701, Jun 2006.

CHAPTER

4

IDENTIFYING THE PERCEIVED LOCAL PROPERTIES OF NETWORKS RECONSTRUCTED FROM BIASED RANDOM WALKS

Identifying the perceived local properties of networks reconstructed from biased random walks. Lucas Guerreiro, Filipi N. Silva, Diego R. Amancio. *Information Sciences.* Current status: submitted.

4.1 Context

Our previous two works (Chapters 2 and 3) have covered the analysis of global behaviors of knowledge acquisition in complex networks. Henceforth, we also intended to understand the role played by properties in local contexts. Therefore, this paper investigates how properties of local nodes may indicate the behavior of the complex network as a whole. Our goal here is to infer how short can a sequence (i.e. walk) be in order to give information of the entire network. Thus, we propose a series of experiments that analyze the influence of sequence sizes and properties of these "partial networks". Nonetheless, the chosen random walk play an important role on such inference, and such agents are also studied in this paper.

In order to achieve our goal, we consider the impacts of biased random walks, which have been vastly studied in the network science field (AMIT; PARISI; PELITI, 1983; KIM; PARK; YOOK, 2016; LIMA *et al.*, 2018a; WANG; YANG, 2019). As previous stated, however, in our proposed study we have explored the context of partial reconstructions, which led to understanding how the information propagates among neighbors and how it influences the general conceptions of the network.

Summarizing the above discussion, the main motivation of this paper was to identify whether the perception on local properties can indicate the behavior of global properties.

4.2 Contributions

The main contributions of this paper can be divided into the framework itself and the results of the paper. The framework presented a perspective on the local properties obtained from a co-occurrence reconstruction of partial sequences. Thus, it can be generalized for different tasks on the analysis of sequences and knowledge acquisition.

Meanwhile, the results obtained from our experiments have unveiled the impact of different dynamics strategies on the recovering of network properties. Moreover, the proposed analysis allowed us to identify the role that the random walkers play on both short and long walks, which can pave the way that further studies may select which walking strategy to use for each different setting and situation.

Identifying the perceived local properties of networks reconstructed from biased random walks

Lucas Guerreiro[1], Filipi N. Silva[2] and Diego R. Amancio[1]

[1]*Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil*

[2]*Indiana University Network Science Institute, Bloomington, Indiana 47408, USA*

(Dated: November 15, 2022)

Abstract

Many real-world systems give rise to a time series of symbols. The elements in a sequence can be generated by agents walking over a networked space so that whenever a node is visited the corresponding symbol is generated. In many situations the underlying network is hidden, and one aims to recover its original structure and/or properties. For example, when analyzing texts, the underlying network structure generating a particular sequence of words is not available. In this paper, we analyze whether one can recover the underlying local properties of networks generating sequences of symbols for different combinations of random walks and network topologies. We found that the reconstruction performance is influenced by the bias of the agent dynamics. When the walker is biased toward high-degree neighbors, the best performance was obtained for most of the network models and properties. Surprisingly, this same effect is not observed for the clustering coefficient and eccentric, even when large sequences are considered. We also found that the true self-avoiding displayed similar performance as the one preferring highly-connected nodes, with the advantage of yielding competitive performance to recover the clustering coefficient. Our results may have implications for the construction and interpretation of networks generated from sequences.

I. INTRODUCTION

Many real-world phenomena are characterized by discrete series of events or decisions happening in succession [3, 34]. This includes how users navigate through websites or social media, how language is written and spoken, music, city navigation, and even people's everyday decisions. In most cases, however, only a limited amount of information is available to infer the rules and the mechanisms driving the generative processes behind these phenomena. For instance, from the perspective of a social media user, the observed content may be limited to their own interests, political positions, friends' preferences and what is being suggested by a recommendation algorithm. Such content normally only constitutes a small fragment of what is present in the complete social media platform. These aspects are often linked with the emergence of biases leading to polarization, formation of echo chambers, and other social phenomena like the friendship paradox [8]. In another example, because of limited individuals' capacity, resources and available personnel, scientists or research groups adopt different strategies to choose the focus of their research among all the possible problems. Such a strategy could favor exploitation over exploration (or vice versa), a decision that could potentially impact the collective discovery process [33]. These examples raise the question on how well aspects of the inherent (generative) process are truly recovered through limited or biased information.

Since many complex systems have been successfully represented by networks (i.e., by the intricate relationships among their components) [1, 10, 23, 36], it is possible to study the aforementioned question in terms of how well the characterization of such structures changes according to the limited information observed through certain dynamics. In particular, for the case of networks, different behaviors of dynamics can be simulated through random walk heuristics. In such systems, an agent performs a walk in a network and reconstructs it based on the set of visited nodes and edges. This process has been investigated, in particular for the case of the knowledge acquisition process, in which pieces of knowledge are learned by agents walking across a network representing knowledge. Previous works [21] have found that it is possible to determine characteristics of the inherent model by only looking at features of the partially reconstructed networks. This indicates that different combinations of network topology and dynamics can lead to potentially different observed features in the generated sequences. In this work, we address the problem of checking how similar are the

observed features of partially reconstructed networks compared to the original structure.

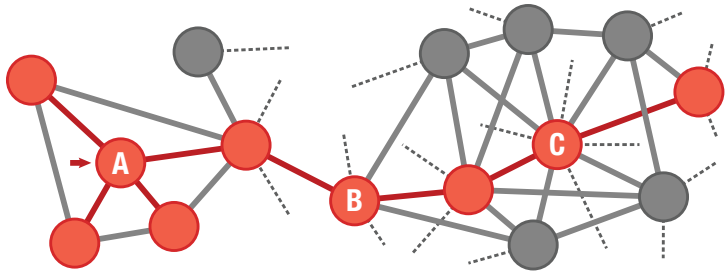


FIG. 1. Example of a subgraph (highlighted in red) representing the limited information observed by a random walk starting at node A.

We approach the problem of reconstructing networks from limited information by employing different types of random walks performed by non-interacting agents. Each agent simulates an individual with limited information and stores a subgraph of the original network reconstructed by co-adjacency. Figure 1 shows an example network in which a random walk was performed starting at node A. This simulates, for instance, a user in social media navigating across different profiles. Given the user's limited information, they may think that node A is the one with most connections in the network, in contrast to the correct answer: C. This is because the agent visited A's neighborhood, while B's and C's neighbors were not. Other properties such as clustering coefficient and centrality measures are also not correctly recovered through this walk. The potential to recover network properties may also depend on the network topology itself. For instance, an irregular network with heterogeneous degrees and high density may be more challenging to navigate than a regular network, since in the first case, hubs may be visited more frequently, potentially leaving regions of low degree poorly explored. As a consequence, the recovered network characteristics may be different and biased compared to those from the original network.

In this work, we study how well network characteristics – such as node degree, clustering coefficient, etc – can be recovered from reconstructions based on finite sequences. We explore the effects of different strategies to generate the sequences, including biased [21] and true self-avoiding [2, 22] random walks. First, we generate sequences of nodes based on the progression of visited nodes given by an agent dynamics. Next, the sentences are used to reconstruct independently a network based on co-adjacency. Network properties for both the original and reconstructed networks are obtained and compared via correlation. We also vary the length of the sequences to simulate different levels of limited information. For this

analysis, we considered real-world networks and realizations of traditional network models in addition to a community-based model, the LFR [24].

Our results indicate that the choice of dynamics employed to generate the sequences has an influence on the correlation values between the recovered and original network properties. The reconstruction performance depends, for instance, if the dynamics are biased by node degree. When highly connected nodes are preferable to be visited (RWD), we achieve the best performance in recovering network properties for most of the considered networks and properties, with the exception of clustering coefficient and eccentricity. In those cases, even by considering the long sequences, it still reaches low values of correlation. On the other hand, for the case that the random walk dynamics avoids highly connected nodes (RWID), we see the worst performance among the considered dynamics. However, it is able to recover the clustering coefficient with similar performance as other dynamics. In addition to that, we explore three other types of random walks, the unbiased random walk (RW) and two self-avoiding strategies, known as true self-avoiding walk [2, 22], one based on edges, which avoids passing through already visited edges(TSAW-edge), and another based on nodes (TSAW-node). TSAW-edge displayed similar performance as the RWD approach, but with no problems in recovering the clustering coefficient. We discuss these results in detail and the potential implications in section IV.

Finally, we also check if the community structure can be recovered from the partial information stored in sequences. This is accomplished by comparing the detected communities' membership of the original networks (or planted for the LFR models) with those from the reconstructed versions. The results seem to depend strongly on the network topology, with mixed patterns across different mixing coefficients of the LFR and real networks. Nonetheless, TSAW-edge and RW display the best performance in that task.

II. RELATED WORKS

The process of random walkers exploring complex network topologies has already been studied by several works [3, 12, 20, 26]. In the context of knowledge acquisition, the sequence of visited nodes in random walks is to recover the set of nodes in the network [3, 12]. In [3], the authors investigated how different agents walking over the network can reconstruct the network topology. In the proposed multi-agent random walk, the true self-avoiding and Lévy

flight-based dynamics outperformed other walk strategies in terms of efficiency in discovering new nodes. Surprisingly, the study also showed that fine-tuning the parameters controlling the agent dynamics had little effect on the global knowledge acquisition performance.

The study conducted in [20] focused on the knowledge acquisition task when several network topologies and agent dynamics are used in a single-agent context. This study found that the true self-avoiding dynamics had the best performance over different settings in discovering nodes in the network. The degree-biased had the slowest learning curve in the experiments. The study has also demonstrated that higher average degrees provide a faster learning rate.

While several studies focused on the knowledge (nodes) acquisition problem [12, 26], the study conducted in [21] used a machine learning approach to recover both the network topology and agent dynamics generating a sequence of symbols. To train the supervised classifiers, sequences of visited nodes were mapped into (reconstructed) networks via the co-occurrence strategy. Then, six different network properties were used to create features describing the observed reconstructed networks. Sixteen different combinations of network topology and agent dynamics were considered to generate sequences. The study revealed that it is possible to recover both the topology and dynamics with high accuracy, provided that the sequence (i.e the random walk) length is not too short. The accuracy of identification increased with the observed sequence length. When less than 20% of the whole network was discovered, both the topology and dynamics were recovered with an accuracy higher than 86% in a supervised classification scenario with 16 classes.

