LEUPHANA
UNIVERSITÄT LÜNEBURG

Robustness of Centrality Measures

Von der Fakultät Wirtschaftswissenschaften
der Leuphana Universität Lüneburg
zur Erlangung des Grades

Doktor der Naturwissenschaften
— Dr. rer. nat. —

genehmigte Dissertation von
Christoph Martin

geboren am 12. Juli 1988 in Hamburg

Eingereicht am:	27.01.2020
Mündliche Verteidigung (Disputation):	03.06.2020

Erstbetreuer und Erstgutachter:	Prof. Dr. Peter Niemeyer
Zweitgutachter:	Prof. Dr. Burkhardt Funk
Drittgutachter:	Prof. Dr. Frank W. Takes

Die einzelnen Beiträge des kumulativen Dissertationsvorhabens sind oder werden wie folgt veröffentlicht:

Martin, C. (2018). The Impact of Partially Missing Communities on the Reliability of Centrality Measures. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), Complex Networks & Their Applications VI (pp. 41–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-72150-7_4.

Martin C., and Niemeyer P. (2019). Influence of measurement errors on networks: Estimating the robustness of centrality measures. Network Science 7, 180–195. https://doi.org/10.1017/nws.2019.12.

Martin, C., & Niemeyer, P. (in press). On the impact of network size and average degree on the robustness of centrality measures. Network Science.

Martin, C., & Riebeling, M. (2020). A process for the evaluation of node embedding methods in the context of node classification. ArXiv E-Prints, arXiv:2005.14683.

Veröffentlichungsjahr: 2020

Veröffentlicht im Onlineangebot der Universitätsbibliothek unter der URL:
http://www.leuphana.de/ub

It is not enough to be in the right place at the right time.
You should also have an open mind at the right time.

Paul Erdős

Danksagung

An dieser Stelle bedanke ich mich besonders bei meinem Doktorvater, Herrn Professor Dr. Peter Niemeyer, für die Motivation, die Unterstützung, die sehr hilfreichen, teilweise auch kritischen Diskussionen und das mir entgegengebrachte Vertrauen. Ohne seine Hartnäckigkeit und Geduld wäre diese Dissertation nicht zustande gekommen.

Darüber hinaus gilt auch den folgenden Personen mein besonderer Dank für ihre große Unterstützung während der Arbeit an meiner Dissertation: Professor Dr. Burkhardt Funk, Professor Dr. Frank Takes, Vincent Bremer, Dennis Becker, Dr. Martin Stange, Dr. Thomas Hansmann, Sebastian Mair, Felix Krieger, meiner Familie und meinen Freunden.

Ebenfalls bedanke ich mich bei der Leuphana Universität Lüneburg für die Unterstützung und insbesondere für mein Promotionsstipendium.

Abstract

Network analysis methods have long been used in the social sciences. About 25 years ago, these methods gained popularity in various other domains and many real-world phenomena have been modeled using networks. Well-known examples include (online) social networks, economic networks, web graphs, metabolic networks, infrastructure networks, and many more.

Technological development made it possible to store and process data on a scale not imaginable decades ago — a development that also includes network data. A particular characteristic of network data is that, unlike standard data, the objects of interest, called nodes, have relationships to (possibly all) other objects in the network. Collecting empirical data is often complicated and cumbersome, hence, the observed data are typically incomplete and might also contain other types of errors. Because of the interdependent structure of network data, these errors have a severe impact on network analysis methods.

This cumulative dissertation is about the impact of erroneous network data on centrality measures, which are methods to assess the position of an object, for example a person, with respect to all other objects in a network. Existing studies have shown that even small errors can substantially alter these positions. The impact of errors on centrality measures is typically quantified using a concept called robustness.

The articles included in this dissertation contribute to a better understanding of the robustness of centrality measures in several aspects. It is argued why the robustness needs to be estimated and a new method is proposed. This method allows researchers to estimate the robustness of a centrality measure in a specific network and can be used as a basis for decision making. The relationship between network properties and the robustness of centrality measures is analyzed. Experimental and analytical approaches show that centrality measures are often more robust in networks with a larger average degree. The study of the impact of non-random errors on the robustness suggests that centrality measures are often more robust if missing nodes are more likely to belong to the same community compared to missingness completely at random. For the development of imputation procedures based on machine learning techniques, a process for the evaluation of node embedding methods is proposed.

Contents

Contents ix

List of Figures xi

List of Tables xiii

1 Introduction 1
1.1 Network model, network data, and measurement errors . . . 2
1.2 Methods . . . 7
1.2.1 Centrality measures . . . 8
1.2.2 Random graph models . . . 10
1.3 Inaccurate network data and centrality measures . . . 12
1.4 Contribution . . . 14
1.5 Conclusions . . . 17
1.6 References . . . 18

2 The impact of partially missing communities on the reliability of centrality measures 27
2.1 Introduction . . . 29
2.2 Methods & experimental setup . . . 31
2.2.1 Data . . . 32
2.2.2 Quantifying measurement errors and reliability . . . 33
2.2.3 Experimental setup . . . 35
2.2.4 Statistical analysis . . . 35
2.3 Results . . . 36
2.4 Discussion . . . 40
2.5 References . . . 41

3 Influence of measurement errors on networks: Estimating the robustness of centrality measures 45
3.1 Introduction . . . 47

3.2 Basic concepts . . . 49
3.3 How to estimate the robustness . . . 52
3.4 Experiments . . . 53
3.4.1 Experiments based on random graphs . . . 54
3.4.2 Experiments based on real-world networks . . . 58
3.4.3 Discussion of standard deviations . . . 67
3.5 Summary and recommendations . . . 70
3.6 References . . . 72

4 On the impact of network size and average degree on the robustness of centrality measures 77
4.1 Introduction . . . 79
4.2 Methods . . . 82
4.2.1 Centrality measures . . . 82
4.2.2 Error mechanisms . . . 83
4.2.3 Robustness of centrality measures . . . 83
4.2.4 Random graph models . . . 84
4.3 Experiments with empirical networks . . . 85
4.3.1 Experimental setup and data . . . 85
4.3.2 Observations for empirical networks . . . 87
4.4 Experiments with random graphs . . . 90
4.4.1 Experiments with constant average degree . . . 90
4.4.2 Experiments with increasing average degree . . . 92
4.5 Analysis and discussion of the degree . . . 96
4.5.1 Analytical approach for the robustness . . . 97
4.5.2 Expressions for the robustness and error types . . . 98
4.5.3 The case of constant average degree . . . 100
4.5.4 The case of increasing average degree . . . 101
4.6 Conclusions and final remarks . . . 104
4.7 References . . . 107

5 A process for the evaluation of node embedding methods in the context of node classification 113
5.1 Introduction . . . 115
5.2 Node embeddings . . . 116
5.3 A process for the comparison of node embedding methods . 117
5.4 Case study for the comparison process . . . 120
5.5 Conclusions . . . 125
5.6 References . . . 127

List of Figures

1.1 Components of a network model 3
1.2 Example for the impact of erroneous network data on centrality measures. 6
(a) Error-free network 6
(b) Observed, erroneous network 6
(c) Most central nodes 6

2.1 Results for the CRG models 37

3.1 Example for an error mechanism as a random graph 50
3.2 Results for ER graphs 56
3.3 Results for BA graphs 57
3.4 Results for the betweenness centrality in real-world networks 59
3.5 Results for the closeness centrality in real-world networks . . 60
3.6 Results for the degree centrality in real-world networks . . . 61
3.7 Results for the eigenvector centrality in real-world networks 62
3.8 Results for the PageRank in real-world networks 63
3.9 Results for the additional experiment (Dolphin network) . . 65
3.10 Results for the additional experiment (Protein network) . . . 66

4.1 Examples for the behavior of the robustness in empirical networks 90
4.2 Results for the robustness of centrality measures in BA graphs 91
4.3 The behavior of the robustness in ER graphs with increasing network size 92
4.4 Configuration model results for one network (CA-HepTh) . . 94
4.5 Robustness of the degree centrality and ratio between the variance of the degree and the variance of the degree change 103

5.1 Setup for the hyperparameter selection 119

5.2 Impact of the dimensionality on the performance for the Facebook network. 123
5.3 Results for the impact of an increasing amount of training data on the classification performance. 124

List of Tables

2.1 Statistics for graphs generated by the random graph configurations . . . 33
2.2 Results for the model in Equation (2.2) . . . 38
2.3 Results for the model in Equation (2.3). . . . 39

3.1 Statistics of real-world networks . . . 58
3.2 Mean sd of estimation and the sd of true robustness for the experiments with real-world networks . . . 68
3.3 Mean sd of estimation and the sd of true robustness for the experiments with random graphs . . . 69

4.1 Descriptive statistics about the networks used in the study . 86
4.2 Mean and standard deviation of the robustness of centrality measures in empirical networks . . . 88
4.3 Robustness of the centrality measures in empirical networks 88
4.4 Empirical networks: rank correlation between global measures and the average robustness. . . . 89
4.5 The rank correlations between the average degree and the robustness (BA and ER graphs) . . . 93
4.6 Results for the experiments using the configuration model . 95

5.1 Summary of networks used in the case study. . . . 121
5.2 Results for the experiments in the case study. . . . 122
5.3 Results for the hyperparameter analysis of node2vec. . . . 123

Introduction

People frequently say that every human knows every other human through six edges. The paper “An Experimental Study of the Small World Problem” by Travers and Milgram (1969) is often cited to support this claim. In this experiment, random people living in the US (in Nebraska or Boston) were asked to send documents to a person, located in Boston, if they know him on a first-name basis. If they do not know him, they should send the documents to an acquaintance who knows the target or has again an acquaintance who might knows the target. The average length of completed chains was 5.2, a figure often referred to. This study can be viewed from a network perspective. Persons are nodes, edges indicate an acquaintance (two people know each other on a first-name basis). The task was to forward the documents in such a way that they would arrive at their destination as quickly as possible.

Only 64 of 296 document folders reached the target person. Thus, this study suffered from a severe case of actor non-response. Therefore, it is questionable whether the actual average chain length in this network is really 5.2. This problem was discussed extensively by Travers and Milgram (1969). The authors have provided arguments that the actual length might be above or below this figure. When this study is referenced, however, this aspect is usually ignored.[1]

Similar to this well-known and frequently cited study, most network studies struggle with inaccurate network data. This dissertation addresses this issue and studies the influence of erroneous data on network analysis methods. The methods investigated in this dissertation are so-called centrality measures which assess the position of an object, for example a person, with respect to all other objects in the network. These measures could, for

[1] This type of experiment was conducted by several researchers in other contexts to investigate the small world problem. In online social networks, for example, a shorter average distance between the actors can be observed (Leskovec and Horvitz, 2008; Boldi and Vigna, 2012).

example, be used to analyze why some letters in the Milgram experiment reached the target and some did not. One might suspect that the successful letters were sent by people who are more central in the contact network.

This chapter serves as an introduction ("accompanying paper") to this cumulative dissertation and is structured as follows. The concept of network models as well as sources and impact of errors in the context of network data are introduced in Section 1.1. Methods and concepts used in articles that are part of this dissertation and which are necessary to discuss existing studies are explained in Section 1.2. Definitions of the robustness as well as the methodology and findings of existing studies are presented in Section 1.3. The main objectives of this dissertation and the scientific contribution of the individual articles are presented in Section 1.4 and Section 1.5 concludes this introduction.

1.1 Network model, network data, and measurement errors

Networks "occur" in various domains. Prominent examples include (online) social networks, the Internet, trade networks, transportation networks, and many more (Newman, 2003). In the last two decades, a myriad of research articles, textbooks, and popular science books have been dedicated to networks (e.g., Barabási (2002); Watts (2003); Newman (2010); Christakis and Fowler (2010); Barabási (2016)). It is often claimed that networks are there or occur naturally. This perspective, however, is quite simplified. Although phenomena from different disciplines may be viewed from a network perspective, networks are not "just there". The abstraction of the phenomenon as a network must be done deliberately and explicitly. A network consists of a finite set of objects and pairwise relationships between these objects (Wasserman and Faust, 1994; Butts, 2009). For the abstraction, it is therefore essential to specify what the nodes are, what the relations (edges) between the nodes are, and how these relations are measured. Often, these relationships are modeled as a (possibly directed) binary relation. In this case, the strength of the relationships is not reflected in the model. Thus, it is crucial to recall that the mere feasibility of modeling a phenomenon using a network does not imply that the network model is a suitable representation. By choosing an inadequate representation, the conclusions may be invalid (Butts, 2009).

The research field "Network Science" is focused on the study of network models. Network models consist of two parts, as shown in Figure 1.1, the network concept and the network data. (Brandes et al., 2013). As briefly

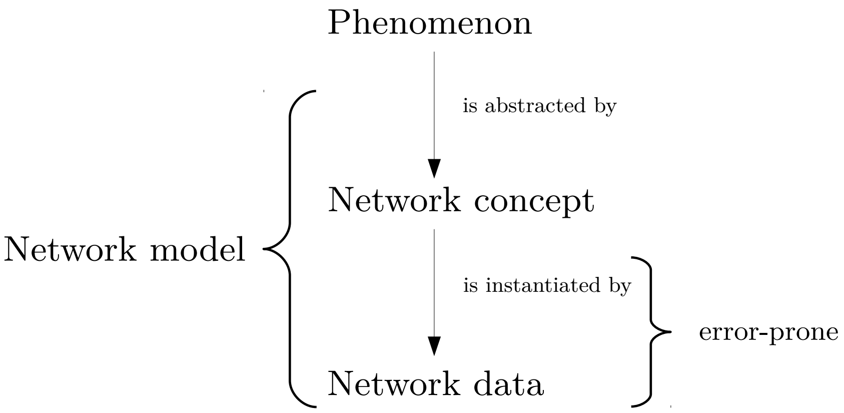


Figure 1.1: Illustration of the components of a network model (adapted from Brandes et al. (2013)).

mentioned above, the procedure for the abstraction of the phenomenon is part of the network concept. The network concept determines, on the one hand, which objects are represented by the nodes. As an example, in the case of a social network, the objects could be persons. On the other hand, the network concept also specifies which relationships are represented by the edges. In the case of the social network, these could be similarities between actors, social relations, interactions, or flows (Borgatti et al., 2009). The representation step determines how specific instances of abstraction are created. It describes how the network data are created, for example, through a survey. This step yields a specific network, which can subsequently be analyzed.

As an example of the application of the network model, assume that we are interested in how information spreads within a company. To investigate this phenomenon, employees could be modeled as nodes in a social network. As a pairwise relationship between employees, the interaction “x and y talk regularly about professional topics” could be chosen. To collect a specific instance of this social network, the following approach is possible. All employees who are present on a particular date are asked to list which other employees they talk to regularly (defined as on average once a week) about company-related topics. The result of this survey is a snapshot of the social network at that point in time.

In empirical studies, the data collection takes place during the instantiation of the network concept. Data collection is inherently error-prone, and thus it is difficult to get exactly the desired data. In the above example, these are all the statements of all employees about their communication behavior regarding all other employees of the company. A variety of errors can occur in this case. For example, it is likely that employees are not present on the day of the survey or that they cannot recall all contacts. In a network based on this data, the corresponding nodes and edges would be missing.

Errors in data collection are always a problem. With network data, errors often have even stronger effects on subsequent data analyses than they do in the case of standard data, due to the network topology. With standard data, it is usually assumed that the individual observations are independent of each other. This is not the case for network data. The existence of an edge typically depends on the existence of other (possibly incident) edges. If, for example, one edge is missing, other edges cannot be explained by this edge (Brandes et al., 2013).

In the following, we illustrate the effect that even small errors in network data can have on subsequent analysis. Figure 1.2a shows a hypothetical error-free network (adapted from (Borgatti et al., 2006)), the network that could be created from entirely error-free data. Let's assume this is the network mentioned in the example above, the social network of a company's employees. Figure 1.2b shows the network created from almost the same data with one small difference: during data collection, the edge between node 4 and 5 was not observed which has severe consequences. Possible reasons could be that actors 4 and 5 have forgotten to name each other in the survey or intentionally do not want to disclose their relationship.

The most apparent difference between those two networks is that the error-free network is connected, whereas the erroneous network consists of two components. It follows that analyses of these two networks yield very different outcomes. In the error-free network, every node can reach every other node in a finite number of steps. Information from a single node can spread throughout the entire network. In contrast, in the observed network, one would come to the conclusion that the actors that correspond to the nodes in either of the two components operate in two different realms and that information available in one group cannot reach the other group.

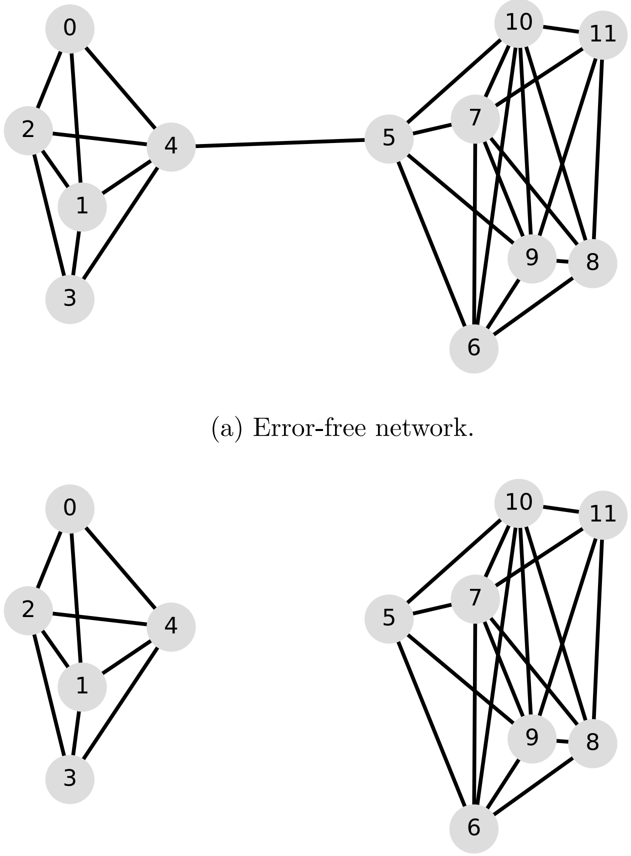
There is also a substantial difference between the two networks when analyzing individual nodes. Centrality measures, described in more detail in Section 1.2.1, map real numbers to all nodes in a network, and thus induce a ranking on the nodes. These rankings are commonly used to compare the positions of individual nodes in the network. There exist numerous centrality measures. In the following, popular measures are used as examples. The table in Figure 1.2c lists the nodes with the highest rank, induced by five centrality measures in the error-free and the observed, erroneous network.

For four of the five measures, the results differ between the error-free and the erroneous network. For betweenness, closeness, and PageRank, the most central node is a different one. In the case of eigenvector centrality, the most central node shares the first rank with two other nodes. Solely in the case of degree centrality, the most central nodes do not change. This example highlights the drastic effects that even minor errors in data collection can have.

In the following, different types and sources of errors in the context of network data are discussed. Errors in data collection always pose a problem in empirical studies. Previous work on errors and networks has focused on errors that occur during surveys but are also applicable to other types of data collection. For example, the "Boundary Specification Problem" (Laumann et al., 1983), i.e., the question for which entities exactly data should be collected, the discrepancy between perceived and actual exchanges between actors, and the problem of informant inaccuracy (Bernard et al., 1984). In addition, temporal aspects of relationships must also be considered, since, for example, the intensity of friendships changes over time (Marsden, 1990). In network studies, it is also common for actors not to respond or not to be available, and thus the relations originating from this actor cannot be recorded (Stork and Richards, 1992).

Frequently, it is not possible to collect data for all nodes and edges of a network. The network might be too large, the survey of all actors is impractical, or the population of entities is not completely known. The latter is, for example, the case with web graphs. In such cases, nodes and edges of the network are sampled. Sampled network data may also be interpreted as erroneous since they are incomplete. Depending on the sampling procedure, sampling can lead to nodes or edges to be missing randomly. However, this must not necessarily be the case (Leskovec and Faloutsos, 2006; Hu and Lau, 2013). Network measures are almost always influenced by sampling. In the case of random node sampling, it might be expected that, for example, the degree distribution of the sampled network is in the same class of distributions as the degree distribution of the entire network. However, this is only the case for classical random graphs, but not for other types of networks (Stumpf and Wiuf, 2005; Bliss et al., 2014; Advani and Malde, 2018).

In the following, the structure of some selected network studies is briefly described to illustrate which types of phenomena are modeled by networks and why data collection errors often occur in these types of studies. Francis et al. (2016) constructed a social network based on the genetic relatedness of oral commensal bacteria collected from the oral microflora of the actors. Ellis et al. (2017) constructed the social network of the southern resident



(a) Error-free network.

(b) Observed, erroneous network.

Centrality measure	Most central nodes in the error-free network	Most central nodes in the observed, erroneous network
Betweenness	5	10, 9, 7
Closeness	5	10, 9, 7
Degree	10, 9, 7	10, 9, 7
Eigenvector	10	10, 9, 7
PageRank	4	10, 9, 7

(c) Most central nodes induced by five centrality measures in the erroneous and the error-free network.

