DOI: 10.15439/2023F4965
ISSN 2300-5963 ACSIS, Vol. 37

Towards modelling and analysis of longitudinal social networks

Jens Dörpinghaus*†, Vera Weil*‡, Martin W. Sommer§
* Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
† University of Koblenz, Germany,
Email: jens.doerpinghaus@bibb.de, https://orcid.org/0000-0003-0245-7752
‡ Department of Mathematics and Computer Science, University of Cologne, Germany,
Email: weil@cs.uni-koeln.de
§ Argelander-Institut für Astronomie, Bonn, Germany

***Abstract*—There are currently several approaches to managing longitudinal data in graphs and social networks. All of them influence the output of algorithms that analyse the data. We present an overview of limitations, possible solutions and open questions for different data schemas for temporal data in social networks, based on a generic RDF-inspired approach that is equivalent to existing approaches. While restricting the algorithms to a specific time point or layer does not affect the results, applying these approaches to a network with multiple time points requires either adapted algorithms or reinterpretation. Thus, with a generic definition of temporal networks as one graph, we will answer the question of how we can analyse longitudinal social networks with centrality measures. In addition, we present two approaches to approximate the change in degree and betweenness centrality measures over time.**

I. Introduction

SOCIAL network analysis (SNA) is an important part of the social sciences and has been used in both theory and practice for several decades. It is important to understand social interactions and networks and how they affect society. In the last few years, there has been a growing interest in the use of social networks in the historical sciences. In religious studies, especially narrative studies and theology, social networks have recently received considerable attention.

Scholars have always seen SNA as part of the humanities, and in recent years there has been a rapid increase in the use of methods from the digital humanities, which includes the humanities and computer science.

Most works indicate that the described data and source problems are one of the greatest hurdles [1]. Although some preliminary work on how missing data influences a network has been carried out [2], there are still several open questions regarding the stability of social networks with respect to missing and additional data. The main question is: Can we still use the same algorithms, if we know that the data are incomplete? The need to work with temporal data makes an answer to this question even more urgent.

The three main research questions of this paper are thus:

- How can we model longitudinal social networks in one graph in the most generic way possible? (RQ1)
- How can we analyse longitudinal social networks with centrality measures? (RQ2)
- Can we approximate the change of centrality measures over time? (RQ3)

These questions cannot be answered without discussing the data schema for temporal data. Therefore, RQ1 is dedicated to the efficient storage of temporal data in a social network. While most entities such as actors and locations have a given lifetime, organisations or functions may have predecessors and successors. In other words: When an entity is detached, what relationships exist, and how can we manage their lifetimes? How can algorithms track and use these temporal data? RQ2 also contains several sub-questions: If a network G contains data for different time points $t_1, ..., t_n$, can we still apply analysis methods, e.g. centrality measures or community detection, that were originally developed for a particular time point? Or do we need to reinterpret the results or adapt the algorithms? Answering these questions is key to understanding the algorithmic challenges of temporal data in social network analysis.

This paper is divided into five sections. After this introduction, we give an overview of related work and the background of this research. We focus on historical network analysis (HNA) because it helps to highlight the challenges and is the natural habitat for longitudinal networks. Our methodological approach is described in the third section, where we discuss the modelling of longitudinal social networks, and their analysis. The fourth section is dedicated to the experimental results. Our conclusions are presented in the final section.

II. Related Work

Modeling temporal or longitudinal data in SNA is a well-known problem [3]. Temporal data lead to complex network structures and Lemercier stated in 2015: “There is no one best way for the analysis or even description of such multi-dimensional data” [4]. There are several modeling challenges, for example with synchronous and asynchronous events or relations, see [5]. Several methods have been proposed, for example, modeling with stream graphs [6], [7], Markov chains [8], [9], with network snapshots [10], or with a discrete set of time points that may contain snapshots. Most of these approaches are equivalent [11]. However, no single graphtheoretic definition currently covers all these approaches. This can be identified as the first gap in research.

Scientists are not only careful about how to model temporal networks, but also how to analyze them: "Traditional analyses of temporal networks have addressed mostly pairwise interactions, where links describe dyadic connections among individuals" [12]- Concetti et al. thus introduced "temporal hypergraphs" to address this challenge. Other researchers proposed visual analysis [13], pattern search [14], or probabilistic discrete temporal models [15]. Centrality measures, widely used in SNA, are also challenging in temporal networks. Some researchers have proposed definitions of temporal closeness, betweenness, and eigenvector centrality, see [16], [17], [18]. However, these definitions remain limited to the underlying graph topology, e.g. Sizemore et al. [18] work with a contact sequence where nodes remain static. In addition, the natural extension of centrality to groups and classes [19], [20] is usually omitted. Other authors propose MLI based on network embedding and machine learning (ML) [21]. In general, ML approaches are widely used in dynamic networks, not only in temporal networks, see [22]. However, these approaches – although providing significant insights on the networks – are not comparable to the results of centrality measures, which makes them difficult to reproduce. Thus, directly related to the first research gap – the lack of a generic definition of temporal networks – is the second gap: How can algorithms track and use this temporal data, and how does this affect the analysis of networks, e.g., with centrality measures?

These issues may be due to the fact that several aspects of knowledge graphs and the semantic web are not widely perceived in the SNA community. They have only recently been brought together [23]. Barats et al. conclude in 2020: FAIR data, a topic directly related to knowledge graphs, "remains a theoretical discussion rather than a shared practice in the field of humanities and social sciences." [24] Thus, our work will try to address the research questions using knowledge graphs.