In [25], the authors analyzed how network properties (e.g. average degree) evolve as the network sample size grows. If a network property is unstable for all sample sizes then it does not represent the network very well; however, if the property does not change as the sample size grows, then the property is considered a good representation of the network. The main contribution of this work is therefore a methodology to quantify if any network property is robust regarding the network size used in the experiments. In networks formed from sequences, it means that unstable properties may vary depending on the sequence length used to form the networks. This means that when recovering local properties, the original value of the property may only be recovered if the sample and original network sizes are consistent. However, one may still find a correlation between values observed in sampled and original networks for unstable metrics.

The study conducted in [23] investigated a teaching-learning perspective using complex networks. In the adopted representation, facts are graph nodes and the relationship or underlying connections between two facts are represented by edges. The study aimed to probe how students learn contents from linear algebra textbooks by considering the nodes exploration process simulating the human memory characteristics during the learning process. Among the main findings, the authors reported that human memory limitation plays a special role in long-term information retention effectiveness and problem-solving creativity.

The relationship between knowledge representation and complex networks has also been studied elsewhere. In [12], knowledge is acquired when nodes and edges are visited by random walkers. Different from other approaches, the experiments considered free and conditional transition edges. While the former is commonly used in most of the works, the latter allows new nodes to be accessed only when certain criteria are met. In this case, the main criteria consist in visiting a subset of nodes in order to make a new node accessible. The author analyzed the knowledge acquisition performance via hierarchical complex networks [11], which are explored via traditional random walks and variations biased toward new links. The study showed that the biased random walks are slower to acquire knowledge in the conditional exploration scenario.

While most of the related works tried to recover the set of nodes or identify the dynamics and topology generating a sequence, here we focused on a different network perspective. We studied if the properties of the reconstructed networks are consistent with the ones of the original networks. This is a relevant topic since in many scenarios one does not have access to the original networks generating a sequence of symbols.

III. METHODOLOGY

In this section, we discuss the proposed methodology. The main purpose of this paper is to analyze whether the local properties of reconstructed and original networks are correlated. The methodology can be divided into the following 4 main steps, which are summarized below. The steps are also illustrated in Figure 2.

- *Original networks*: here we used network models to represent different network topologies. Examples of models include random and geographical networks. We also used examples of real-world networks modeling e.g. social and biological complex systems.

- *Network dynamics*: in many real-world situations, network data is only available as a sequence of symbols [14]. Sequences can be generated by an agent walking over the network via different rules.

- *Network reconstruction and properties extraction*: the observed sequence generated in the previous step is used to reconstruct. Several properties of the obtained networks are then extracted.

- *Correlation analysis*: the properties of the original and reconstructed networks are compared. The properties are compared in terms of network metrics (e.g. clustering coefficient) and, in modular networks, the partitions representing the network communities are compared.

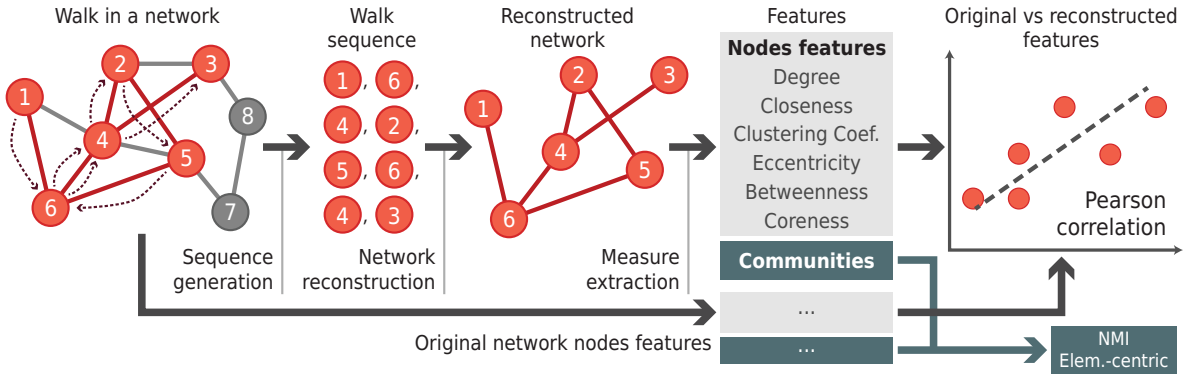


FIG. 2. Schematics of the methodology. First, we look into the original network and annotate the concerned properties for each node. In the next step, we iterate over the network with the desired dynamics in order to generate a sequence of symbols. In the third step, we reconstruct the network using the discovered nodes and edges, and we annotate the node's properties in this reconstructed network. Finally, we build correlations among the reconstructed and original nodes by comparing each node i in the reconstructed network to its respective node in the original network.

A. Original networks

We considered the most common and diverse topology models in the literature [30]. Similar to previous studies [21], our analysis focused on networks with $N = 5,000$ nodes and average degree $\langle k \rangle = 4$. The following models were used here:

- *Erdős-Rényi (ER)*: this is the traditional random network model. The probability π of two nodes being linked by an edge is constant [16, 17].

- *Barabási-Albert (BA)*: this model implements a scale-free topology [5]. Different from ER networks, the BA model is based on a growth model, since at each step new nodes are included in the network. The probability π_{ij} of a new node i to connect to an old node j with k_j links is proportional to k_j:

$$\pi_{ij} = \frac{k_i}{\sum_j k_j}. \tag{1}$$

- *Waxman (Wax)*: this topology is an implementation of a geographic network. The first step consists in randomly placing each node in a two-dimensional plane. A link between two nodes i and j is provided by a probability formulation that decays exponentially with the geographical distance between the nodes [20, 38].

- *Modular Networks (LFR)*: this model [24] creates networks with nodes clustered into network communities. The main parameters used to construct modular networks are the number of communities (n_C), the exponent for the degree sequence in the network (t_1), the minus exponent for the community size distribution (t_2), and the mixing parameter (μ), which measures how well defined communities are. Lower values of mixing value lead to well-defined network communities. We used the following parameters to construct the networks: $n_C = 5$, $t_1 = 3$, $t_2 = 0$ and $\mu = \{0.05, 0.2, 0.8\}$. Similar values have been used in related works [3, 20, 21],

We also conducted our experiments in real networks modeling diverse complex systems:

- *Facebook*: this network comprises social relationships among Facebook employees. The network comprises 320 nodes [18].

- *Power Grid*: this is a classic geographical network modeling the US Western States Power Grid. Nodes represent transforms or power relay points, while edges are power lines. The network comprises 4,941 nodes [37].

- *Econ-Poli*: we have used Economics Poli network which contains 3,915 nodes and presents behavior on interconnected economic agents [32].

- *Web-EPA*: we also have used the Web-EPA network with the size of 4,271 nodes and implements information in web level for hyperlinks across the internet that link to the *www.epa.gov* website [32].

- *Bio (DM-CX)*: The Bio (DM-CX) is a biological real network that represents co-occurrences on *Drosophila melanogaster* fly pairs of genes acquired from the FlyNet repository. The network comprises 4,032 nodes [6, 32].

- *socfb-JohnsHopkins55*: this is a social network representing Facebook connections inside the Johns Hopkins community. The network comprises 5,180 nodes [32].

B. Network dynamics

Agent dynamics have been used in a wide variety of network-based studies, including epidemic spreading and knowledge acquisition analysis [3, 20, 26]. In this work, agent dynamics are used to explore the network and generate a sequence of visited nodes. The sequence of nodes is assumed to be the information available for network reconstruction []. We have used 5 well-known walks:

- *Traditional Random Walk (RW)*: in the traditional random walk the agent selects the next node to visit randomly among its neighbors. The probability of the agent moving from node i to node j is $p_{ij} = 1/k_i$, where k_i represents the degree of i-th node.

- *Degree-biased Random Walk (RWD)*: this random walk considers the degree of the neighbor nodes when selecting the next node to be visited by the agent. The probability of visiting a node is proportional to its degree:

$$p_{ij} = \frac{k_j}{\sum_{l \in \Gamma_i} k_l}, \tag{2}$$

  where Γ_i is the set comprising the neighbors of i.

- *Inverse of the Degree-biased Random Walk (RWID)*: Similar to the RWD dynamics, the RWID walk uses the degree of the neighborhood when defining the probabilities. However, in this dynamics, the agent prefers to visit the nodes with smaller degrees, i.e.

$$p_{ij} = \frac{k_j^{-1}}{\sum_{l \in \Gamma_i} k_l^{-1}}. \tag{3}$$

- *True Self-avoiding Random Walk on nodes (TSAW-node)*: in this random walk, the agent avoids the nodes that were already visited [2, 22]. Therefore, in this network, the

nodes not yet visited are preferred to be visited, which works in favor of the network exploration. Let f_j be the frequency that j has been visited. The mechanism to avoid already visited nodes is encoded according to:

$$p_j = \frac{e^{-\lambda f_j}}{\sum_{l \in \Gamma_i} e^{-\lambda f_l}}. \tag{4}$$

- *True Self-avoiding Random Walk on edges (TSAW-edge)*: similarly to the traditional TSAW dynamics, in this random walk there is an avoiding bias, however in this implementation, the agent avoids edges previously visited instead of nodes, as implemented in works related to network exploration [3, 20]. The probability transition is computed as

$$p_{ij} = \frac{e^{-\lambda f_{ij}}}{\sum_{l \in \Gamma_i} e^{-\lambda f_{il}}}, \tag{5}$$

  In both versions of the true self-avoiding random walks, we are using $\lambda = \ln 2$, as in related works [3, 20, 21].

C. Network reconstruction and properties extraction

The sequences generated by the random walks are used to reconstruct the networks. In our experiments, we probed how the sequence length affects the properties of the reconstructed networks. We investigated the results for the following set of sequence length w: $\{100, 200, 400, 500, 600, 800, 1000, 2000, 5000, 20000, 50000\}$. The reconstruction is performed by recreating the edges observed in the sequence. This procedure is equivalent to the co-occurrence approach usually employed in network analysis, where two symbols are linked whenever they are adjacent in the sequence. When analyzing texts using network science, the co-occurrence approach is widely employed [1, 9, 27]. For each combination of network topology, agent dynamics and walk size, we considered 20 sequence realizations.