Figure 1.2: Example for the impact of erroneous network data on centrality measures. Although only one edge has not been observed, the most central nodes change.

killer whale population. Photographs of whales were analyzed, whales that were observed closely together were considered to be part of the same group. Fischer et al. (2018) proposed the use of cost-effective GPS devices to collect movement data of wildlife. In their study, device malfunctions were reported for 8.2% of the 110 sampling locations. Leecaster et al. (2016) collected the contact duration students in a middle school in Utah using wireless proximity sensors. The sensor's signal could be interfered by clothes, bodies, or objects such as tables. A signal about 20 seconds long was defined as an interaction. Wood (2017) created a drug trafficking collaboration network based on a Government's Sentencing Memorandum, a summary of testimonies, intercepted phone calls, and other evidence.

Another major area in which network data are collected is the study of protein-protein interactions (ppi), which are conducted to, for instance, improve the understanding and treatment of diseases. (De Las Rivas and Fontanillo, 2010; Schwartz et al., 2009; Guimera and Sales-Pardo, 2010). Although there are a number of procedures to obtain ppi data, the accuracy varies strongly between those procedures (von Mering et al., 2002).

This selection of studies already shows that in network studies, it is often hardly possible due to the design of the study or the nature of the data to be collected, to collect or measure the corresponding data without errors. An additional factor for studies in which the relationships are not themselves binary is that a threshold value for dichotomizing of the edges must be specified if the network model does not permit weighted edges.

1.2 Methods

This section explains basic concepts and methods that are necessary for the discussion of the research questions in the next section and that are used in the articles included in this dissertation.

A graph $G(V, E)$ consists of a node set V and an edge set E, $E \subseteq V \times V$. We denote the number of nodes in G by N and the number of edges by M. In the following, we assume that G is undirected, unweighted, and simple, i.e., it does not contain loops nor multiple edges. For many of the concepts used, however, there are also versions for directed graphs. The adjacency matrix of a graph is denoted by A, where $A_{i,j} = 1$ if there is an edge between node v_i and v_j (i.e., $(v_i, v_j) \in E(G)$) and 0 otherwise. The neighborhood of a node u is $N(u) = \{v : (u, v) \in E(G)\}$. It is the set of nodes that are connected to u (Newman, 2003).

1.2.1 Centrality measures

Numerous measures for analyzing networks exist. Based on their scale, these measures can be categorized into macro-, meso-, and microscale measures (Zanin et al., 2016). Macroscale measures consider the network as one entity, for example, the number of nodes or edges and the diameter. Mesoscale measures include concepts that relate to groups of nodes within the network, such as modularity or motifs. Microscale measures are concepts that focus on individual nodes, though, they are often aggregated like, for example, the number of connections a node has.

Centrality measures map a real number to every node in a network, implying a ranking on the nodes, and are invariant under isomorphisms. They solely depend on the network structure and not on, for example, additional information about the nodes (Koschützki et al., 2005). Since the object of centrality measures are the individual nodes, centrality measures belong to the category of microscale measures. Centrality measures are used in a variety of fields. To illustrate, a few examples are outlined hereafter.

Page et al. (1999) developed one of the most famous centrality measures, the PageRank, to rank websites and order search engine results. It is also used by Google Search. Kiss and Bichler (2008) applied centrality measures to customer networks to identify influencers for marketing campaigns. Banerjee et al. (2013) proposed a centrality measure to identify people who are important for the spread of information through social networks and applied it to study the diffusion of microfinance in small villages in India. In centrality-based targeted vaccination, centrality measures are used to identify the most efficient targets for vaccination strategies and thus manipulate the network structure (Wang et al., 2016). In the study of biological systems, e.g., gene regulatory networks, centrality measures are also used to improve the understanding of those systems (Koschützki and Schreiber, 2008).

In the following, the five centrality measures used in the articles included in this dissertation are described in more detail: betweenness, closeness, degree, eigenvector centrality, and the PageRank. By $c_G(u)$, we denote the centrality value for a specific node u in a graph G w.r.t. a centrality measure c. If the context permits, we do not explicitly mention the graph and u, v, w are in V.

The degree centrality is a neighborhood-based centrality, it is defined as the number of neighbors of a node:

$$\text{degree}(u) = |N(u)|. \tag{1.1}$$

The eigenvector centrality and the PageRank are both feedback measures (Koschützki et al., 2005). They are defined recursively, the centrality value of a node depends on the centrality values of its neighbors. If G is connected, then the eigenvector centrality of a node u defined by the unique solution to

$$\mathrm{evc}(u) = \frac{1}{\lambda} \sum_{v \in N(u)} \mathrm{evc}(v), \tag{1.2}$$

where λ is the largest eigenvalue of A (Bonacich, 1987). This measure was first described by Landau (1895). The existence of a unique solution follows from the Perron-Frobenius theorem (Newman, 2003).

The PageRank is defined as the unique solution to

$$\mathrm{PageRank}(u) = d \sum_{v \in N(u)} \frac{\mathrm{PageRank}(v)}{\mathrm{degree}(v)} + (1 - d), \tag{1.3}$$

with d as damping factor (usually 0.85) (Brin and Page, 1998). Originally introduced for directed graphs, this concept is also applicable for undirected graphs. One of the main differences between these two measures is that, in the case of the eigenvector centrality, all neighbors of a node receive the total centrality value of this node. In contrast, in the case of the PageRank, neighbors of a node only receive a fraction of the node's centrality value, which depends on the total number of neighbors of this node.

For the calculation of the eigenvector for these two centrality measures, eigenvector centrality and PageRank, the power iteration method is commonly used. A vector of length N is initialized with random values and multiplied with the adjacency matrix. In every successive iteration, the result is again multiplied with the adjacency matrix until the result converges (Langville and Meyer, 2005).

Betweenness and closeness are both path-based centralities which depend on the shortest paths between pairs of nodes in a network. Let $\sigma_{v,w}(u)$ denote the number of shortest paths between nodes v and w which contain u, $\sigma_{v,w}$ the total number of shortest paths between v and w, and $\mathrm{dist}(u, v)$ the distance, the length of the shortest path, between node u and v. A component of a graph is a maximal subgraph in which the nodes can reach each other, i.e., the distance between those nodes is finite. A graph is connected if it consists of one component, i.e., the pairwise distance between all nodes in the graph is finite (Diestel, 2017).

The betweenness centrality measures on how many shortest paths a node occurs. The idea behind it is that entities that appear more frequently on these paths are more central since, for example, more information passes through them as it spreads through the network (Freeman, 1977):

$$\text{betweenness}(u) = \sum_{v \neq u \neq w} \frac{\sigma_{v,w}(u)}{\sigma_{v,w}}, \tag{1.4}$$

The closeness centrality measures how "close" one node is to all other nodes in the network: the closer, the more central. The closeness is quantified using the distance (Bavelas, 1950; Freeman, 1978):

$$\text{closeness}(u) = \frac{1}{\sum_v \text{dist}(u, v)} \tag{1.5}$$

Note that this definition assumes that the graph is connected since the distance between nodes in different components is not defined.

For both, the calculation of the betweenness and the closeness for all nodes in a graph, the all-pairs shortest path problem has to be solved. Using the Floyd-Warshall Algorithm (Floyd, 1962), this can be done in $\mathcal{O}(N^3)$ if there are no negative cycles in the graph. For an unweighted graph, the runtime can be reduced to $\mathcal{O}(MN)$ (Brandes, 2001). Despite this reduction in complexity, the exact calculation of these centrality measures often takes too much time for larger networks leading to the calculation being impracticable.

1.2.2 Random graph models

Random graph models have been used extensively since the 1950s. Among mathematicians, the term random graph is often used synonymously for the Erdős-Rényi (ER) random graph model discussed below (Solomonoff and Rapoport, 1951; Erdős and Rényi, 1959; Gilbert, 1959; Bollobás, 2001; Newman, 2003). For this model, various properties and effects have been shown as, for example, the emergence of a giant component. In addition, it has laid the foundation for further studies on percolation, for example, for the resilience of networks and diffusion processes (Callaway et al., 2000; Moore and Newman, 2000; Newman, 2002).

Many properties that occur in real-world network do, however, not occur in ER graphs. In particular, these graphs have a completely different degree distribution than real-world networks and there are no communities. This model has, however, initiated the development of other, more sophisticated random graph models and the characterization of growth processes that yield networks that capture characteristics of real-world networks, especially effects like a scale-free degree distribution or the small-world property.

For example, ER graphs have a low average distance, a feature observed in many real-world networks, but nodes in ER graphs show low local clustering. To address this issue, Watts and Strogatz (1998) have proposed a random graph model which incorporates parts of ER graphs. They start with a regular lattice of nodes and randomly rewire some of the edges. This model yields graphs which show both, low average distance and high local clustering. These types of networks are called small-world networks.

Exponential random graphs models are another approach for identifying characteristics that explain the formation of edges and thus explain the observed network. Those characteristics can be structural or covariates related to nodes or edges. These models yield a probability distribution over all graphs (Holland and Leinhardt, 1981; Strauss, 1986; Robins et al., 2007).

Studies about the robustness of centrality measures commonly utilize random graph models for multiple reasons. The number of real-world networks available is limited and their properties cannot be altered, hence, it is difficult to generalize results found for these individual networks. Random graph models allow the repeated generation of networks and their properties, e.g. network size, density, or clustering, to be controlled. Thus, researchers can study the relationships between those properties and the robustness of centrality measures (Frantz et al., 2009). In addition to simulation-based studies, some random graph models can be accessed analytically (Platig et al., 2013). A brief description of models that are frequently used for the study of the robustness of centrality measures — also in this dissertation — is given below.

The Erdős-Rényi random graph model introduced by Solomonoff and Rapoport (1951), Erdős and Rényi (1959), and Gilbert (1959) has two parameters: the number of nodes n and the edge probability p. Since all node pairs are connected with the same probability (p), the degree distribution of the nodes in this model follows a binomial distribution. Note that the existence of an edge is entirely independent of the existence of other edges.

In contrast, the Barabási-Albert model is based on the idea of preferential attachment (Simon, 1955; Price, 1976; Barabási and Albert, 1999). Consequently, the probability that a new node will connect to an existing node is proportional to the degree of the existing node. This model also has two parameters. In addition to the number of nodes n, the parameter m specifies the number of connections that a new node makes to existing nodes. Due to this generation process, the degree distribution of the nodes in graphs generated by this model follows a power-law distribution.

The configuration model is a method to create random graphs based on existing degree sequences (Newman et al., 2001). In this model, there are no other parameters apart from the degree sequence. First, an empty graph on n nodes is created (n is given by the degree sequence). Next, every node u receives degree(u) stubs (here, degree(u) is the desired degree of node u). Finally, pairs of stubs are chosen and connected with equal probability. This procedure might results in graphs with multiple edges and loops.

1.3 Inaccurate network data and centrality measures

As illustrated the example in Section 1.1, centrality measures can be strongly influenced by erroneous network data. This dissertation contributes to the understanding of the robustness of centrality measures.

Although the two topics sound similar, studies on the robustness of networks have a different focus than studies about the robustness of centrality measures. The subject of studies on the robustness of networks is the question how the functionality of a network as a whole is influenced by, for example, the removal of nodes (see Albert et al. (2000); Callaway et al. (2000); Holme et al. (2002); Klau and Weiskircher (2005); Cohen et al. (2000)). If the term robustness is used in this work without further specification, then the term always refers to the robustness of centrality measures.

Existing studies on the robustness of centrality measures are analyzed in order to be able to outline the main focus of this dissertation. Primarily, the emphasis is on how the robustness is quantified. Findings of existing studies are summarized, and gaps in research are identified. To discuss the robustness, we first have to introduce two terms. The error-free network is the network which we are actually interested in, it is the network constructed from correct, error-free network data and is typically not available. The observed network is the network that is constructed from the data actually collected.

There exist different approaches to measure the robustness, described below are the commonly used ones. There are approaches in which only a portion of the nodes are considered and approaches in which all nodes are considered. The first type includes top-k and overlap metrics. In the case of top-k measurements, the frequency with which the most central node or nodes in the error-free network occur among the k most central nodes in the observed network is considered (Borgatti et al., 2006; Frantz et al., 2009; Tsugawa and Ohsaki, 2015; Erman and Todorovski, 2015; Frantz and Carley, 2017). For overlap measurements, it is calculated how many of the most central nodes in the error-free network also occur among the most central nodes in the observed network. The fraction of nodes to be considered (e.g., 10% of the most central nodes) is a parameter that has to be specified (Borgatti et al., 2006; Frantz et al., 2009; Tsugawa and Ohsaki, 2015; Ufimtsev et al., 2016).

The most common way to quantify the robustness is to consider the centrality values of all nodes using a correlation measure. The correlation between the values for the nodes in the error-free and observed networks is calculated, considering only the values for nodes contained in both networks. Depending on the study, the Pearson correlation (Bolland, 1988; Costenbader and Valente, 2003; Borgatti et al., 2006; Smith and Moody, 2013; Silk et al., 2015; Smith et al., 2017) or a rank correlation (Kim and Jeong, 2007; Wang et al., 2012; Erman and Todorovski, 2015; Niu et al., 2015; Schulz, 2016; Holzmann et al., 2019) is used. In contrast to correlations measures, an additional parameter has to be chosen in case of the top-k and overlap metrics, which influences the results (Ufimtsev et al., 2016).

Most studies on the robustness of centrality measures are simulation-based. This approach has changed little since Bolland (1988) and consists, roughly spoken, of the following steps: A network is considered to be error-free and different types of errors are applied to that network. This process yields a simulated observed network. Centrality measures are calculated for both networks and the robustness is calculated using one of the approaches outlined above. The underlying network is either based on empirical network data or is generated using a random graph model. The errors are often simulated in such a way that the error affects all nodes or edges with equal probability.

The robustness of a centrality measures depends on many influencing factors, in particular, the type and extent of the error, the centrality measure under consideration, the network topology, and the way the robustness is measured. Deriving universal statements about the robustness of centrality measures is, therefore, problematic — apart from the observations that the measures are less robust, the larger the error is, and that the degree centrality is usually more robust than other measures. Often, the results of different studies are inconclusive. For example, the question of whether centrality measures are more robust in larger networks. While Borgatti et al. (2006) and Niu et al. (2015) do not observe an association between network size and robustness, the observations by Costenbader and Valente (2003); Wang et al. (2012); Smith et al. (2017) are ambiguous in that regard. In contrast, Silk et al. (2015) and Lee and Pfeffer (2015) find that centrality measures in larger network might be more robust.

Results in simulation studies depend strongly on the specific characteristics of the data. In addition, empirical studies are often limited in terms of the number of networks studied, as the number of available networks is limited.

There are studies that pursue an analytical approach to the robustness (Ghoshal and Barabási, 2011; Platig et al., 2013; Tsugawa and Ohsaki, 2015; Holzmann et al., 2019). However, this also poses major challenges and is often only made possible by many assumptions. In most cases, the results are therefore limited to certain classes of networks, types of errors, and centrality measures.

The concept of robustness presents a separate challenge for practical application. The calculation of robustness requires that the error-free network is known. However, this is not the case — if this were the case, the robustness would not have to be calculated. There are studies proposing methods to compensate for the effects of measurement errors. For example, Butts (2003) developed Bayesian models that incorporate false positive edges and false negative edges yielding a probability distribution over the edges. Huisman (2009) studied imputation procedures and their practicality in the context of network-level measures. Kim and Leskovec (2011) introduced an expectation-maximization algorithm to recover missing nodes and edges. However, these studies did not investigate whether these methods also improve the results for centrality measures or whether they can be used to estimate their robustness.

1.4 Contribution

The overview of the previous section indicates a need for further research on the topic of robustness of centrality measures. Hence, the main focus of this dissertation is to understand

(1) the impact of (primarily) non-random errors on the robustness of centrality measures,

(2) how network properties are related to the robustness of centrality measures, and

(3) how the robustness can be calculated without knowledge of the error-free network.

These issues are addressed by the four interrelated articles that are part of this dissertation.

The impact of partially missing communities on the reliability of centrality measures

In Martin (2018), Chapter 2, we address (1). The paper investigates the reliability of centrality measures when missing nodes are likely to belong to the same community. We study the behavior of five commonly used centrality measures in uniform and scale-free networks in various error scenarios. We find that centrality measures are generally more reliable when missing nodes are likely to belong to the same community than in cases in which nodes are missing uniformly at random. In scale-free networks, the betweenness centrality becomes, however, less reliable when missing nodes are more likely to belong to the same community. Moreover, centrality measures in scale-free networks are more reliable in networks with stronger community structure. In contrast, we do not observe this effect for uniform networks. Our observations suggest that the impact of missing nodes on the reliability of centrality measures might not be as severe as the literature suggests.

The role of network size for the robustness of centrality measures

In Martin and Niemeyer (in press), Chapter 4, we address (2). Previous studies have observed that the robustness mainly depends on the network structure, the centrality measure, and the type of error. Previous findings regarding the influence of network size on robustness are, however, inconclusive.

Based on twenty-four empirical networks, we investigate the relationship between global network measures, especially network size and average degree, and the robustness of the degree, eigenvector centrality, and PageRank. We demonstrate that, in the vast majority of cases, networks with a higher average degree are more robust.

For random graphs, we observe that the robustness of Erdős-Rényi networks decreases with an increasing average degree, whereas with Barabási-Albert networks, the opposite effect occurs: with an increasing average degree, the robustness also increases.

As a first step into an analytical discussion, we prove that for Erdős-Rényi networks of different size but with the same average degree, the robustness of the degree centrality remains stable.

Influence of measurement errors on networks: Estimating the robustness of centrality measures

In Martin and Niemeyer (2019), Chapter 3, we address (3). Previous studies have dealt either with the general effects of measurement errors on centrality measures or with the treatment of erroneous network data. As discussed above, the robustness concept relies on knowing both the error-free and the observed network, and thus it is not possible to calculate the robustness when only the observed network is available — which is usually the case. In this paper, we propose a method for estimating the impact of measurement errors on the reliability of a centrality measure, given the measured network and assumptions about the type and intensity of the measurement error. This method allows researchers to estimate the robustness of a centrality measure in a specific network and can, therefore, be used as a basis for decision making.

In our experiments, we apply this method to random graphs and real-world networks. We observe that our estimation is, in the vast majority of cases, a good approximation for the robustness of centrality measures. Beyond this, we propose a heuristic to decide whether the estimation procedure should be used. We analyze, for specific networks, why the eigenvector centrality is less robust than, for example, the PageRank. Finally, we give recommendations on how our findings can be applied to future network studies.

A process for the evaluation of node embedding methods in the context of node classification

In Martin and Riebeling (2020), Chapter 5, we also address (3). In Martin and Niemeyer (2019), we argue that imputation techniques should be improved, especially with regard to using centrality measures on the imputed network. Since there have been considerable advances in the field of machine learning in recent years, it is logical that these methods are used for imputation methods in the future. In order to be able to use network data with common machine learning methods, an embedding for the network data has to be found first. In order to evaluate which embedding method is suitable, we develop a framework to compare node embedding methods with each other. This paper is, therefore, the preliminary work for the development of imputation procedures on the basis of machine learning techniques.

Node embedding methods find latent lower-dimensional representations which are used as features in machine learning models. In the last few years, these methods have become extremely popular as a replacement for manual feature engineering.

Since authors use various approaches for the evaluation of node embedding methods, existing studies can rarely be efficiently and accurately compared. We address this issue by developing a process for a fair and objective evaluation of node embedding procedures w.r.t. node classification.

This process supports researchers and practitioners to compare new and existing methods in a reproducible way.

We apply this process to four popular node embedding methods and make valuable observations. With an appropriate combination of hyperparameters, good performance can be achieved even with embeddings of lower dimensions, which is positive for the run times of the downstream machine learning task and the embedding algorithm. Multiple hyperparameter combinations yield similar performance. Thus, no extensive, time-consuming search is required to achieve reasonable performance in most cases.

1.5 Conclusions

The articles that are part of this dissertation contribute to a better understanding of the robustness of centrality measures. The role of the individual articles was highlighted in the previous section. The two journal-length articles constitute the main contribution. Martin and Niemeyer (2019), Chapter 3, introduces a new method for the estimation of the robustness and has been published in Network Science. Martin and Niemeyer (in press), Chapter 4, provides new results on the relationship between robustness and the average degree and is currently under review in Network Science. Preliminary results of this study have been published at a conference (Martin and Niemeyer, 2020).

Based on the findings of this dissertation, there are several possibilities for further research. For example, it would be interesting to investigate how non-random errors occur in network data and how they can be explained, for example, by covariates of nodes or edges. In the same setting, more detailed results on the effects of non-random errors and the joint occurrence of several types of errors would be valuable. Furthermore, more sophisticated random graph models could be used to further investigate the relationship between network properties and the robustness in simulation studies and analytically.

1.6 References

A. Advani and B. Malde. Credibly Identifying Social Effects: Accounting for Network Formation and Measurement Error. *Journal of Economic Surveys*, 32(4):1016–1044, 2018.

R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(July):378–382, 2000.

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. The diffusion of microfinance. *Science*, 341(6144), 2013.

A. Barabási. *Linked: The New Science of Networks*. Perseus Pub., 2002.

A.-L. Barabási. *Network Science*. Cambridge University Press, 2016.

A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(October):509–512, 1999.

A. Bavelas. Communication patterns in Task-Oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

H. R. Bernard, P. Killworth, D. Kronenfeld, and L. Sailer. The Problem of Informant Accuracy - The Validity of Retrospective Data. *Annual Review of Anthropology*, 13:495–517, 1984.

C. A. Bliss, C. M. Danforth, and P. S. Dodds. Estimation of global network statistics from incomplete data. *PLoS ONE*, 9(10):1–18, 2014.

P. Boldi and S. Vigna. Four degrees of separation, really. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pages 1222–1227, 2012.