III. METHOD

We will use a definition of a knowledge graph that combines the approaches of [14], [23]:

Definition 1 (Temporal Social Network)**.** *A* Social Network *is a graph $G = (V, E, \mathcal{T})$ with vertices (nodes) $v \in V$, edges (relations) $e \in E$ and a time domain $\mathcal{T} = \{t_0, ..., t_k\}$ where $t_i \in \mathbb{R}$ and $t_i < i_j \ \forall i < j$. Every node and edge may exist at one or multiple intervals of timepoints*

$$[t_s, t_e] = \{x \in \mathcal{T} : t_s \leq x \leq t_e; t_s, t_e \in \mathcal{T}\}$$

denoted by $t(v)$ and $t(e)$. Thus, $t : V \cup E \to I \subseteq \mathbb{R}$. We denote the graph G at time t by

$$G^t = (V^t, E^t), \text{ where}$$

$$V^t = \{v \in G \,|t \in t(v)\}, \ E^t = \{v \in E \,|t \in t(e)\},$$

so that

$$\bigcup_{t \in \mathcal{T}} G^t = G.$$

Both edges and vertices are part of previously well-defined categories, $V \subseteq C_1 \cup C_2 \cup ... \cup C_n$ and $E \subseteq R_1 \cup R_2 \cup ... \cup R_m$.

Is is important to notice, that – in contrast to other definitions, e.g. [25] – both edges and nodes are temporal. Unless otherwise noted, we assume that G is an undirected graph. We will now present examples of the notation introduced above.

Each vertex $v \in V$ has a lifetime $t(v)$. In general, any edge connected to v may only exist for times $t \in t(v)$. But this rule is not strict. For example, we can define categories for successors T_s and predecessors T_p, so that these edges can indicate a predecessor of a certain position at any time. For these edges we set $t(e) = \emptyset$, they are 'timeless'. In addition, v can be part of several categories, e.g. it can be an actor $v \in C_a$ and a politician $v \in C_p$. Thus, our approach can combine static and temporal information.

We will now prove that this definition is equivalent to stream graphs:

Theorem 1. *The temporal social network defined in 1 is equivalent to the concept of a stream graph introduced by Latapy, Magnien and Viard in [6] for discrete time instants T.*

Proof. "$\Rightarrow$" Let $G = (V', E, \mathcal{T})$ be a temporal social network as defined in Definition 1. We create a stream graph as follows: First, we can set the discrete time instants T to the time domain $\mathcal{T}$, thus $T = \mathcal{T}$. In addition, both node set are equal, thus $V = V'$.

The set of temporal nodes, $W \subseteq T \times V$, can be constructed as

$$W = \{(t(v), v) \forall v \in V\}.$$

The set of links $E \subseteq T \times V \otimes V$ can be constructed by

$$E = \{(t(e), e_1, e_2) \forall e = (e_1, e_2) \in E\}.$$

However, if $t(e) = \emptyset$, we define $t(e) = [\min_{t \in T}, \max_{t \in T}]$.

"$\Leftarrow$" Let S be a stream graph as defined by [6] with discrete time instants T, the node set V, a temporal node set $W \subseteq T \times V$ and a temporal edge set $E \subseteq T \times V \otimes V$.

We create a temporal social network $G = (V', E, \mathcal{T})$ as follows: Again, we the discrete time instants and nodes are equal and we set $\mathcal{T} = T$, $V' = V$. For each set of presence time $w = (t, (t, v)) \in W$ we define $t(v) = [\min t, \max t]$ and the same for edges $e = (t, (t, e_1, e_2)) \in E$. □

As we can see, the only difficulties are those edges and vertices that are 'timeless'. However, extending their interval to $\mathcal{T}$ models their behaviour in the intended way. It is quite easy to see that both approaches are also equivalent to models using snapshots of time points [21]. For a detailed overview we refer to [11].

Thus, Definition 1 is well aligned with other approaches. However, it is also compatible with semantic web approaches and makes it easier to integrate analysis approaches. We will now move on to modelling longitudinal social networks with semantic web technologies.

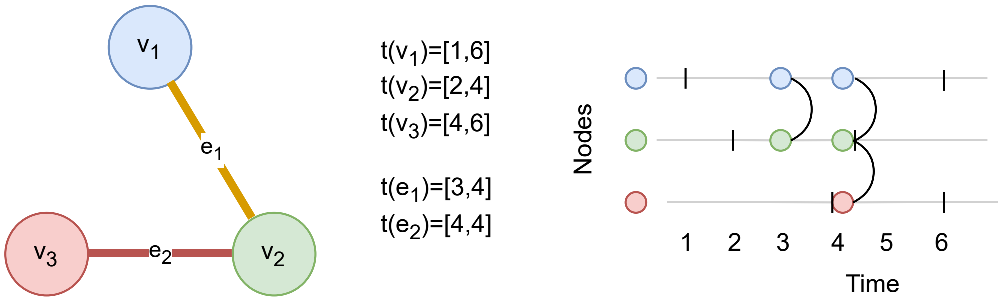

Fig. 1. Illustration of the graph in example 2 with a definition of lifetimes in the middle and a visualisation of the lifetime of edges and the sequence of edges over time (right).

A. Modelling longitudinal social networks

The initial definition of a social network in [23] corresponds to the definition of a knowledge graph. In particular, the categories for nodes $C_1, ..., C_n$ and edges $R_1, ..., R_m$ can be modelled using RDF classes. So we need to add time intervals to nodes and edges. To do this, Hobbs and Pan introduced the time ontology, see [26], [27]. Here they use a function *duration*: Intervals × TemporalUnits to express intervals. We can set $\mathit{duration}(v) = t(v)$ and $\mathit{duration}(e) = t(e)$ for any node $v \in V$ and edge $e \in E$.

Thus, any social network according to the knowledge graph definition in [23] can be easily transformed into a temporal social network, where time is modelled as a property of nodes and edges.

Example 2. *Consider the graph $G = (V, E, \mathcal{T})$ in figure 1 with $V = \{v_1, v_2, v_3\}$ and $E = \{e_1, e_2\}$ and a set of time intervals $t(v_1) = [1, 6]$, $t(v_2) = [2, 4]$, $t(v_3) = [4, 6]$, $t(e_1) = [3, 4]$ and $t(e_2) = [4, 4]$. They also provide a visualisation according to [18]: We visualise time by plotting a sequence of edges on a time scale. However, we extend the latter approach by adding information about the lifetime of nodes.*

In this case, each lifetime can be mapped according to the temporal duration.

It is worth noting that the general knowledge graph definition of a social network is open to adding a variety of additional data while maintaining the general graph structure. Thus, it is useful for modelling not only temporal social networks, but also any other temporal data, e.g. disease models.