Once the network is reconstructed, our aim is to analyze if relevant properties can be recovered. To characterize the networks, we used well-known network metrics, including local, quasi-local and global metrics. We computed the degree, clustering coefficient, closeness, betweenness eccentricity and coreness centrality of the networks [13, 30]. All metrics are defined for unweighted and undirected networks. For networks with modular structure, we

also detect the communities using the Leiden method [35].

D. Correlation analysis

In order to evaluate how well the structural properties are preserved in the reconstructed networks, we evaluated the correlation of the observed metrics for the same node. For each node $i^{(R)}$ we measure a network property $\mu(i^{(R)})$ (e.g. degree or clustering coefficient). The same property is also measured for the corresponding node in the original network, i.e. $\mu(i^{(O)})$. Finally, we measured the correlation between $\mu(i^{(R)})$ and $\mu(i^{(O)})$, for all nodes of the reconstructed network.

Since we are also interested in the modular structure of networks, we also compared the similarity of partitions in the original and reconstructed network via normalized mutual information (NMI) [28, 30]. Higher values of normalized mutual information mean that the partitions of the original and reconstructed network are similar.

IV. RESULTS AND DISCUSSION

In Section IV A, we analyze if the local properties of the original networks are preserved for distinct biased random walks. In Section IV B, the efficiency in recovering local properties is analyzed in the context of the knowledge acquisition task [3].

A. Efficiency in recovering the original properties

In our first analysis, we intended to probe whether the properties of the reconstructed networks are consistent with the properties of the original ones, according to the procedure described in Figure 2. The scatter plot of the node degree observed in original and reconstructed networks is shown in Figure 3 for the five agent dynamics in the LFR network (with mixing parameter $m = 0.05$). Each column represents a different sequence length (chosen proportionally to the original network length), while lines are different agent dynamics. For each subpanel, we show in the x- and y-axis the node degree observed in the reconstructed and original networks, respectively. We also show in each subpanel the Pearson and Spearman correlations (C_p and C_s, respectively).

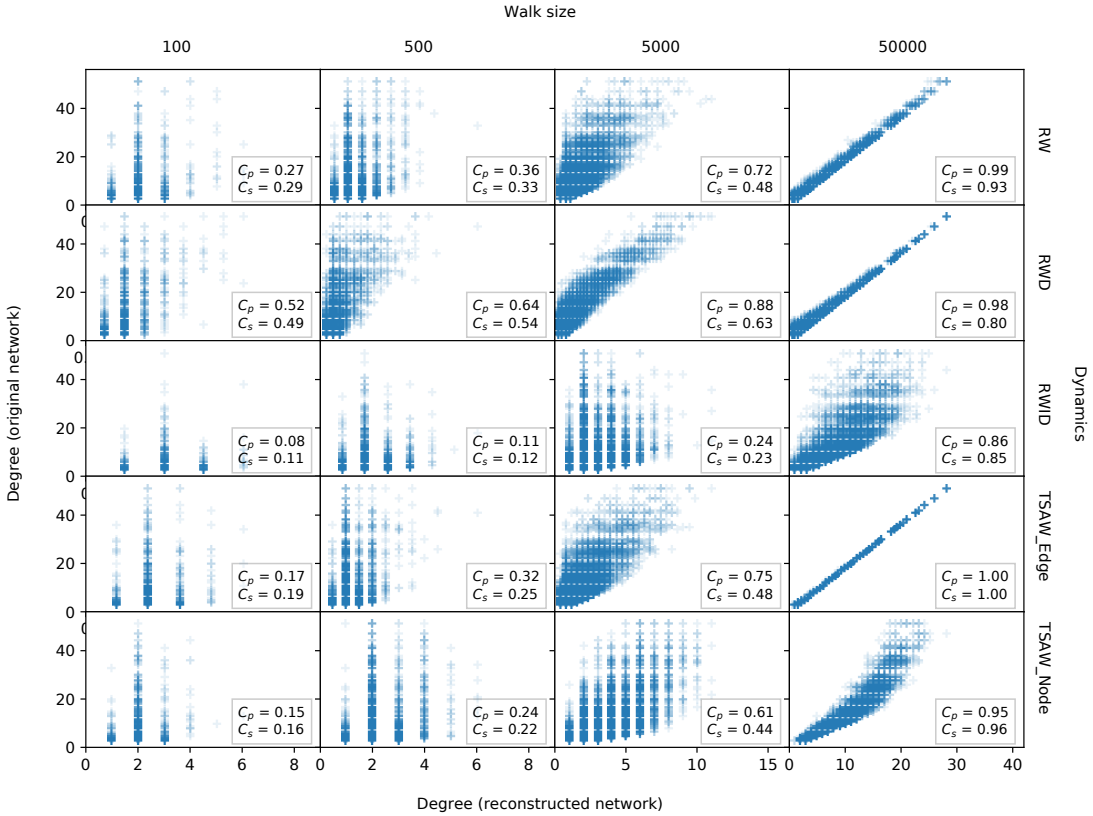


FIG. 3. Scatter-plot of node degree obtained in original vs. reconstructed LFR networks (with mixing parameter $m = 0.05$). Each subpanel corresponds to a different configuration of agent dynamics and walks length (w). The walk length is distributed between 100 and 50,000 steps.

We notice that for small sample sizes ($w = 100$ and $w = 500$) the node degree property is not well represented since the agent did not acquire enough information to create an accurate representation of the original network. This might be an explanation of previous results showing that networks reconstructed via very short walks are not consistent with their original topological nature [21]. Interestingly, larger sample sizes not necessarily imply high correlation values. Considering the RWID agent dynamics and $w = 5,000$ steps, the Pearson correlation, C_p, reaches only 0.24, even though higher values were observed for the other considered dynamics. Considering the same walk length, we found $C_p = 0.88$ for the RWD walk. This means that the node degree is consistent (i.e. linearly correlated) with the ones observed in the original networks.

For large values of w, i.e. long sequences, one should expect that most of the network structure (nodes and links) is retrieved by the agents [3]. Therefore, the correlations should

reach high values. In fact, this is observed in the TSAW Edge dynamics ($C_p = 1$). This means that this particular TSAW is not only efficient in knowledge acquisition, but it also captures the local connectivity [3]. RWD outperformed RW in this particular network. Finally, it is clear that after $5,000$ steps, the RWID walk may recover relevant information regarding node connectivity but it does not perform as well as the other dynamics.

While in Figure 3 we focused on a scatter-plot analysis of a single network model, the scatter plot for other network models are similar to the one shown for the other LFR networks (results not shown). The behavior of the correlations in different agent dynamics and network topologies is shown in Figure 4. The figure illustrates the evolution of the Pearson correlation for distinct sequence lengths. We also show the NMI, comparing the structure of communities found in the LFR original and reconstructed networks.

The results in Figure 4 reveal that the RWID dynamics displayed the lowest correlation values in almost all scenarios. This is especially true in networks where hubs play a prominent role, as is the case of BA and LFR networks. Therefore, avoiding hubs in networks where hubs are relevant causes a distortion in network metrics observed in reconstructed networks. Conversely, the RWD dynamics displayed competitive performance, even in networks with no evident presence of hubs (see e.g. degree in ER networks). The efficiency of RWD is more evident in BA and LFR networks, even in short sequences. The RW dynamics displayed a performance that is similar to the TSAW Edge walk. A major difference was found only for particular metrics, e.g. the eccentricity in ER networks for long sequences. Interestingly, we found that major differences in performance can be found for different versions of the TSAW walk. This is the case of the betweenness in BA networks. For walk lengths larger than $2,000$ steps, the true self-avoiding rule applied on edges turned out to be more efficient than the same rule applied on nodes.

The efficiency of metrics recovery has a minor dependency on the walking strategy when considering the clustering coefficient, coreness and the NMI metrics. For the particular case of networks with community structure (i.e. LFR networks), we noticed that there is no evident differences in performance in networks with high values of mixing parameter. The differences in performance are only evident for well-defined communities, i.e. for networks with mixing parameter lower than 0.20. However, the NMI reaches roughly 0.50 even after long walks. In well-defined communities (mixing parameter = 0.05), we found that the structure can indeed be recovered. However, a large number of steps is still required to