J. M. Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253, 1988.

B. Bollobás. *Random Graphs*. 2001.

P. Bonacich. Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136, 2006.

S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

U. Brandes. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, 25(2):163–177, 2001.

U. Brandes, G. Robins, A. McCranie, and S. Wasserman. What is network science? *Network Science*, 1(01):1–15, 2013.

S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

C. T. Butts. Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks*, 25(2):103–140, 2003.

C. T. Butts. Revisiting the foundations of network analysis. *Science*, 325 (5939):414–416, 2009.

D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

N. Christakis and J. Fowler. *Connected: The Amazing Power of Social Networks and How They Shape Our Lives.* HarperCollins Publishers, 2010.

R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85(21):4626–4628, Nov. 2000.

E. Costenbader and T. W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, 2003.

J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, June 2010.

R. Diestel. *Graph Theory.* Springer Graduate Texts in Mathematics. Springer-Verlag, Reinhard Diestel, 5 edition, 2017.

S. Ellis, D. W. Franks, S. Nattrass, M. A. Cant, M. N. Weiss, D. Giles, K. C. Balcomb, and D. P. Croft. Mortality risk and social network position in resident killer whales: Sex differences and the importance of resource abundance. *Proceedings of the Royal Society B: Biological Sciences*, 284 (1865):20171313, Oct. 2017.

P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6: 290–297, 1959.

N. Erman and L. Todorovski. The effects of measurement error in case of scientific network analysis. *Scientometrics*, 104(2):453–473, 2015.

M. Fischer, K. Parkins, K. Maizels, D. R. Sutherland, B. M. Allan, G. Coulson, and J. Di Stefano. Biotelemetry marches on: A cost-effective GPS device for monitoring terrestrial wildlife. *PLOS ONE*, 13(7):e0199617, July 2018.

R. W. Floyd. Algorithm 97: Shortest path. *Commun. ACM*, 5(6):345, June 1962.

S. S. Francis, M. M. Plucinski, A. D. Wallace, and L. W. Riley. Genotyping Oral Commensal Bacteria to Predict Social Contact and Structure. *Plos One*, 11(9):e0160201, 2016.

T. L. Frantz and K. M. Carley. Reporting a network's most-central actor with a confidence level. *Computational and Mathematical Organization Theory*, 23(2):301–312, June 2017.

T. L. Frantz, M. Cataldo, and K. M. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, 2009.

L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35, 1977.

L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

G. Ghoshal and A.-L. Barabási. Ranking stability and super-stable nodes in complex networks. *Nature communications*, 2:394, 2011.

E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30 (4):1141–1144, 1959.

R. Guimera and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, Apr. 2010.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.

P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han. Attack Vulnerability of Complex Networks. *Phys. Rev. E*, 65(5 Pt 2):56109, 2002.

H. Holzmann, A. Anand, and M. Khosla. Delusive PageRank in Incomplete Graphs. In L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, and L. M. Rocha, editors, *Complex Networks and Their Applications VII*, pages 104–117, Cham, 2019. Springer International Publishing.

P. Hu and W. Lau. A survey and taxonomy of graph sampling. *arXiv.org*, pages 1–34, 2013.

M. Huisman. Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, 10(1):1–29, 2009.

M. Kim and J. Leskovec. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SIAM International Conference on Data Mining*, pages 47–58, 2011.

P. J. Kim and H. Jeong. Reliability of rank order in sampled networks. *European Physical Journal B*, 55(1):109–114, 2007.

C. Kiss and M. Bichler. Identification of influencers — Measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008.

G. W. Klau and R. Weiskircher. Robustness and Resilience. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations*, pages 417–437. Springer Berlin Heidelberg, 2005.

D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2:193–201, 2008.

D. Koschützki, K. Lehmann, and L. Peeters. Centrality Indices. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations*, pages 16–61. Springer Berlin Heidelberg, 2005.

E. Landau. Zur relativen Wertbemessung der Turnierresultate. *Deutsches Wochenschach*, 11:192–202, 1895.

A. N. Langville and C. D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM review*, 47(1):135–161, 2005.

E. O. Laumann, P. V. Marsden, and D. Prensky. The Boundary Specification Problem in Network Analysis. In R. Burt and M. Minor, editors, *Applied Network Analysis*, pages 18–34. Sage Publications, 1983.

J.-S. Lee and J. Pfeffer. Robustness of Network Centrality Metrics in the Context of Digital Communication Data. *Proceedings of the 48th Hawaii International Conference on System Sciences*, 2015.

M. Leecaster, D. J. A. Toth, W. B. P. Pettey, J. J. Rainey, H. Gao, A. Uzicanin, and M. Samore. Estimates of social contact in a middle school based on self-report and wireless sensor data. *PLOS ONE*, 11(4), 2016.

J. Leskovec and C. Faloutsos. Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006.

J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM.

P. V. Marsden. Network data and measurement. *Annual Review of Sociology*, 16(1):435–463, 1990.

C. Martin. The Impact of Partially Missing Communities on the Reliability of Centrality Measures. In C. Cherifi, H. Cherifi, M. Karsai, and M. Musolesi, editors, *Complex Networks & Their Applications VI*, pages 41–52. Springer International Publishing, 2018.

C. Martin and P. Niemeyer. Influence of measurement errors on networks: Estimating the robustness of centrality measures. *Network Science*, 7(2): 180–195, 2019.

C. Martin and P. Niemeyer. The role of network size for the robustness of centrality measures. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, editors, *Complex Networks and Their Applications VIII*, pages 40–51, Cham, 2020. Springer International Publishing.

C. Martin and P. Niemeyer. On the impact of network size and average degree on the robustness of centrality measures. *Network Science*, in press.

C. Martin and M. Riebeling. A Process for the Evaluation of Node Embedding Methods in the Context of Node Classification. *ArXiv E-Prints*, arXiv:2005.14683, 2020.

C. Moore and M. E. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000.

M. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64, 2001.

M. E. J. Newman. *Networks: An Introduction.* OUP Oxford, 2010.

Q. Niu, A. Zeng, Y. Fan, and Z. Di. Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131, 2015.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab / Stanford InfoLab, Nov. 1999.

J. Platig, E. Ott, and M. Girvan. Robustness of network measures to link errors. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6), 2013.

D. d. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306, 1976.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

J. Schulz. Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, 107(3):1283–1298, 2016.

A. S. Schwartz, J. Yu, K. R. Gardenour, R. L. Finley, Jr, and T. Ideker. Cost-effective strategies for completing the interactome. *Nature methods*, 6(1):55–61, Jan. 2009.

M. J. Silk, A. L. Jackson, D. P. Croft, K. Colhoun, and S. Bearhop. The consequences of unidentifiable individuals for the analysis of an animal social network. *Animal Behaviour*, 104:1–11, 2015.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4): 425–440, 1955.

J. A. Smith and J. Moody. Structural Effects of Network Sampling Coverage I: Nodes Missing at Random. *Social Networks*, 35(4), 2013.

J. A. Smith, J. Moody, and J. H. Morgan. Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48:78–99, 2017.

R. Solomonoff and A. Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, June 1951.

D. Stork and W. D. Richards. Nonrespondents in Communication Network Studies: Problems and Possibilities. *Group & Organization Management*, 17(2):193–209, 1992.

D. Strauss. On a general class of models for interaction. *SIAM review*, 28 (4):513–527, 1986.

M. P. Stumpf and C. Wiuf. Sampling properties of random graphs: The degree distribution. *Physical Review E*, 72(3):036118, 2005.

J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969.

S. Tsugawa and H. Ohsaki. Analysis of the Robustness of Degree Centrality against Random Errors in Graphs. In *Studies in Computational Intelligence*, volume 597, pages 25–36. 2015.

V. Ufimtsev, S. Sarkar, A. Mukherjee, and S. Bhowmick. Understanding Stability of Noisy Networks through Centrality Measures and Local Connections. *CoRR*, 2016.

C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, May 2002.

D. J. Wang, X. Shi, D. A. McFarland, and J. Leskovec. Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409, 2012.

Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao. Statistical physics of vaccination. *Physics Reports*, 664:1–113, 2016.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

D. Watts. *Six Degrees: The Science of a Connected Age.* Science (W.W. Norton). Norton, 2003.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.

G. Wood. The structure and vulnerability of a drug trafficking collaboration network. *Social Networks*, 48:1–9, 2017.

M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, and S. Boccaletti. Combining complex networks and data mining: Why and how. *Physics Reports*, 635:1–44, 2016.

The impact of partially missing communities on the reliability of centrality measures

Christoph Martin

This paper has been published as

Martin, C. (2018). The Impact of Partially Missing Communities on the Reliability of Centrality Measures. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), Complex Networks & Their Applications VI (pp. 41–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-72150-7_4.

The layout has been revised.

Abstract

Network data is usually not error-free, and the absence of some nodes is a very common type of measurement error. Studies have shown that the reliability of centrality measures is severely affected by missing nodes. This paper investigates the reliability of centrality measures when missing nodes are likely to belong to the same community. We study the behavior of five commonly used centrality measures in uniform and scale-free networks in various error scenarios. We find that centrality measures are generally more reliable when missing nodes are likely to belong to the same community than in cases in which nodes are missing uniformly at random. In scale-free networks, the betweenness centrality becomes, however, less reliable when missing nodes are more likely to belong to the same community. Moreover, centrality measures in scale-free networks are more reliable in networks with stronger community structure. In contrast, we do not observe this effect for uniform networks. Our observations suggest that the impact of missing nodes on the reliability of centrality measures might not be as severe as the literature suggests.

2.1 Introduction

Centrality measures are commonly used in network analysis. Network data is, however, rarely error-free. Some parts of a network are often not recorded correctly. For example, nodes could be missing due to the non-response effect or the boundary specification problem (Kossinets, 2006). Some users in social networking services may have restrictive privacy settings and thus their profiles are not accessible or the number of API requests might be limited (Rezvanian and Meybodi, 2015). In addition, this type of error does commonly occur when bio-logging is used to collect interaction data (Silk, 2017).

Studies have found that the reliability of centrality measures is often severely compromised by missing nodes (Smith et al., 2017; Smith and Moody, 2013; Boldi et al., 2013; Wang et al., 2012; Frantz et al., 2009; Kim and Jeong, 2007; Borgatti et al., 2006; Kossinets, 2006; Costenbader and Valente, 2003; Bolland, 1988). These studies agree that higher levels of error lead to lower reliability. The exact extent of the impact of missing nodes on the reliability does, however, depend on a variety of factors. Analyzing Erdős-Rényi networks, Borgatti et al. (2006) observed that centrality measures behaved similarly. Considering different types of empirical networks and random graphs, studies found that the reliability of centrality measures strongly depends on the type of network. For example, Smith et al. (2017) found that centrality measures are more reliable in larger, more centralized networks. Closeness centrality was been reported to be more reliable than betweenness and degree centrality (Kim and Jeong, 2007). Moreover, Boldi et al. (2013) observed that social networks are more robust to missing nodes than web graphs. Despite their important findings, previous studies have mostly focused on the case where nodes are missing uniformly at random. A notable exception is Smith et al. (2017). In this study, the authors investigated the effect of missing nodes in cases in which the probability that a node is missing depends on their centrality. They found that the reliability is worse when more central nodes are missing.

Since nodes in a network are interconnected, it seems obvious that the behavior of nodes in a community will, at least to some extent, determine whether other nodes from that community can be observed or not (Smith et al., 2017). For example, some groups within a network may have concerns about the collection of their data and therefore collectively refuse to participate in a survey or adopt strict policies regarding the use of their data in social networking services. In animal research, subgroups of a population may be able to avoid being trapped and tagged and are therefore missing in the resulting network. Despite the multitude of possible scenarios in which this type of measurement error may occur, it has not been considered in previous research.

In this study, we investigate how reliable centrality measures are when missing nodes are likely to belong to the same community, i.e., the probability that a node is missing depends on which other nodes are missing.[1] In particular, we examine whether this type of measurement error has a stronger or smaller impact on the reliability than the purely random absence of nodes. We use two random graph models to answer this question for uniform and scale-free networks. These models enable us to analyze the influence of the community structure on the reliability.

Our results suggest that centrality measures are more reliable when missing nodes are likely to belong to the same community than in cases in which nodes are missing uniformly at random. In scale-free networks, however, the betweenness centrality becomes less reliable when missing nodes are more likely to belong to the same community. Moreover, in scale-free networks, centrality measures are more reliable in networks with stronger community structure. In contrast, we do not observe this effect for uniform

[1] Despite the similar wording, this work does not address the reliability or robustness of networks. For an extensive overview about the robustness of networks see Havlin and Cohen (2010) and Barabási (2016).

networks. In addition to presenting these findings, we introduce a novel approach which we will refer to as "community bias". This approach allows us to simulate different levels of measurement error and enables us to study their impact on the reliability of centrality measures.

2.2 Methods & experimental setup

We denote an undirected, unweighted graph by G and the vertex set of a graph G by $V(G)$. In all graphs that we consider in this study, every node belongs to some community. We denote the nodes that belong to community j in graph G by $V_j(G)$. We use the terms graph and network interchangeably. A centrality measure c is a real-valued function that assigns centrality values to all nodes in a graph and is invariant to structure-preserving mappings, i.e., centrality values depend solely on the structure of a graph. External information (e.g., node or edge attributes) have no influence on the centrality values (Koschützki et al., 2005). Similarly to Martin and Niemeyer (2017), we denote the centrality value for node $u \in V(G)$ by $c_G(u)$ and the centrality values for all nodes in G $(u_1, u_2, \ldots, u_n)$ by the vector $c(G) := (c(u_1), \ldots, c(u_n))$.

The following centrality measures are used in this study: closeness centrality (Freeman, 1978), betweenness centrality (Freeman, 1977), degree centrality, eigenvector centrality (Bonacich, 1987), and the PageRank (Brin and Page, 1998).

There are multiple definitions of communities in networks and, depending on the context, some are more appropriate than others. In this study, a community is a subgraph where each of its vertices is more strongly attached to vertices in that subgraph than to vertices in any other subgraph (Hu et al., 2008; Fortunato and Hric, 2016). Hence, the fraction of edges that a node has to other nodes which are not part of its community (compared to the total number of edges that are connected to this node) is an indicator of the strength of the community structure ("community strength") in a network. The lower this ratio, the stronger the community structure. We can quantify the strength of the community structure by calculating the modularity of a graph with respect to a mapping which maps the nodes to communities (Newman and Girvan, 2004).

Some community definitions allow communities to overlap. We focus on non-overlapping communities. For a more detailed discussion of the definition of communities in networks see Wasserman and Faust (1994), Boccaletti et al. (2006), and Fortunato (2010).

2.2.1 Data

To investigate the effect of community structure on the reliability of centrality measures, we use two random graph models. There are two main reasons to use synthetic graphs. When using random graphs models, we know the ground-truth mapping from nodes to communities, i.e., we know which nodes belong to the same community. In contrast, for real-world networks we might not know if there are communities at all (Fortunato and Barthélemy, 2007). Moreover, the random graph models enable us to vary the strength of the community structure (as described above) and thus gives us the opportunity to study the effect of the community strength on the reliability of centrality measures explicitly.

In the clustered random graph (CRG) model, n nodes are partitioned into k sets. Nodes in the same set belong to the same community. With probability p_{intra}, edges are created between nodes in the same community. Edges between nodes that are not in the same community are created with probability p_{inter}. This model was originally introduced by (Girvan and Newman, 2002) to benchmark community detection algorithms. Since the number of edges from a node to other nodes in the same community and the number of edges from a node to nodes in other communities both follow a binomial distribution (with different parameters though), this model is conceptually close to Erdős-Rényi graphs (Erdős and Rényi, 1959; Garbers et al., 1990).

We use two configurations of the CRG model, one with weaker community structure (CRG_{weak}) and one with stronger community structure ($\text{CRG}_{\text{strong}}$). In both configurations, we set $n = 1000$ and $k = 25$. For the CRG_{weak} configuration, we set p_{intra} to 0.1 and p_{inter} to 0.01. For the $\text{CRG}_{\text{strong}}$ configuration, we set p_{intra} to 0.2 and p_{inter} to 0.005.

The second model is the Lancichinetti-Fortunato-Radicchi (LFR) model as described in Lancichinetti et al. (2008) and Staudt et al. (2017). According to this model, the distribution of the node degrees and distribution of the community sizes both follow a power-law distribution. The degree distribution and the community size distribution in empirical networks can often be described by a power-law distribution (Leskovec et al., 2008; Clauset et al., 2009). Hence, graphs generated by this models share various characteristics with real-world networks.

For the degree distribution, we use an average degree of 10, a maximum degree of 50, and an exponent of -2. For the community size distribution, we use minimum community size of 5, maximum community size of 100, and an exponent of -2. The mixing parameter μ determines the fraction of neighbors of each node that do not belong to the node's own community.

Table 2.1: Statistics for graphs generated by the random graph configurations that are used in this paper. Numbers are mean values based on 100 realizations. Standard deviations are listed in parentheses. “Clustering” denotes the average clustering coefficient and “Communities” denotes the number of communities.

	$\text{CRG}_{\text{strong}}$	CRG_{weak}	$\text{LFR}_{\text{strong}}$	LFR_{weak}
Nodes	1000	1000	1000	1000
Edges	6380 (86)	6798 (78)	5028 (129)	5022 (126)
Diameter	5.0 (0.1)	4.9 (0.3)	6.0 (0.0)	5.0 (0.2)
Communities	25 (0.0)	25 (0.0)	69 (7.9)	68 (8.6)
Clustering	0.08 (0.003)	0.02 (0.001)	0.18 (0.017)	0.03 (0.002)
Modularity	0.581 (0.006)	0.253 (0.006)	0.534 (0.007)	0.149 (0.007)

Again, we use two configurations of the LFR model. One with $\mu = 0.8$ and thus a weaker community structure (LFR_{weak}) and one with $\mu = 0.4$ and thus a stronger community structure ($\text{LFR}_{\text{strong}}$). In addition, we use a third variation of this model for the second part of our experiments. Here we use the same parameters as described above, but vary the mixing parameter μ from 0.15 to 0.95 in steps of 0.05. We denote these 17 configurations by $\text{LFR}_{\text{varying}}(\mu)$.

In both of these random graph models, the centrality measures are usually correlated. However, this is not problematic since centrality measures in real-world networks are also often correlated with each other (Valente et al., 2008). Various properties of graphs that are generated by the random graph configurations used in this paper are listed in Table 2.1.

2.2.2 Quantifying measurement errors and reliability

Modeling measurement errors As discussed in the introduction, in a variety of scenarios, it is reasonable to assume that missing nodes are not independent of each other. In fact, missing nodes might belong to the same community and thus the absence of nodes is “biased” towards communities. Here we describe a novel approach to model this type of measurement error.

To simulate that $\lceil \alpha \cdot |V(G)| \rceil$ nodes are missing from a graph G, we create a copy of G that we denote by G' and proceed as follows:

1. First, we choose a community from which one node will be removed. We denote this community by j. We can enumerate the communities since the random graph models provide us the mapping from the nodes to the communities.

Let $P(j)$ be the probability that community j will be selected. Then

$$P(j) \propto [1 + missing(j)]^{\lambda} \tag{2.1}$$

if there are still nodes in the graph that belong to community j. Here, $missing(j)$ denotes the number of nodes that belong to community j and have already been removed from the graph G' and λ is a non-negative real number which determines the strength of the "community bias". If all nodes of community j have already been removed from the graph, $P(j) = 0$.

2. Next, we randomly choose a node from $V_j(G)$ that has not yet been removed from G' and remove it from G'
3. We repeat this process until $\lceil \alpha \cdot |V(G)| \rceil$ nodes have been removed from G'.

This procedure has two parameters. The intensity of the simulated measurement error is controlled by α, the fraction of nodes that are removed from the graph. The extent of the bias of missing nodes to belong to the same community ("community bias") is controlled by λ. If $\lambda = 0$, then there is no community bias and all nodes have the same probability to be removed from the graph (independently of already missing nodes) if all communities are of the same size. Otherwise $P(j)$ has to be reweighted w.r.t. the community sizes. For $\lambda > 0$, nodes are more likely to be removed from communities where nodes have already been removed. For large values of λ, the community that gets chosen in the first iteration usually gets chosen again and again until all nodes from that community are removed from the graph. In this case, entire communities are essentially removed successively.

Quantifying the reliability Network data is usually affected by measurement errors, as discussed in the introduction (e.g., some actors are missing). Hence, we seek to reveal the reliability of centrality values that are calculated based on the erroneous network data. To quantify this reliability of centrality measures, we use the Kendall tau-b rank correlation coefficient τ (Kendall, 1945). Rank correlations are commonly used to evaluate the ramifications of network modifications on centrality measures because researchers are often interested in the ranking of nodes derived from centrality measures rather than in the actual centrality values (Kim and Jeong, 2007; Wang et al., 2012; Lee and Pfeffer, 2015).

Let G be the "error-free" graph, G' an erroneous version of G which is affected by some type of measurement error, and c a centrality measure.

We define the reliability of the centrality measure c with respect to G, G', and the type of measurement error as $\tau(c(G), c(G'))$ (Martin and Niemeyer, 2017). Similar to existing studies, we only consider entries in $c(G)$ and $c(G')$ which correspond to nodes that do exist in G and G' (Kim and Jeong, 2007; Wang et al., 2012). Moreover, we only consider nodes that are in the largest connected component of the particular graph. (We observed, however, that almost all graphs in our experiments were connected.) For reasons of brevity, we only write τ for the reliability of a centrality measure c when G, G', and c are apparent from the context.

