DOI: 10.15439/2022F43
ISSN 2300-5963 ACSIS, Vol. 32

Centrality Measures in multi-layer Knowledge Graphs

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

***Abstract*—Knowledge graphs play a central role for linking different data which leads to multiple layers. Thus, they are widely used in big data integration, especially for connecting data from different domains. Few studies have investigated the questions how multiple layers within graphs impact methods and algorithms developed for single-purpose networks, for example social networks. This manuscript investigates the impact on the centrality measures of graphs with multiple layers compared to a those measures in single-purpose graphs. In particular, (a) we develop an experimental environment to (b) evaluate two different centrality measures – degree and betweenness centrality – on random graphs inspired by social network analysis: small-world and scale-free networks. The presented approach (c) shows that the graph structures and topology has a great impact on its robustness for additional data stored. Although the experimental analysis of random graphs allows us to make some basic observations we will (d) make suggestions for additional research on particular graph structures that have a great impact on the stability of networks.**

I. Introduction

KNOWLEDGE graphs have been shown to play an important role in recent knowledge mining and discovery, for example in the fields of digital humanities, life sciences or bioinformatics. They also include single purpose networks (like social networks), but mostly they contain also additional information and data, see for example [1], [2], [3]. Thus, a knowledge graph can be seen as a multi-layer graph comprising different data layers, for example social data, spatial data, etc. In addition, scientists study network patterns and structures, for example paths, communities or other patterns within the data structure, see for example [4]. Very few studies have investigated the questions how multiple layers within graphs impact methods and algorithms developed for single-purpose networks, see [5]. This manuscript investigates the impact of a growing part of other layers on centrality measures in a single-purpose graph. In particular, we develop an experimental environment to evaluate two different centrality measures – degree and betweenness centrality – on random graphs inspired by social network analysis: small-world and scale-free networks.

This paper is divided into five sections. The first section gives a brief overview of the state of the art and related work. The second section describes the preliminaries and background. We will in particular introduce knowledge graphs and centrality measures. In the third section, we present the experimental setting and the methods used for this evaluation. The fourth section is dedicated to experimental results and the evaluation. Our conclusions are drawn in the final section.

II. Preliminaries

The term *knowledge graph* (sometimes also called a *semantic network*) is not clearly defined, see [6]. In [7], several definitions are compared, but the only formal definition was related to RDF graphs which does not cover labeled property graphs. As another example, [8] gives a definition of knowledge graphs limited to the definition of important features. Knowledge graphs were introduced by Google in 2012, when the Google Knowledge Graph was published on the use of semantic knowledge in web search, see https://blog.google/products/search/introducing-knowledge-graph-things-not/. This is a representation of general knowledge in graph format. Knowledge graphs also play an important role in the Semantic Web and are also called semantic networks in this context.

Thus, a *knowledge graph* is a systematic way to connect information and data to knowledge. It is thus a crucial concept on the way to generate knowledge and wisdom, to search within data, information and knowledge. Context is the most important topic to generate knowledge or even wisdom. Thus, connecting knowledge graphs with context is a crucial feature.

Definition 1 (Knowledge Graph)**.** *We define a knowledge graph as graph $G = (E, R)$ with entities $e \in E = \{E_1, ..., E_n\}$ coming from formal structures E_i like ontologies.*

The relations $r \in R$ can be ontology relations, thus in general we can say every ontology E_i which is part of the data model is a subgraph of G indicating $O \subseteq G$. In addition, we allow inter-ontology relations between two nodes e_1, e_2 with $e_1 \in E_1$, $e_2 \in E_2$ and $E_1 \neq E_2$. In more general terms, we define $R = \{R_1, ..., R_n\}$ as a list of either inter-ontology or inner-ontology relations. Both E as well as R are finite discrete spaces.

Every entity $e \in E$ may have some additional metainformation which needs to be defined with respect to the application of the knowledge graph. For instance, there may be several node sets (some ontologies, some actors (like employees or stakeholders, for example), locations, ...) $E_1, ..., E_n$ so that $E_i \subset E$ and $E = \cup_{i=1,...,n} E_i$. The same holds for R when several context relations come together such as ”is relative of”, ”has business affiliation”, ”has visited”, etc.

By using formal structures within the graph, we are implicitly using the model of a labeled property graph, see [9] and [10]. Here, nodes and edges form a heterogeneous set. Nodes and edges can be identified by using a single or multiple labels, for example using $\lambda : E \rightarrow \Sigma$, where Σ denotes a set of labels. We need to mention that both concepts are equivalent, since graph databases use the concept of labeled property graphs.

Here, our experimental setting is – without loss of generality – settled in social network analysis (SNA). It is quite obvious that a social network containing actors may easily be extended with other data, for example spacial data (e.g. locations, rooms, towns, countries), or social groups (e.g. companies, clubs), or any other information (e.g. information data about actors). Once a social network is built, we may start to ask questions like “How many friends does actor X have?” or “To how many groups does actor Y belong?”. The mathematical formulation of these questions would be “What is the degree of node X?” and “How many communities C_i can be found such that $Y \in C_i$?”. The mathematical foundations in this and the following sections are based on the works of [11] and [12] unless otherwise noted.

