

A comparative analysis of knowledge acquisition performance in complex networks

Lucas Guerreiro¹, Filipi N. Silva² and Diego R. Amancio¹

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: July 24, 2020)

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.

I. INTRODUCTION

Many real-world systems can be naturally represented by sequences corresponding to chains of events or transitions between states, including human actions [9], machine workflow [36], scientists mobility [21] and language [14]. 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 [5].

Network science has been employed to represent a great variety of complex systems [2, 7, 17, 18, 20, 27]. In recent studies, complex networks have displayed the potential to represent the space of transitions between states for many types of systems [14, 24, 25]. 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 [6], in which nodes represent knowledge that is 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 [33] 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 [17] in contrast to more well-defined structures, such as regular graphs, as a consequence, even simple stochastic dynamics can lead to intricate sequences [5].

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 coverage 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 coverage 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 coverage 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.

II. RELATED WORKS

Random walks (RW) have been studied in many networked applications [8, 13, 15, 32]. 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 [5, 6, 23, 28, 39]. In [6], 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 [38]. 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) [4, 41] – tends to be covered faster than the network borders.

In [28] 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 [5]. In [5], 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 [16] yielded the best performance. This finding is compatible with the idea that language is optimized for transmission [11], since knitted networks are representations of co-occurrence language networks [12, 31, 37, 40].

The study reported in [25] 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 [39]. 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 [23] 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.

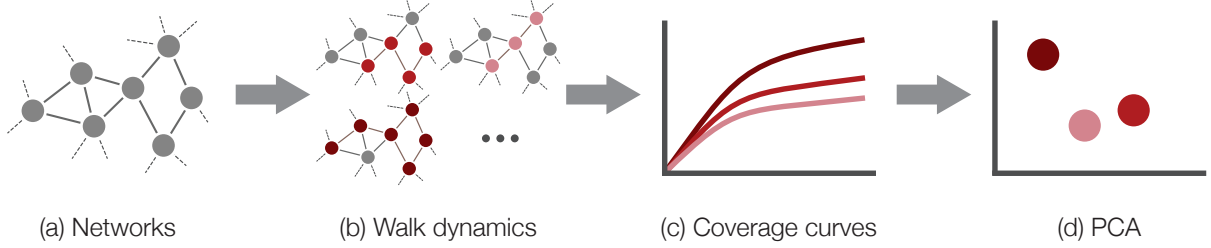


FIG. 1. Methodology employed to analysis the behavior of learning curves. In (a), we selected different network topologies. In (b), dynamics based on variations of random walks were considered to explore the networks. In (c), we obtain the learning curves describing how many nodes are discovered as the 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.

III. METHODOLOGY

The main objective of this paper is to compare the efficiency of different walking strategy to discover new nodes in the network. We compare well known random walk strategies in different 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 Figure 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 III A.
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 III B.
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 III C).
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 III D.

A. 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 small-world networks, adding the characteristic to have all the nodes with similar degrees, i.e., 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 is 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:

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

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 [26]. 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 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 Figure 2. The visualizations were generated using the *Networks3d* software [35]. It is clear that for different models the nodes with highest degrees (orangish nodes) are distributed in different ways.