2.2.3 Experimental setup

For all random graph configurations that are described in Section 2.2.1, we study the impact of erroneous data collection on the reliability of centrality measures as follows:

1. Generate a graph according to the random graph configuration (e.g., LFR_{weak}) and denote it by G.

2. Apply the remove node procedure (Section 2.2.2) with parameters α and λ to G and denote the resulting modified graph by G'.

3. Finally, calculate the reliability of the centrality measures $\tau(c(G), c(G'))$ as described above (Section 2.2.2).

For our experiments, we use the following parameters: As centrality measures c we use betweenness, closeness, degree, eigenvector centrality, and PageRank. As the fraction of nodes that are removed from the graph, we use values of α ranging from 0.025 to 0.5 in steps of 0.025. To control the extent of the community bias (the likelihood that missing nodes belong to the same community), we use values of λ ranging from 0 to 3 in steps of 0.5. For all combinations of these parameter values, we perform the experiment 100 times.

The NetworkKit library (Staudt et al. (2016), v4.3) is used for graph generation and calculation of centrality measures. The NetworkX library (Hagberg et al. (2008), v1.11) is used for various graph modifications.

2.2.4 Statistical analysis

In addition to a visual inspection, we use two linear models to investigate the relationship between the reliability of centrality measures and the error level, the community bias, and the strength of communities.

To analyze the results for the configurations $\mathrm{CRG}_{\mathrm{weak}}$, $\mathrm{CRG}_{\mathrm{strong}}$, $\mathrm{LFR}_{\mathrm{weak}}$, and $\mathrm{LFR}_{\mathrm{strong}}$, we use the following model:

$$\tau = \beta_0 + \beta_1^{i,j} \cdot \sqrt{\alpha} + \beta_2^{i,j} \cdot \sqrt{\alpha} \cdot \lambda + \epsilon \tag{2.2}$$

With i and j as indices for the centrality measure and graph configuration, respectively. This allows us to have different coefficients for each centrality measure and graph configuration. The error term is denoted by ϵ.

To analyze the results of the experiments regarding the $\mathrm{LFR}_{\mathrm{varying}}(\mu)$ models (with μ ranging from 0.15 to 0.95 in steps of 0.05), we use the following model:

$$\tau = \beta_0 + \beta_1^{i,j} \cdot \sqrt{\alpha} + \beta_2^{i,j} \cdot \sqrt{\alpha} \cdot \mu + \beta_3^{i,j} \cdot \sqrt{\alpha} \cdot \lambda + \beta_4^{i,j} \cdot \sqrt{\alpha} \cdot \mu \cdot \lambda + \epsilon \tag{2.3}$$

With i and j as indices for the centrality measure and graph configuration, respectively. The error term is denoted by ϵ.

We use the square root function to take into account observations from previous studies which have revealed a non-linear relationship between missing nodes and reliability (Smith and Moody, 2013; Smith et al., 2017). Moreover, our experiments have shown that these models provide a better fit to the data than models which do not use this transformation.

2.3 Results

As outlined in the introduction, network data is often affected by measurement errors and in many cases, it is reasonable to assume that there is some dependency between the nodes that are missing. The goal of this study is to investigate how reliable centrality measures are when network data is incomplete and the missing nodes are likely to belong to the same community.

In general, our results suggest that centrality measures are more reliable when missing nodes are biased to belong to the same community. Moreover, for scale-free networks (LFR model) we observe that centrality measures are more reliable in networks with stronger community structure. However, we also observe that, in scale-free networks, the betweenness centrality becomes less reliable with increasing bias.

Figure 2.1 illustrates the results for the graphs generated by the CRG models. For better visibility, the plot only contains results for $\lambda \in \{0, 2\}$ and $\alpha \in \{0.1, 0.3, 0.5\}$. The bottom and top of the boxes indicate the first and third quartiles, respectively. The thick line within the box indicates the

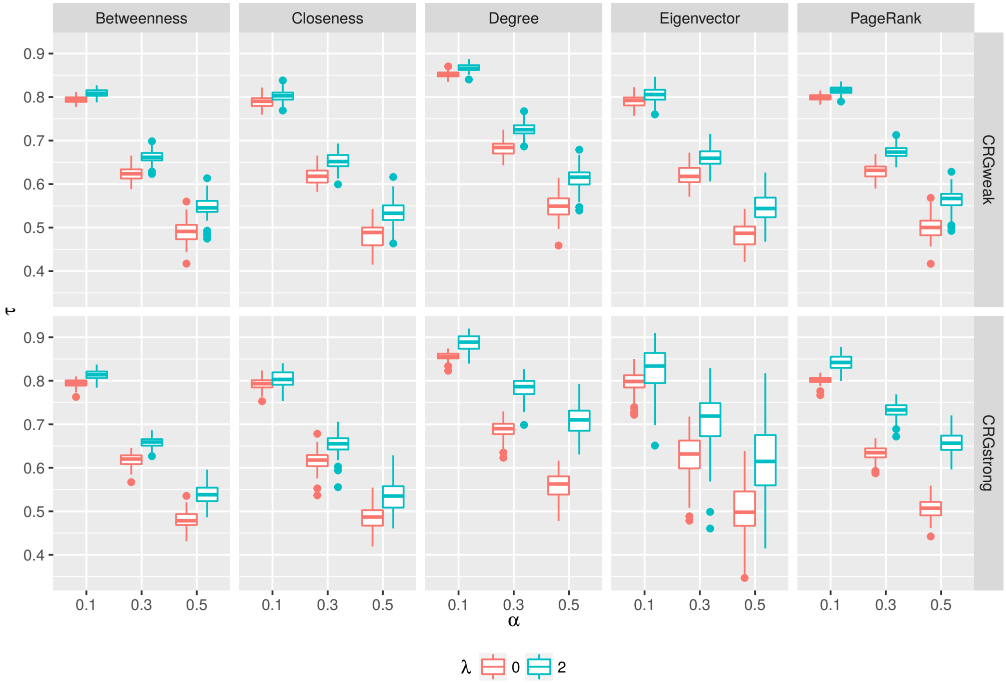


Figure 2.1: The figure shows results for the CRG models. For better visibility, the plot only contains results for $\lambda \in \{0, 2\}$ and $\alpha \in \{0.1, 0.3, 0.5\}$. The bottom and top of the boxes indicate the first and third quartiles, respectively. The thick line within the box indicates the median.

Table 2.2: Results for the model in Equation (2.2). Standard errors are listed in parentheses. All coefficients are highly significant (p-value < 0.001). Intercept: 1.034 (1.9E-04), adjusted R^2: 0.938.

	CRGstrong		CRGweak	
	$\sqrt{\alpha}$	$\sqrt{\alpha} \cdot \lambda$	$\sqrt{\alpha}$	$\sqrt{\alpha} \cdot \lambda$
Betweenness	-0.783	0.039	-0.766	0.034
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Closeness	-0.813	0.046	-0.782	0.032
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Degree	-0.698	0.099	-0.663	0.039
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Eigenvector	-0.868	0.110	-0.786	0.040
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
PagerRank	-0.777	0.094	-0.755	0.039
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
	LFRstrong		LFRweak	
	$\sqrt{\alpha}$	$\sqrt{\alpha} \cdot \lambda$	$\sqrt{\alpha}$	$\sqrt{\alpha} \cdot \lambda$
Betweenness	-0.567	0.002	-0.580	0.005
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Closeness	-0.340	-0.031	-0.494	-0.008
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Degree	-0.372	0.031	-0.413	0.004
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
Eigenvector	-0.362	-0.016	-0.490	-0.010
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)
PagerRank	-0.491	0.038	-0.535	0.006
	(9.6E-04)	(5.0E-04)	(9.6E-04)	(5.0E-04)

median. As can be seen, even small levels of error result in a considerable drop (ranging from 0.1 to 0.2) of the reliability. Moreover, all centrality measures are generally more reliable when the missing nodes belong to the same community ($\lambda = 2$) compared to uniform node missingness ($\lambda = 0$). This effect is more noticeable in cases with stronger community structure (CRG$_{\text{strong}}$). We also notice that the variance of the reliability increases with increasing error level. It is particularly high for the eigenvector centrality and lowest for the degree centrality.

For a more detailed analysis of the relationship between the bias of missing nodes (controlled by λ) and the reliability, we use the model shown in Equation (2.2) from Section 2.2.4. (We also performed our analyses using more robust methods (i.e., weighted linear regression and quantile

Table 2.3: Results for the model in Equation (2.3). Standard errors are listed in parentheses. All coefficients are highly significant (p-value < 0.001). Intercept: 1.037 (1.0E-04), adjusted R^2: 0.873.

	$\sqrt{\alpha}$	$\sqrt{\alpha} \cdot \lambda$	$\sqrt{\alpha} \cdot \mu$	$\sqrt{\alpha} \cdot \mu \cdot \lambda$
Betweenness	-0.606	-0.008	0.028	0.021
	(6.4E-04)	(3.4E-04)	(1.0E-03)	(5.7E-04)
Closeness	-0.239	-0.046	-0.311	0.045
	(6.4E-04)	(3.4E-04)	(1.0E-03)	(5.7E-04)
Degree	-0.356	0.062	-0.067	-0.071
	(6.4E-04)	(3.4E-04)	(1.0E-03)	(5.7E-04)
Eigenvector	-0.298	-0.020	-0.230	0.014
	(6.4E-04)	(3.4E-04)	(1.0E-03)	(5.7E-04)
PagerRank	-0.444	0.065	-0.117	-0.070
	(6.4E-04)	(3.4E-04)	(1.0E-03)	(5.7E-04)

regression) and these results are consistent with the results reported in this section.) The coefficient and standard error estimates for this model are listed in Table 2.2. These results confirm our previous observation for the CRG configurations: higher community bias is related to higher reliability (interaction term $\sqrt{\alpha} \cdot \lambda$). For the LFR$_{\text{strong}}$ configurations, this effect only occurs for the degree centrality and the PageRank. For the closeness centrality, we observe the opposite effect. For the betweenness and eigenvector centrality, the coefficients are small and the effect is negligible. The coefficients of the interaction term regarding the LFR$_{\text{weak}}$ model are significant but small, the effect is hardly noticeable. Comparing the CRG and the LFR model, the effect of λ on the reliability is usually stronger in graphs generated by one of the CRG models.

To analyze the impact of the community strength (controlled by μ) on the reliability of centrality measures, we use the model shown in Equation (2.3). The coefficient and standard error estimates for this model are listed in Table 2.3. The coefficients for $\sqrt{\alpha}$ and $\sqrt{\alpha} \cdot \mu$ are in good agreement with the results of the first model (Equation (2.2)).

All centrality measures except the betweenness centrality become more reliable with increasing strength of the community structure (indicated by $\sqrt{\alpha} \cdot \mu$). The contrary is true for the betweenness centrality. The results also show (indicated by $\sqrt{\alpha} \cdot \mu \cdot \lambda$) that, in case of betweenness, degree centrality and PageRank, a bias of missing nodes towards community amplifies the previously mentioned effect. The contrary is true for the closeness centrality, though the effect is small. In case of the eigenvector centrality, the effect is negligible.

2.4 Discussion

Networks are complex, and it is hard to collect network data without missing any nodes or edges. Previous studies have shown that missing nodes can severely affect the reliability of centrality measures. Most studies focus, however, on cases in which nodes are missing uniformly at random. Yet in a variety of scenarios, it is reasonable to assume that missing nodes may belong to the same community.

In this study, we investigated the reliability of centrality measures when network data is incomplete and the missing nodes are likely to belong to the same community. In addition, we introduced a novel approach, called "community bias", which allows researchers to simulate different levels of measurement error.

In our experiments on uniform and scale-free networks, we observed that centrality measures are more reliable when missing nodes are likely to belong to the same community compared to those cases in which nodes are missing uniformly at random. In scale-free networks, the betweenness centrality, however, becomes less reliable with increasing bias. Moreover, in these networks, centrality measures are also more reliable in networks with stronger community structure. In contrast, we did not observe this effect for uniform networks.

To the knowledge of the author, this is the first study which examines the effect that missing nodes have if their absence depends on the underlying community structure. A direct comparison to other studies is therefore difficult. In contrast to the present study, Niu et al. (2015) found that the biased manipulation of networks has more severe consequences than a uniformly random manipulation. It is important to note here that the manipulations in Niu et al. (2015) were applied to the edges; nodes were, however, not considered. Our study shows that an increasing bias is associated with higher reliability, which is a novel finding. If there are legitimate reasons to assume that nodes that have not been observed during the data collection are more likely to belong to the same community, the impact of missing nodes on the reliability of centrality measures might not be as severe as previous studies have suggested.

These findings are encouraging. Although graphs generated by the LFR model share many properties with real-world networks, it would be interesting to see results based on empirical data as well as results for larger networks. We are going to investigate these cases in our future studies. Furthermore, future work may investigate other types of interdependencies between missing nodes, for example, based on node attributes.

2.5 References

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5):175–308.

Boldi, P., Rosa, M., and Vigna, S. (2013). Robustness of social and web graphs to node removal. *Social Network Analysis and Mining*, 3(4):829–842.

Bolland, J. M. (1988). Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182.

Borgatti, S. P., Carley, K. M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136.

Brin, S. and Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Clauset, A., Rohilla Shalizi, C., and J Newman, M. E. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Frantz, T. L., Cataldo, M., and Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Garbers, J., Promel, H., and Steger, A. (1990). Finding clusters in VLSI circuits. In *1990 IEEE International Conference on Computer-Aided Design. Digest of Technical Papers*, pages 520–523. IEEE Comput. Soc. Press.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15.

Havlin, S. and Cohen, R. (2010). *Complex networks: structure, robustness and function.*

Hu, Y., Chen, H., Zhang, P., Li, M., Di, Z., and Fan, Y. (2008). Comparative definition of community and corresponding identifying algorithm. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(2).

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251.

Kim, P. J. and Jeong, H. (2007). Reliability of rank order in sampled networks. *European Physical Journal B*, 55(1):109–114.

Koschützki, D., Lehmann, K., and Peeters, L. (2005). Centrality Indices. In Brandes, U. and Erlebach, T., editors, *Network Analysis: Methodological Foundations*, pages 16–61. Springer Berlin Heidelberg.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(4).

Lee, J.-s. and Pfeffer, J. (2015). Robustness of Network Centrality Metrics in the Context of Digital Communication Data. *Proceedings of the 48th Hawaii International Conference on System Sciences.*

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, pages 695–704.

Martin, C. and Niemeyer, P. (2017). Estimating the sensitivity of centrality measures w.r.t. measurement errors. *arXiv.org e-prints*, (arxiv.org/abs/1704.01045).

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):26113.

Niu, Q., Zeng, A., Fan, Y., and Di, Z. (2015). Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131.

Rezvanian, A. and Meybodi, M. R. (2015). Sampling social networks using shortest paths. *Physica A*, 424:254–268.

Silk, M. J. (2017). The next steps in the study of missing individuals in networks: a comment on Smith et al. (2017). *Social Networks.*

Smith, J. A. and Moody, J. (2013). Structural Effects of Network Sampling Coverage I: Nodes Missing at Random. *Social networks*, 35(4).

Smith, J. A., Moody, J., and Morgan, J. H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48:78–99.

Staudt, C. L., Hamann, M., Safro, I., Gutfraind, A., and Meyerhenke, H. (2017). Generating scaled replicas of real-world complex networks. *Studies in Computational Intelligence*, 693:17–28.

Staudt, C. L., Sazonovs, A., and Meyerhenke, H. (2016). NetworKit: A tool suite for large-scale complex network analysis. *Network Science*, 4(4):508–530.

Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. (2008). How correlated are network centrality measures? *Connections*, 28(1):16–26.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press.

Influence of measurement errors on networks: Estimating the robustness of centrality measures

Christoph Martin and Peter Niemeyer

This paper has been published as

Martin C., and Niemeyer P. (2019). Influence of measurement errors on networks: Estimating the robustness of centrality measures. Network Science 7, 180–195. https://doi.org/10.1017/nws.2019.12.

The layout has been revised.

Abstract

Most network studies rely on a measured network that differs from the underlying network which is obfuscated by measurement errors. It is well known that such errors can have a severe impact on the reliability of network metrics, especially on centrality measures: a more central node in the observed network might be less central in the underlying network.

Previous studies have dealt either with the general effects of measurement errors on centrality measures or with the treatment of erroneous network data. In this paper, we propose a method for estimating the impact of measurement errors on the reliability of a centrality measure, given the measured network and assumptions about the type and intensity of the measurement error. This method allows researchers to estimate the robustness of a centrality measure in a specific network and can, therefore, be used as a basis for decision making.

In our experiments, we apply this method to random graphs and real-world networks. We observe that our estimation is, in the vast majority of cases, a good approximation for the robustness of centrality measures. Beyond this, we propose a heuristic to decide whether the estimation procedure should be used. We analyze, for certain networks, why the eigenvector centrality is less robust than, amongst others, the PageRank. Finally, we give recommendations on how our findings can be applied to future network studies.

3.1 Introduction

Measurement errors in network data are a central problem in the field of network analysis, as virtually all empirical network data are affected by some kind of measurement error. Previous research has shown that these errors often have a major impact on the results of network measures, especially on centrality measures (Costenbader and Valente, 2003; Smith and Moody, 2013). For example, a more central node in the measured (erroneous) network might be less central in the hidden (unobserved, error-free) network.

Currently, most applied network studies only report that measurement errors might have affected the data collection (e.g., due to the absence of actors on the day of the survey (Wang et al., 2016), due to the study design, or due to the quality of some external data source (Fischer et al., 2018)). Most of the time, however, the impact of these measurement errors on centrality measures are not quantified. This might be due to the fact that there is currently no established way to estimate the impact of these of measurement errors on centrality measures. In this paper, we present a method to approximate this impact, given a measured network and some hypotheses about the underlying error mechanism.

Current research can be divided into two main categories: impact studies and treatment studies.

In impact studies, researchers have investigated the impact that different types of measurement errors have on the reliability of centrality measures in the case of random graphs (Borgatti et al., 2006; Frantz et al., 2009; Wang et al., 2012) and real-world networks (Costenbader and Valente, 2003; Kim and Jeong, 2007; Wang et al., 2012; Smith and Moody, 2013; Platig et al., 2013; Silk et al., 2015; Niu et al., 2015; Lee and Pfeffer, 2015), cf. Smith et al. (2017) for an extensive survey.

These studies provide guidelines for researchers on how to design future studies (e.g., what kind of measurement error might be especially harmful in a given scenario) and suggestions on which centrality measure might be more reliable in a given scenario (Smith and Moody, 2013). Unfortunately, it is difficult to identify general patterns for the reliability of centrality measures based on network metrics. As common sense suggests, centrality measures become less reliable with an increasing level of error, but the particular relationship between error level and reliability is highly dependent on the type of measurement error, the centrality measure, and the network structure (Wang et al., 2012).

There are also studies which address how to treat erroneous network data, in order to reconstruct the unknown true network. Such treatments can, for example, be used to reconstruct partially observed networks or to estimate the network statistics of the underlying network (Butts, 2003; Huisman, 2009; Handcock and Gile, 2010; Kim and Leskovec, 2011; Frantz and Carley, 2017; Wang et al., 2016; Newman, 2018; Krause et al., 2018; Žnidaršič et al., 2018).

Our contribution connects these two areas. We propose a method for estimating the impact of measurement errors on the reliability of a centrality measure, given the measured network and assumptions about the type and intensity of the measurement error. This method allows researchers to measure the robustness of a centrality measure in a specific network and can, therefore, be used as a basis for decision making — for example, to decide whether the centrality values are reliable enough for the purposes of the study or whether one of the aforementioned treatment procedures should be applied. One of the strengths of this method is that the estimates are easy to calculate. Simply explained, we apply the assumed error mechanism (e.g., random removal of 10% of the edges) several times, independently, to the measured network and suggest the mean impact of this procedure as the estimate for the measurement error between the unknown true network and the measured network.

We test this method in various simulation scenarios based on random graphs and real-world networks as well. We find that the estimation works in many cases, especially at lower error levels (e.g., 10% missing edges or vertices). At higher error levels (e.g. 30% missing edges or vertices) the estimation still works for degree centrality and PageRank. For sparse or small networks the situation is more challenging, especially in the case of eigenvector centrality.

The rest of this paper is organized as follows: we formalize the concepts of robustness and error mechanisms in Section 3.2. The estimation method is presented in Section 3.3, and the experiments are described and discussed in Section 3.4. A summary and concluding recommendations can be found in Section 3.5.

3.2 Basic concepts

Let G be an undirected, unweighted, finite graph with vertex set $V(G)$ and edge set $E(G)$. A centrality measure c is a real-valued function that assigns centrality values to all nodes in a graph and is invariant to structure-preserving mappings, i.e., centrality values depend solely on the structure of a graph. External information (e.g., node or edge attributes) has no influence on the centrality values (Koschützki et al., 2005). We denote the centrality value for node $u \in V(G)$ by $c_G(u)$ and the centrality values for all nodes in G $(u_1, u_2, \dots, u_n)$ by the vector $c(G) := (c(u_1), \dots, c(u_n))$.

The following centrality measures are used in this study: closeness centrality, betweenness centrality (Freeman, 1978), degree centrality, eigenvector centrality (Bonacich, 1987), and the PageRank (damping factor 0.85) (Brin and Page, 1998). All centrality measures are calculated using the igraph library (version 0.7.1, Csardi and Nepusz (2006)).