In general, we define a *Graph* $G = (V, E)$ with a set of edges or vertices V – these are actors, locations or any other nodes in the network – and edges E, which describe the relations between nodes. The number of nodes $|V|$ is usually denoted with n. Given two nodes $s =$Simon and $j =$Jerusalem we may add an edge or relation (s, j) between both describing for example, that Simon is or was in Jerusalem. Then we say s and j are *connected* or they are *neighbors*. The *neighborhood* of a vertice v is denoted with $N(v)$ and describes all nodes connected to v. If we are interested in the size of this neighborhood we calculate the node *degree* given by $deg(v) = |N(v)|$.

The neighborhood thus gives information about the connectedness of an actor in the network. This can be useful to illustrate the direct influence of an actor within the complete network, especially for actors with a high node degree. But it is obvious that the amount of relations does not necessarily give a good idea on their quality or how we could use these relations. While the node degree is often used as a measure to create random graphs, it is in general not a good measure in order to analyze particular actors in networks, see [13].

Nevertheless, the *degree centrality* for a node $v \in V$ is given by

$$dc(v) = \frac{deg(v)}{n-1}$$

The output value ranges between 0 and 1 and gives a reference to the direct connections. As discussed, it omits all indirect relations and in particular the node’s position in the network.

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

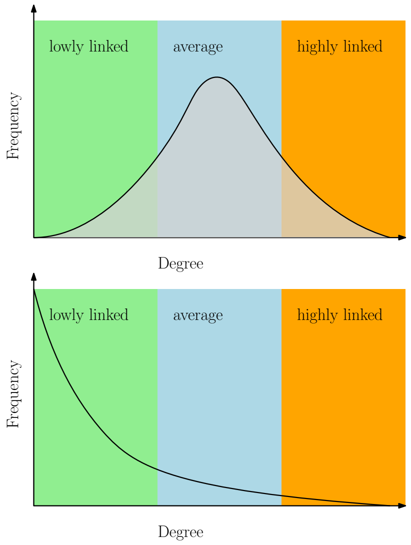

Fig. 1. Top: In random networks the degree distribution follows a given random distribution. Here, most nodes are average linked and an equal number of nodes is lowly and highly linked. Bottom: Real networks often follow other or even no standard random distribution. Here, a scale-free distribution is shown: Most nodes are lowly linked whereas only very few notes are highly linked.

Definition 3 (Small World Network [14])**.** *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 $k \gg 1$.*

In any case, the *degree distribution* provides us with information about the network structure since we can distinguish between sparsely and densely connected networks. While [13] suggests statistical analysis to compute the correlation between attributes of the network and the density of nodes, this will not work for the small networks and the missing statistical values. In any case, although scale-free networks are not an universal characteristic for real-world networks, we might use this approach to get a first overview about the network itself. Random graphs, like the Erdős–Rényi networks, follow a Poisson distribution. Scale-free networks, inspired by real-world social networks, follow a power law. See Figure 1 for two examples of a random graph and a more common distribution in real word networks.

We will now discuss one more property to evaluate nodes and their position in the networks. These properties can be used to calculate statistical parameters, so-called *centrality measures*, cf. [15] and [16]. They answer the question “Which nodes in this network are particularly significant or important?”.

Betweenness analyzes critical connections between nodes and thus gives an indication of individuals that can change the flow of information in a network. This measure is based on paths in a network:

> Much of the interest in networked relationships comes from the fact that individual nodes benefit (or suffer) from indirect relationships. Friends might provide access to favors from their friends, and

> information might spread through the links of a network.[13]

A *path* p in a graph $G = (V, E)$ 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 $i \in \{1, \ldots, t-1\}$. The length $|p|$ of the path p is the total number of edges – not nodes. Thus $|p| = t-1$. The path p links the starting node v_1 and an ending node v_t. In a path, no crossings are allowed, thus $v_i \neq v_j$ for all $i, j \in \{1, ..., t\}$. If all properties of a path are met except that the beginning and the end vertex are the same – that is, $v_1 = v_t$ – we denote this set as a *circle*.

Betweenness centrality was first introduced by [17][1] and considers other indirect connections, see [19]. Given a node v, it calculates the number $P_v(k, j)$, that is, the number of all shortest paths in a network for all beginning and ending nodes $k, j \in V$ that pass through v. If $P(k, j)$ denotes the total number of paths between k and j, the importance of v is given by the ratio of both values. Thus the betweenness centrality according to [13] is given by

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

where n denotes the number of the vertices in the graph. This parameter allows an analysis of the critical links and how often a node lies on such a path. This centrality measure thus answers the questions whether a node can change the flow of information in a network or whether it is a bridge between other nodes, see [19].

While betweenness assumes network flows to be like packages flowing from a starting point to a destination, other measures consider multiple paths: For example, the so-called *eigenvector centrality* – introduced by [20] – measures the location of directly neighboring nodes in the network. For the eigenvector centrality, we "count walks, which assume that trajectories can not only be circuitous, but also revisit nodes and lines multiple times along the way."[21] This measure not only classifies the direct possibility to influence neighbors, but also ranks the indirect possibility to influence the whole network. For a detailed mathematical background we refer to [13].