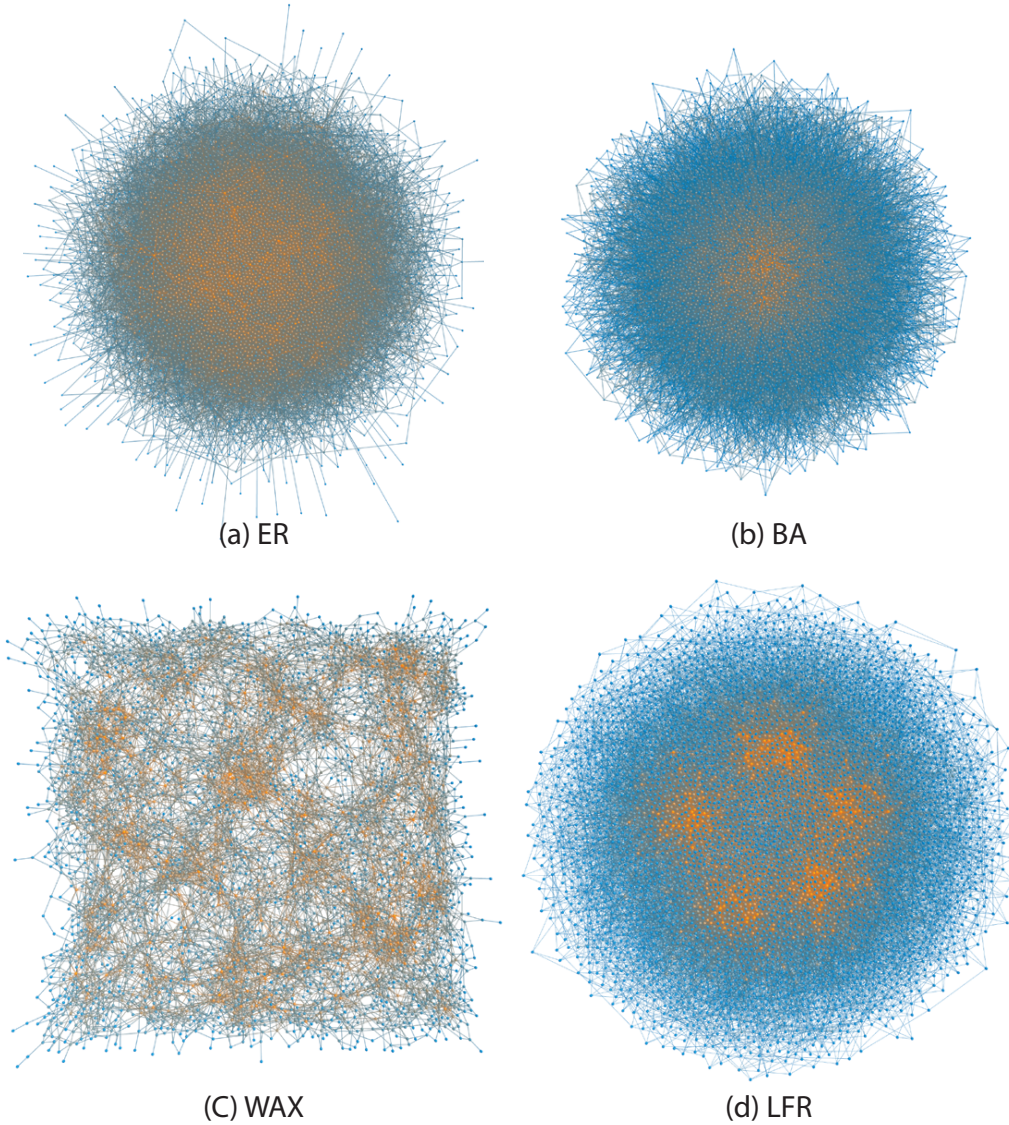


FIG. 2. Force-directed visualizations of the considered network models. Different colors correspond to different node degrees. The visualizations were generated using the *Networks3d* software [35].

B. Network dynamics

In order to recover the symbols from these models we have worked with the following walk dynamics: traditional random walk (RW) [29], random walk biased by degree (RWD) [10], random walk biased by the inverse of the degree (RWID) [10], and true self-avoiding walk (TSAW) [3, 24]. These walks have been widely employed to study the dynamics of learning curves in the last few years [5, 6, 28]. 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 $v_j \in \Gamma_i$:

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

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 :

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

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

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

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.

C. 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 [19]. For each pair of network and random walk dynamics, we considered 5,000 iterations (steps). Learning curves are then obtained as the fraction of the total number of *different* nodes visited after a given number of steps.

The dynamics observed by visiting sequentially network nodes has an analogy with the process of generating written texts [5]. If we consider that, at each step, a symbol is generated to represent that the current node has been visited, after 5,000 steps we have a sequence of symbols (i.e. a text) comprising 5,000 words. The learning curve can thus be seen as the vocabulary 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 [30], the learning curve observed in network discovery processes tends to follow a different pattern [6].

D. Principal Component Analysis

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 i -th position of the vector represents the fraction of nodes visited after the i -th

step. Because such a representation of curves yields several strongly correlated features, we use Principal Component Analysis (PCA) [22] to remove possible correlations. In fact, as we shall show, two dimensions of the PCA analysis 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 [34].

IV. 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 networks and recorded the coverage 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 Figure 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 [6]. 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.