Let G and G' be two graphs and c a centrality measure. A pair of nodes $u, v \in V(G) \cap V(G')$ and $u \neq v$ is called concordant w.r.t. c if both nodes have distinct centrality values and the order of u and v is the same in c(G) and c(G'), i.e., either $c_G(u) < c_G(v)$ and $c_{G'}(u) < c_{G'}(v)$ or $c_G(u) > c_G(v)$ and $c_{G'}(u) > c_{G'}(v)$. A pair of nodes is called discordant if both nodes have distinct centrality values and the order of u and v in c(G) differs from the order of u and v in c(G'), i.e., either $c_G(u) < c_G(v)$ and $c_{G'}(u) > c_{G'}(v)$ or $c_G(u) > c_G(v)$ and $c_{G'}(u) < c_{G'}(v)$. Ties are neither concordant nor discordant.

A random graph consists of a finite set of graphs Ω equipped with a function P that assigns a probability to every graph in this set (Bollobás and Riordan, 2002).

Network data can be influenced by a variety of different measurement errors. Wang et al. (2012) categorized measurement errors into six groups: false negative nodes and edges, false positive nodes and edges, and false aggregation and disaggregation. For example, when 10% of the edges are missing in the measured network data, the graph constructed from this observed data suffers from false negative edges.

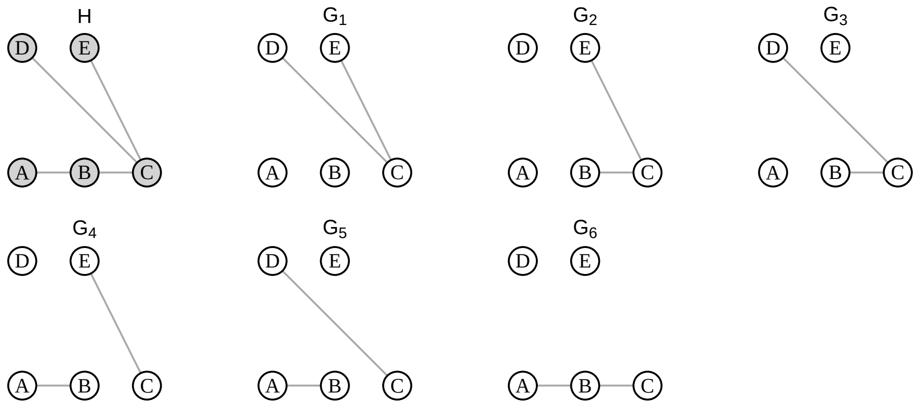


Figure 3.1: In this example, φ is defined as the error mechanism "50% of all edges are missing uniformly at random". Hence, $\varphi(H)$ is a random graph with possible outcomes $\Omega = \{G_1, G_2, \ldots, G_6\}$ and $P(G_i) = \frac{1}{6}$.

To describe measurement errors, we introduce the notion of an error mechanism. An error mechanism φ is a procedure that describes measurement errors that may occur during the data collection (e.g., 50% of the edges are missing, at random). For a given graph G, the error mechanism $\varphi(G)$ is defined as a random graph. The outcomes of $\varphi(G)$ are the graphs that result from G by applying the given error-procedure, and each of these graphs is equipped with the probability of occurrence. To illustrate this concept, consider the graphs shown in Figure 3.1. The initial graph is denoted by H (drawn in the upper-left corner). We assume that we know the error mechanism that compromises the data collection. For this example, we assume that the error mechanism φ is edges missing uniformly at random with an error level of 50%. All graphs in the set of possible outcomes for this random graph $\Omega = \{G_1, G_2, \ldots, G_6\}$ are also shown in Figure 3.1. In this example, the probability function is $P(G_i) = \frac{1}{6}$, all graphs in Ω occur with the same probability. However, this concept is not limited to a uniform distribution.

In general, error mechanisms can rely on node or edge attributes. In this study, we focus on four common error mechanisms that do not depend on external attributes:

1. Nodes missing uniformly at random (rm nodes): A fraction of nodes (and all edges connected to these nodes) is missing in the measured network. All nodes have the same probability to be missing in the measured network.

2. Edges missing uniformly at random (rm edges unif.): A fraction of edges is missing in the measured network. All edges have the same probability to be missing in the measured network.

3. Edges missing proportionally (rm edges prop.): A fraction of edges is missing in the measured network. The probability that an edge is missing in the measured network is proportional to the sum of the degree values of the endpoints.

4. Spurious edges (add edges): The measured network contains too many edges. Every non-existing edge has the same probability to be erroneously observed.

Let G and G' denote graphs on the same vertex set and c a centrality measure. To measure the robustness of c w.r.t. these two graphs, we use Kendall's tau ("tau-b") rank correlation coefficient (Kendall, 1945). Correlations are commonly used to measure the robustness of centrality measures. Like existing studies, we also used rank correlations to minimize the influence of outliers (Kim and Jeong, 2007; Lee and Pfeffer, 2015; Wang et al., 2012).

We calculate the robustness ρ for a centrality measure c with respect to G and G' as follows:

$$\rho_c(G, G') = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_t) * (n_c + n_d + n_{t'})}} \tag{3.1}$$

With n_c as the number of concordant pairs and n_d as the number of discordant pairs w.r.t. the order given by $c(G)$ and $c(G')$. Ties only in $c(G)$ are denoted by n_t, ties only in $c(G')$ by $n_{t'}$.[1] It follows straightforwardly that the values of $\rho_c(G, G')$ are in the interval $[-1, 1]$.

[1] It may occur that $V(G') \neq V(G)$. In these cases, we only consider entries in $c(G)$ and $c(G')$ that correspond to nodes that are in both graphs (G and G'). This is a common approach for the comparison of graphs on different vertex sets (Wang et al., 2012).

Let us apply this concept to a graph illustrated in Figure 3.1. Assume that we have measured the graph labeled as G_6 and that we are interested in the robustness of the degree centrality. Then, the degree centrality values are $deg(H) = (1, 2, 3, 1, 1)$ and $deg(G_6) = (1, 2, 1, 0, 0)$. Based on the degree values, we can calculate the robustness of the degree centrality with respect to G_6 and H: $\rho_{deg}(G_6, H) = 0.53$.

If there are no ties in $c(G)$ and $c(G')$, then $\rho_c(G, G') = \frac{n_c - n_d}{n_c + n_d}$ which is Goodman and Kruskal's rank correlation coefficient γ (Goodman and Kruskal, 1954). In this special case, $\frac{\rho_c(G,G')+1}{2}$ is the probability that two nodes with distinct centrality values, randomly chosen from the common vertex set of G and G', have the same order in $c(G)$ and c(G'), that is, they are concordant.

3.3 How to estimate the robustness of centrality measures

In network studies, the measured network data often contain sampling errors (Leecaster et al., 2016; Schulz, 2016; Wang et al., 2016). But in general, the authors of such studies have no tools to describe the impact of sampling errors on the network measures (e.g., centrality measures) they apply. In general, the assumptions made about sampling errors are mentioned in the limitations, but they are not considered as part of the network model.

The robustness concept as introduced in Section 3.2 helps researchers to describe this impact: given the measured network M, the (unknown) hidden network H, and a centrality measure c, the robustness $\rho_c(H, M)$ measures the impact of the sampling error on the centrality values of the nodes in the measured network. Thus, the robustness can be used to measure the reliability of a centrality measure with respect to sampling errors. We call $\rho_c(H, M)$ the "true robustness".

As the hidden network H is not known, the true robustness cannot be computed explicitly. In this section, we propose a method for the estimation of the true robustness based on the measured network M. Moreover, we provide an example for the application and demonstrate how the estimation results can be evaluated.

Our estimation approach is based on the observation that, given a graph G, a centrality measure, and some error procedure, in many experiments the robustness is nearly proportional to the error intensity. That is, removing 20% of the edges, randomly, has about twice as much impact on the centrality measure as removing 10% of the edges (Borgatti et al., 2006; Frantz et al., 2009; Wang et al., 2012). Now let G' denote the graph resulting from G by removing 10% of the edges, and let G'' denote the graph resulting from G' by removing 10% of the remaining edges. One possible explanation for the observed linearity could be that the robustness with respect to G and G' is close to the robustness with respect to G' and G", which is $\rho_c(G, G') \sim \rho_c(G', G'')$.

If we apply this idea to our definition of the true robustness and take into account our notion of error mechanism (as random graphs), we yield the following estimation:

$$\hat{\rho}_c(M, H) := E(\rho_c(M, \varphi(M))). \tag{3.2}$$

Since $\varphi(M)$ is a random graph, $\rho_c(M, \varphi(M))$ is a random variable; hence we use the expected value of this expression as the estimate for the robustness. In practice, this value is computed by sampling.

3.4 Experiments

In this section, the efficiency of the estimation method is analyzed under different conditions using two types of simulation experiments. First, experiments based on random graphs are conducted, followed by experiments based on real-world networks. To control for the true robustness, in all experiments we start with a given *hidden* network H and construct the *measured* network M by applying the error mechanism to the hidden network. As part of these experiments, we calculate the following values for each run:

True robustness: For every single experiment we compute the true robustness $\rho_c(M, H)$.

Estimated robustness: For every single experiment we compute $\hat{\rho}_c(M, H)$, as introduced in Section 3.3.

Mean and standard deviation of true robustness: For every series of experiments (i.e., fixed centrality measure, fixed error mechanism, fixed random type or initial real-work network), we report the mean value and the standard deviation of all corresponding true robustness values. Note that large values indicate that the true robustness very much depends on the specific choice of removed/added vertices/edges. In such cases, estimating the true robustness given just the measured network and the error mechanism is hardly possible. In this sense, a large standard deviation of true robustness is a good indicator for ill-conditioned estimation problems. Note that these values cannot be computed without knowing the hidden network.

Mean standard deviation of estimation: By definition, the estimated robustness is an expected value. Here we compute the mean of all corresponding standard deviations. Note that this value by construction is very closely related to the standard deviation of true robustness, but it can be computed without knowing the hidden network.

Mean estimation error: For every series of experiments (i.e., fixed centrality measure, fixed error mechanism, fixed random type, or initial real-work network), we compute the mean absolute difference between true robustness and estimated robustness.

Implicit error: As a kind of benchmark for the estimation error, we compute the robustness error that would occur if we ignored the impact of the measurement error on the centrality measure. In this case we would consider the correlation between the centralities of hidden and measured network to be 1. That is, for every series of experiments we define the implicit error as (1− mean true robustness).

3.4.1 Experiments based on random graphs

As a first step to validate whether the proposed methods yield useful results, we apply the four error mechanisms (node missing uniformly, edges missing uniformly, edges missing proportional, and spurious edges) to Erdős-Rényi graphs (ER graph) (Erdős and Rényi, 1959) and Barabási-Albert graphs (BA graph) (Barabási and Albert, 1999) and estimate the corresponding robustness. For every error mechanism, we consider two cases: a moderate scenario of 10% error level and a more intense scenario with 30% error level. For all combinations of centrality measures and error mechanisms, we perform the experiment described below 1,000 times.

1. We generate a random graph and denote it by H. This graph represents the (error-free) hidden network. We use two types of random graphs:

    (a) an ER graph with 100 nodes and edge probability 0.2 and [2]

    (b) a BA graph with 100 nodes (parameter $m = 11$, undirected).

[2] Our experiments have shown that the choice of p has little influence on the main results associated with this section. Hence we will only consider the case of $p = 0.2$.

2. We choose a graph from $\varphi(H)$ and denote it by M. This graph represents the measured network that is affected by measurement errors. For evaluation purposes, the true robustness $\rho_c(H, M)$ is calculated and denoted by ρ.

3. Based on the measured network M, we estimate the true robustness $\hat{\rho}_c(M, H)$.

The results for the random graphs are shown in Figure 3.2 and Figure 3.3. Every panel shows the results for the five centrality measures under the influence of one of the four error mechanisms with either 10% or 30% intensity.

The blue bar indicates the mean estimation error of the 1,000 simulation runs. The red bar indicates the implicit error as defined at the beginning of this section. Note again that the implicit error is closely related to the mean true robustness since it is defined as (1 − implicit error). That is, a long red bar indicates a weak correlation between the centrality vector in measured and hidden networks and vice versa. The length of the bars indicates the mean values and the error bars the corresponding standard deviations.

Now let us first focus on the impact of different error mechanisms on the true robustness (1 − red bar). For example, the true robustness of the betweenness in an ER graph under the influence of 10% spurious edges (1st panel in Figure 3.2) is 0.78 with a standard deviation (sd) of 0.03.

For ER graphs (Figure 3.2), within the two error levels, there are only small differences regarding the influence of the error mechanisms on the centrality measures. The degree is the most robust measure in this setting.

With an error level of 30%, the robustness is always lower than in the respective cases with 10%. The standard deviation is also higher. In contrast to the cases with 10%, at 30% the absence of edges depending on the edge degree leads to lower robustness when compared the other error mechanisms. These findings are conclusive with Borgatti et al. (2006) and Frantz et al. (2009). The homogeneity of the results regarding the different centrality measures is not surprising given the high correlation between the centrality measures in the case of ER random graphs (Valente et al., 2008).

For BA graphs, we make similar observations. Higher error levels lead to lower robustness. In contrast to ER graphs, however, degree centrality is not always the most robust here. We notice that the standard deviation is not as homogeneous as in the ER experiments (e.g., eigenvector centrality).

Regarding estimation errors, the pattern is the same for both graph types. The estimation error is always small compared to the implicit error. With fixed intensity the difference between the error types and the centrality measures is small. The estimates at 10% error level have a smaller error

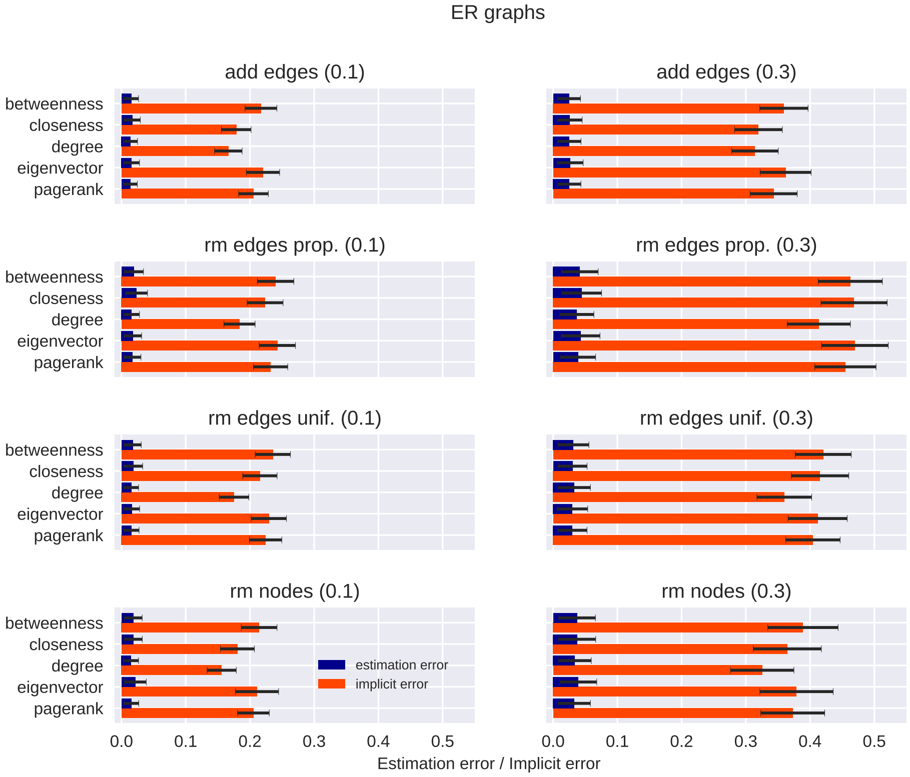


Figure 3.2: The figure shows the results for **ER graphs**. The blue bar indicates the mean absolute error of the estimation $(|\rho - \hat{\rho}|)$ for 1,000 simulation runs. We call this error the estimation error. The red bar indicates the error that is caused by the flawed data collection. This error would be accepted if the influence of the measurement error were to be ignored when analyzing the network $(1 - \rho)$. We call this value the implicit error. The length of the bars indicates the mean values and the error bars the standard deviations. In the case of ER graphs the behavior is homogeneous: the true robustness depends primarily on the error intensity. The estimation errors are consistently low.

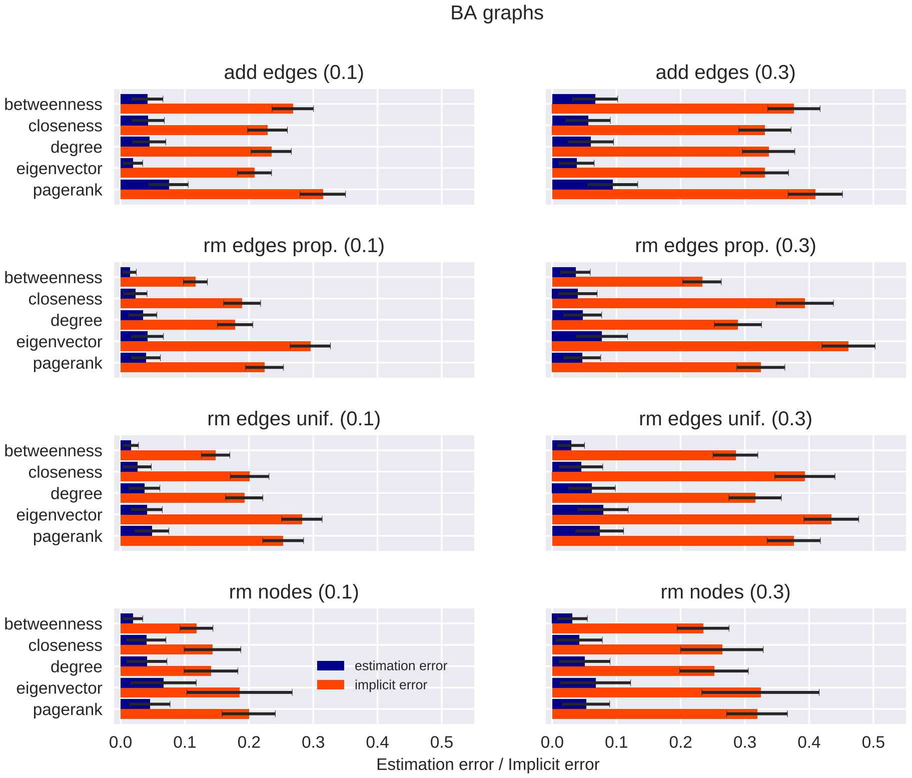


Figure 3.3: The figure presents the results for **BA graphs**. The centrality measures have varying reactions to the different error mechanisms. The eigenvector centrality variation is noticeably high when nodes are removed.

than at 30%. The standard deviation is homogeneous. The estimates are most accurate for spurious edges, worst for missing nodes and missing edges proportional to the edge degree.

For BA graphs (Figure 3.3) the results are more heterogeneous. Although the true robustness is at about the same level as for ER graphs, the estimation errors vary strongly depending on the choice of centrality measure and error mechanism. If nodes are missing, the estimates are usually poorest and the variance highest.

3.4.2 Experiments based on real-world networks

Table 3.1: Statistics of real-world networks. If the original network is not connected, we consider only the largest connected component.

Network	Nodes	Edges	Clustering	Density	Diameter	Source
Dolphin	62	159	0.3029	0.0841	8	Lusseau et al. (2003)
Jazz	198	2,742	0.6334	0.1406	6	Gleiser and Danon (2003)
Protein	1,458	1,948	0.1403	0.0018	19	Jeong et al. (2001)
Hamsterster	1,788	12,476	0.1655	0.0078	14	Kunegis (2013)

Next, we apply our methods from Section 3.3 to real-world networks to investigate the suitability of these methods for practical application. We choose four networks from different domains and thus different structural properties to get an impression of how these methods perform on real data (see Table 3.1 for descriptive statistics). As before we use our proposed methods to estimate the robustness of five centrality measures under the influence of four error mechanisms and two error intensities. For every combination of network, centrality measure, and error mechanism, the experimental setup is as follows:

1. Due to the very nature of the hidden networks, we cannot access them. Hence, for the sake of our experiments, we treat the real-world network as the error-free hidden network H. (This is a common approach used in existing studies about the robustness of centrality measures (Wang et al., 2012)).

2. To simulate erroneous data collection, we choose a graph from $\varphi(H)$ and denote it by M. This graph represents the measured network that is affected by measurement errors. For evaluation purposes, the true robustness $\rho_c(H, M)$ is calculated and denoted by ρ.

3. Based on the measured network M, we estimate the true robustness $\hat{\rho}_c(M, H)$.

For every combination, we perform this experiment 1,000 times.