Less popular measures are Katz prestige, and Bonacich's measure, see [13]. It has been shown that these measures are closely related, see [22].

III. Method

We evaluate the degree centrality and betweenness centrality on random graphs. First, we consider Scale-Free Networks with n nodes, see [13]. Moreover, [23] 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 purpose, we scale the number of nodes n and use $\alpha = 0.41$, $\beta = 0.54$, and $\gamma = 0.05$. We chose this random graph model since it is generic and feasible for computer simulations for measuring and evaluation purposes, see [24], [25].

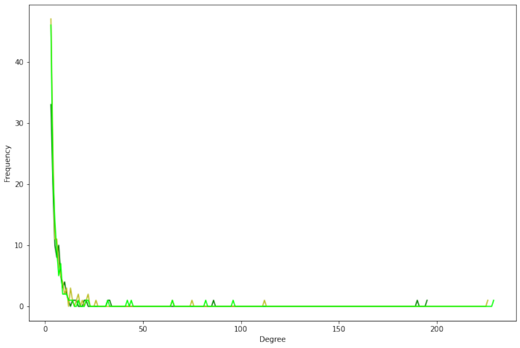

Fig. 2. Frequency of nodes with a given degree for three random Scale-Free Networks with $n = 150$ nodes.

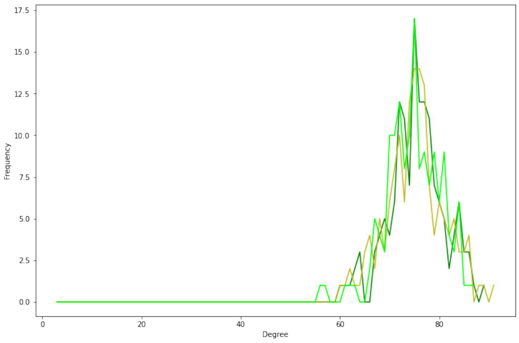

Fig. 3. Frequency of nodes with a given degree for three Newman-Watts-Strogatz small-world random graph with $n = 500$ nodes.

Figure 2 shows the frequency of nodes (y-axis) with a particular degree (x-axis) for three random networks with $n = 150$ nodes. Compared to Figure 2, Figure 1 clearly shows the scale-free distribution, in which many nodes have a small degree and only few nodes have a very large degree: most nodes are hence lowly linked. Thus these small-degree nodes lead to a few communities which are highly connected.

The second random graph uses a fixed degree distribution and is widely known as Newman-Watts-Strogatz small-world random graph [26]. The algorithm to create such as graph takes a number of nodes n, the number of k nearest neighbors that form a ring topology and the probability p for adding a new edge. A small-world graph contains only small average paths and thus has a small diameter, see [13]. Some studies like [27] study the relation between scale-free and small-world networks, in particular the relationship between the average

[1] Initially introduced for symmetric relations – undirected graphs – it was extended to asymetric relations – directed graphs – by [18].

path length and local clusterings. In general, it is possible to generate scale-free networks with small-world attributes, see [28].

Figure 3 shows the frequency of nodes with a given degree for three random networks with $n = 500$ nodes. Compared to Figure 1, Figure 3 clearly shows the Poisson distribution with many nodes having an average degree. Together with Figure 2 it also illustrates the “long tail” of the scale-free distribution, see [13].

We will now evaluate how graph structures and in particular measures change when additional information are stored in extra layers. We partition a graph into an uncolored part that contains the ‘original’ data and into a part with blue nodes in which novel ‘extra’ data stored. These blue nodes simulate one or more new layers in the knowledge graph. One could imagine a graph in which every node represents a scientist in a social network, and two persons are connected whenever they are tied in the network (e.g. friends, collaborators, etc.). We now want to add more information to our graph by adding blue nodes. Every blue node represents a specific conference. Two blue nodes are connected whenever the conferences address - at least partly - the same community. A scientist is connected to a conference whenever they attended the workshop. The original graph is here the set of scientists, the blue nodes (the conferences) form a new layer, in which the extra data is stored.

Thus, given a random graph $G = (V, E)$, a next step comprises a probability p_b for blue nodes which leads to a graph G with blue nodes $B \subset V$. We first compute the centrality measures for all nodes in $V \setminus B$ in the graph $G = (V, E)$. Then we compute those measures for all nodes in $G \setminus B$, this time in the Graph $G \setminus B = (V \setminus B, E)$. Thus, we have two vectors $c_1, c_2 \in \mathbb{R}^n$ where here, n is the number of nodes in $V \setminus B$. We denote c_i by $c_i = \left(c_i^1, c_i^2, c_i^3, ...\right)$.

While comparing two vectors, we are interested in two values. The first one is the total number of misordered elements, that is, the total number of positions on which the elements differ from each other. The second value that we compute in order to compare two vectors is the number of moved elements. For this we count those elements that have a different predecessor and / or successor in the first vector compared to the second one.