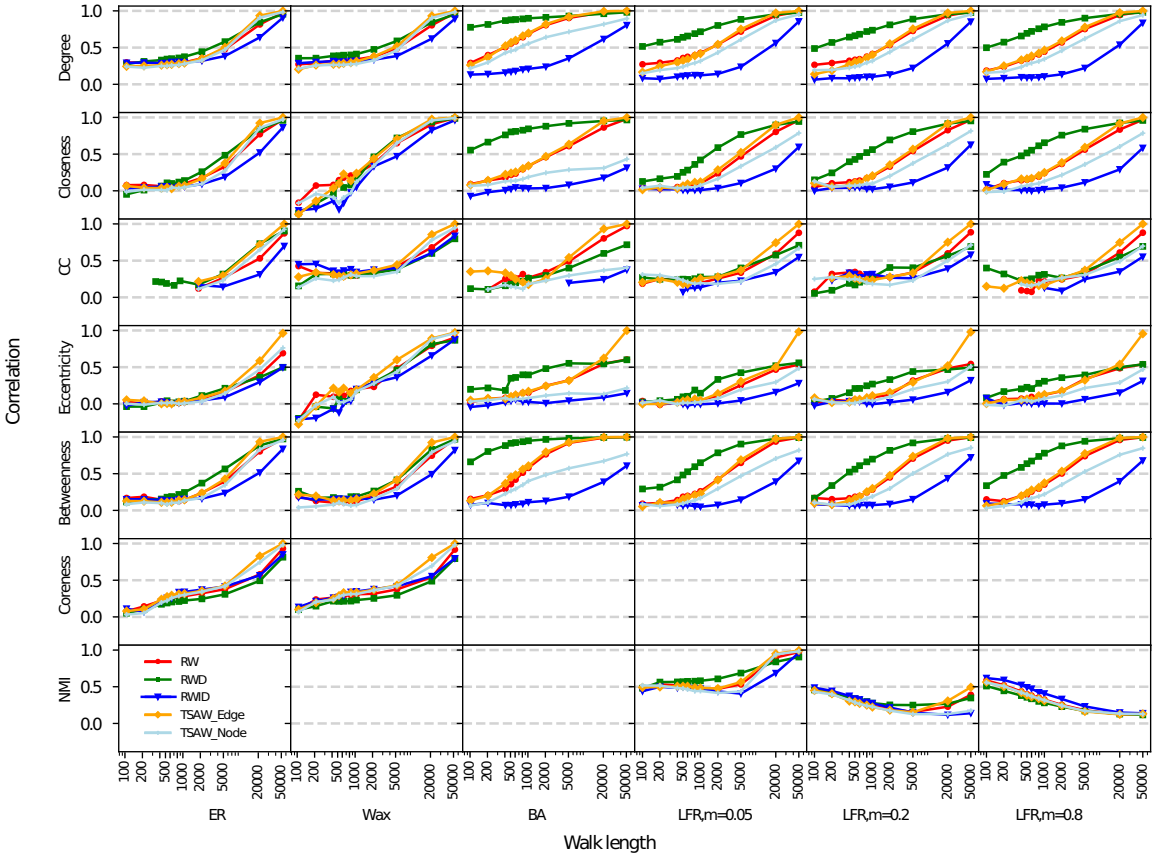


FIG. 4. Evolution of the efficacy in recovering network metrics in reconstructed networks. The x-axis represents the walk length used to generate the sequence, and the y-axis is the Pearson correlation for local metrics obtained in the original vs. reconstructed networks. The last column is the NMI, which is used to compare the partitions in networks with community structure. While the coreness is not defined in BA and LFR networks, the NMI is computed only for networks with community structure.

achieve high performance.

In Figure 5 we show the efficiency of the recovery for real-world networks obtained from systems of different disciplines (as described in Section III A). We observe the same overall behavior for both RWD and TSAW Edge dynamics. In most cases, the RWD outperforms other approaches for short sequences, while the TSAW dynamics performs better when longer sequences are considered. We can also notice that, again, the RWID dynamics had the worst comparable performance in terms of correlation over the original network for most properties.

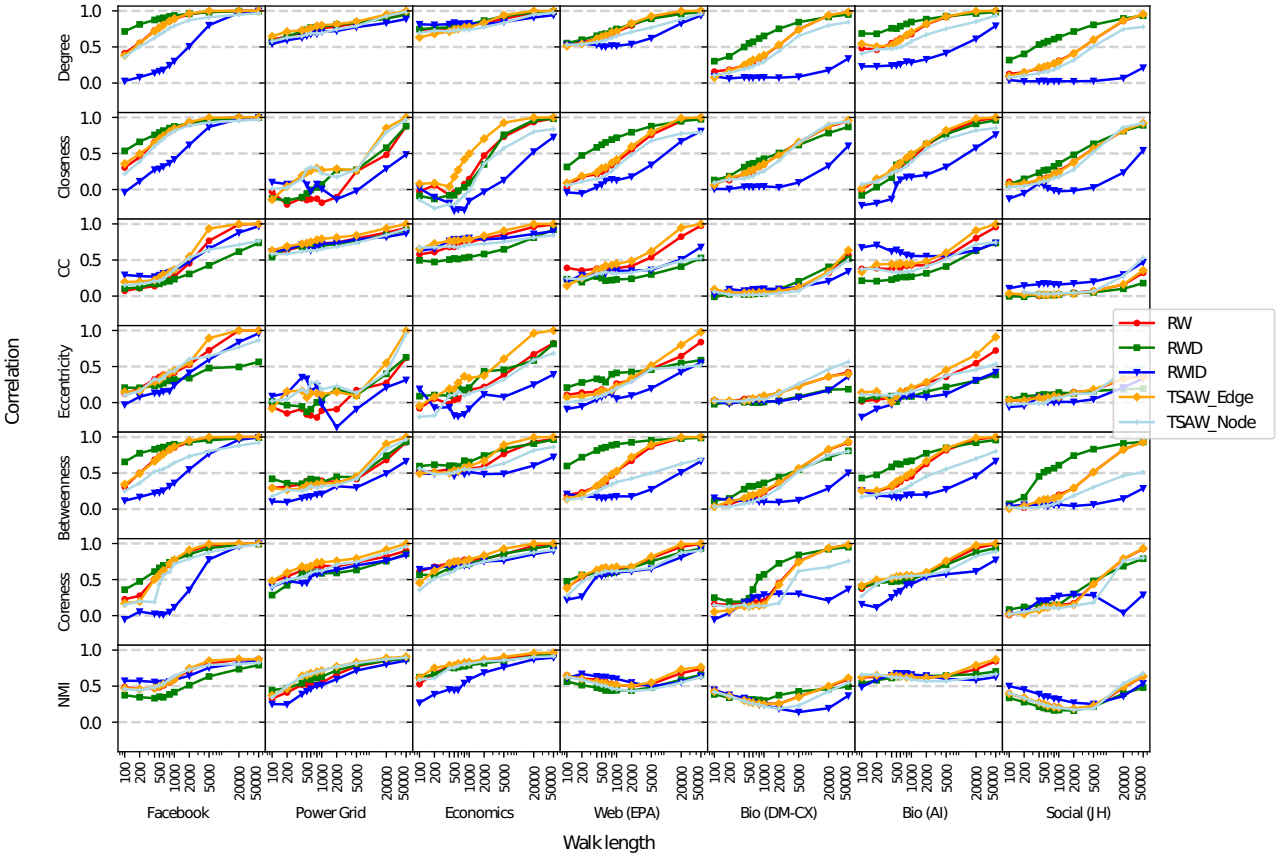


FIG. 5. Pearson correlation for the selected real networks. Each data point represents the correlation of the same nodes for the original and reconstructed network from sequences acquired by dynamics for a given walk length, w ($100 \leq w \leq 50000$). The last column represents the average NMI of each model obtained with the Leiden algorithm.

When comparing the efficiency of properties recovery across different networks, the Economics and Power Grid networks have almost all of the considered properties recovered with high correlation for sequences comprising more than 2,000 nodes. Conversely, for some properties in both Bio (DM-CX) and Social (JH) networks, 50,000 steps were not sufficient to recover the original metrics with high efficiency. This is the case of the clustering coefficient, eccentricity and coreness. Concerning the different network properties, the community structure could be recovered with efficiency only for three networks (according to the NMI metric). Interestingly, we can see that neither walk outperformed the others substantially regarding the recovery of network structure.

B. Efficiency in recovering network properties and knowledge acquisition

While in the previous section we focused on analyzing the recovery of network properties, here we also consider the knowledge acquisition performance as an additional feature of the random walks [20]. In the *knowledge acquisition* task, each node is considered as a piece of knowledge, and the performance metric corresponds to the fraction of the total number of nodes that have been discovered in the reconstructed network in comparison with the original network [20]. In this context, we analyze whether the reconstructed network is a good representation of the original ones in a twofold fashion: (i) the correlation of the properties of the discovered nodes; (ii) the computation of how many nodes from the original networks have been recovered.

In Figure 6, in the x-axis, each point represents the knowledge acquired by the walkers for a given sequence length L, i.e. we show the fraction of unique nodes discovered for that sequence length. In the y-axis, we show the correlation of properties obtained for the same value of L, according to the methodology described in Section III C. One may notice that, in most scenarios, the RWD dynamics improved the recovery correlation as more nodes are discovered. In particular scenarios, high correlation values are reached in the first steps of the walk, however, many more steps are required to discover a significant portion of the network (see e.g. the closeness metric for the BA network). We also note that the knowledge acquisition and correlation performance may increase with a similar speed – this is the case e.g. of the clustering coefficient in BA networks). As for the RWID dynamics, we observe that, in many cases, even when a large portion of the network is discovered, a low correlation is found (see e.g. the closeness centrality in BA networks). As for the TSAW, in most cases, the edge-based version presented a higher correlation when the same amount of nodes was discovered. This is evident, for example, when recovering the betweenness in BA networks. When half of the network is discovered, the edge-based version presents an almost perfect correlation, while the node-based version only achieves a correlation value close to 0.50.

We have also analyzed both correlation and knowledge acquisition relationships in real-world networks. The results are shown in Figure 7. In the Facebook network, for a fixed amount of discovered nodes, the highest correlation is mostly achieved with RWD dynamics. Both clustering and eccentricity metrics This result is compatible with the behavior of BA networks. Surprisingly, in the Power Grid, Economics and BIO (AI) networks, there is