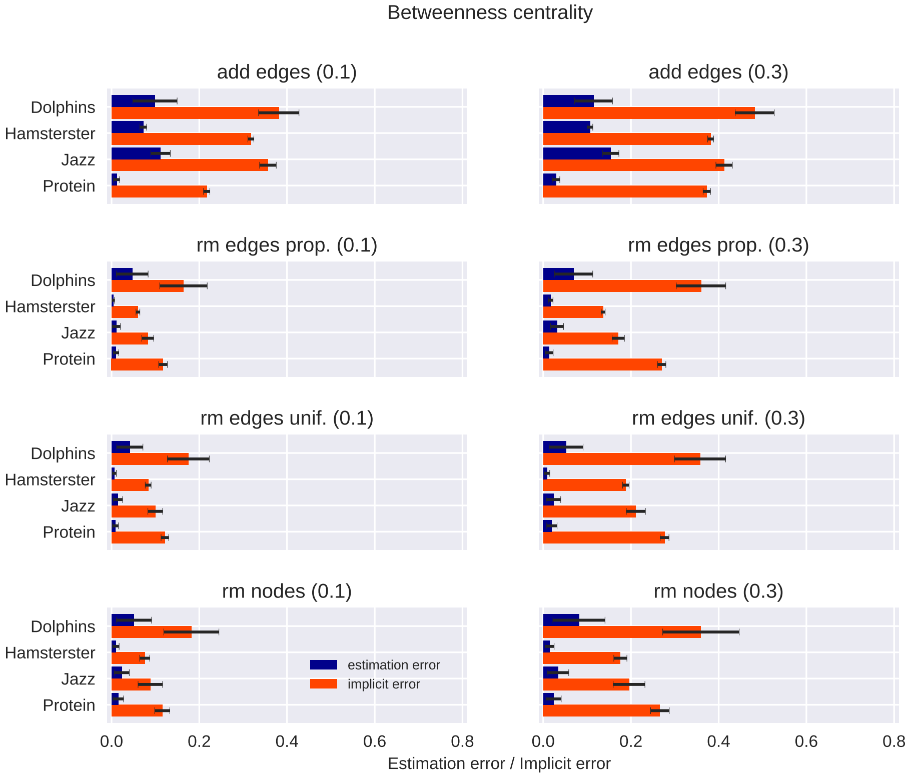


Figure 3.4: The figure displays the results for the **betweenness centrality** in real-world networks. The true robustness depends on the network, type, and intensity of the error. The estimates are good in most cases, but the error type spurious edges leads to increased estimation errors.

The results for real-world networks are shown in Figures 3.4–3.8. In summary, the results are promising. The estimation error is always below, in most cases far below, the implicit error. With a lower error level (10%), the error of estimation is often very low (mean estimate error values below 0.03).

The robustness of the centrality values is usually strongly dependent on the respective network structure and the type of error, where a higher error level always reduces the robustness. Regardless of the error type, degree

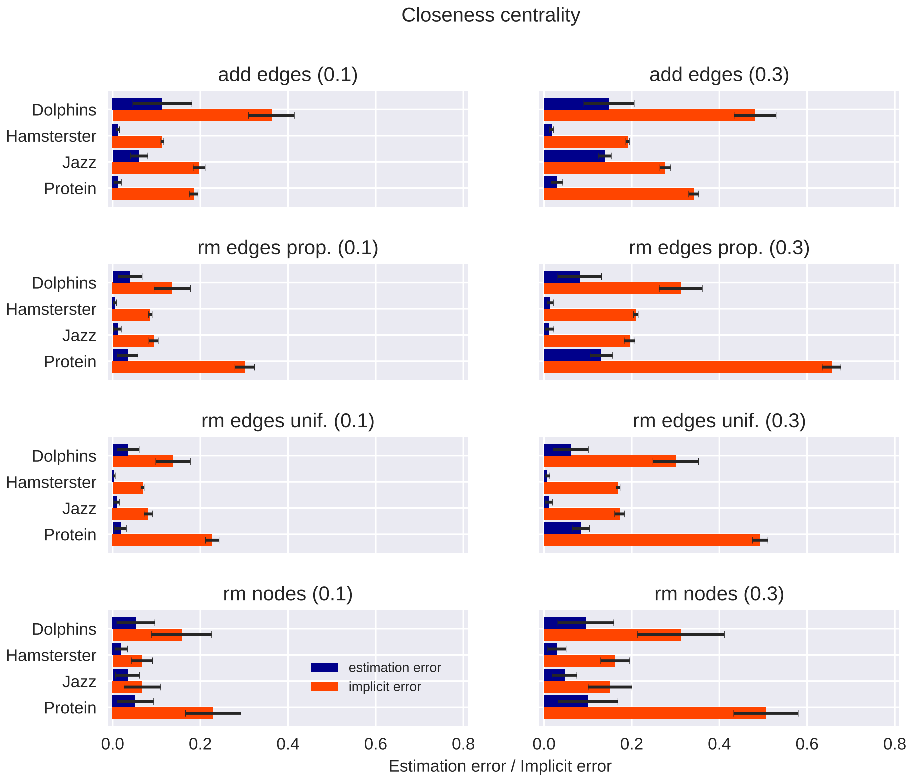


Figure 3.5: The figure shows the results for the **closeness centrality** in real-world networks. The true robustness depends on the network, type, and intensity of the error. The estimates are good in most cases, especially for low error intensities. For higher intensities the errors are higher, especially in case of the Dolphin and Protein networks.

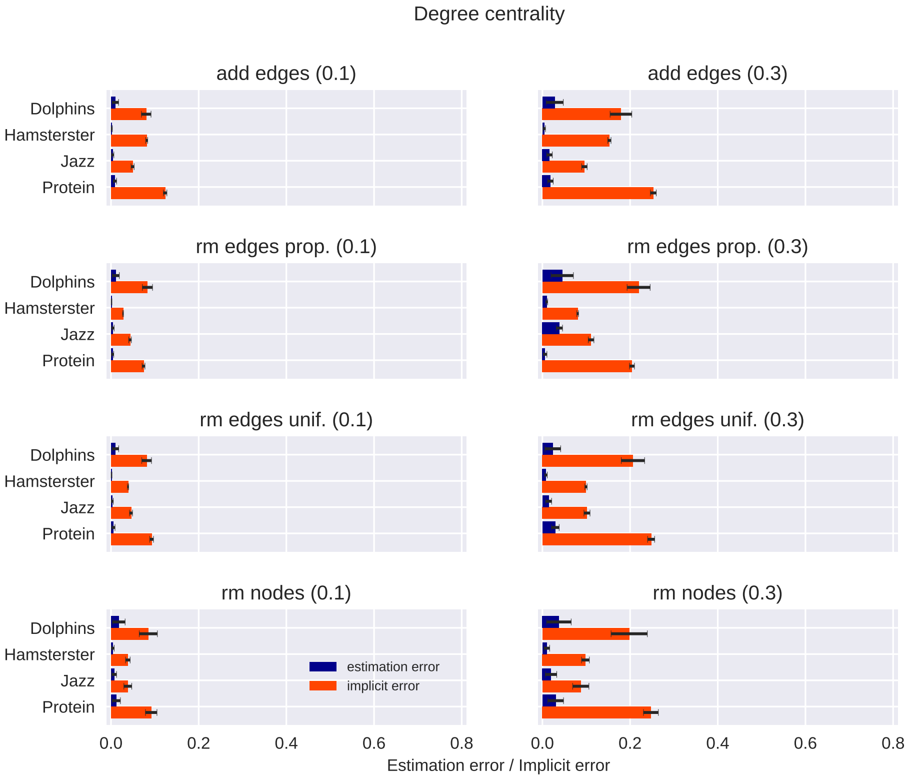


Figure 3.6: The figure illustrates the results for the **degree centrality** in real-world networks. As expected, the robustness in this case is the highest and the estimation error the smallest.

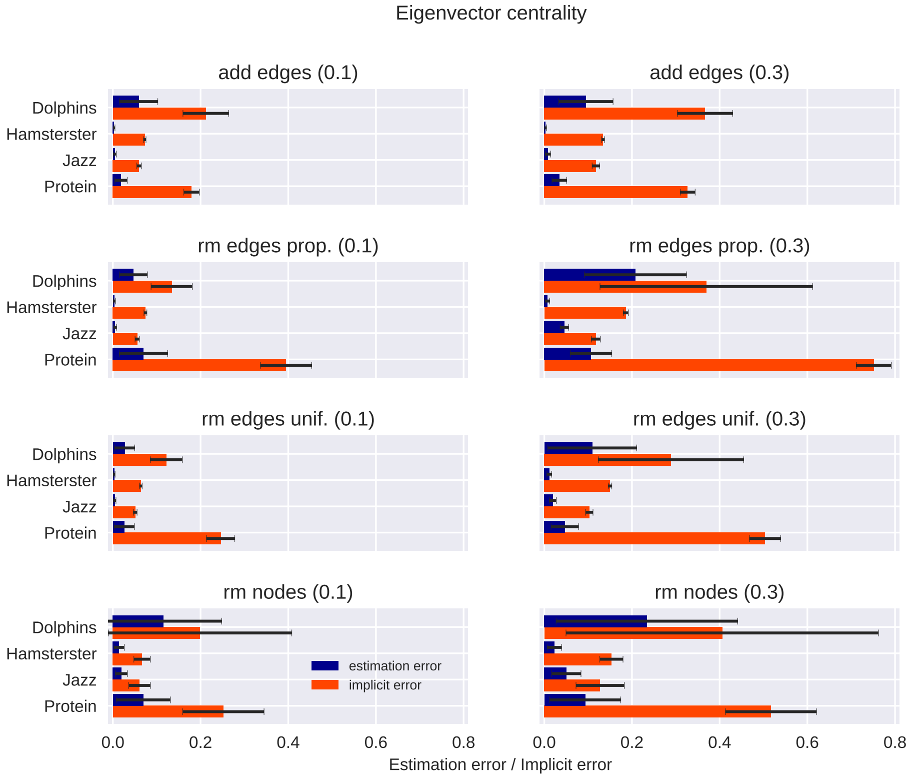


Figure 3.7: The figure shows the results for the **eigenvector centrality** in real-world networks. The results can be divided into two groups. In the Jazz and Hamsterster networks the robustness is high and the estimation error low. In the Protein and Dolphin networks both values are considerably worse, and the fluctuation of both values is higher. This effect is discussed in detail in Section 3.4.

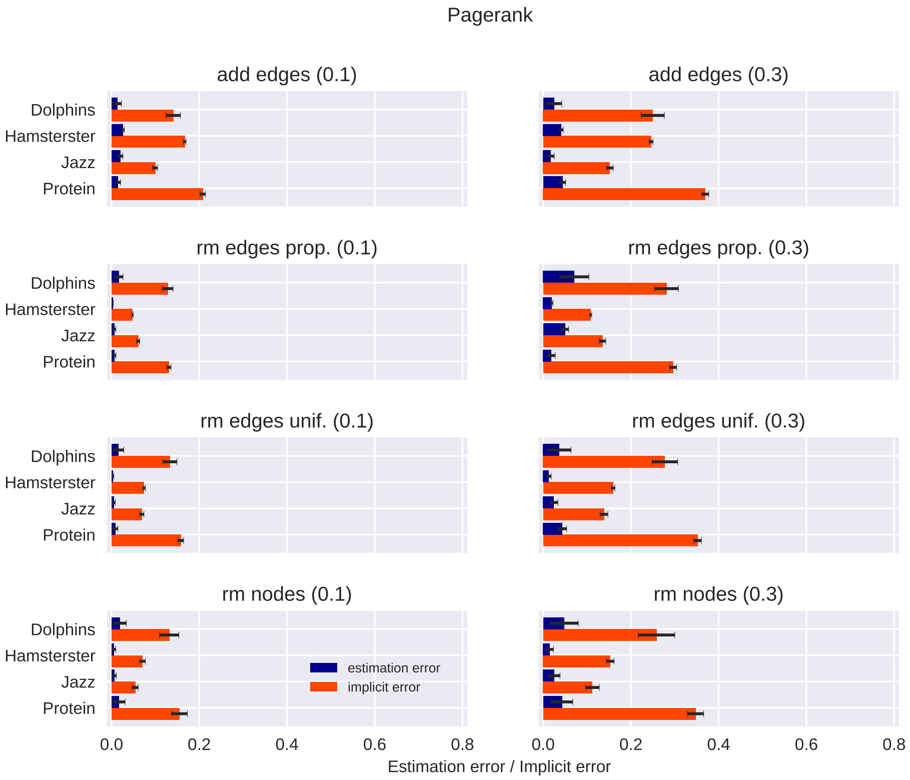


Figure 3.8: The figure presents the results for the **PageRank** in real-world networks. The true robustness depends on the network, type, and intensity of the error. Although robustness is reduced by increasing the error level, estimation errors are relatively low for both intensities.

centrality and PageRank are most robust, while eigenvector centrality is most sensitive to measurement errors.

Degree centrality (Figure 3.6) is robust in the case of all studied networks, most values are in regions below 0.1. However, in the Protein and Dolphin networks, the values are considerably lower (visible in the figure by the higher implicit error) than in the other two networks, in all error types and intensities. The estimation errors are also low (mean values at 0.1 error level mostly below 0.02). The results for PageRank (Figure 3.8) are similar, it is usually robust but somewhat more sensitive than the degree centrality. The difference between the implicit error and the estimation error is equally large. Also, for betweenness (Figure 3.4) and closeness (Figure 3.5), the implicit error is greater than the estimation error, but the difference between them is smaller compared to degree and PageRank. In addition, the standard deviation is higher for both the estimation error and the robustness, especially for the error types and missing nodes and additional edges. The latter has a particularly strong influence on the betweenness. This might be due to the fact that by adding edges (randomly), many new abbreviations are created between the nodes and thus the high diameters (especially in the Hamsterster and Protein networks) are reduced.

The overall impression of the results is most heterogeneous for the eigenvector centrality (Figure 3.7). The difference between the Jazz and Hamsterster networks, on one, and the Protein and Dolphin networks, on the other hand, is most noticeable in this case. While implicit error for the first mentioned networks is comparable to the other centrality measures (except the error mechanism missing nodes), for the other two networks the robustness is low and the standard deviation high (very high for the Protein network). This effect is particularly strong when nodes are missing or edges are missing proportionally to the edge degree. The estimation error varies particularly strongly, especially in cases where the robustness has a high standard deviation.

In our experiments with real-world networks, we have observed that in some cases concerning eigenvector centrality very high implicit errors as well as estimation errors occur. Therefore we want to take a more detailed look at one of these cases. For this purpose we have again performed an experiment (Dolphin network, 30% nodes missing randomly, eigenvector centrality). We track the robustness and the percentage of the total eigenvector centrality that is associated with the removed nodes. The results of this experiment are shown in Figure 3.9. These results demonstrate that the eigenvector centrality of the removed nodes has a strong influence on the robustness. Especially if the removed nodes have a large share of the

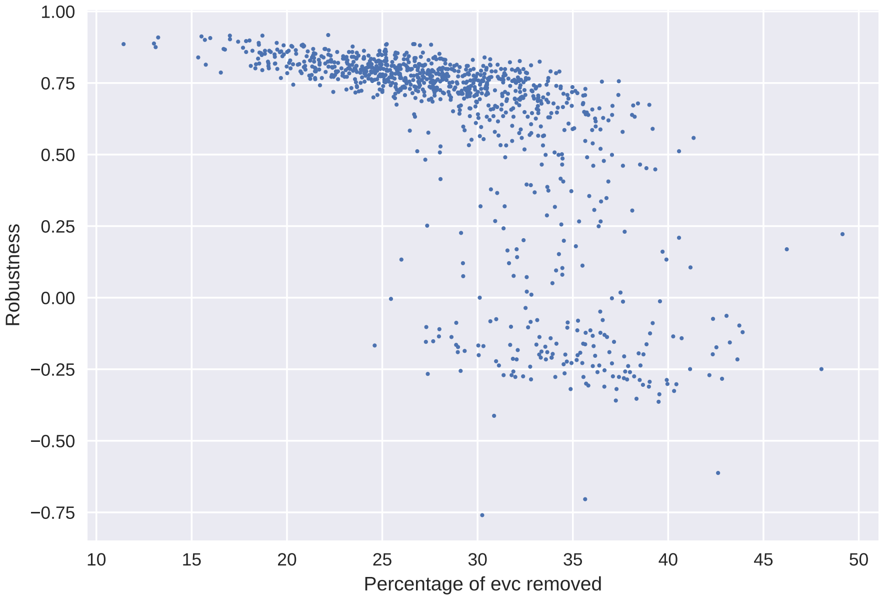


Figure 3.9: The results for the additional experiment regarding the Dolphin network (robustness of eigenvector centrality when 30% of the nodes are removed randomly) are shown in the figure. These results demonstrate that the eigenvector centrality of the removed nodes has a strong influence on the robustness. Especially if the removed nodes have a large share of the total eigenvector centrality (> 30%), the order of the nodes (based on the eigenvector centrality) has little to do with the order of the nodes in the original network.

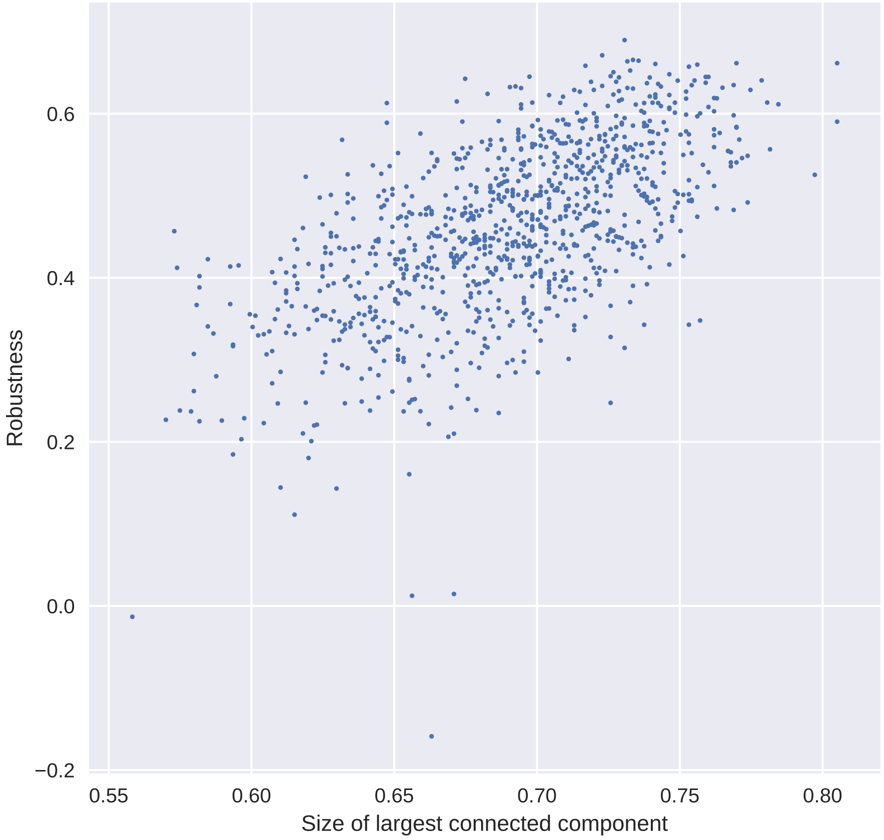


Figure 3.10: The results for the additional experiment regarding the Protein network (robustness of eigenvector centrality when 30% of the nodes are removed randomly) are presented in the figure. We observe that the lower the size of the largest connected component in the modified network, the lower the robustness (recall, the number of nodes removed is constant in this experiment).

total eigenvector centrality ($> 30\%$), the order of the nodes (based on the eigenvector centrality) has little to do with the order of the nodes in the original network.

There are two main reasons for this observation. First, the eigenvector values in the Dolphin network are unequally distributed among the nodes. Of the 62 nodes, 12 (approx. 19%) hold more than 50% of the eigenvector centrality (we call these nodes "high evc nodes").

Furthermore, removing a node with high eigenvector centrality has a high impact on all adjacent nodes in contrast, for example, to the degree centrality and PageRank, where the removal of a hub has a small effect on all adjacent nodes.[3]

The Dolphin network also exhibits another effect. With 62 nodes this network is relatively small. The number of "high evc nodes" removed by the error mechanism is binomial distributed ($n = 12$, $p = 0.3$). Due to the low number of trials in this distribution, the number of possible values of the random variable has a relatively broad range. For example, the events that only one "high evc node" is removed and six "high evc nodes" are removed have approximately the same probability of about 7%. However, the effects of these two events on the robustness differ drastically. In this case, the robustness of the centrality measure depends primarily on the outcome of the random experiment.

In the Protein network, the eigenvector centrality is also unevenly distributed. Here 41 of 1,458 nodes (2.8%) hold 50% of the eigenvector centrality. Furthermore, this network is relatively sparse (1,948 edges). Therefore, we will carry out the aforementioned experiment again with this network. This time we track the size of the largest connected component of the modified network in addition to the robustness of the eigenvector centrality. The results are shown in Figure 3.10. From this, we can see that the lower the size of the largest connected component, the lower the robustness (recall, the number of nodes removed is constant in this experiment).

3.4.3 Discussion of standard deviations

As discussed at the beginning of this section, the problem of guessing the true robustness based on some measured network and knowledge about the underlying error mechanism is ill-conditioned if the true robustness depends

[3] Removing a high evc node reduces the unnormalized evc-value of all neighbors by the high evc-value of the removed node. In the case of the degree centrality, the centrality of all neighbors is reduced by 1. Similarly, in the case of PageRank, the unnormalized centrality of each neighbor is reduced by the PageRank of the hub divided by its (high) degree.

Table 3.2: The mean sd of estimation (e) and the sd of true robustness (t) for the experiments with real-world networks are listed in the table. The values have been multiplied by 100 for better readability.