Example III.1. *Let $c_1 = [1, 2, 3, 4, 5]$, $c_2 = [5, 3, 2, 1, 4]$ and $c_3 = [1, 5, 2, 3, 4]$. If c_1 is the original ordering, we see that c_2 has a totally different order. In c_3 the entry 5 is moved, but the rest of the list is unchanged, although still 4 elements are on the wrong location. Hence, the number of misordered elements in c_1 compared to c_2 is 5. The number of moved elements is 5 and 1.*

To identify both errors, we first define function e:

$$e(i, c_1, c_2) = \begin{cases} 0 & c_1^i = c_2^i \\ 1 & c_1^i \neq c_2^i \end{cases}$$

That is, $e(i, j, c_1, c_2) = 1$ if the element on the ith position of c_1 differs from the element on the jth position in c_2. To shorten notation, we write $e(i, c_1, c_2)$ whenever $i = j$.

	ϵ	ϵ_N	ϵ	ϵ_N	ϵ	ϵ_N
Scale-Free	$n = 150$		$n = 300$		$n = 500$	
Mean	0.95	0.46	0.97	0.47	0.98	0.48
Small-World	$k = 4$		$k = 8$		$k = 50$	
Mean	0.97	0.97	0.97	0.96	0.95	0.96

TABLE I
MEAN VALUES FOR DEGREE CENTRALITY ERRORS.

Let x be an element contained in every c_u, $u \in \mathbb{N}$. Then $p(x, c_u)$ denotes the predecessor of element x in c_u and $s(x, c_u)$ denotes the successor of x in c_u. If x is the first element in c_u, then $p(x, c_u) = \emptyset$. If x is the last element of c_u, then $s(x, c_u) = \emptyset$. With these definitions, we define e_N:

$$e_N(x, c_1, c_2) = \begin{cases} 1 & \text{if } p(x, c_1) = \emptyset \text{ and } s(x, c_1) \neq s(x, c_2), \\ & \text{or } s(x, c_1) = \emptyset \text{ and } p(x, c_1) \neq p(x, c_2), \\ & \text{or } s(x, c_1) \neq s(x, c_2) \text{ and } p(x, c_1) \neq p(x, c_2), \\ 1/2 & \text{if } s(x, c_1) \neq s(x, c_2) \text{ and } p(x, c_1) = p(x, c_2), \\ & \text{or } s(x, c_1) = s(x, c_2) \text{ and } p(x, c_1) \neq p(x, c_2), \\ 0 & otherwise. \end{cases}$$

In other words, we consider the predecessor of an element in c_1 and check if this element is still a predecessor of this element in c_2, and analyse analoguously the successor of an element.

With this, we define two error measures ϵ and ϵ_N:

$$\epsilon(c_1, c_2) = \sum_{i=1}^{n} e(i, c_1, c_2)$$

$$\epsilon_N(c_1, c_2) = \sum_{x \in c_1} e_N(x, c_1, c_2)$$

Example III.2. *Let’s reconsider Example III.1: Recall that $c_1 = [1, 2, 3, 4, 5]$, $c_2 = [5, 3, 2, 1, 4]$ and $c_3 = [1, 5, 2, 3, 4]$. Then, $\epsilon(c_1, c_2) = 5$ and $\epsilon_N(c_1, c_2) = 5$. Moreover, $\epsilon(c_1, c_3) = 4$ and $\epsilon_N(c_1, c_3) = 2.5$.*

We will now analyze different scenarios to evaluate the impact of additional blue nodes on a scale-free and a small-world network.

IV. RESULTS

A. Degree Centrality

The Degree Centrality was evaluated with errors ϵ and ϵ_N for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$, see Figure 4) and Newman-Watts-Strogatz small-world random graphs ($n = 150$, $k \in \{4, 8, 50\}$, see Figure 5). The mean values are given in Table I.

Here, we see that the Small-World graph has a very high error rate for both ϵ and ϵ_N even for small p_B. In particular, the values are rather constant, no matter what value was chosen. In addition, the graph topology for different values of k has only very little impact on the error rate. Thus, even small changes

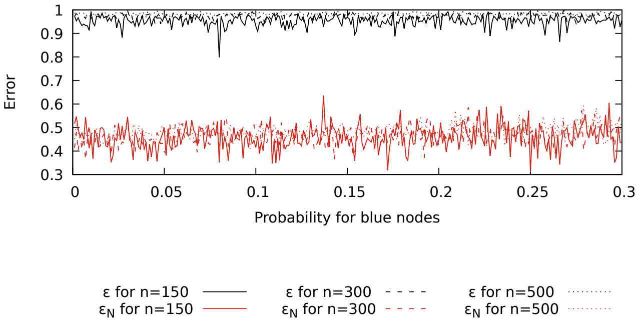

Fig. 4. Degree Centrality errors for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$) for different values of p_B between 0 and 0.3.