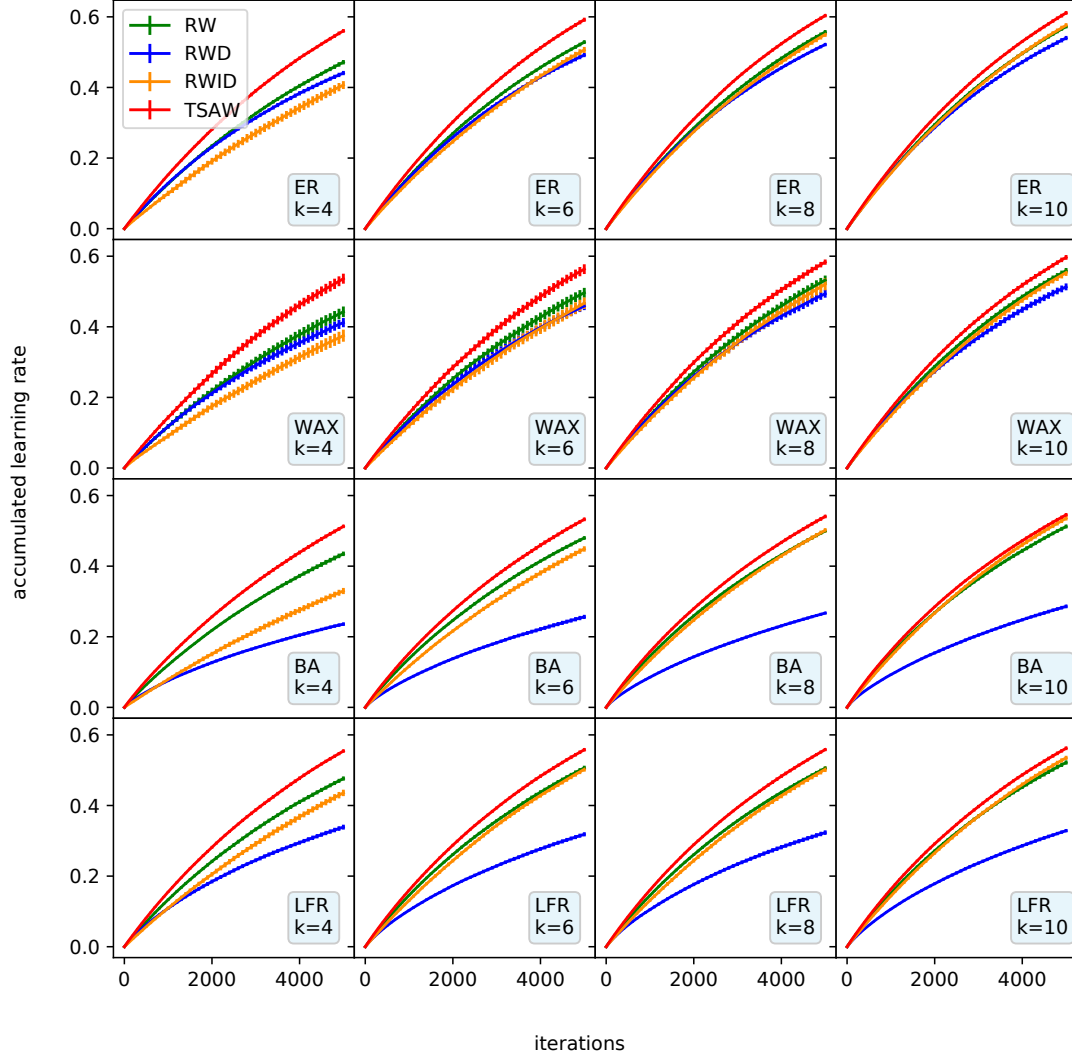


FIG. 3. Learning curves for $N = 5,000$ nodes and the models ER, BA, WAX and LFR. Each row and column correspond to different network topologies and average degrees, respectively.

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.