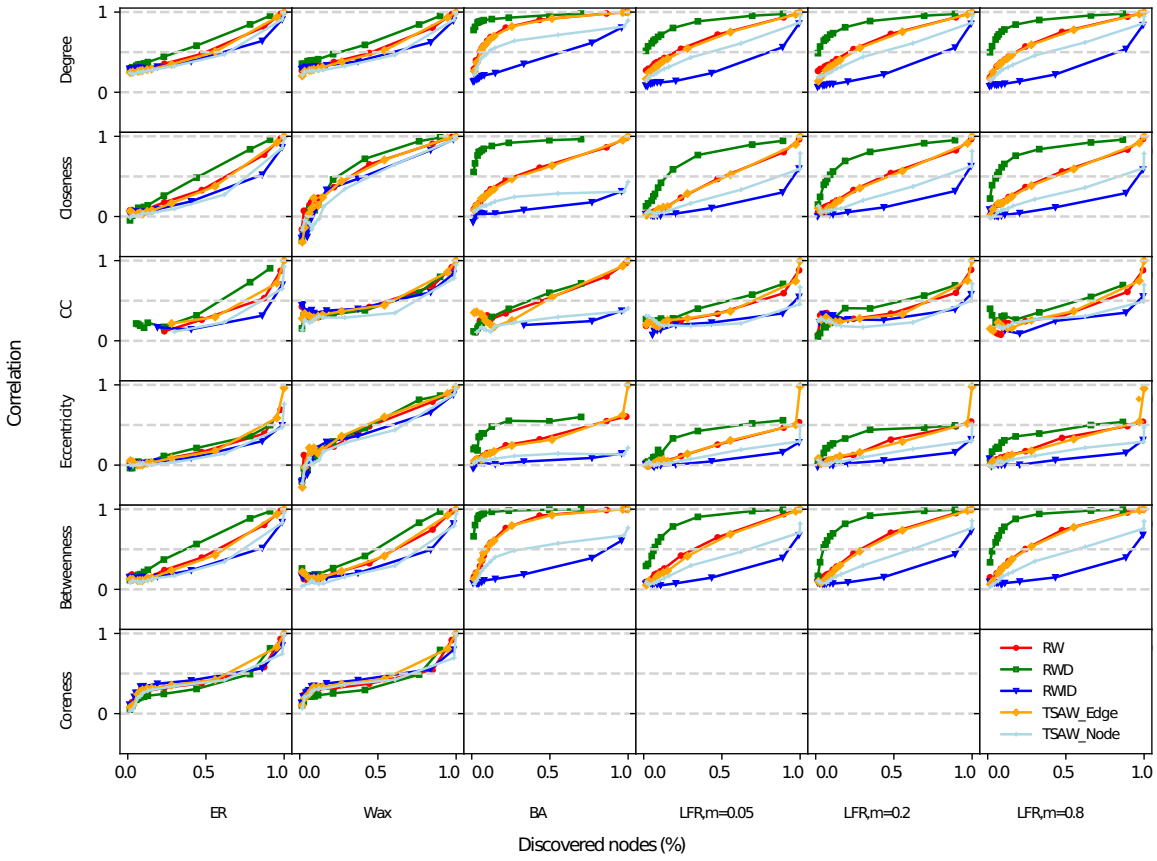


FIG. 6. Efficiency in recovering network metrics in for the network models. Each data point represents the correlation of local network metrics between the original and reconstructed network from sequences acquired by dynamics for a given walk length, where the x-axis represents the knowledge acquired for each walk length (w).

no evident difference in the behavior observed for distinct random walks for most of the considered metrics. The RWD again seems to provide the highest values of correlation when the same number of nodes is discovered for shortest paths-dependent metrics (closeness and betweenness) in the Web network. Finally, we note that RWD also achieved the highest accuracy for the degree, closeness and betweenness in the social network. Finally, in almost all metrics and networks, we again observed that the TSAW-edge strategy is more efficient in recovering nodes' properties than the nodes-based counterpart.

All in all, the results indicate that the RWD agent dynamics performs equally or better than the other walk dynamics when recovering local metrics when the same fraction of nodes

is discovered. We should keep in mind, however, that the RWD is outperformed by other random walk strategies in the knowledge acquisition task [3, 20]. In other words, while the RWD is efficient in recovering nodes' properties, it takes longer to discover new nodes. Another interesting finding is that many of the real-world networks displayed a behavior that is similar to the one observed for the respective models. This is the case of the Facebook network, which displayed a behavior consistent with BA networks.

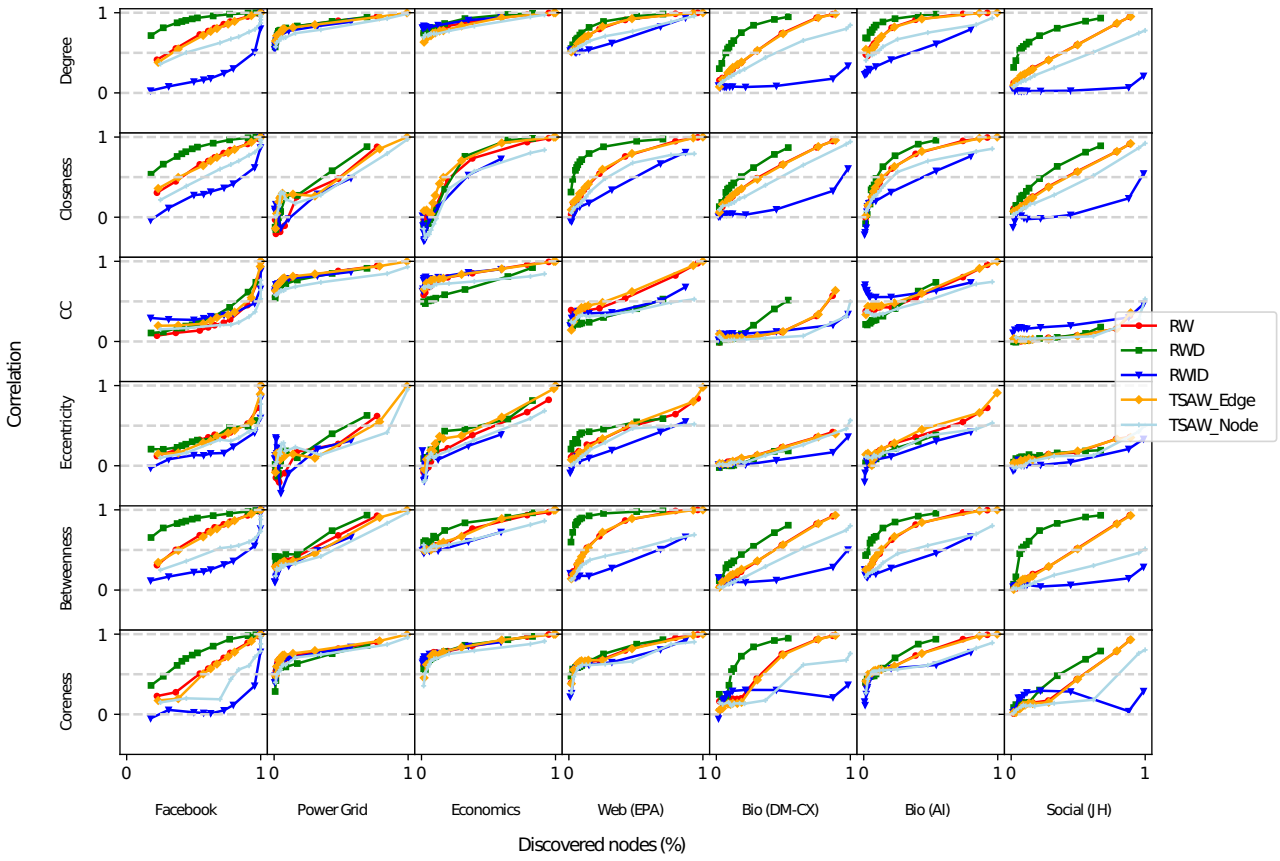


FIG. 7. Pearson correlation for the selected real networks. Each data point represents the correlation of the same nodes for the original and reconstructed network from sequences acquired by dynamics for a given walk length, where the x-axis represents the knowledge acquired for each walk length, w, where $100 \leq w \leq 50000$.

V. CONCLUSION

In the current paper, we proposed a framework to identify the efficiency in recovering network metrics arising from the reconstructed structure generated by a sequence of symbols. Networks were reconstructed using the well-known co-occurrence approach. The efficiency in recovering the network structure was evaluated by comparing reconstructed and original networks via correlation of network metrics. The analysis included four network models and six real networks. Five different random walks were evaluated. Here we focused on analyzing the ability to recover network metrics as many networked-based applications depend on the accurate representation of network topology [4, 29].

Our experiments revealed that long walks do not necessarily yield a high correlation between original and reconstructed networks. We also found that the TSAW Edge dynamics achieved a high correlation for most of the experiments. Surprisingly, while having a similar strategy to select neighbors, the TSAW Node dynamics did not achieve competitive correlations. In modular networks, the walking strategy based on avoiding hubs did not achieve competitive performance in particular network models (e.g. RWID for all considered values of mixing parameter). Conversely, the RWD dynamics, performed well in most scenarios, especially when the size of the sequence size used in the reconstructed network was typically lower than $2,000$ nodes. Such behavior was similar for model and real networks.

We also analyzed the interplay between network reconstruction and knowledge acquisition performance. The experiments demonstrated that the RWD outperforms other dynamics with regard to network metrics recovery efficiency. However, this random walk discovers new nodes slower than others. In other words, discovering nodes faster may not reflect in a good local network representation via network metrics. Finally, we also noted that the true self-avoiding walking – an efficient metric in the knowledge acquisition task – might have distinct behavior in recovering network metrics depending on which network elements are avoided. We found that avoiding visited edges is more efficient in network metrics recovery than avoiding nodes, according to the true self-avoiding rule.

In general, for shorter walks, the performance in recovering the properties of the original networks can vary substantially depending on its architecture and type of walk dynamics. In addition, for some combinations of networks, walks and metrics; even longer walks can lead to low correspondence between the real and observed properties. Such results indicate

that biases can be easily formed depending on the walk dynamics, length of the sequences, and topological characteristics of the networks.

A potential application of this work is understanding how recommendation algorithms (in social media or content platforms) impact in the perceived knowledge of the network. For instance, we found that clustering coefficient was not reliably recovered from network reconstructions based on the RWD dynamics. This suggests that when new content is recommended to users based on their number of views (or number of links to them, similar to the RWD dynamics), this may lead to the misleading notion that related contents (local) are not interconnected among themselves but only through hubs.

Another application is understanding the different user behaviors in click-streams [31] data. Such a type of data covers sequences of web access or actions taken in by users in a online platform or across the whole internet. Users may navigate across content by using different strategies, which could potentially be identified by the patterns of the reconstructed networks. A similar approach could be used to understand the foraging process of researchers in science [15], i.e., the different strategies they use to perform or seek for new experiments, research questions and theories. This can be accomplished by considering researchers as agents walking across a knowledge space made from publications [7].

While this paper focused on a global recovery strategy, in future studies we intend to analyze whether different parts of the network are more easily recovered. In addition, we also intend to analyze if other reconstruction methods lead to improved reconstruction accuracy. The results could lead to potential new approaches to model sequences as complex networks, with potential implications in applications relying on co-occurrence approaches [19]. In addition to that, future work can also focus on walks directly inspired by content recommendation strategies commonly used in media content platforms, which can lead to new ways to diversify content for the users or to mitigate biases in these platforms.