B. Temporal graph structures

Similar to the approaches of [28], [18] we can study time-respecting structures in a graph. However, definition 1 of temporal social networks makes it easier to generalise graph structures as it keeps the generic definition of a graph.

A *path* p in a graph $H = (V, E)$ is a set of vertices $v_1, ..., v_t$, $t \in \mathbb{N}$, for example written as

$$p = [v_1, ..., v_t],$$

where $(v_i, v_{i+1}) \in E$ for $i \in \{1, \ldots, t-1\}$. However, to track the meaning of time in a temporal social network $G = (V, E, \mathcal{T})$, we define p^t, which is a path p that exists at time t. In turn, we define $t(p)$ as the interval of time in which the path p exists in G.

Unless otherwise noted, we use G for a temporal social network $G = (V, E, \mathcal{T})$ and H for any undirected graph.

We can add this generic notation for other structures as well. For example, we denote the time-respecting degree of a node v by $d^t(v)$. In this way, we get a series of *temporal degree centrality measures* (TDC) for a node $v \in V$ denoted by

$$dc^t(v) = \frac{d^t(v)}{n-1}.$$

In addition, we can analyse the *temporal degree distribution* which tells us about the network structure since we can distinguish between sparsely and densely connected networks.

Betweenness centrality (BC) was first introduced by [29] and considers other indirect links, see [30]. Given a node v, $bc(v)$ is defined as

$$bc(v) = \sum_{k \neq j, v \neq k, v \neq j} \frac{P_v(k, j)}{P(k, j)} \cdot \frac{2}{(n-1)(n-2)},$$

that is, we compute the number of all shortest paths $P_v(k, j)$ in a network for all starting and ending nodes $k, j \in V$ that pass through v. Let $P(k, j)$ denote the total number of shortest paths between k and j. Then the importance of v is given by the ratio of the two values of P_v and P. Again, for any time $t \in \mathcal{T}$ we may set $P_v^t(k, j)$ and $P^t(k, j)$ accordingly, such that

$$bc^t(v) = \sum_{k \neq j, v \neq k, v \neq j} \frac{P_v^t(k, j)}{P^t(k, j)} \cdot \frac{2}{(n-1)(n-2)}$$

defines the series of *temporal betweenness centrality* (TBC). This definition is similar to that of [18], who, however, used the concept of fastest paths.

We will proceed similarly with *closeness centrality* (CC). Given a node $i \in V$ we can compute the average distance between the first and other nodes $j \in V$ with $\sum_{j \neq i} d(i, j)$, where $d(i, j)$ denotes the length of a shortest path between i and j. Then, according to [31], we can compute closeness-centrality as follows:

$$cl(v) = \frac{n-1}{\sum_{u \in V} d(u, v)}.$$

Again, with a definition of $d^t(i,j)$ for the length of a shortest path at time $t \in \mathcal{T}$ at hand, we can define *temporal closeness centrality* (TCC) as

$$cl^t(v) = \frac{n-1}{\sum_{u\in V} d^t(u,v)}.$$

However, these definitions are currently not more than a containment of well-known centrality measures on time snapshots of the temporal social network. They allow an interpretation of these snapshots, comparable to static social networks, and they provide a series of centrality measures that can be interpreted as the progression of these measures over time.

For social networks, perceiving the world with as few snapshots as possible is most feasible. Other approaches, e.g. defining paths closely so that they could split up from one time to another, if the interval is so small that an event lasts less, is often necessary to model traffic [16]. Social interaction, on the other hand, does usually change on the basis of longer lasting events. This is a crucial observation, because computing temporal paths with increasing timestamps from one node to the next is computationally hard, see [25].

While interdisciplinary approaches are available, applications from humanities and in particular historical networks research lead to a different perspective on data. For example, a closed organization may still have an influence on parts of the network or may be referred to later. However, with our novel approach, we will evaluate the behavior of analysis methods like centrality measures and community detection and discuss limitations and challenges for further research.

C. Random graphs

For further analysis, we rely on random graphs. The *degree distribution* provides us with information about the network structure since we can distinguish between sparsely and densely connected networks. In social network analysis (SNA), the following two graphs are widely considered:

Definition 2 (Scale-Free Network)**.** *A network is scale-free if the fraction of nodes with degree s follows a power law $s^{-\alpha}$, where $\alpha > 1$.*

Definition 3 (Small World Network [32])**.** *Let $G = (V, E)$ be a connected graph with n nodes and average node degree k. Then G is a small-world network if $k \ll n$ and $k \gg 1$.*

[33] introduced a widely used graph model with three random parameters $\alpha + \beta + \gamma = 1$. These values define probabilities and thus define attachment rules to add new vertices between either existing or new nodes. This model allows loops and multiple edges, where a loop denotes one edge where the endvertices are identical, and multiple edges denote a finite number of edges that share the same endvertices. Thus, we convert the random graphs to undirected graphs. For testing putposes, we scale the number of nodes n and use $\alpha = 0.41$, $\beta = 0.54$, and $\gamma = 0.05$. This random graph model is generic and feasible for computer simulations for measuring and evaluation purposes, see [34], [35].

One of the core concepts important in social network research is the graph diameter $D(G)$. From the 1960s on, it was widely discusses whether the average path length of social networks is near six, see [36]. However, there is an ongoing discussion on this issue, see for example [37], [38]. However, it was shown that in a scale-free network the diameter is always lower than $\log(n)$, and if the fixed number m of earlier vertices is larger than 1, in general the diameter is lower than $\frac{\log(n)}{\log\log(n)}$, see [39]. Here, n describes not only the number of steps to create the random graph, but also the number of nodes in the graph. While the connection between a particular graph and a particular diameter is quite complex, see [40], we can rely on these bounds. For small-world random graphs we find [41] the almost surely upper bound $D(G) \leq \frac{72}{p}\log^2 n$ while [42] proved the diameter is usually bound by $\log(n)$.

The diameter of a scale-free graph is in general quite low, while in small-world graphs it is bound by $\log(n)$. However, we may expect random graphs to have a different behavior from real-world social networks. Thus, for some of the following proofs we will assume that $D(G) \leq 5$.