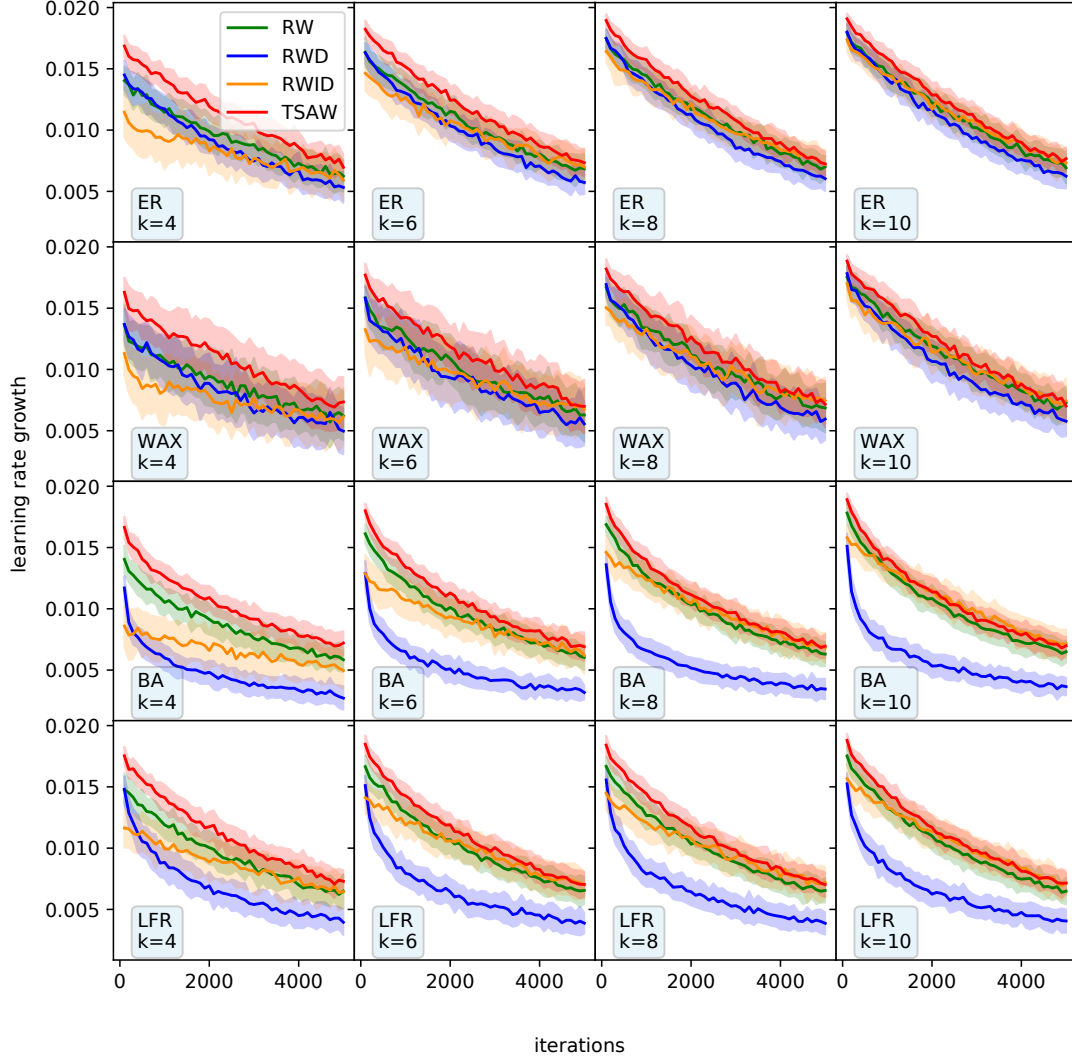


FIG. 4. Learning rates for the considered models and $N = 5,000$. Each curve indicates the growth of the number of discovered nodes across the simulation epochs.

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 (or rates of growth) of the cumulative discovery curves. Figure 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.

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 Figure 4) at epochs 100 iterations apart (see Section III D).

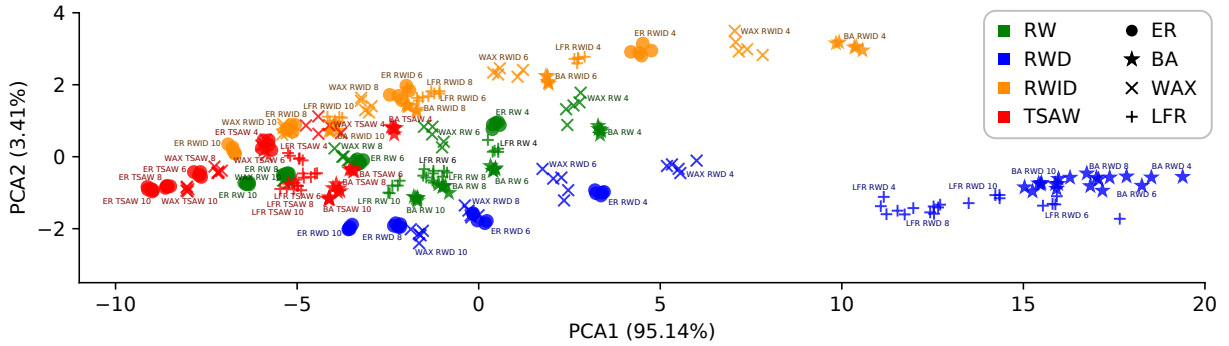


FIG. 5. PCA results for ER, BA, WAX, and LFR for $N = 5,000$ nodes. Each instance represents a learning curve obtained for a specific pair of network topology and agent dynamics. Interestingly, in some cases, different combinations of topology/dynamics can lead to similar learning curves.