ACKNOWLEDGMENTS

Diego R. Amancio acknowledges financial support from CNPq (grant no. 311074/2021-9) and São Paulo Research Foundation (FAPESP grant no. 2020/06271-0).

---

[1] C. Akimushkin, D. R. Amancio, and O. N. Oliveira Jr. On the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications*, 495:49–58, 2018.

[2] D. J. Amit, G. Parisi, and L. Peliti. Asymptotic behavior of the "true" self-avoiding walk. *Phys. Rev. B*, 27:1635–1645, Feb 1983.

[3] H. F. Arruda, F. N. Silva, L. F. Costa, and D. R. Amancio. Knowledge acquisition: A complex networks approach. *Information Sciences*, 421:154 – 166, 2017.

[4] H. F. Arruda, F. N. Silva, V. Q. Marinho, D. R. Amancio, and L. F. Costa. Representation of texts as complex networks: a mesoscopic approach. *Journal of Complex Networks*, 6(1):125–144, 2018.

[5] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[6] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[7] K. Börner, F. N. Silva, and S. Milojević. Visualizing big science projects. *Nature Reviews Physics*, 3(11):753–761, 2021.

[8] G. T. Cantwell, A. Kirkley, and M. Newman. The friendship paradox in real and model networks. *Journal of Complex Networks*, 9(2):cnab011, 2021.

[9] H. Chen, X. Chen, and H. Liu. How does language change as a lexical network? an investigation based on written chinese word co-occurrence networks. *PloS one*, 13(2):e0192545, 2018.

[10] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[11] L. F. Costa. Voronoi and fractal complex networks and their characterization. *International Journal of Modern Physics C*, 15(01):175–183, 2004.

[12] L. F. Costa. Learning about knowledge: A complex network approach. *Physical Review E*, 74(2):026103, 2006.

[13] K. Das, S. Samanta, and M. Pal. Study on centrality measures in social networks: a survey. *Social network analysis and mining*, 8(1):1–11, 2018.

[14] H. F. de Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. Connecting network science and information theory. *Physica A: Statistical Mechanics and its Applications*, 515:641–648, 2019.

[15] M. Dubova, A. Moskvichev, and K. Zollman. Against theory-motivated experimentation in science, Jun 2022.

[16] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[17] P. Erdös and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[18] M. Fire and R. Puzis. Organization mining using online social networks. *Networks and Spatial Economics*, 16(2):545–578, Jun 2016.

[19] E. M. Grames, A. N. Stillman, M. W. Tingley, and C. S. Elphick. An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, 10(10):1645–1654, 2019.

[20] L. Guerreiro, F. N. Silva, and D. R. Amancio. A comparative analysis of knowledge acquisition performance in complex networks. *Information Sciences*, 555:46 – 57, 2021.

[21] L. Guerreiro, F. N. Silva, and D. R. Amancio. Recovery of network topology and dynamics via sequence characterization. *arXiv preprint arXiv:2206.15190*, 2022.

[22] Y. Kim, S. Park, and S.-H. Yook. Network exploration using true self-avoiding walks. *Physical review. E*, 94 4-1:042309, 2016.

[23] A. A. Klishin, N. H. Christianson, C. S. Q. Siew, and D. S. Bassett. Learning dynamic graphs, too slow, 2022.

[24] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[25] M. Latapy and C. Magnien. Complex network measurements: Estimating the relevance of observed properties. In *IEEE Infocom 2008 - The 27th Conference on Computer Communications*, pages 1660–1668, 2008.

[26] T. S. Lima, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. The dynamics of knowledge acquisition via self-learning in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(8):083106, 2018.

[27] J. Machicao, E. A. Corrêa Jr, G. H. Miranda, D. R. Amancio, and O. M. Bruno. Authorship attribution based on life-like network automata. *PloS one*, 13(3):e0193703, 2018.

[28] A. F. McDaid, D. Greene, and N. Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

[29] A. Mehri, A. H. Darooneh, and A. Shariati. The complex networks approach for authorship attribution of books. *Physica A: Statistical Mechanics and its Applications*, 391(7):2429–2437, 2012.

[30] F. Menczer, S. Fortunato, and C. A. Davis. *A first course in network science*. Cambridge University Press, 2020.

[31] E. Olmezogullari and M. S. Aktas. Pattern2vec: Representation of clickstream data sequences for learning user navigational behavior. *Concurrency and Computation: Practice and Experience*, 34(9):e6546, 2022.

[32] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[33] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015.

[34] A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett. Knowledge gaps in the early growth of semantic feature networks. *Nature human behaviour*, 2(9):682–692, 2018.

[35] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, Mar 2019.

[36] J. Wang and N. Yang. Dynamics of collaboration network community and exploratory innovation: The moderation of knowledge networks. *Scientometrics*, 121(2):1067–1084, 2019.

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

[38] B. Waxman. Routing of multipoint connections. *IEEE J. Sel. Area Comm*, 1:286–292, 01 1988.

CHAPTER

5

CONCLUSION

Complex networks is an already consolidated interdisciplinary area, which has been explored for countless fields, such as transportation (HáZNAGY *et al.*, 2015), biology (COSTA *et al.*, 2007), logistics (BECKER; WAGNER-KAMPIK, 2021), energy systems (SALEH; ESA; MOHAMED, 2018), and scientometrics (AMANCIO; JR; COSTA, 2015). Although this research area goes back to 300 years as it is being used since the XVIII century, the interest in complex systems has been increasing over the past decades. Visualizing daily tasks as complex networks (or graphs) is natural to human beings, and therefore methods and applications of complex networks optimizations can lead to improvements in many real-world problems (NEWMAN, 2003).

In the context of theoretical and applicability of complex networks, the knowledge acquisition area has been gaining attention and standing out due to power of the comprehension provided to the network science field (ARRUDA *et al.*, 2017; LIMA *et al.*, 2018b; ARRUDA *et al.*, 2019). Researches is this area, such as the work from Arruda *et al.* (2017) and Arruda *et al.* (2019) allowed developments to understand how agents behave when navigating within topics or concepts, for example. Therefore, studies in this area are a hot topic in complex networks and motivated us during this work in order to propose methods to move this area further.

The main focus of this thesis was on the analysis of the resulting sequences from network exploration by random walkers. Firstly, we dove into the understanding of the behavior of dynamics in network topologies and analyzed the impact on knowledge acquisition performance. This analysis allowed us to understand the behavior of one of the properties of networks exploration: the node discovery rate. Later, we extended our research in the search of a framework that uses machine learning methods to identify the generating topology and dynamics of a sequence. We could successfully prove that it is possible to recover the structures in a limited range and demonstrated the impact of sequence sizes for this task. Moreover, we could find the most relevant properties in the recovery task. Finally, we have explored the reconstruction step further and analyzed how biased dynamics represent the network's properties. For this study,

we observed how different sequence sizes may be enough to understand the entire network properties. Therefore, such study may indicate whether a local view of a large network shows "real" information of the network or if such perspective can be biased by the local visualization.

The work presented in this thesis can shed light to other studies in the theory of network science or in applications for common problems. Among the relevant applications that may arise from the studies discussed in this work, we can cite those in natural language processing, such as, for instance, the text correction task. In this case, we can consider a network that comprises the words as nodes, and those words that are connected have an edge linking them. Therefore, we can look to a text and when two words that are not linked in the network appear, we can indicate a more probable word based on the visited nodes context in the presented structure.

We consider our research as complete in relation to our initial objectives, however there is plenty to discover in this exciting field for future works. In relation to the recovery of network topologies and dynamics, we see that our work has proved the possibility of such classification, but a theoretical proof could also be taken beyond our empirical analysis. This work could also be taken further by adding other network, e.g. real networks, in order to observe the behavior of different topologies. We also want to find out how other theoretical and more dense networks will behavior, using, for example, larger average degree in the proposed experiments. Another interesting work would be to find an optimal classifier which could recover structures for most situations. We also want to further analyze whether the reconstruction method may impact on the recovery process for the situations in Chapter 3 and Chapter 4. Further, we can also understand the impact in embeddings, considering our methods can bring novel frameworks for such cases. Lastly, one may find an optimal walker by combining the most effective characteristics for each situation; in this case, such dynamics could adapt during the walk process given a property-related task it might want to achieve.

BIBLIOGRAPHY

AMANCIO, D. R.; JR, O. N. O.; COSTA, L. da F. Topological-collaborative approach for disambiguating authors' names in collaborative networks. **Scientometrics**, v. 102, n. 1, p. 465–485, Jan 2015. ISSN 1588-2861. Available: <https://doi.org/10.1007/s11192-014-1381-9>. Citation on page 95.

AMARAL, L. A. N.; SCALA, A.; BARTHÉLÉMY, M.; STANLEY, H. E. Classes of small-world networks. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 97, n. 21, p. 11149–11152, 2000. ISSN 0027-8424. Available: <https://www.pnas.org/content/97/21/11149>. Citation on page 23.

AMIT, D. J.; PARISI, G.; PELITI, L. Asymptotic behavior of the "true" self-avoiding walk. **Phys. Rev. B**, American Physical Society, v. 27, p. 1635–1645, Feb 1983. Available: <https://link.aps.org/doi/10.1103/PhysRevB.27.1635>. Citations on pages 24 and 69.

ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. Connecting network science and information theory. **Physica A: Statistical Mechanics and its Applications**, v. 515, p. 641 – 648, 2019. ISSN 0378-4371. Available: <http://www.sciencedirect.com/science/article/pii/S0378437118313438>. Citations on pages 24, 29, 43, 95, and 106.

ARRUDA, H. F.; SILVA, F. N.; COSTA, L. F.; AMANCIO, D. R. Knowledge acquisition: A complex networks approach. **Information Sciences**, v. 421, p. 154 – 166, 2017. ISSN 0020-0255. Available: <http://www.sciencedirect.com/science/article/pii/S0020025517309295>. Citations on pages 24, 25, 29, 43, 95, and 106.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. ISSN 0036-8075. Available: <https://science.sciencemag.org/content/286/5439/509>. Citations on pages 24 and 104.

BARAT, K.; CHAKRABARTI, B. K. Statistics of self-avoiding walks on random lattices. **Physics Reports**, Elsevier, v. 258, n. 6, p. 377–411, 1995. Citation on page 29.