D. Analysing networks

For a detailed overview of centrality measures, we can consider the series of a particular measure, e.g. a generic c (centrality, e.g. which could refer to closeness or betweenness centrality), which is basically a vector in $\mathbb{R}^{|\mathcal{T}|}$:

$$\widetilde{c}(v) = \left(c^{t_1}(v), ..., c^{t_{|\mathcal{T}|}}(v)\right).$$

Note that $c^{t_i}(v) = \emptyset$ if $t_i \notin t(v)$. We define

$$|\widetilde{c}(v)| = \sum_{i\in\{1,...,|\mathcal{T}|,\, c^{t_i}(v)\neq\emptyset\}} l(t_{i-1}, t_i)|,$$

where $l(t_{i-1}, t_i)$ defines the length of time elapsed between two times t_{i-1} and t_i. For $x \in V$ or $x \in E$ we set

$$l(x) = \sum_{i\in\{1,...,|\mathcal{T}|,\, c^{t_i}(x)\neq\emptyset\}} l(t_{i-1}, t_i)$$

as the lifespan of x. However, if all times are equally distributed, this simplifies to

$$|\widetilde{c}(v)| = |\mathcal{T}| - |\{x \in \widetilde{c}(v) \,|\, x = \emptyset\}|.$$

This allows us to calculate the *average temporal centrality* of a node v over its lifetime as

$$\overline{c}(v) = \frac{1}{|\widetilde{c}(v)|} \sum_{t\in\mathcal{T}, c^t(v)\neq\emptyset} c^t(v).$$

However, for a proper analysis of centrality measures over time, we should also consider the *temporal centrality* of a node v:

$$A\left(c\left(v\right)\right) = \sum_{i\in\{1,...,|\mathcal{T}|,\, c^{t_i}(v)\neq\emptyset\}} c^{t_i}(v) l(t_{i-1}, t_i).$$

Again, for evenly distributed time, $l(t_{i-1}, t_i) = 1$ and

$$A\left(c\left(v\right)\right) = \sum_{t\in\mathcal{T}, c^t(v)\neq\emptyset} c^t(v).$$

We can also normalise this measure by life span as *normalised temporal centrality* to compare the centrality measure over time within a life span:

$$A'(c(v)) = \frac{1}{l(v)} \sum_{i \in \{1,\dots,|\mathcal{T}|,\, c^{t_i}(v) \neq \emptyset\}} c^{t_i}(v) l(t_{i-1}, t_i).$$

In section IV we will discuss several working examples and offer an interpretation of these values in light of the current state of research on degree and betweenness centrality.

First, we consider how a centrality measure evolves over time. Since we need to plot this for n nodes, we consider a heatmap visualisation that bins the number of nodes in a given interval. Next, we can plot the average centrality measure at a particular time and the average centrality over all time points, as we show in Figure 2.

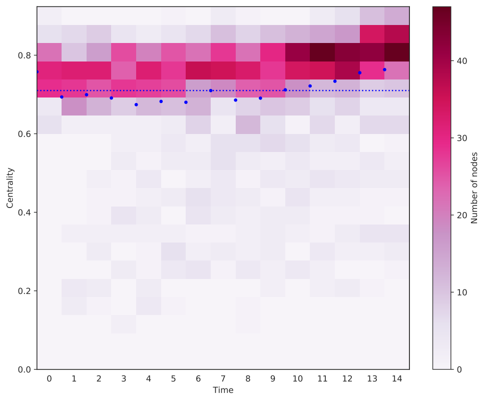

Fig. 2. Illustration of the distribution of a centrality measure over time, grouped into 20 bins between 0 and 1, as a heatmap. The blue horizontal line refers to the overall average centrality, while the blue dots refer to the average degree at a given time. This illustrates the degree centrality for $\mathcal{G}^s(100, 15, 0.1)$.

This figure gives us a good overview of how many nodes are below and above the average centrality at a given time, and whether the network at a given time is special for the scenario. To analyse and compare a particular node with this overall picture, we can plot $\widetilde{c}(v)$ and $\overline{c}(v)$, as we show in Figure 3

Some [17] considered calculating and plotting $\widetilde{c}(v)$, [16] added probabilities. Thus, in addition to the classical approach (e.g. [18]), $\widetilde{c}(v)$ and $\overline{c}(v)$ allow the study of static centrality measures at a time $t \in \mathcal{T}$, comparing the individual centrality value of a particular node with the average node degree and the distribution of node degrees. In addition, by plotting the series of centrality over time, we can compare the temporal centrality measures within a given interval or across the entire timeline. While some general measures, such as average temporal centrality, have been studied previously [3], their interpretation remains vague. If networks change significantly over time, this value is not comparable.

E. Approximating the changes over time

Let $\mathfrak{G}^p = \{G_1, \dots G_\iota\}$ be a series of graphs and $p \in \mathbb{R}$ with $0 \leq p \leq 1$ and

$$|(V(G_i) \cup V(G_i+1)) \setminus (V(G_i) \cap V(G_i+1))| \leq p|V(G_i)|,$$

$$|(E(G_i) \cup E(G_i+1)) \setminus (E(G_i) \cap E(G_i+1))| \leq p|E(G_i)|,$$

for $i \in \{1, \dots, \iota - 1\}$. Thus, $\mathfrak{G}^p$ is a series of graphs with a fixed set of differences and changes from one to the other.

Now we can approximate the changes over time, or the error in the centrality measures that can occur due to these changes. Unless otherwise noted, we will consider $\mathfrak{G}^p = \{G_1, \dots G_\iota\}$.