The obtained data projection, shown in Figure 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.

The profile of the PCA axes in the original space, shown in Figure 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 Figure 6. 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 (presenting 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.

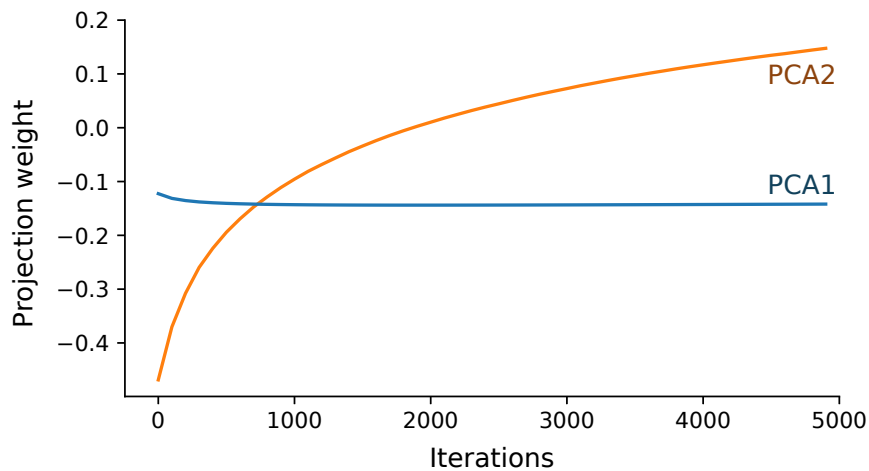


FIG. 6. Projection profiles of PCA1 and PCA2 axes along the original space.

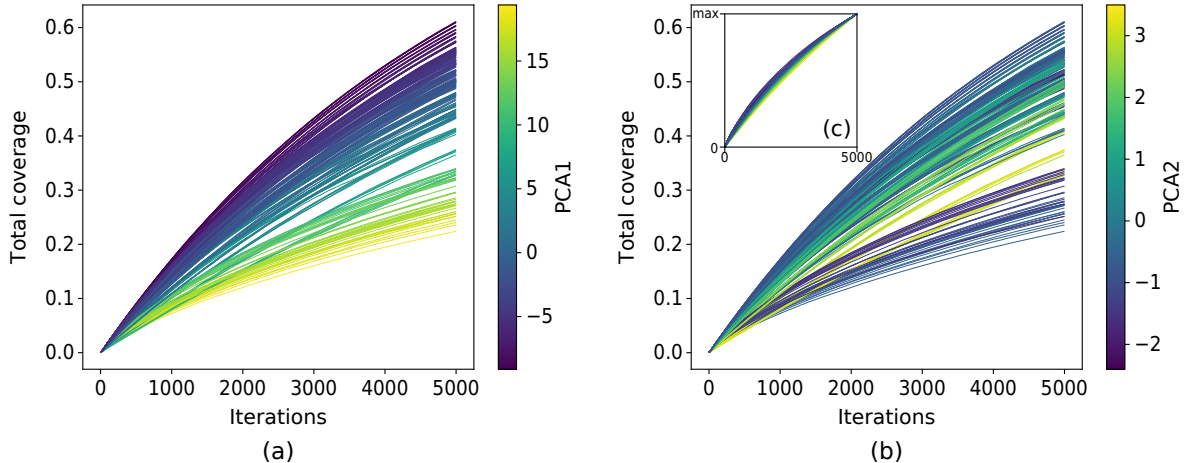


FIG. 7. Averaged learning curves for all the considered configurations. The color of each curve indicates the PCA1 (a) or PCA2 (b). The insight (c) shows all the curves normalized by their respective maximum value.

V. 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 [5, 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 Principal Component Analysis. 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 loss 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).

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.