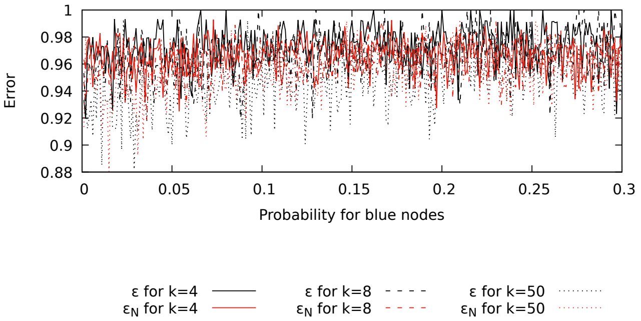

Fig. 5. Degree Centrality errors for Newman-Watts-Strogatz small-world random graph ($n = 150$, $k \in \{4, 8, 50\}$) for different values of p_B between 0 and 0.3.

in the graph structure (a very small value for p_B) have a great impact on the degree centrality. Since Small-World graphs have a high level of local clustering, the random exclusion of blue nodes will most likely effect not only one cluster, but also other clusters. This changes not only the position, but also the ordering of node degrees.

A different scenario occurs when considering Scale-Free graphs. Again we see a very high error rate for ϵ, even for small p_B. The values for ϵ_N are usually near to .5 (mean values 0.46, 0.47, 0.48). Neither the graph size n nor the value for p_B has an impact on these errors. Here, we see the scale-free distribution: the blue nodes do change the position of the degree centrality, but while they also change the ordering within clusters, they do not affect the complete ordering due to the longer distance between nodes.

B. Betweenness Centrality

The Betweenness Centrality was evaluated with errors ϵ and ϵ_N for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$, see Figure 6) and Newman-Watts-Strogatz small-world random graphs ($n = 150$, $k \in \{4, 8, 50\}$, see Figure 7). The mean values are given in Table II.

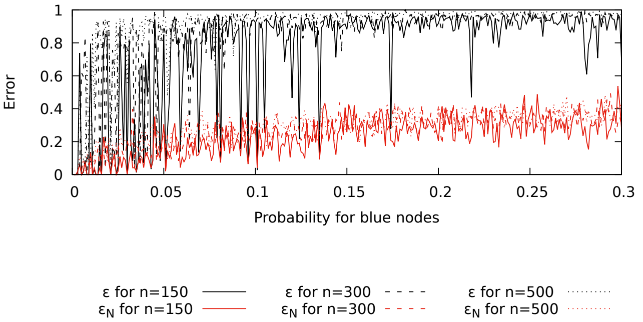

Fig. 6. Betweenness Centrality errors for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$) for different values of p_B between 0 and 0.3.

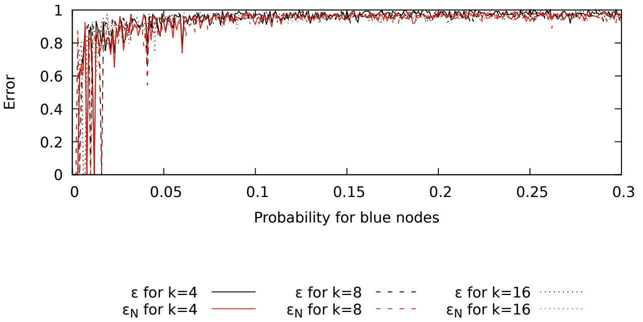

Fig. 7. Betweenness Centrality errors for Newman-Watts-Strogatz small-world random graph ($n = 150$, $k \in \{4, 8, 50\}$) for different values of p_B between 0 and 0.3.

	ϵ	ϵ_N	ϵ	ϵ_N	ϵ	ϵ_N
Scale-Free	$n = 150$		$n = 300$		$n = 500$	
Mean	0.77	0.23	0.87	0.27	0.91	0.29
Small-World	$k = 4$		$k = 8$		$k = 50$	
Mean	0.94	0.92	0.94	0.92	0.94	0.93

TABLE II
MEAN VALUES FOR BETWEENNESS CENTRALITY ERRORS.

Betweenness centrality (see Figure 6) in scale-free graphs is very much influenced by the choice for p_B. Again, the total error ϵ becomes very high although there are several outliers. More interesting is again the ordering error ϵ_N: although the error increases with a rising value of p_B, it remains very low. Again, the number of nodes n has only very little impact on the error measures.

Here, again, the Small-World graph has a very high error rate for both ϵ and ϵ_N although not for very small p_B, see Figure 7. In particular, we may find a boundary p'_B so that the values are rather constant for $p_B > p'_B$. Again, the graph topology for different values of k has only very little impact on the error rate. Thus, even small changes in the graph structure (a very small value for p_B) have a great impact on the betweenness centrality. Thus, the random choice of blue nodes again destroys the structures of local clustering which will most likely effect not only one cluster, but also other clusters.

We will now consider two graph structures to take a closer look at their impact on the error measures.