Theorem 3. *Let $i \in \{1, \dots \iota - 1\}$ so that $v \in V(G_i)$ and $v \in V(G_{i+1})$. Then it holds that*

$$dc^{i+1}(v) \geq \frac{d^i(v) - p|V(G_i)|}{|V(G_i)| - 1 + p|V(G_i)|},$$

$$dc^{i+1}(v) \leq \frac{d^i(v) + p|V(G_i)|}{|V(G_i)| - 1 - p|V(G_i)|}.$$

Proof. We know that

$$dc^i(v) = \frac{d^i(v)}{|V(G_i)| - 1}.$$

However, due to the definition of $\mathfrak{G}^p$, we know that at most $p|V(G_i)|$ new connections from v to other nodes can exist in G_{i+1} or may be lost. Thus, in G_{i+1} it holds that

$$d^i(v) - p|V(G_i)| \leq d^{i+1}(v) \leq d^i(v) + p|V(G_i)|.$$

In addition, we know that for G_{i+1}

$$|V(G_i)| - p|V(G_i)| \leq |V(G_{i+1})| \leq |V(G_i)| + p|V(G_i)|$$

holds. Hence the claim follows. □

For betwenness centrality, we define

$$\sigma = |N(G_i)|p,$$

$$\epsilon = \min\{D(G_i)^2, 2|V(G_i)|p\},$$

where $D(G)$ is the diameter of G. We will prove two lemmata to obtain a bound for $bc^{i+1}(v)$ for $v \in V$.

Lemma 4. *Let $i \in \{1, \dots \iota - 1\}$ so that $v \in V(G_i)$ and $v \in V(G_{i+1})$. Then,*

$$P_v(k,j)\frac{1}{\sigma} \leq P_v^{i+1}(k,j)$$

holds.

Proof. All shortest paths between $k, j \in V(G_i)$ have the same length $\mathfrak{l} \leq D(G_i)$. For $D(G_i) \leq 5$, $\mathfrak{l} = \delta(v)$ holds: If $D(G_i) = 3$, k, j must both be adjacent to v. If $D(G_i) = 4$, we say k must be adjacent to v and $\nu \in \mathbb{N}^+$ nodes exist which are adjacent to j and v, which implies $\delta(v)$ paths. If $D(G_i) = 5$, $\nu \in \mathbb{N}^+$ nodes exist which are adjacent to j and v, and $\mu \in \mathbb{N}^+$ nodes exist which are adjacent to k and v, which implies $\delta(v)$ paths.

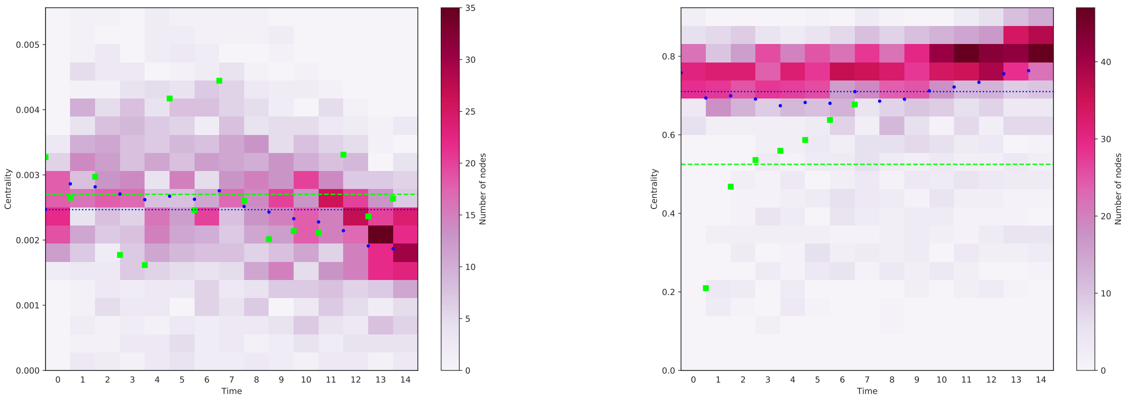

Fig. 3. Illustration of the distribution of a centrality measure over time, grouped into 20 bins between 0 and 1, as a heatmap. The blue horizontal line refers to the overall average centrality, while the blue dots refer to the average degree at one point in time. Both figures show $\widetilde{c}(v)$ and $\overline{c}(v)$ (green dots and horizontal line, respectively) for two different nodes. Left: This node exists over all 15 time points and usually shows that the betweenness centrality varies a lot. Right: This node exists from time 1 to 7 and has an increasing degree centrality value. The network is based on $\mathcal{G}^s(100, 15, 0.1)$.

Let us assume that a maximum of edges and nodes will be removed from G_i towards G_{i+1} and a maximum number of them is adjacent to v. Then, at most $|N(G_i)|p$ edges and neighbours of v can be removed in G_{i+1} which, in turn, removes one possible shortest path between k, j over v. Thus, $P_v^{i+1}(k,j)$ cannot have more than $P_v(k,j)\frac{1}{|N(G_i)|p} = P_v(k,j)\frac{1}{\sigma}$ of the initial paths through v. □

Lemma 5. *Let $i \in \{1, \ldots \iota - 1\}$ so that $v \in V(G_i)$ and $v \in V(G_{i+1})$. Then,*

$$P_v^{i+1}(k,j) \le \begin{cases} P_v(k,j)\epsilon & P_v(k,j) > 0 \\ D(G_i)^2\epsilon & P_v(k,j) = 0 \end{cases}$$

holds.

Proof. As shown in the proof of Lemma 4, all shortest paths between $k, j \in V(G_i)$ have the same length $\mathfrak{l} \le D(G_i)$ and for $D(G_i) \le 5$, $\mathfrak{l} = \delta(v)$ holds.

Let us assume that a maximum number of edges and nodes will be added to G_{i+1}. This is at maximum $2|V(G_i)|p$. However, no more than $D(G_i) \cdot D(G_i) = D(G_i)^2$ paths between k and j may exist if $P_v(k,j) > 0$. Thus,

$$P_v^{i+1}(k,j) \le P_v(k,j)\min\{D(G_i), |V(G_i)|p\} = P_v(k,j)\epsilon$$

holds.

If $P_v(k,j) = 0$, we know that no more than $D(G_i)^2$ paths may exist at all. Thus,

$$P_v^{i+1}(k,j) \le P_v(k,j)\min\{D(G_i), |V(G_i)|p\} = P_v(k,j)\epsilon$$

holds. □

Theorem 6. *Let $i \in \{1, \ldots \iota - 1\}$ so that $v \in V(G_i)$ and $v \in V(G_{i+1})$. Then,*

$$bc^i(v)\epsilon \le bc^{i+1}(v) \le bc^i(v)\frac{1}{\sigma}$$

holds.