BECKER, T.; WAGNER-KAMPIK, D. Complex networks in manufacturing and logistics: A retrospect. In: ____. **Dynamics in Logistics: Twenty-Five Years of Interdisciplinary Logistics Research in Bremen, Germany**. Cham: Springer International Publishing, 2021. p. 57–70. ISBN 978-3-030-88662-2. Available: <https://doi.org/10.1007/978-3-030-88662-2_3>. Citation on page 95.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics Reports**, v. 424, n. 4, p. 175 – 308, 2006. ISSN 0370-1573. Available: <http://www.sciencedirect.com/science/article/pii/S037015730500462X>. Citation on page 25.

BONAVENTURA, M.; NICOSIA, V.; LATORA, V. Characteristic times of biased random walks on complex networks. **Phys. Rev. E**, American Physical Society, v. 89, p. 012803, Jan 2014. Available: <https://link.aps.org/doi/10.1103/PhysRevE.89.012803>. Citation on page 105.

BOYER, D.; SOLIS-SALAS, C. Random walks with preferential relocations to places visited in the past and their application to biology. **Phys. Rev. Lett.**, American Physical Society, v. 112, p. 240601, Jun 2014. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.112.240601>. Citation on page 23.

CHEN, Q.; CHANG, H.; GOVINDAN, R.; JAMIN, S. The origin of power laws in internet topologies revisited. In: **Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies**. [S.l.: s.n.], 2002. v. 2, p. 608–617 vol.2. ISSN 0743-166X. Citation on page 23.

COMIN, C. H.; PERON, T.; SILVA, F. N.; AMANCIO, D. R.; RODRIGUES, F. A.; COSTA, L. F. Complex systems: Features, similarity and connectivity. **Physics Reports**, v. 861, p. 1–41, 2020. Citation on page 29.

COSTA, L. da F. Learning about knowledge: A complex network approach. **Phys. Rev. E**, American Physical Society, v. 74, p. 026103, Aug 2006. Available: <https://link.aps.org/doi/10.1103/PhysRevE.74.026103>. Citations on pages 23, 25, and 106.

COSTA, L. da F.; COSTA, F.; RODRIGUES, F.; CRISTINO, A. Complex networks: The key to systems biology. **Genetics and Molecular Biology**, v. 31, 12 2007. Citation on page 95.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Available: <https://aclanthology.org/N19-1423>. Citation on page 25.

ERDÖS, P.; RÉNYI, A. On random graphs I. **Publicationes Mathematicae Debrecen**, v. 6, p. 290, 1959. Citations on pages 24 and 103.

ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. In: **PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES**. [S.l.: s.n.], 1960. p. 17–61. Citation on page 103.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. **SIGCOMM Comput. Commun. Rev.**, Association for Computing Machinery, New York, NY, USA, v. 29, n. 4, p. 251–262, Aug. 1999. ISSN 0146-4833. Available: <https://doi.org/10.1145/316194.316229>. Citation on page 23.

GROVER, A.; LESKOVEC, J. Node2vec: Scalable feature learning for networks. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 855–864. ISBN 9781450342322. Available: <https://doi.org/10.1145/2939672.2939754>. Citation on page 25.

HERRERO, C. Self-avoiding walks and connective constants in clustered scale-free networks. **Physical Review E**, v. 99, 01 2019. Citation on page 29.

HOSSU, D.; HUMAILA, H.; MOCANU, S.; SARU, D. Complex networks to model the economic globalization process. **IFAC Proceedings Volumes**, v. 42, n. 25, p. 62 – 67, 2009. ISSN 1474-6670. 12th IFAC Workshop on Supplementary Ways for Improving International Stability.

Available: <http://www.sciencedirect.com/science/article/pii/S147466701530015X>. Citation on page 23.

HáZNAGY, A.; FI, I.; LONDON, A.; NEMETH, T. Complex network analysis of public transportation networks: A comprehensive study. In: **2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)**. [S.l.: s.n.], 2015. p. 371–378. Citation on page 95.

ITO, T.; CHIBA, T.; OZAWA, R.; YOSHIDA, M.; HATTORI, M.; SAKAKI, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 98, n. 8, p. 4569–4574, 2001. ISSN 0027-8424. Available: <https://www.pnas.org/content/98/8/4569>. Citation on page 23.

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A.-L. The large-scale organization of metabolic networks. **Nature**, v. 407, n. 6804, p. 651–654, 2000. ISSN 1476-4687. Available: <https://doi.org/10.1038/35036627>. Citation on page 23.

KIM, Y.; PARK, S.; YOOK, S.-H. Network exploration using true self-avoiding walks. **Phys. Rev. E**, American Physical Society, v. 94, p. 042309, Oct 2016. Available: <https://link.aps.org/doi/10.1103/PhysRevE.94.042309>. Citations on pages 24 and 69.

LIMA, T. S.; ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. The dynamics of knowledge acquisition via self-learning in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing LLC, v. 28, n. 8, p. 083106, 2018. Citations on pages 29, 43, and 69.

LIMA, T. S.; ARRUDA, H. F. de; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. d. F. The dynamics of knowledge acquisition via self-learning in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, v. 28, n. 8, p. 083106, 2018. Available: <https://doi.org/10.1063/1.5027007>. Citations on pages 95 and 106.

LOVÁSZ, L. Random walks on graphs: A survey. In: Miklós, D.; Sós, V. T.; Szőnyi, T. (Ed.). **Combinatorics, Paul Erdős is Eighty**. Budapest: János Bolyai Mathematical Society, 1996. v. 2, p. 353–398. Citations on pages 24 and 105.

MEERSCHAERT, M. M.; SCALAS, E. Coupled continuous time random walks in finance. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 370, n. 1, p. 114–118, 2006. Citation on page 29.

MILGRAM, S. The small world problem. **Psychology Today**, v. 2, p. 60–67, 1967. Citation on page 103.

NEWMAN, M. **Networks: An Introduction**. USA: Oxford University Press, Inc., 2010. ISBN 0199206651. Citation on page 23.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 2003. Available: <https://doi.org/10.1137/S003614450342480>. Citations on pages 23 and 95.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14**, ACM Press, 2014. Available: <http://dx.doi.org/10.1145/2623330.2623732>. Citation on page 25.

PODANI, J.; OLTVAI, Z. N.; JEONG, H.; TOMBOR, B.; BARABÁSI, A.-L.; SZATHMÁRY, E. Comparable system-level organization of archaea and eukaryotes. **Nature Genetics**, v. 29, n. 1, p. 54–56, 2001. ISSN 1546-1718. Available: <https://doi.org/10.1038/ng708>. Citation on page 23.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 105, n. 4, p. 1118–1123, 2008. ISSN 0027-8424. Available: <https://www.pnas.org/content/105/4/1118>. Citation on page 25.

SALEH, M.; ESA, Y.; MOHAMED, A. Applications of complex network analysis in electric power systems. **Energies**, v. 11, n. 6, 2018. ISSN 1996-1073. Available: <https://www.mdpi.com/1996-1073/11/6/1381>. Citation on page 95.

SCOTT, J. **Social Network Analysis: A Handbook**. 2. ed. SAGE Publications, 2000. ISBN 9780761963394. Available: <https://books.google.com.br/books?id=Ww3_bKcz6kgC>. Citation on page 23.

TANG, J.; QU, M.; WANG, M.; ZHANG, M.; YAN, J.; MEI, Q. Line: Large-scale information network embedding. In: **Proceedings of the 24th International Conference on World Wide Web**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2015. (WWW '15), p. 1067–1077. ISBN 9781450334693. Available: <https://doi.org/10.1145/2736277.2741093>. Citation on page 25.

TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. **SOCIOMETRY**, v. 32, n. 4, p. 425–443, 1969. Citation on page 103.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Available: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citation on page 25.

WANG, J.; YANG, N. Dynamics of collaboration network community and exploratory innovation: The moderation of knowledge networks. **Scientometrics**, Springer, v. 121, n. 2, p. 1067–1084, 2019. Citation on page 69.

WASSERMAN, S.; FAUST, K. **Social Network Analysis: Methods and Applications**. [S.l.]: Cambridge University Press, 1994. (Structural Analysis in the Social Sciences). Citation on page 23.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998. ISSN 1476-4687. Available: <https://doi.org/10.1038/30918>. Citations on pages 24 and 103.

WAXMAN, B. M. Routing of multipoint connections. **IEEE Journal on Selected Areas in Communications**, v. 6, n. 9, p. 1617–1622, Dec 1988. ISSN 1558-0008. Citation on page 104.

APPENDIX

A

COMPLEX NETWORKS AND KNOWLEDGE ACQUISITION

The field of complex networks has been thoroughly investigated and applied in recent years, although we can date its beginning back in 1736 when the mathematician Leonhard Euler solved the problem of the "Seven Bridges of Königsberg". The name of the problem refers to the bridges existent in a city in Prussia, which has created a rumor between the cities' citizens that one could cross all the seven bridges that divided two islands in a large complex not repeating any already crossed bridge. Euler was interested by this problem, and solved it using what later has been known as the origin of the graph theory, that is closely related to the nowadays so-called complex networks.

Since its start, the field has received great attention, especially during the last 20 years due to advances in computer networks, transportation and telecommunications. Complex networks have been used to represent and analyze various complex systems, such as ecological, biological, social and technological. Hence, the importance to determine and identify their aspects regarding its structure among other particularities.

This appendix starts by discussing the characteristics of complex networks, as well as their main properties and structures, then we dive into a generalization of the complex networks problem: knowledge acquisition.