C. Cliques

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$. The nodes in $G \setminus B = (V \setminus B, E)$ are denoted by $v_1, \ldots, v_{n-|B|}$ while the nodes in B are denoted by $v_{n-|B|+1}, \ldots, v_n$. We further assume that $G \setminus B$ is still connected. Let $dc(G)$ be the vector containing the degree centrality measures for all nodes v in G in descending order, where - after the computation of $dc(v)$ for all $v_1, \ldots, v_n \in V(G)$ - the values for all $v \in B$, that is, v_i with $i = n - |B| + 1, \ldots, n$, are deleted. Hence,

$$dc(G) = \big(dc(v_1), dc(v_2), ..., dc(v_{n-|B|})\big)$$

with $dc(v_j) \geq dc(v_{j+1})$ for all $j \in \{1, \ldots, n - |B|\}$.

Let $bc(G)$ be the vector containing the betweenness centrality measures for all nodes in G in descending order, where - after the computation of $bc(v)$ for all $v_1, \ldots, v_n \in V(G)$ - the values for all $v \in B$, that is, v_i with $i = n - |B| + 1, \ldots, n$, are deleted. That is,

$$bc(G) = (bc(v_1), bc(v_2), ..., bc(v_n - |B| + 1))$$

with $bc(v_j) \geq bc(v_{j+1})$ for all $j \in \{1, \ldots, n - |B|\}$. Let $p_{dc}(v)$ respectively $p_{bc}(v)$ be the position of node v in the vector $dc(G)$ respectively $bd(G)$. When it is clear from the context which vector is meant, we omit the index and simply write $p(v)$.

We may now prove some very basic observations on how a single blue node may influence the different error measures ϵ and ϵ_N, given that the blue node is part of a cluster in G. Here, with a cluster or a clique we denoate a complete subgraph of G.

Lemma IV.1. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then*

$$\epsilon(dc(G), dc(G \setminus B)) \leq n - 1 - \min_{v \in N(u)} p_{dc}(v)$$

holds.

Proof. Let $a_1 = dc(G)$ and $a_2 = dc(G \setminus B)$. The only nodes which are affected by a decreasing degree centrality are those in the neighborhood $N(u)$ of the blue node u, since for $v \in N(u)$, only one node in the neighborhood of v is removed in $G \setminus B$ compared to G. Thus,

$$a_1^{p(v)} = a_2^{p(v)} - 1 \;\; \forall v \in N(u)$$

holds. Observe that $\min_{v \in N(u)} p_{dc}(v)$ denotes the smallest position in $dc(G)$ of a node in $N(u)$ (that is, the highest ranked neighbor of u in $dc(G)$). All nodes in $dc(G)$, that are higher ranked are not affected by the deletion of u. Recall that $dc(G)$ only has $n - |B| = n - 1$ entries. Thus, at most $n - 1 - \min_{v \in N(u)} p_{dc}(v)$ nodes change their position in $dc(G \setminus B)$ compared to $dc(G)$. □

We can rely on the same basic observations for the error measure ϵ_N:

Lemma IV.2. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then*

$$\epsilon_N(dc(G), dc(G \setminus B)) \leq k - 1$$

holds.

Proof. Let $a_1 = dc(G)$ and $a_2 = dc(G \setminus B)$. Again, the only nodes which are affected by a decreasing degree centrality are those in the neighborhood of the blue node, that is the set $v \in N(u)$. Here, only one node in the neighborhood of these nodes is removed in $G \setminus B$. While the internal order of all nodes in $G \setminus \{C_k \setminus \{u\}\}$ does not change and the internal order of the $k - 1$ nodes in $C_k \setminus \{u\}$ remains untouched as well, at most the $k - 1$ nodes in $C_k \setminus \{u\}$ are shifted to a certain degree to the right, since their value in $dc(G \setminus B)$ decreased compared to $dc(G)$. Every vertex in $C_k \setminus \{u\}$ hence contributes at most 1 to the sum computed in $\epsilon_N(dc(G), dc(G \setminus B))$, which leads to the upper bound $k - 1$. □

The herefore stated lemma explains why this error increases for small-world networks: The node degree is high and a lot of local clusters exist.

Since betweenness centrality is also affected by the global structure of the graph, counting all shortest paths, the situation is slightly different.

Lemma IV.3. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then*

$$\epsilon(bc(G), bc(G \setminus B)) \leq \begin{cases} 0 & if\ d(u) = k - 1 \\ \sum_{w \neq y} P_u(w, y) & otherwise. \end{cases}$$

Proof. **Case 1** $d(u) = k - 1$: In this case, u only lies on shortest paths between u and any node in $G \setminus \{u\}$, since $N(u) = C_k \setminus \{u\}$. That is, every shortest path through C_k ignores u, since u is only connected to nodes within $C_k \setminus \{u\}$, see Figure 8.