Proof. Recall that

$$bc(v) = \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v(k,j)}{P(k,j)} \cdot \frac{2}{(n-1)(n-2)}.$$

We have already shown the following two inequalities with lemmata 4 and 5:

$$P_v(k,j)\frac{1}{\sigma} \le P_v^{i+1}(k,j) \le P_v(k,j)\epsilon$$

Thus, with Lemma 4 we can show:

$$\begin{aligned} bc^{i+1}(v) &= \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v^{i+1}(k,j)}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &\le \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v(k,j)\frac{1}{\sigma}}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &= \frac{1}{\sigma} \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v^{i+1}(k,j)}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &= bc^i(v)\frac{1}{\sigma} \end{aligned}$$

Similarly, with Lemma 5 we can show:

$$\begin{aligned} bc^{i+1}(v) &= \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v^{i+1}(k,j)}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &\ge \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v(k,j)\epsilon}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &= \epsilon \sum_{k \ne j, v \ne k, v \ne j} \frac{P_v^{i+1}(k,j)}{P^{i+1}(k,j)} \cdot \frac{2}{(n-1)(n-2)} \\ &= bc^i(v)\epsilon \end{aligned}$$

□

We will now continue with an experimental setting showing the results of these bounds.

IV. Experimental results

We evaluate the degree centrality and betweenness centrality on random graphs, see Section III-C. First, we consider scale-free networks with n nodes, see [31]. With this, we create a series of random Graphs $\mathcal{G}^s(n,i,p)$ which creates one initial scale-free network with n nodes and $i-1$ more random graphs with a probability of $p/2$ for each node and edge to be deleted and $p/2$ for each node and edge to be deleted and a new one created. In addition, we consider scale-free networks and create a series of random Graphs $\mathcal{G}^w(n,i,p)$ which starts with one initial small world network with n nodes and $i-1$ more random graphs with a probability of $p/2$ for each node and edge to be deleted and $p/2$ for each node and edge to be deleted and a new one created.

We will evaluate both degree centrality and betweenness centrality on the following four random graph series:

- $\mathcal{G}^s(50,15,p)$, $p \in \{0.15, 0.05\}$
- $\mathcal{G}^w(50,15,p)$, $p \in \{0.15, 0.05\}$
- $\mathcal{G}^s(150,15,p)$, $p \in \{0.15, 0.05\}$
- $\mathcal{G}^w(150,15,p)$, $p \in \{0.15, 0.05\}$

For evaluation purposes, we select several nodes and display the distribution of the centrality measure over time and the approximation of the changes over time.

A. Degree centrality

We present an evaluation of sample nodes in Figures 4-7. We show the upper and lower bounds for degree centrality introduced in Theorem 3.

First, small world random graphs are shown in Figures 4 and 5. Here the bounds on degree centrality are quite tight, but get worse for larger p. We can make a similar observations for scale-free networks in Figures 6 and 7.

Thus, the bounds introduced in Theorem 3 work well for small p and provide overall good results for estimating the evolution of degree centrality for the next time step when p is known.

B. Betweenness centrality

We will now consider the upper and lower bounds for betwenness centrality introduced in Theorem 6. We present a selected evaluation of small-world graphs in Figures 8 and 9. For the small-world graph in Figure 8 (left), the node has a lifetime between 0 and 5, but a centrality measure of zero. This figure shows how the upper bound approximates $D(G_i)^2$. For larger p in Figure 9, the node for $n=50$ has a lifetime between 3 and 10. Compared to Figure 8, a higher value of p results in even less sharp bounds. For the larger small-world network, neither the upper nor lower bounds are sharp, although the upper bound tends to be even worse.

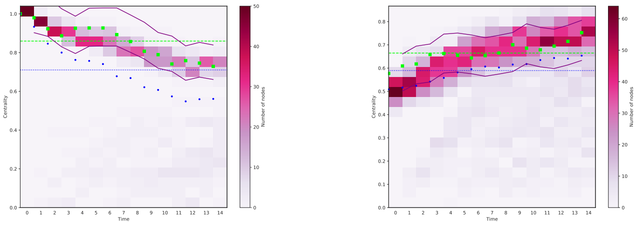

Fig. 4. $\mathcal{G}^w(n,15,p)$, $p=0.05$ with $n=50$ (left) and $n=150$ (right).

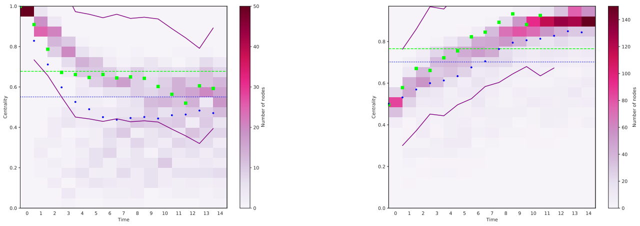

Fig. 5. $\mathcal{G}^w(n,15,p)$, $p=0.15$ with $n=50$ (left) and $n=150$ (right).

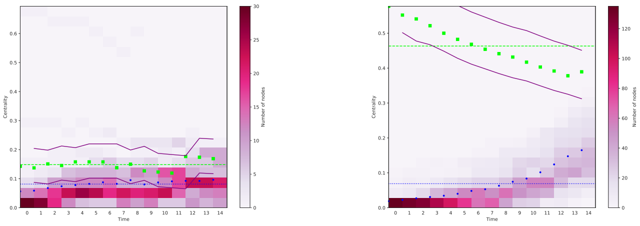

Fig. 6. $\mathcal{G}^s(n,15,p)$, $p=0.05$ with $n=50$ (left) and $n=150$ (right).

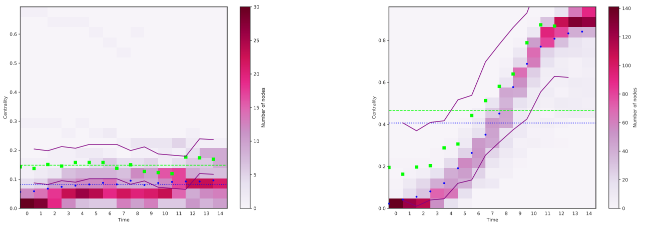

Fig. 7. $\mathcal{G}^s(n,15,p)$, $p=0.15$ with $n=50$ (left) and $n=150$ (right).