			Betw		Clos		Deg		Evc		Page	
			t	e	t	e	t	e	t	e	t	e
Network	Error	Level										
Dolphin	add edges	10%	4.61	3.66	5.24	3.71	1.10	1.18	5.21	3.62	1.65	1.61
		30%	4.45	3.90	4.79	3.76	2.47	2.69	6.32	4.21	2.65	2.89
	rm edges prop.	10%	5.41	4.73	4.15	4.32	1.16	1.18	4.71	10.24	1.24	1.60
		30%	5.64	6.66	4.91	7.17	2.63	3.31	24.25	24.94	2.73	3.47
	rm edges unif.	10%	4.77	4.76	3.95	4.27	1.11	1.14	3.66	5.47	1.61	1.71
		30%	5.84	6.18	5.20	6.43	2.66	2.73	16.59	16.37	2.92	3.07
	rm nodes	10%	6.28	7.23	6.88	7.81	2.07	2.15	20.94	21.56	2.21	2.51
		30%	8.71	10.39	9.93	14.07	4.14	5.42	35.64	30.70	4.17	5.70
Hamst	add edges	10%	0.69	0.57	0.33	0.28	0.22	0.22	0.28	0.27	0.32	0.29
		30%	0.67	0.55	0.40	0.37	0.37	0.38	0.34	0.33	0.45	0.42
	rm edges prop.	10%	0.44	0.43	0.42	0.38	0.11	0.11	0.33	0.34	0.17	0.18
		30%	0.44	0.62	0.53	0.47	0.20	0.23	0.58	0.58	0.24	0.29
	rm edges unif.	10%	0.67	0.64	0.36	0.36	0.14	0.14	0.30	0.31	0.31	0.33
		30%	0.75	0.69	0.46	0.47	0.25	0.28	0.41	0.46	0.42	0.50
	rm nodes	10%	1.15	1.18	2.42	2.47	0.52	0.54	1.88	1.94	0.66	0.69
		30%	1.49	1.78	3.30	3.77	0.89	1.10	2.65	3.30	0.89	1.17
Jazz	add edges	10%	1.90	1.43	1.36	0.94	0.35	0.38	0.52	0.50	0.55	0.47
		30%	1.87	1.48	1.23	0.82	0.63	0.74	0.87	0.86	0.77	0.79
	rm edges prop.	10%	1.36	1.10	1.07	1.02	0.31	0.36	0.50	0.54	0.36	0.43
		30%	1.41	1.85	1.22	1.30	0.65	1.04	1.04	2.56	0.73	1.15
	rm edges unif.	10%	1.69	1.73	0.97	0.97	0.35	0.35	0.42	0.44	0.49	0.56
		30%	2.18	2.31	1.13	1.14	0.69	0.79	0.85	0.99	0.89	1.11
	rm nodes	10%	2.80	2.88	4.18	4.25	0.90	0.95	2.47	2.76	0.70	0.81
		30%	3.62	4.32	4.99	5.77	1.85	2.31	5.51	8.02	1.57	2.26
Protein	add edges	10%	0.72	0.66	0.98	0.82	0.40	0.37	1.77	1.48	0.63	0.53
		30%	0.85	0.79	1.13	0.86	0.67	0.68	1.70	1.05	0.79	0.69
	rm edges prop.	10%	1.00	0.85	2.23	1.67	0.33	0.32	5.85	5.74	0.49	0.52
		30%	0.96	1.10	2.13	1.94	0.57	0.57	3.99	10.87	0.80	0.69
	rm edges unif.	10%	0.88	0.90	1.54	1.49	0.42	0.39	3.23	3.53	0.66	0.63
		30%	1.00	1.09	1.76	1.45	0.80	0.64	3.57	10.24	0.88	0.87
	rm nodes	10%	1.72	1.78	6.33	6.19	1.31	1.33	9.28	9.42	1.81	1.80
		30%	2.12	2.60	7.35	6.65	1.70	2.01	10.39	11.70	1.86	1.93

Table 3.3: The mean sd of estimation (e) and the sd of true robustness (t) for the experiments with random graphs are listed in the table. The values have been multiplied by 100 for better readability.

Network		Barabási				Erdős-Rényi			
		10%		30%		10%		30%	
		t	e	t	e	t	e	t	e
Centrality	Error								
Betweenness	add edges	3.23	2.17	4.07	2.92	2.48	1.96	3.76	3.23
	rm edges prop.	1.87	1.63	3.01	2.76	2.84	2.05	5.00	4.02
	rm edges unif.	2.22	1.99	3.49	3.13	2.75	1.95	4.37	3.41
	rm nodes	2.56	2.59	4.04	4.86	2.79	2.36	5.47	5.56
Closeness	add edges	3.10	2.27	4.07	3.03	2.31	1.79	3.72	3.10
	rm edges prop.	2.88	2.45	4.46	3.59	2.80	2.29	5.16	4.44
	rm edges unif.	3.01	2.40	4.68	3.40	2.68	2.14	4.47	3.68
	rm nodes	4.42	4.57	6.42	8.60	2.65	2.47	5.30	5.70
Degree	add edges	3.14	2.28	4.08	3.01	2.15	1.75	3.63	3.12
	rm edges prop.	2.76	1.84	3.67	2.75	2.43	1.91	4.92	4.19
	rm edges unif.	2.90	1.99	4.09	2.92	2.31	1.76	4.26	3.40
	rm nodes	4.20	4.04	5.34	5.88	2.27	1.96	4.94	5.18
Eigenvector	add edges	2.65	1.85	3.72	2.80	2.59	1.91	3.97	3.21
	rm edges prop.	3.14	2.18	4.14	3.33	2.83	2.24	5.19	4.77
	rm edges unif.	3.16	2.08	4.27	3.12	2.75	2.05	4.61	3.88
	rm nodes	8.20	7.66	9.14	9.29	3.36	3.08	5.71	6.55
Pagerank	add edges	3.55	2.33	4.24	2.99	2.31	1.76	3.67	3.03
	rm edges prop.	2.95	1.89	3.74	2.61	2.67	1.91	4.77	3.90
	rm edges unif.	3.18	2.09	4.15	2.86	2.53	1.79	4.24	3.25
	rm nodes	4.15	4.15	4.73	5.50	2.47	2.03	4.98	5.10

very much on the specific choice of added/removed vertices/edges — that is, the standard deviation of true robustness is large. As illustrated in Table 3.2 and Table 3.3 large values (sd true robustness > 0.15) are observed only for the combinations of eigenvector centrality and Dolphin networks. The same tables indicate that the mean sd of the estimated robustness is a very good indicator for a large sd of true robustness. In fact, the sd of the estimated robustness has large values (sd estimated robustness > 0.15) for exactly the same combinations. While the sd of true robustness is not accessible without knowledge of the hidden network (and hence cannot serve as a heuristic to decide whether to apply our method), the sd of the estimated robustness can be computed, given the measured network and the error mechanism. Therefore, we recommend using the suggested estimation for true robustness only if the sd of estimated robustness is small.

3.5 Summary and recommendations

Errors in network data are a ubiquitous problem in network analysis. Even though the reliability of centrality measures has been studied extensively in the literature, there is no method available that allows researchers to estimate the reliability of centrality measures in the case of imperfect measured data.

In the first part of this study, we proposed such a method for estimating the impact of measurement errors on the reliability of a centrality measure, given the measured network and assumptions about the type and intensity of the measurement error. To check the applicability of this method we have conducted a series of simulation experiments based on random graphs and real-world networks as well.

Regarding the robustness of random graphs and real-world networks, the results are conclusive with existing studies (Borgatti et al., 2006; Frantz et al., 2009). Moreover, we have observed that the (measurable) standard deviation of the estimated robustness is a good indicator for the (not measurable) standard deviation of the true robustness. Our results provide compelling evidence that our proposed estimation method is a suitable technique for the estimation of the robustness of centrality measures.

Based on these findings, we would like to offer the following recommendations for network studies, where centrality measures are analyzed and hypotheses about underlying measurement errors are available:

- Researchers should compute the estimated robustness and the corresponding standard deviation for all relevant centrality measures (Python code[4] can be downloaded and easily extended for specific centrality measures and error mechanisms).
- For those centrality measures where the computed standard deviation is large (> 0.15), the concept of true robustness is not appropriate as it is expected to depend very much on the specific added/removed edges/vertices. In all other cases, we recommend to report the estimated robustness in order to contribute to a better assessability of the conclusions of the study, which are based on centrality measures.
- If there are different centrality measures to choose from that are equally suitable for the study, the estimated robustness can be used as a further selection criterion.
- The estimated robustness can also be used as a basis for deciding whether to apply treatment procedures, such as imputation (Huisman, 2009; Wang et al., 2016; Krause et al., 2018).

Future studies should analyze the stability of the presented estimation approach with respect to the assumptions on the underlying error mechanism. If 20% missing edges (randomly) are assumed, while actually 15% of the vertices are missing proportional to vertex degree, what impact does this imprecise error-assumption have on the estimated true robustness? To extend the approach to directed networks, suitable error mechanisms have to be analyzed. While the current study focuses on robustness with respect to certain node level metrics, it might be interesting to see corresponding results for other metrics, such as the estimation of the most central node (Frantz and Carley, 2017).

Conflicts of interest

The authors have nothing to disclose.

[4] `https://github.com/crsqq/EstimatingCentralityRobustness`

3.6 References

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(October):509–512.

Bollobás, B. and Riordan, O. (2002). Mathematical results on scale-free random graphs. *Handbook of Graphs and Networks: From the Genome to the Internet*, pages 1–38.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182.

Borgatti, S. P., Carley, K. M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136.

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks*, 25(2):103–140.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307.

Csardi, G. and Nepusz, T. (2006). The igraph Software Package for Complex Network Research. *InterJournal, Complex Systems*, (1695):1–9.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Fischer, M., Parkins, K., Maizels, K., Sutherland, D. R., Allan, B. M., Coulson, G., and Di Stefano, J. (2018). Biotelemetry marches on: A cost-effective GPS device for monitoring terrestrial wildlife. *PLOS ONE*, 13(7):e0199617.

Frantz, T. L. and Carley, K. M. (2017). Reporting a network's most-central actor with a confidence level. *Computational and Mathematical Organization Theory*, 23(2):301–312.

Frantz, T. L., Cataldo, M., and Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Gleiser, P. M. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, 10(1):1–29.

Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251.

Kim, M. and Leskovec, J. (2011). The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SIAM International Conference on Data Mining*, pages 47–58.

Kim, P. J. and Jeong, H. (2007). Reliability of rank order in sampled networks. *European Physical Journal B*, 55(1):109–114.

Koschützki, D., Lehmann, K., and Peeters, L. (2005). Centrality Indices. In Brandes, U. and Erlebach, T., editors, *Network Analysis: Methodological Foundations*, pages 16–61. Springer Berlin Heidelberg.

Krause, R. W., Huisman, M., Steglich, C., and Sniiders, T. A. (2018). Missing Network Data A Comparison of Different Imputation Methods. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 159–163.

Kunegis, J. (2013). KONECT - The koblenz network collection. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web.*

Lee, J.-S. and Pfeffer, J. (2015). Robustness of Network Centrality Metrics in the Context of Digital Communication Data. *Proceedings of the 48th Hawaii International Conference on System Sciences.*

Leecaster, M., Toth, D. J. A., Pettey, W. B. P., Rainey, J. J., Gao, H., Uzicanin, A., and Samore, M. (2016). Estimates of social contact in a middle school based on self-report and wireless sensor data. *PLOS ONE*, 11(4).

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54(4):396–405.

Newman, M. E. J. (2018). Network structure from rich but noisy data. *Nature Physics*.

Niu, Q., Zeng, A., Fan, Y., and Di, Z. (2015). Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131.

Platig, J., Ott, E., and Girvan, M. (2013). Robustness of network measures to link errors. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6).

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, 107(3):1283–1298.

Silk, M. J., Jackson, A. L., Croft, D. P., Colhoun, K., and Bearhop, S. (2015). The consequences of unidentifiable individuals for the analysis of an animal social network. *Animal Behaviour*, 104:1–11.

Smith, J. A. and Moody, J. (2013). Structural Effects of Network Sampling Coverage I: Nodes Missing at Random. *Social Networks*, 35(4).

Smith, J. A., Moody, J., and Morgan, J. H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48:78–99.

Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. (2008). How correlated are network centrality measures? *Connections*, 28(1):16–26.

Wang, C., Butts, C. T., Hipp, J. R., Jose, R., and Lakon, C. M. (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to Add Health. *Social Networks*, 45:89–98.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409.

Žnidaršič, A., Ferligoj, A., and Doreian, P. (2018). Stability of centrality measures in valued networks regarding different actor non-response treatments and macro-network structures. *Network Science*, 6(01):1–33.

On the impact of network size and average degree on the robustness of centrality measures

Christoph Martin and Peter Niemeyer

This paper is forthcoming as

Martin, C., & Niemeyer, P. (in press). On the impact of network size and average degree on the robustness of centrality measures. Network Science.

Preliminary results of this study have been published as

Martin, C., & Niemeyer, P. (2020). The role of network size for the robustness of centrality measures. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), Complex networks and their applications VIII (pp. 40–51). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36687-2_4.

The layout has been revised.

Abstract

Measurement errors are omnipresent in network data. Most studies observe an erroneous network instead of the desired error-free network. Thus, subsequent analyses are based on erroneous data. It is well known that such errors can have a severe impact on the reliability of network metrics, especially on centrality measures: a central node in the observed network might be less central in the underlying, error-free network. The robustness is a common concept to measure these effects and is often defined as the correlation between the centrality values in the observed and the underlying network.

Studies have shown that the robustness primarily depends on the centrality measure, the type of error (e.g., missing edges or missing nodes), and the network topology (e.g., tree-like, core-periphery). Previous findings regarding the influence of network size on the robustness are, however, inconclusive.

In this paper, we present empirical evidence and analytical arguments indicating that there exist arbitrary large robust networks as well as arbitrary large non-robust networks, and that the average degree is more suitable to explain the robustness than the network size. We demonstrate that, in the vast majority of cases, networks with a higher average degree are more robust.

For the case of the degree centrality and Erdős-Rényi (ER) graphs, we present explicit formulas for the computation of the robustness, mainly based on the joint distribution of node degrees and degree-changes which allow us to analyze the robustness for ER graphs with a constant average degree or increasing average degree.

4.1 Introduction

Networks are used to model various real-world phenomenons. Typical use cases are (online) social networks, web graphs, protein-protein interaction networks, infrastructure networks, and many more (Newman, 2003). Networks are, however, sensitive to errors in the data underlying the network. The reasons for such errors are manifold. When collecting data for a social network, for example, actors may be missing on the day of the survey or the number for the nomination of possible friends may be limited by the survey questionnaire (Wang et al., 2016). The collection of protein-protein interaction data is, depending on the method used, inevitably associated with uncertainty, which is consequently also part of the network constructed from this data (De Las Rivas and Fontanillo, 2010). When creating co-authorship or citation networks, authors or papers can be included multiple times or not at all, for example, due to incorrect spelling (Erman and Todorovski, 2015; Schulz, 2016). All these errors affect the outcome of network analysis methods and thus, the conclusions that depend on these methods (Marsden, 1990; Kossinets, 2006).

In the field of network analysis, centrality measures are commonly used to analyze the position of nodes in a network. These measures map a real number to every node in the network which can be used to rank the nodes. It is well known that errors in the network data can have a severe impact on the reliability of centrality measures. For example, the best-ranked actor might actually not be the best in the erroneous network. We measure this impact using the concept of robustness of centrality measures, which is the correlation between the centrality values in the error-free and the erroneous network.[1] Previous studies have used the Pearson correlation to measure the robustness (Bolland, 1988; Costenbader and Valente, 2003; Borgatti et al., 2006). Like most recent studies, we use a rank correlation (Kim and Jeong, 2007; Wang et al., 2012; Holzmann et al., 2019; Martin and Niemeyer, 2019). The effects of errors on the robustness of centrality measures depend on several variables, e.g., the type of centrality measure, the type and extent of the error, the network topology (e.g., tree-like, core-periphery), and how we measure the robustness (Frantz et al., 2009; Smith and Moody, 2013). Few studies have addressed the issue of robustness of centrality measures from an analytical perspective. Ghoshal and Barabási (2011) investigated the existence of super-stable nodes w.r.t. degree and PageRank. Platig et al. (2013) investigated the joint occurrence of missing and false links. Tsugawa and Ohsaki (2015) adapted this approach to measuring the robustness focusing on most central nodes.

In this article we investigate the robustness of empirical networks that vary in size and structure. Existing studies are inconclusive about the relationship between network size and robustness. No relationship between size and robustness is noticeable in the empirical part of Niu et al. (2015). In Costenbader and Valente (2003) and Borgatti et al. (2006), the authors observed that larger network size could be related to both, higher and lower robustness, depending on the network structure. In Wang et al. (2012), the smaller network is usually more robust than the larger one. In contrast, Smith and Moody (2013) noticed that larger networks are frequently more

[1] Although the two topics sound similar, studies on the robustness of networks have a different focus than studies about the robustness of centrality measures. The subject of studies on the robustness of networks is the question, how the functionality of a network as a whole is influenced by, for example, the removal of nodes (see Albert et al. Albert et al. (2000) or Callaway et al. Callaway et al. (2000)). If the term robustness is used in this work without further specification, then the term always refers to the robustness of centrality measures.

robust. For a comprehensive review of the existing work on the robustness of centrality measures, we refer to Smith et al. (2017). To the best of our knowledge, however, there exist no studies that explicitly analyze the relationship between average degree and robustness and previous studies have mostly been concerned about smaller networks (approx. less than 1000 nodes). This raises the question whether the concept of robustness of centrality measures is at all relevant in the context of larger networks. In contrast to existing studies, we specifically investigate the relationship between the size as well as the average degree and the robustness of centrality measures. In addition, we provide analytical results for this relationship based on the interpretation of the robustness as a probability.

To examine these contrary observations regarding the network size and the robustness in greater detail, we proceed as follows: First, we investigate the robustness of the degree, the eigenvector centrality and the PageRank in 24 empirical networks coming from diverse domains (Section 4.3). We hardly observe any association between network size and robustness, but a high correlation between average degree and robustness. This observation holds for all considered centrality measures and error types that involve removing nodes or edges. We further investigate the effect of network size on the robustness using the Erdős-Rényi (ER) and the Barabási-Albert (BA) random graph model (Section 4.4). For both models, we observe that robustness is independent of network size if the average degree remains constant. If the average degree increases, then centrality measures in BA graphs become more robust, in contrast to ER graphs. We also make these observations in our experiments with the configuration model where random graphs are generated based on the degree distributions of the empirical networks.

In Section 4.5, we introduce an analytical approach for the robustness. We derive explicit expressions for the robustness of the degree centrality in ER graphs. We use these expressions to prove that for ER networks of different size but with the same average degree, the robustness of the degree centrality remains stable. As a consequence, there exist robust and non-robust networks of varying sizes, at least w.r.t. the degree centrality. We also provide arguments, based on the variance of the degree and the variance of the degree change, as to why the robustness increases or decreases with increasing average degree, depending on the type of network.

4.2 Methods

A graph $G(V, E)$ consists of a node set V and an edge set E, $E \subseteq \binom{V(G)}{2}$. We denote the number of nodes in G by N and the number of edges by M. All graphs considered in this paper are undirected, unweighted, and simple, i.e., they do not contain loops nor multiple edges. The adjacency matrix of a graph is denoted by A, where $A_{i,j} = 1$ if there is an edge between node v_i and v_j (i.e., $\{v_i, v_j\} \in E(G)$) and 0 otherwise. The neighborhood of a node u is $N(u) = \{v : \{u, v\} \in E(G)\}$. It is the set of nodes that are connected to u. The degree is the number of connections that a node has, $\text{degree}(u) = |N(u)|$. The degree of an edge is the sum of the degree values of the end nodes, $\text{degree}(\{u, v\}) = \text{degree}(u) + \text{degree}(v)$. We denote the degree sequence of a graph G by $\text{ds}(G)$.

4.2.1 Centrality measures

Centrality measures map a real number to every node in the graph and thus imply a ranking on the nodes. These measures solely depend on the structure of the graph and not on, for example, additional information about the nodes (Koschützki et al., 2005). By $c_G(u)$ we denote the centrality value for a specific node u in a graph G w.r.t. a centrality measure c. If the context permits, we do not explicitly mention the graph. The vector of centrality values for all nodes in G is defined as $c(G) = (c_G(v_1), \ldots, c_G(v_N))$.

The most straightforward centrality measure is the degree centrality which was already discussed above, $\text{degree}(u) = |N(u)|$. The eigenvector centrality and the PageRank are both feedback measures. They are defined recursively, the centrality value of a node depends on the centrality values of its neighbors. If G is connected, then the eigenvector centrality of a node u defined by the unique solution to $\text{evc}(u) = \frac{1}{\lambda} \sum_{v \in N(u)} \text{evc}(v)$, where λ is the largest eigenvalue of A (Bonacich, 1987). The unique solution to $\text{PageRank}(u) = d \sum_{v \in N(u)} \frac{\text{PageRank}(v)}{\text{degree}(v)} + (1 - d)$ defines the PageRank; with d as damping factor (in our case 0.85) (Brin and Page, 1998). Originally introduced for directed graphs, this concept is also applicable for undirected graphs. One of the main differences between these two measures is that, in case of the eigenvector centrality, all neighbors of a node receive the total centrality value of this node. In contrast, in case of the PageRank, neighbors of a node only receive a faction of the nodes centrality value, which depends on the total amount of neighbors of this node.