Thus, the number of shortest paths in G that include u is $n - 1$, that is,

$$\sum_{w \neq y, w \neq u, y \neq u} P_u(w, y) = n - 1$$

holds. Hence, for $v \in G \setminus B$,

$$\sum_{w \neq y, w \neq u, y \neq u} P_v^{G \setminus B}(w, y) = \sum_{w \neq y, w \neq u, y \neq u} P_v^{G}(w, y) - 1$$

where P^G denotes a shortest path in the graph G. In other words: There exists a value $b \in \mathbb{R}$ such that for every $v \in G \setminus B$,

$$bc(G)^{p(v)} = bc(G \setminus B)^{p(v)} - b.$$

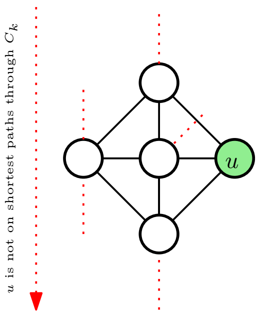

Fig. 8. If $d(u) = k - 1$, u lies only on shortest paths between u and any other node in $G \setminus \{v\}$. Any shortest path through C_k ignores u, since u is only connected to nodes within C_k.

Hence, the ordering of the values in $bc(G)$ compared to $bc(G \setminus B)$ does not change and the considered diffence is 0.

Case 2 $d(u) \geq k$: In this case, u lies on

$$s = \sum_{w \neq y, w \neq u, y \neq u} P_u(w, y)$$

shortest paths within the graph. Thus, removing this node will affect all these paths and thus the ordering of at most s nodes will be changed. Hence $\epsilon(bc(G), bc(G \setminus B)) \leq s$. □

Here, we considered highly connected blue nodes within clusters and their impact on the error measures. Usually, external data may not only be added to such dense structures but also to single nodes. Thus, we will now discuss the impact of nodes with degree 1.

D. Nodes with Degree 1

We now consider the special case in which the only existing blue node has degree 1.

Lemma IV.4. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$. Then*

$$\epsilon\left(dc\left(G\right), dc\left(G \setminus B\right)\right) \leq p_{dc}(v)$$

holds.

Proof. Since $d(u) = 1$, there is only one node v in $N(u)$. This node is on position $p_{dc}(v)$ in $dc(G)$ and it is the only node affected in $G \setminus B$. Thus, at most $p_{dc}(v)$ nodes are affected. □

A similar observation can be made for ϵ_N:

Lemma IV.5. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $|N(u)| = 1$, that is, $d(u) = 1$. Then*

$$\epsilon_N(dc(G), dc(G \setminus B)) \leq 2$$

holds.

Proof. Let $a_1 = dc(G)$ and $a_2 = dc(G \setminus B)$. If $d(u) = 1$, there is only one node v in $N(u)$. This node is on position $p_{dc}(v)$ in a_1 and it is the only node affected in $G \setminus B$. Thus, either it is at the same position in a_2 or on a different one which will affect the nodes on position $p_{dc}(u) - 1$ and $p_{dc}(u) + 1$ in a_1 and $p_{dc}(u) - 1$ and $p_{dc}(u) + 1$ in a_2. Thus

$$\epsilon_N(dc(G), dc(G \setminus B)) \leq 2$$

holds. □

While degree centrality is a local centrality measure, betweenness is a global measure. Since u lies only on shortest paths between u and any other node in the graph, we can make the following observation:

Lemma IV.6. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$ and let*

$$t = \max\left\{n - 1, \sum_{w \neq y} P_v(w, y)\right\}.$$

Then

$$\epsilon(bc(G), bc(G \setminus B)) \leq t$$

holds.

Proof. Let $a_1 = bc(G)$ and $a_2 = bc(G \setminus B)$.

Since $d(u) = 1$, there is only one particular node $v \in N(u)$. Thus, all shortest paths containing u will include v. The highest impact of removing u will hence be on v. Moreover, $n - 1$ shortest paths that include u exist in $G \setminus B$. Thus, the first part of the maximum holds.

In general, $s = \sum_{w \neq y} P_v(w, y)$ shortest paths in G include v. Thus, removing u will also change the betweenness for s nodes and the second part of the maximum holds. □

The evaluation of ϵ_N will—in general—decrease the worst-case scenario:

Lemma IV.7. *Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$ and let*

Let

$$z = \max\left\{2 \sum_{w \neq y} P_u(w, y), 2 \sum_{w \neq y} P_v(w, y)\right\}.$$

Then

$$\epsilon_N(bc(G), bc(G \setminus B)) \leq z$$

holds.

Proof. Let $a_1 = bc(G)$ and $a_2 = bc(G \setminus B)$. Since $d(u) = 1$, there is only one particular node $v \in N(u)$. Thus, all shortest paths containing u will include v. Moreover, $n - 1$ such paths exist and the highest impact of removing u will be on v. But since we are interested in the ordering of nodes, the total number of reordered entries in a_2 may be just a factor. We can estimate this factor with the total number of shortest paths containing u which is $\sum_{w \neq y} P_u(w, y)$.

In general, again, $s = \sum_{w \neq y} P_v(w, y)$ shortest paths in G include v. Thus, removing u will also change the betweenness for s nodes and the second part of the maximum holds. □

These error estimations are not sharp. In addition, if the size of B increases, it will be even more challenging to specify the error rates. But together with our experimental results, these estimations offer us a first impression of problematic graph structures having a great impact on ϵ and ϵ_N.