For the scale-free random networks in Figures 10 and 11 the situation is similar. However, the heatmap shows that most nodes have small betweenness centrality values, while there are many outliers. In Figure 10 we see that again the lower bound is sharper than the upper bound. However, for $n=50$ we see an example that shows that in some cases the upper bound is suitable to estimate the change over time. Comparing these results to the results shown in Figure 11 again highlights that these bounds get less precise for larger p.

Thus, the upper and lower bounds for betwenness centrality introduced in theorem 6 are not suitable for estimating change over time in any situation. However, the lower bound tends to be sharper than the upper bound, where the behaviour is sometimes unpredictable.

V. Discussion and Outlook

Several approaches exist to manage longitudinal data in networks. All of them bias the output of algorithms analyzing the data. We presented an overview on limitations, possible solutions and open questions to different data schemas for temporal data in social networks based on a generic RDF-inspired approach. In this way, we answered out first research question: How can we model longitudinal social networks in one graph as generic as possible? While not the primary focus of our work, this approach allows the integration of further data from the semantic web making results and approaches directly available for social networks.

We also discussed a second research question. How can we analyse longitudinal social networks with centrality measures? While limiting algorithms to one particular time point or layer does not influence the output, applying them to a network comprising multiple time points does either need adjusted algorithms or reinterpretation. We presented a solution for adjusted approaches and could show that if a network G contains data for different time points $t_1, ..., t_n$, we can still apply centrality measures that were originally developed for a particular time point. We proposed the concepts of average temporal centrality and temporal centrality as core concepts to analyse the temporal development of centrality over the given time, together with a novel representation to compare an individual node against the whole graph. Indeed, we need to reinterpret these results and adapt algorithms. However, while our approach works for all centrality measures, we only considered betwenness centrality and degree centrality and more research needs to consider other centrality measures and methods like community detection. Answering these questions is key to understanding the algorithmic challenges of temporal data in social network analysis.

Our third question was concerned whether we can approximate the change of centrality measures over time. We presented upper and lower bounds for betwenness and degree centrality. However, these bounds need a prior knowledge of the change ratio p between different time points. With an increasing value of p, these bounds become less sharp. More research needs to focus on different types of bounds, in particular for other centrality measures. In addition, a detailed analysis of graph substructures having an influence on the temporal behavior of centrality measures might be fruitful, in particular if p is unknown.

However, rewriting algorithms to analyse longitudinal social networks and the re-interpretation of existing measures and algorithms demands discussion between different scientific domains. Therefore, our paper is also a plea for more interdisciplinary exchange, in particular between mathematics, computer science, social sciences and the humanities.

References

[1] J. Leidwanger, C. Knappett, P. Arnaud, P. Arthur, E. Blake, C. Broodbank, T. Brughmans, T. Evans, S. Graham, E. S. Greene *et al.*, "A manifesto for the study of ancient mediterranean maritime networks," *Antiquity*, vol. 88, no. 342, 2014.

[2] S. de Valeriola, "Can historians trust centrality?" *Journal of Historical Network Research*, vol. 6, no. 1, 2021.

[3] P. Holme and J. Saramäki, "A map of approaches to temporal networks," *Temporal network theory*, pp. 1–24, 2019.

[4] C. Lemercier, "Taking time seriously. how do we deal with change in historical networks?" in *Knoten und Kanten III. Soziale Netzwerkanalyse in Geschichts- und Politikforschung*. Transcript, 2015, pp. 183–211.

[5] S. Lehmann, "Fundamental structures in temporal communication networks," *Temporal Network Theory*, pp. 25–48, 2019.

[6] M. Latapy, T. Viard, and C. Magnien, "Stream graphs and link streams for the modeling of interactions over time," *Social Network Analysis and Mining*, vol. 8, pp. 1–29, 2018.

[7] M. Latapy, C. Magnien, and T. Viard, "Weighted, bipartite, or directed stream graphs for the modeling of temporal networks," *Temporal Network Theory*, pp. 49–64, 2019.

[8] T. P. Peixoto and M. Rosvall, "Modelling temporal networks with markov chains, community structures and change points," *Temporal network theory*, pp. 65–81, 2019.

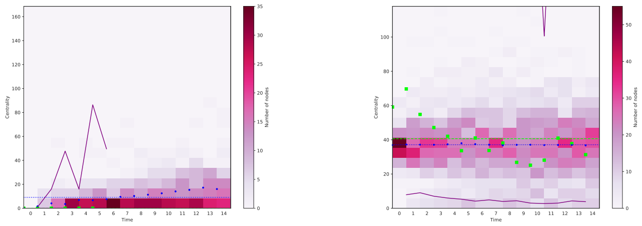

Fig. 8. $\mathcal{G}^w(n, 15, p)$, $p = 0.05$ with $n = 50$ (left) and $n = 150$ (right).

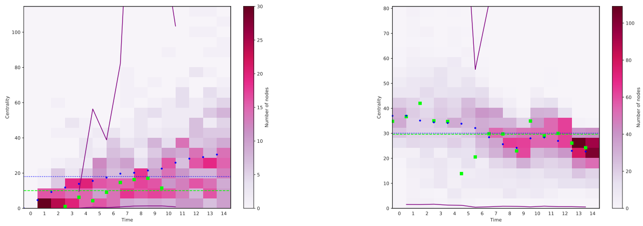

Fig. 9. $\mathcal{G}^w(n, 15, p)$, $p = 0.15$ with $n = 50$ (left) and $n = 150$ (right).

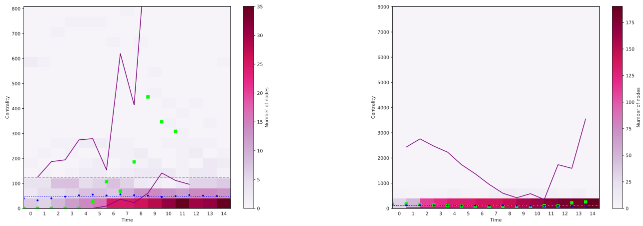

Fig. 10. $\mathcal{G}^s(n, 15, p)$, $p = 0.05$ with $n = 50$ (left) and $n = 150$ (right).