- [1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, Sep 2001.
- [2] D. R. Amancio, O. N. Oliveira Jr, and L. d. F. Costa. Topological-collaborative approach for disambiguating authors' names in collaborative networks. *Scientometrics*, 102(1):465–485, 2015.
- [3] D. J. Amit, G. Parisi, and L. Peliti. Asymptotic behavior of the "true" self-avoiding walk. *Phys. Rev. B*, 27:1635–1645, Feb 1983.
- [4] H. F. Arruda, L. F. Costa, and D. R. Amancio. Using complex networks for text classification: Discriminating informative and imaginative documents. *EPL (Europhysics Letters)*, 113(2):28007, 2016.
- [5] 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.
- [6] 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.
- [7] K. Ban, M. Perc, and Z. Levnajić. Robust clustering of languages across wikipedia growth. *Royal Society open science*, 4(10):171217, 2017.
- [8] K. Barat and B. K. Chakrabarti. Statistics of self-avoiding walks on random lattices. *Physics Reports*, 258(6):377–411, 1995.
- [9] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.
- [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] R. F. Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

- [12] N. Castro and M. Stella. The multiplex structure of the mental lexicon influences picture naming in people with aphasia. *Journal of Complex Networks*, 7(6):913–931, 2019.
- [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, V. Q. Marinho, and D. R. Amancio. Semantic flow in language networks discriminates texts by genre and publication date. *Physica A: Statistical Mechanics and its Applications*, page 124895, 2020.
- [15] 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.
- [16] L. F. Costa. Knitted complex networks. *arXiv: 0711.2736*, 2007.
- [17] L. 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.
- [18] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007.
- [19] L. F. Costa and G. Travieso. Exploring complex networks through random walks. *Physical Review E*, 75(1):016102, 2007.
- [20] A. S. da Mata. Complex networks: a mini-review. *Brazilian Journal of Physics*, July 2020.
- [21] C. Franzoni, G. Scellato, and P. Stephan. Foreign-born scientists: mobility patterns for 16 countries. *Nature biotechnology*, 30(12):1250–1253, 2012.
- [22] F. L. Gewers, G. R. Ferreira, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. Principal component analysis: A natural approach to data exploration. *arXiv:1804.02502*, 2018.
- [23] C. Herrero. Self-avoiding walks and connective constants in clustered scale-free networks. *Physical Review E*, 99, 01 2019.
- [24] Y. Kim, S. Park, and S.-H. Yook. Network exploration using true self-avoiding walks. *Phys. Rev. E*, 94:042309, Oct 2016.
- [25] I. T. Koponen and M. Nousiainen. Concept networks in learning: finding key concepts in learners’ representations of the interlinked structure of scientific knowledge. *Journal of Complex Networks*, 2(2):187–202, 02 2014.
- [26] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community

- detection algorithms. *Phys. Rev. E*, 78:046110, Oct 2008.
- [27] Z. Levnajić. Derivative-variable correlation reveals the structure of dynamical networks. *The European Physical Journal B*, 86(7):298, 2013.
 - [28] 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.
 - [29] 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.
 - [30] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf’s law leads to heaps’ law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12), 2010.
 - [31] V. Q. Marinho, G. Hirst, and D. R. Amancio. Authorship attribution via network motifs identification. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 355–360. IEEE, 2016.
 - [32] 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.
 - [33] J. R. Norris. *Markov chains*. Number 2. Cambridge university press, 1998.
 - [34] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PLoS ONE*, 14(1), 2019.
 - [35] F. N. Silva, D. R. Amancio, M. Bardosova, L. F. Costa, and O. N. Oliveira. Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, 10(2):487 – 502, 2016.
 - [36] M. Sipser. Introduction to the theory of computation. *ACM Sigact News*, 27(1):27–29, 1996.
 - [37] M. Stella and A. Zaytseva. Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth. *PeerJ Computer Science*, 6:e255, 2020.
 - [38] D. K. Sutantyó, S. Kernbach, P. Levi, and V. A. Nepomnyashchikh. Multi-robot searching algorithm using lévy flight and artificial potential field. In *2010 IEEE Safety Security and Rescue Robotics*, pages 1–6. IEEE, 2010.
 - [39] P. Thagard. *Conceptual revolutions*. Conceptual revolutions. Princeton University Press, Princeton, NJ, US, 1992.

- [40] J. V. Tohalino and D. R. Amancio. Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539, 2018.
- [41] B. A. N. Travençolo and L. F. Costa. Accessibility in complex networks. *Physics Letters A*, 373(1):89–95, 2008.