We could show that blue nodes in clusters have a great influence on both ϵ and ϵ_N while those nodes with a small neighborhood have a rather small influence on ϵ_N. This gives a first idea why in general scale-free networks are more robust regarding ϵ_N. The degree centrality is only influenced by local structures but in general the errors are higher while the betweenness centrality is in general more complex and the results of this paper can only give some hints, but further research needs to be done here.

V. Discussion and Outlook

This paper investigates the impact on two particular centrality measures of graphs with multiple layers compared to single-purpose graphs. We presented an experimental environment to evaluate two different centrality measures – degree and betweenness centrality – on random graphs inspired by social network analysis: small-world and scale-free networks. The result clearly shows that the graph structures and topology has a great impact on its robustness for additional data stored. In particular, we could identify nodes with a high node degree and closely connected communities or clusters as problematic for reordering the centrality measures. Thus, we could show that small-world networks are rather less robust than scale-free networks.

Although the experimental analysis of random graphs allows us to make some basic observations, we could also present some very preliminary error approximations for two cases: A node within a cluster C_k and a node v with $d(v) = 1$. These results underline the experimental results. We need to mention that a lot of research needs to be done in this field, because we only considered degree and betweenness centrality.

In particular, we can identify the following questions for further research: Is it possible to find good error approximations for larger sets of blue nodes B? How do ϵ and ϵ_N behave on any given node $v \in B \subset V$ with $d(v) = m$? What are (other) graph structures that have a great impact on the stability of networks for degree, betweenness and other centralities?

To sum up, it is valid to extend single-purpose networks with data from other sources. In particular, we considered random social networks as a basis. Thus, extending social networks with other information layers is possible, although it will change the behavior of measurements like network centrality. The effect highly depends on the given graph structure. More interdisciplinary research is needed to investigate the impact on real-world data within the context of humanities. In addition, further research needs to be done on the robustness of other centrality measures.

References

[1] D. Suárez, J. M. Díaz-Puente, and M. Bettoni, “Risks identification and management related to rural innovation projects through social networks analysis: A case study in spain,” *Land*, vol. 10, no. 6, p. 613, 2021.

[2] L. M. Berhan, A. L. Adams, W. L. McKether, and R. Kumar, “Board 14: Social networks analysis of african american engineering students at a pwi and an hbcu–a comparative study,” in *2019 ASEE Annual Conference & Exposition*, 2019.

[3] C. Rollinger, “Amicitia sanctissime colenda,” *Freundschaft und soziale Netzwerke in der Späten Republik*, 2014.

[4] J. Dörpinghaus and A. Stefan, “Knowledge extraction and applications utilizing context data in knowledge graphs,” in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019, pp. 265–272.

[5] G. Rossetti, S. Citraro, and L. Milli, “Conformity: A path-aware homophily measure for node-attributed networks,” *IEEE Intelligent Systems*, vol. 36, no. 1, pp. 25–34, 2021.

[6] D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, and A. Wahler, *Introduction: What Is a Knowledge Graph?* Cham: Springer International Publishing, 2020, pp. 1–10. [Online]. Available: https://doi.org/10.1007/978-3-030-37439-6_1

[7] L. Ehrlinger and W. Wöß, “Towards a definition of knowledge graphs.” *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. , no. 48, 2016.

[8] H. Paulheim, “Knowledge graph refinement: A survey of approaches and evaluation methods,” *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[9] M. A. Rodriguez and P. Neubauer, “The graph traversal pattern,” in *Graph data management: Techniques and applications*. IGI Global, 2012, pp. 29–46.

[10] ——, “Constructions from dots and lines,” *Bulletin of the American Society for Information Science and Technology*, vol. 36, no. 6, pp. 35–41, 2010.

[11] R. Diestel, *Graphentheorie*. Berlin: Springer, 2012, vol. 4. Auflage, korrigierter Nachdruck 2012.

[12] J. Matoušek, J. Nešetřil, and H. Mielke, *Diskrete Mathematik*. Berlin: Springer, 2007.

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

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

[15] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[16] P. J. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*, ser. Structural Analyses in the Social Sciences, 27. Cambridge: University Press, 2005, vol. .

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

[18] D. R. White and S. P. Borgatti, “Betweenness centrality measures for directed graphs,” *Social networks*, vol. 16, no. 4, pp. 335–346, 1994.

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

[20] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” *Journal of mathematical sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[21] S. P. Borgatti, “Centrality and network flow,” *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[22] M. Ditsworth and J. Ruths, “Community detection via katz and eigenvector centrality,” *arXiv preprint arXiv:1909.03916*, 2019.

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

[24] 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.

[25] 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.

[26] M. Newman and D. Watts, “Renormalization group analysis of the small-world network model,” *Physics Letters A*, vol. 263, no. 4, pp. 341–346, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0375960199007574

[27] J. Aarstad, H. Ness, and S. A. Haugland, “In what ways are small-world and scale-free networks interrelated?” in *2013 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2013, pp. 1483–1487.

[28] K. Klemm and V. M. Eguiluz, “Growing scale-free networks with small-world behavior,” *Physical Review E*, vol. 65, no. 5, p. 057102, 2002.