A.1 Concepts and Properties

A network can be represented as a graph $G = (V, E)$, where V defines a set of nodes, while E is the set of edges linking nodes in the graph. The connectivity of edges, i.e. how nodes are linked, is the property that differs among several kinds of networks. Moreover, these edges can define a single type of connection in a adjacency matrix A, that is the most common method of representation, with the entries for each node i and j represented as in A.1.

$$a_{ij} = \begin{cases} 1, & i, j \varepsilon E \\ 0, & otherwise \end{cases} \tag{A.1}$$

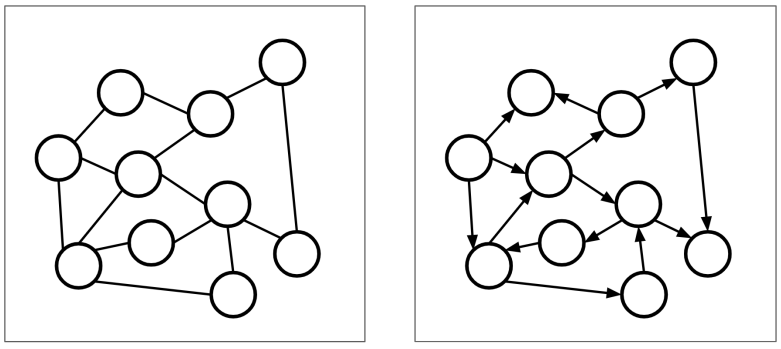
The adjacency matrix can be constructed from different methods. A common way to build these connections is by determining the distance between each pair of nodes, denoting a link if the distance is smaller than a certain threshold.

There are also *weighted networks*, defined by a different adjacency matrix, where the connections are not limited to values 0 or 1, but by the weights (or strength) of the links between each pair of nodes, that are usually between 0 (when there are no edges linking them) and the maximum weight, as represented in A.2.

$$a_{ij} = \{w_{ij} \tag{A.2}$$

The edges in the graph representation of a complex network may also have a direction. In the case of a directed network (Figure 1b), the edge $e = (u, v)$, indicates the path from node u to node v, not necessarily having a connection from v to u unless there is also such an edge $e = (v, u)$. In undirected networks (Figure 1a), an edge $e = (u, v)$ states that there is a path from u to v and from v to u.

Figure 1 – Example of hypothetical undirected and directed networks



(a) Undirected Network (b) Directed Network

Source: Elaborated by the author.

The directionality of a network may not impact on the number of edges of the model, i.e. we might have a directed network G_1 with a set of edges E, and these same edges E composing an undirected network G_2, however the difference would be established by the flow of the network when performing a walk on it, and also in a resulting non symmetrical adjacency matrix.

In network analysis we can observe different sets of properties from the structure. For both directed and undirected networks, we have the node degree property, for example. In undirected graphs we have a single k_i degree of a node i, determined by the number of edges

incident with the node; considering an adjacency matrix A, we have that the degree of a node i referred as in A.3.

$$k_i = \sum_{j \varepsilon V} a_{ij} \tag{A.3}$$

For directed graphs there are two node degrees for each node: inbound and outbound. The same definition of A.3 applies, but we consider the node i in-degree k_{in} as the number of edges directed to the node i and the out-degree k_{out} as the total edges directed from the the node i to other nodes. Therefore the total degree k_i is given by the sum of $k_{in} + k_{out}$. We also have the average degree of a network given by the average degrees of the graph, or as stated in A.4.

$$\overline{k} = \frac{\sum_{ij \varepsilon V} a_{ij}}{n} \tag{A.4}$$

A.2 Network Models

As discussed previously, networks may assume different settings and structures. When randomising a network configuration, we may fall into some common structures that have already been researched and understood, or we may build our network to imitate some already defined structure. In this section we will discuss the most common network models.

A.2.1 Erdös-Rényi model

The Erdös-Rényi Model (ER) (ERDÖS; RÉNYI, 1959; ERDÖS; RÉNYI, 1960) is considered the simplest one, and the most common random network. This type of model produces a random uniform distribution in the network, which may not simulate some real-world phenomena with accuracy, since real situations usually are not uniformly random, but it is much used for modeling artificial networks. To achieve this model we have an initial set of nodes n and a probability p for each pair nodes to link or not. Therefore we will have a network $G(n, m)$ defined by the probability of connecting each pair of nodes $e = (u, v)$, creating a set of edges m.

A.2.2 Watts-Strogatz model

The Watts-Strogatz Model (WS) (WATTS; STROGATZ, 1998) model is the most common structure to reproduce the small-world phenomenon. These networks have been studied since the Milgram's experiment that proved everyone is connected to everyone by up to 6 people (MILGRAM, 1967; TRAVERS; MILGRAM, 1969). In the WS model we take an initial set of n nodes placed in a circle, and we connect each node to its k neighbors, creating an interlinked circular graph, then we iterate through these nodes rewiring them by a probability p. Therefore

we will have a resulting network from the probabilities connections, with a smaller p we will have a more likely circular shape, and larger p will give us other polygonal shapes.

A.2.3 Barabási-Albert model

The Barabási-Albert Model (BA) (BARABÁSI; ALBERT, 1999) is a scale-free model. This network represents many real systems, such as Internet and social networks. The behaviour of these real complex systems have been imitated, considering ER and WS models could not represent some characteristics of these real-world examples such as being fitted by a power-law feature. In this network we start with small set of nodes m_0, and we started to add new nodes to graph and connect them to a set of nodes $m < m_0$ of the nodes of graph with a probability p.

A.2.4 Waxman model

The Waxman Model (WAX) (WAXMAN, 1988) is a random geometric model. This spatial network also may represent social networks, but it naturally incorporates community structures, differently from ER and BA model that do not have such characteristic. Random geometric networks are created by distributing nodes in a spatial model, and those pair of nodes that are separate by a distance smaller than a predefined threshold generate an edge between them. For the WAX networks, the distance between nodes influence in a probability p for the links to exist, which is inversely proportional to the distance between nodes. Therefore, we may have lesser connections than when using a simple geometric model, such feature helps to simulate many real situations, such as wireless connections decay. The probabilities of connections in this model are stated as follows

$$p_{u,v} = \alpha \exp(\frac{-d(u,v)}{\beta L}), \tag{A.5}$$

where u and v are the set of nodes, α and β are the model parameters and $d(u,v)$ is the distance - usually Euclidean - between the pair of nodes.

A.3 Network Dynamics

Given the structures defined on the previous section, our next step on the definition of the networks to get to sequences is the dynamics. Dynamics define how an agent walks on a network. Analogously to how humans function, when taking a path in an unknown environment we may take decisions based in some method, we might, for example, just choose randomly to where go next, or we might avoid some previous steps in order to try to get closer to our destination. That is how dynamics in complex networks work: we choose an algorithm to walk in a given topology.

The simplest dynamics model and one of the most used ones is the Random walk (RW). Random walks have been used in various real-world applications such as discussed by Bonaventura, Nicosia and Latora (2014), Lovász (1996). In a standard random walk dynamics, the agent present in the time t in a node i randomly selects among the neighbors of i which node go in the time $t + 1$, and so on during a determined number of iterations. Therefore, this dynamics is completely random as stated

$$p_{ij} = \frac{1}{k_i}, \tag{A.6}$$

where p_{ij} is the probability of moving from a node i to an adjacent node j, k represents the degree of a node, and Γ_i is the set of nodes connected to the node i.

For different sets of applications, this random walk model has been modified. One example is the Random walk degree biased (RWD), which occurs similarly to the random walk on the selection of the neighbor to visit, however here we take into account the degree of the neighbors and the probability of selecting a node is proportional to the degree of the neighbors, the representation can be seen as

$$p_{ij} = \frac{k_j}{\sum_{l \in \Gamma_i} k_l} \tag{A.7}$$

There is also the idea of preferring to visit those nodes with smaller degrees, which is called the Random walk biased by the inverse of the degree (RWID). In this dynamics, we also take into account the degrees of the neighbors, but those nodes with smaller degrees have more probability to be chosen. The equation below demonstrates how these probabilities are selected

$$p_{ij} = \frac{k_j^{-1}}{\sum_{l \in \Gamma_i} k_l^{-1}} \tag{A.8}$$

There are two other interesting dynamics, the Self-avoiding walk (SAW) and the True self-avoiding walk (TSAW). In the SAW dynamics, the agent does not return to an already visited node; this may take the agent to discover more nodes, but may trap the agent in a dead end node. Therefore, an alternative is the TSAW, in this dynamics we avoid revisiting a node by attributing a smaller probability of visiting a node each time the node receives the agent, in this way we give preference to visit new nodes, but we also avoid trapping the agent. Some authors expand this theory from nodes to edges, therefore we avoid to revisit edges instead of nodes, which we can see as

$$p_{ij} = \frac{\gamma^{-f_e}}{\sum_{j \in Ne} \gamma^{-f_j}}, \tag{A.9}$$

where γ represents a parameter of the dynamics that provides self-avoiding behaviour when its value is greater than 1, and the parameter f_e represents the memory of visits to the edge e.

Furthermore, agents in real applications can be related to the aforementioned dynamics. We can consider, for example, the actions of a researcher looking into papers of a subject. The papers can be mapped as nodes of the network and the citations among the papers will determine the edges of the structure. We notice that the researcher avoids to read the same paper twice, however this is not a restricted rule, i.e. he may go back to the same paper, but he will prefer to read other papers in order to acquire more knowledge on his field. Therefore, the researcher will perform a kind of self-avoiding strategy when reading the papers. In this sense, we can see that all the dynamics used in this thesis can be related to some "real-world" exploration strategy.

A.4 Knowledge Acquisition

As presented in the previous sections, we have the structures, i.e. topology, and the agents set of rules, i.e. dynamics. Some works such as Costa (2006), Arruda *et al.* (2017), Arruda *et al.* (2019), Lima *et al.* (2018b) have explored this combination in terms of knowledge acquisition. Therefore, we can define the nodes as key concepts, for example, and the edges as a relationship between two concepts. In this case, we can consider an agent walking over the network with the goal of discovering as much nodes as possible and as fast as possible. Analogously, the concepts might be in the edges, therefore the objective of agent would be to discover most edges in the structure.

In the context of studies such as the presented by Costa (2006), when the knowledge is being acquired, the dynamics explore a time-series of symbols (ARRUDA *et al.*, 2017). This series is also known as a *sequence* and it comprises information about the network (ARRUDA *et al.*, 2019). Therefore, there is a research field in network science that explores the characteristics of knowledge acquisition in complex networks and that looks to enlight this relationship between topologies, dynamics, and sequences.


ICMC
USP
SÃO CARLOS