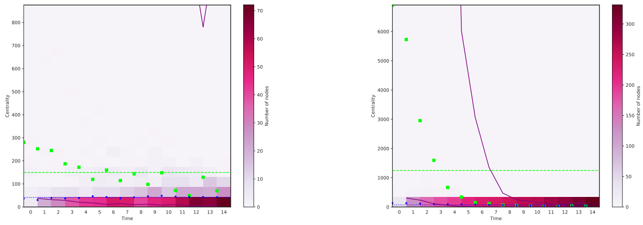

Fig. 11. $\mathcal{G}^s(n, 15, p)$, $p = 0.15$ with $n = 50$ (left) and $n = 150$ (right).

[9] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer, "Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks," *Nature communications*, vol. 5, no. 1, p. 5024, 2014.

[10] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels: Statistical models for time-evolving networks," in *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6*. Springer, 2013, pp. 201–210.

[11] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[12] G. Cencetti, F. Battiston, B. Lepri, and M. Karsai, "Temporal properties of higher-order interactions in social networks," *Scientific reports*, vol. 11, no. 1, p. 7028, 2021.

[13] J. S. Yi, N. Elmqvist, and S. Lee, "Timematrix: Analyzing temporal social networks using interactive matrix-based visualizations," *Intl. Journal of Human–Computer Interaction*, vol. 26, no. 11-12, pp. 1031–1051, 2010.

[14] M. Franzke, T. Emrich, A. Züfle, and M. Renz, "Pattern search in temporal social networks," in *Proceedings of the 21st International Conference on Extending Database Technology*, 2018.

[15] S. Hanneke, W. Fu, and E. P. Xing, "Discrete temporal models of social networks," *Electronic Journal of Statistics*, vol. 4, pp. 585–605, 2010.

[16] R. K. Pan and J. Saramäki, "Path lengths, correlations, and centrality in temporal networks," *Physical Review E*, vol. 84, no. 1, p. 016105, 2011.

[17] D. Taylor, S. A. Myers, A. Clauset, M. A. Porter, and P. J. Mucha, "Eigenvector-based centrality measures for temporal networks," *Multiscale Modeling & Simulation*, vol. 15, no. 1, pp. 537–574, 2017.

[18] A. E. Sizemore and D. S. Bassett, "Dynamic graph metrics: Tutorial, toolbox, and tale," *NeuroImage*, vol. 180, pp. 417–427, 2018.

[19] M. G. Everett and S. P. Borgatti, "The centrality of groups and classes," *The Journal of mathematical sociology*, vol. 23, no. 3, pp. 181–201, 1999.

[20] S. Rasti and C. Vogiatzis, "Novel centrality metrics for studying essentiality in protein-protein interaction networks based on group structures," *Networks*, vol. 80, no. 1, pp. 3–50, 2022.

[21] E.-Y. Yu, Y. Fu, X. Chen, M. Xie, and D.-B. Chen, "Identifying critical nodes in temporal networks by network embedding," *Scientific reports*, vol. 10, no. 1, p. 12494, 2020.

[22] P. Cinaglia and M. Cannataro, "Network alignment and motif discovery in dynamic networks," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 11, no. 1, p. 38, 2022.

[23] J. Dörpinghaus, S. Klante, M. Christian, C. Meigen, and C. Düing, "From social networks to knowledge graphs: A plea for interdisciplinary approaches," *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100337, 2022.

[24] C. Barats, V. Schafer, and A. Fickers, "Fading away... the challenge of sustainability in digital studies." *DHQ: Digital Humanities Quarterly*, vol. 14, no. 3, 2020.

[25] D. Santoro and I. Sarpe, "Onbra: Rigorous estimation of the temporal betweenness centrality in temporal networks," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1579–1588.

[26] J. R. Hobbs and F. Pan, "An ontology of time for the semantic web," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 1, pp. 66–85, 2004.

[27] M. Grüninger, "Verification of the owl-time ontology," in *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*. Springer, 2011, pp. 225–240.

[28] V. Nicosia, J. Tang, C. Mascolo, M. Musolesi, G. Russo, and V. Latora, "Graph metrics for temporal networks," *Temporal networks*, pp. 15–40, 2013.

[29] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[30] T. Schweizer, *Muster sozialer Ordnung: Netzwerkanalyse als Fundament der Sozialethnologie*. Berlin: D. Reimer, 1996.

[31] M. O. Jackson, *Social and Economic Networks*. Princeton: University Press, 2010.

[32] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology*, vol. 105, no. 2, pp. 493–527, 1999.

[33] B. Bollobás, C. Borgs, J. T. Chayes, and O. Riordan, "Directed scale-free graphs." in *SODA*, vol. 3, 2003, pp. 132–139.

[34] B. Bollobás and O. M. Riordan, "Mathematical results on scale-free random graphs," *Handbook of graphs and networks: from the genome to the internet*, pp. 1–34, 2003.

[35] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[36] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[37] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[38] J. S. Kleinfeld, "The small world problem," *Society*, vol. 39, no. 2, pp. 61–66, 2002.

[39] O. Riordan *et al.*, "The diameter of a scale-free random graph," *Combinatorica*, vol. 24, no. 1, pp. 5–34, 2004.

[40] F. Ma, X. Wang, and P. Wang, "Scale-free networks with invariable diameter and density feature: Counterexamples," *Physical Review E*, vol. 101, no. 2, p. 022315, 2020.

[41] L. Gu, H. L. Huang, and X. D. Zhang, "The clustering coefficient and the diameter of small-world networks," *Acta Mathematica Sinica, English Series*, vol. 29, no. 1, pp. 199–208, 2013.

[42] C. Martel and V. Nguyen, "Analyzing kleinberg's (and other) small-world models," in *Proceedings of the twenty-third annual ACM symposium on Principles of distributed computing*, 2004, pp. 179–188.