4.2.2 Error mechanisms

When collecting data, external factors and the selection of the sampling method can lead to inaccurate network data. We use four procedures to model the impact of errors on information about nodes and edges. We call these procedures error mechanisms. They model an error that affects the nodes or edges of a graph. Their inputs are a graph G and a parameter α which controls the intensity of the error. The procedure returns one graph from the set of all possible erroneous versions of the graph G. In this study, we use the following error mechanisms:

add edges (e+): αN edges are added to the graph. If αN is not an integer, then we use the smallest integer that is greater than or equal to αN ($\lceil \alpha N \rceil$). The new edges are chosen uniformly at random from the $\binom{N}{2} - M$ possible edges.

remove edges unif. (e-): $\lceil \alpha N \rceil$ edges are removed from the graph. The edges are chosen uniformly at random from $E(G)$.

remove edges degree (e-(p)) also removes $\lceil \alpha N \rceil$ edges. The edges are, however, chosen with probability proportional to the edge degree (i.e., $P(\{u, v\}) = \frac{\text{degree}(\{u,v\})}{\sum_{e \in E(G)} \text{degree}(e)}$).

remove nodes (n-): $\lceil \alpha N \rceil$ nodes are removed from the graph. The nodes are chosen uniformly at random from $V(G)$.

For a more detailed discussion of error mechanisms as random graphs, see Martin and Niemeyer (2019).

4.2.3 Robustness of centrality measures

To quantify the impact of errors in data collection on centrality measures, we use the concept of robustness, which measures how the ranking of nodes, induced by the centrality measure, changes. For two graphs, G and H, a centrality measure c, and a correlation $corr$, we denote the robustness by $r_{corr,c}(G, H)$, where H is the erroneous graph (a "modified" version of G, i.e., H is on the same node set as G or on a subset of that node set). If G and H are not on the same node set then, similar to Wang et al. (2012), we only consider nodes that exist in both graphs. Since the robustness is defined as a correlation, the values for the robustness of a centrality measure are in $[-1, 1]$.

In the same way as Kim and Jeong (2007) and Holzmann et al. (2019), we use Kendall's τ ("tau-b") rank correlation coefficient (Kendall, 1945) to measure the robustness of centrality measures. In this case, the robustness is defined as follows:

$$r_{\tau,c}(G,H) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_t)(n_c + n_d + n_{t'})}} \tag{4.1}$$

The number of concordant pairs and discordant pairs w.r.t. $c(G)$ and $c(H)$ are n_c and n_d, respectively. A pair of nodes u, v is concordant if $(c_G(v) - c_G(u)) \cdot (c_H(v) - c_H(u)) > 0$ and discordant if $(c_G(v) - c_G(u)) \cdot (c_H(v) - c_H(u)) < 0$. Ties in $c(G)$ (i.e., $c_G(v) - c_G(u) = 0$) are denote by n_t and ties in $c(H)$ (i.e., $c_H(v) - c_H(u) = 0$) are denoted by $n_{t'}$.

Goodman and Kruskal's rank correlation coefficient γ (Goodman and Kruskal, 1954) is closely related to Kendall's τ and the robustness using this measure is defined as follows:

$$r_{\gamma,c}(G,H), = \frac{n_c - n_d}{n_c + n_d}. \tag{4.2}$$

If all pairs are either concordant or discordant w.r.t. the centrality c (i.e., there are no ties), then both measures are equal. Since $(r_{\gamma,c}(G,H) + 1)/2$ can be interpreted as the probability that two randomly chosen nodes have the same order in c_G and c_H w.r.t. c, this measure is more accessible for an analytical perspective and we use it in Section 4.5. For the empirical part, we use Kendall's τ to provide comparability to existing studies.

4.2.4 Random graph models

In this paper, we use the Erdős-Rényi model, the, Barabási-Albert model, and the configuration model.

The Erdős-Rényi random graph model, introduced by Erdős and Rényi (1959), has two parameters: the number of nodes n and the edge probability p. Since all node pairs are connected with the same probability (p), the degree distribution of the nodes in this model follows a binomial distribution.

In contrast, the Barabási-Albert model is based on the idea of preferential attachment. Consequently, the probability that a new node will connect to an existing node is proportional to the degree of the existing node. This model also has two parameters. In addition to the number of nodes n, the parameter m specifies the number of connections that a new node makes to existing nodes. Due to this generation process, the degree distribution of the nodes in graphs generated by this model follows a power-law distribution (Barabási and Albert, 1999).

The configuration model is a method to create random graphs based on existing degree sequences (Newman et al., 2001). In this model there are no other parameters apart from the degree sequence. First, an empty graph on n nodes is created (n is given by the degree sequence). Next, every node u receives degree(u) stubs (here, degree(u) is the desired degree of node u). Finally, pairs of stubs are chosen and connected with equal probability. This procedure might results in graphs with multiple edges and loops. For our study, however, we ignore those (i.e., we work with the simple versions of these graphs).[2]

4.3 Experiments with empirical networks

In this section, we investigate, for empirical networks, the relationship between the robustness of centrality measures on the one side and the corresponding network size and average degree on the other. We consider the following centrality measures: degree, eigenvector centrality and PageRank. Both, the eigenvector centrality and the PageRank are feedback measures and fast to calculate (Koschützki et al., 2005). However, they have rarely been considered simultaneously in previous studies. Since PageRank can be very stable in scale-free networks (Ghoshal and Barabási, 2011), a comparison with the eigenvector centrality is therefore interesting, we will see that both measures behave differently with regard to their robustness. For the calculation of the betweenness and the closeness for all nodes in a graph, the all-pairs shortest path problem has to be solved. The running time for that is at least quadratic (Brandes, 2001). Since the centrality values in the simulation part of this study (described in the following section) have to be recalculated numerous times, these measures are not considered.

4.3.1 Experimental setup and data

For our empirical study, we use all the undirected and unweighted networks available through the Koblenz Network Collection (Kunegis, 2013) at the beginning of 2019. Hence the networks used in this study can be seen as a random sample of networks that stem from different domains and therefore differ from each other in structure and size. The 24 real-world networks and descriptive statistics for them are listed in Table 4.1. As part of the data pre-processing, we have removed any existing loops. If a network consists of several components, we only consider the largest connected component and hence all networks are connected.

[2] We use NetworkX (version 2.2, Hagberg et al. (2008)) to generate random graphs and calculate centrality measures.

Table 4.1: Descriptive statistics about the largest connected components of the networks used in the study. The average degree is abbreviated as $\langle d \rangle$.

Name	Nodes	Edges	$\langle d \rangle$	Density	Transitivity	Source
zachary	34	78	4.6	1.4e-01	2.6e-01	Zachary (1977)
dolphins	62	159	5.1	8.4e-02	3.1e-01	Lusseau et al. (2003)
pdzbase	161	209	2.6	1.6e-02	2.9e-03	Beuming et al. (2005)
jazz	198	2,742	27.7	1.4e-01	5.2e-01	Gleiser and Danon (2003)
vidal	2,783	6,007	4.3	1.6e-03	3.5e-02	Rual et al. (2005)
facebook	4,039	88,234	43.7	1.1e-02	5.2e-01	Leskovec and Mcauley (2012)
CA-GrQc	4,158	13,422	6.5	1.6e-03	6.3e-01	Leskovec et al. (2007)
powergrid	4,941	6,594	2.7	5.4e-04	1.0e-01	Watts and Strogatz (1998)
reactome	5,973	145,778	48.8	8.2e-03	6.1e-01	Joshi-Tope et al. (2005)
CA-HepTh	8,638	24,806	5.7	6.6e-04	2.8e-01	Leskovec et al. (2007)
pgp	10,680	24,316	4.6	4.3e-04	3.8e-01	Boguñá et al. (2004)
CA-HepPh	11,204	117,619	21.0	1.9e-03	6.6e-01	Leskovec et al. (2007)
CA-AstroPh	17,903	196,972	22.0	1.2e-03	3.2e-01	Leskovec et al. (2007)
CA-CondMat	21,363	91,286	8.5	4.0e-04	2.6e-01	Leskovec et al. (2007)
deezer-RO	41,773	125,826	6.0	1.4e-04	7.5e-02	Rozemberczki et al. (2019)
deezer-HU	47,538	222,887	9.4	2.0e-04	9.3e-02	Rozemberczki et al. (2019)
deezer-HR	54,573	498,202	18.3	3.3e-04	1.1e-01	Rozemberczki et al. (2019)
brightkite	56,739	212,945	7.5	1.3e-04	1.1e-01	Cho et al. (2011)
livemocha	104,103	2,193,083	42.1	4.0e-04	1.4e-02	Zafarani and Liu (2009)
petster-cat	148,826	5,447,464	73.2	4.9e-04	1.1e-02	Dünker and Kunegis (2015)
douban	154,908	327,162	4.2	2.7e-05	1.0e-02	Zafarani and Liu (2009)
gowalla	196,591	950,327	9.7	4.9e-05	2.3e-02	Cho et al. (2011)
dblp	317,080	1,049,866	6.6	2.1e-05	3.1e-01	Yang and Leskovec (2012)
petster-dog	426,485	8,543,321	40.1	9.4e-05	1.4e-02	Dünker and Kunegis (2015)

To analyze the effects of different errors on the robustness of centrality measures in the empirical networks, we use a simulation-based experimental procedure. An iteration of the experiment is performed as follows: Starting from a network G (one of the 24 networks listed in Table 4.1) we apply the error mechanism with the intensity α. The resulting modified network is called H. Finally, we calculate the robustness of the centrality measure c: $r_{\tau,c}(G, H)$ (as defined in Section 4.2.3). We repeat this procedure 100 times and compute the mean and the standard deviation of the robustness for each network for all combinations of centrality measure (degree, eigenvector centrality, PageRank), error mechanism (add edges, remove edges uniform, remove edges proportional to the edge degree, and remove nodes), and error level ($\alpha \in \{0.1, 0.2, \dots, 0.5\}$).

4.3.2 Observations for empirical networks

We start with the results aggregated across all networks. Similar to previous studies in this area (as discussed in Section 4.1), we observe that the robustness declines with an increasing level of error. Therefore we will subsequently focus on an error level of $\alpha = 0.2$ since the results for the other error levels yield the same conclusions and the impact of the error level is not our main objective.

When looking at the average across all networks (Table 4.2), degree centrality is always the most robust. For the removal error mechanisms the PageRank is more robust than the eigenvector centrality. In the case of additional edges, the opposite effect can be observed. Regarding the standard deviation, the ranking is constant across all error types, degree centrality varies least, followed by PageRank. The robustness of the eigenvector centrality fluctuates the most, sometimes the standard deviation is two to three times as large as for the first mentioned measures. With regard to the effect of the type of measurement error on robustness, degree centrality and PageRank behave similarly. The absence of edges proportional to the edge degree has the weakest effect, spurious edges the strongest. For eigenvector centrality, on the other hand, the first error type has the strongest influence on the robustness.

As we look at the relationship between robustness and global network measures, we notice that there are both: large networks that are very sensitive to errors (e.g., douban) and small networks that are very robust (e.g., Jazz). The mean values of the robustness for every network are listed in Table 4.3.

In the following, we discuss the relationship between global network measures and robustness in more detail. Table 4.4 lists the rank correlation

Table 4.2: Mean and standard deviation of the robustness of centrality measures in empirical networks, aggregated over all networks in Table 4.1.

Error mechanism	e+		e-		e-(p)		n-	
	mean	sd	mean	sd	mean	sd	mean	sd
Centrality								
Degree	0.85	0.06	0.89	0.05	0.91	0.05	0.90	0.05
Eigenvector	0.75	0.18	0.81	0.11	0.73	0.15	0.79	0.15
PageRank	0.73	0.07	0.82	0.07	0.86	0.07	0.83	0.07

Table 4.3: Mean values of the robustness of centrality measures in empirical networks.

Centrality	Degree				Eigenvector				PageRank			
Error mechanism	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-
Network												
CA-AstroPh	0.89	0.94	0.97	0.94	0.83	0.92	0.89	0.90	0.77	0.88	0.94	0.89
CA-CondMat	0.86	0.90	0.93	0.90	0.79	0.86	0.81	0.82	0.74	0.81	0.88	0.83
CA-GrQc	0.84	0.88	0.94	0.89	0.60	0.77	0.62	0.71	0.73	0.79	0.88	0.81
CA-HepPh	0.82	0.93	0.98	0.93	0.63	0.87	0.92	0.85	0.70	0.86	0.96	0.86
CA-HepTh	0.85	0.88	0.92	0.88	0.63	0.76	0.71	0.70	0.76	0.80	0.86	0.81
brightkite	0.80	0.89	0.95	0.89	0.84	0.80	0.71	0.79	0.68	0.81	0.91	0.80
dblp	0.84	0.88	0.93	0.88	0.71	0.79	0.54	0.74	0.71	0.78	0.84	0.80
deezer_HR	0.91	0.93	0.95	0.93	0.84	0.92	0.89	0.90	0.83	0.88	0.92	0.88
deezer_HU	0.89	0.90	0.91	0.90	0.81	0.85	0.80	0.83	0.80	0.82	0.85	0.83
deezer_RO	0.87	0.88	0.90	0.88	0.77	0.81	0.65	0.79	0.76	0.78	0.84	0.78
dolphins	0.86	0.85	0.85	0.86	0.69	0.79	0.74	0.68	0.79	0.79	0.79	0.80
douban	0.66	0.78	0.82	0.78	0.93	0.71	0.71	0.71	0.50	0.67	0.76	0.71
facebook	0.93	0.96	0.97	0.96	0.47	0.91	0.85	0.88	0.73	0.90	0.92	0.91
gowalla	0.82	0.91	0.94	0.91	0.86	0.81	0.58	0.81	0.68	0.82	0.91	0.83
jazz	0.92	0.93	0.92	0.94	0.91	0.92	0.91	0.91	0.87	0.89	0.90	0.92
livemocha	0.88	0.95	0.96	0.95	0.95	0.90	0.86	0.90	0.78	0.91	0.92	0.92
pdzbase	0.81	0.82	0.84	0.83	0.76	0.61	0.54	0.60	0.68	0.70	0.73	0.71
petster_cat	0.88	0.94	0.85	0.96	0.98	0.87	0.60	0.95	0.75	0.91	0.80	0.91
petster_dog	0.87	0.94	0.90	0.94	0.97	0.92	0.64	0.92	0.71	0.90	0.90	0.91
pgp	0.80	0.87	0.94	0.87	0.67	0.73	0.69	0.69	0.69	0.77	0.88	0.78
powergrid	0.80	0.79	0.80	0.79	0.13	0.59	0.58	0.50	0.70	0.68	0.70	0.69
reactome	0.89	0.96	0.97	0.96	0.59	0.93	0.91	0.91	0.72	0.90	0.94	0.91
vidal	0.83	0.87	0.90	0.87	0.87	0.76	0.63	0.75	0.72	0.78	0.84	0.78
zachary	0.80	0.83	0.85	0.85	0.76	0.67	0.64	0.67	0.72	0.76	0.79	0.80

between the average robustness and the respective values for the global network measures. For all removal error types, the robustness tends to be higher with increasing average degree. We observe almost perfect correlation for cases where edges or nodes are missing uniformly at random and still high correlation values when edges are missing proportional. For the degree centrality, the correlation is also high for the case of spurious edges. For PageRank and eigenvector centrality, this is, however, not the case. While for the transitivity a moderate correlation with the robustness can still be observed, the number of nodes as well as the density are, in most cases, basically uncorrelated with the robustness. This observation may comes rather unexpected since growing networks often show "densification", which means the average degree grows with the number of nodes (Leskovec et al., 2007).

Figure 4.1 shows the behavior of robustness for three groups in each panel in exemplary fashion. The robustness of the eigenvector centrality in case of missing edges (uniformly) is depicted in the first panel. There is a recognizable association, but in this case the variance is higher than in most other cases. The same effect can be observed with the eigenvector centrality also in connection with missing nodes. The middle panel shows the observation for PageRank and add edges. This behavior is typical for all centrality measures under the influence of additional edges, there is no obvious pattern. The lower panel shows the combination PageRank and missing edges uniform. In this case, the relationship between average degree and robustness is most prominent. Robustness is higher when the average degree is also higher. The variance of robustness is also low. This behavior occurs for PageRank and Degree for all cases of missing edges (uniform and proportional) and missing nodes.

Table 4.4: Empirical networks: rank correlation between global measures and the average robustness.

Centrality	Degree				Eigenvector				PageRank			
Error mechanism	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-
Avg. degree	0.77	0.97	0.63	0.98	0.27	0.92	0.52	0.93	0.43	0.96	0.72	0.95
Density	0.18	0.01	0.01	0.03	-0.38	0.02	0.30	-0.11	0.26	0.07	0.00	0.15
Network size	0.02	0.31	0.16	0.29	0.52	0.26	-0.14	0.43	-0.18	0.26	0.22	0.18
Transitivity	0.26	0.23	0.58	0.23	-0.63	0.27	0.50	0.04	0.22	0.16	0.49	0.23

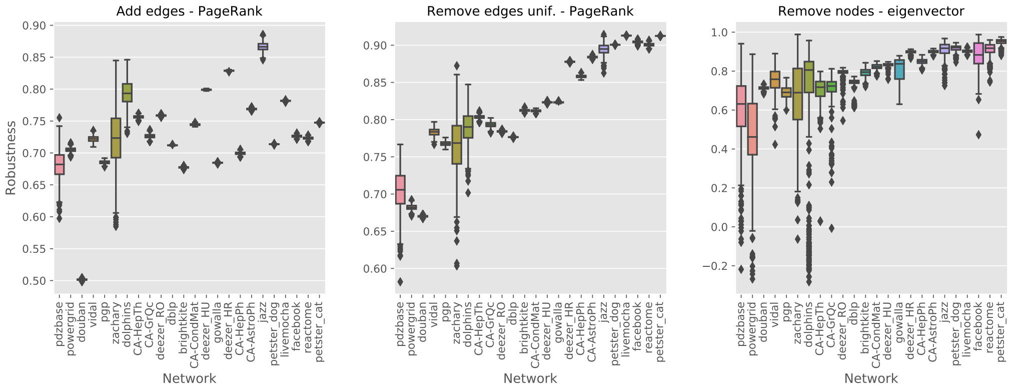


Figure 4.1: Illustrative examples for the three different behaviors of the robustness in empirical networks. The networks are sorted their average degree (ascending). The median robustness is indicated in each box, whiskers are 1.5 times the interquartile range.

4.4 Experiments with random graphs

In Section 4.3, we examined the robustness of 24 empirical networks from different domains. We observed that there exist small and robust as well as large and sensitive networks regarding the reliability of centrality measures. In addition, we have analyzed the relationship between the robustness of centrality measures in these networks with different global network measures. We observed that there is little association between network size and robustness. We found, however, that in many cases the higher the average degree of the network, the higher the robustness. To study this effect in more detail, we conduct further experiments in this section. We use different random graph models to control the average degree and to measure the effects of its change on robustness. For this purpose we choose two different perspectives. In Section 4.4.1, we keep the average degree constant and increase the size of the network. In Section 4.4.2, we control the average degree while keeping the network size constant.

4.4.1 Experiments with constant average degree

In this section, we use the ER model and the BA model to investigate the behavior of the robustness when the average degree is fixed while the network size increases. The experimental setup is similar to that of Section 4.3.1. Instead of using empirical networks, however, we generate ER and BA graphs with an average degree of 10 and a network size

$n \in (100, 500, 1000, 1500, \ldots, 10000)$, which we call G. Then we apply the error mechanism with the intensity α to G which results in the erroneous network H and calculate the robustness of the centrality measure c: $r_{\tau,c}(G, H)$. We repeat this procedure 100 times for the two random graph models and the varying values for the network size for all combinations of centrality measure (degree, eigenvector centrality, PageRank), error mechanism (add edges, remove edges uniform, remove edges proportional to the edge degree, and remove nodes), and error level ($\alpha \in \{0.1, 0.2, \ldots, 0.5\}$).

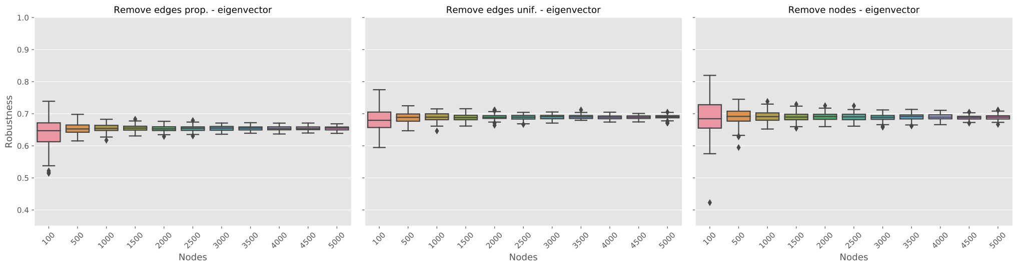


Figure 4.2: Results for the robustness of centrality measures in BA graphs. Here, the network size increases while the average degree remains constant, the error level is 0.2. For the network size values up to 5000 are shown for better readability; for larger values, hardly any changes occur.

The results for the ER graphs are very homogeneous, for all centrality measures and error mechanisms we observe the same behavior: robustness does not change with increasing network size. However, the variance decreases with increasing network size. We limit the discussion to an error level of $\alpha = 0.2$ since the results for the other error levels are conclusive with these results. Figure 4.3 is symptomatic for all other cases. It shows the robustness (ordinate) of the eigenvector centrality when nodes are missing, the network size is shown on the abscissa. It is noticeable here that the variance decreases sharply with the first increases in network size. Above a size of approx. 2000, the change is hardly visible.

For the BA graphs we observe, with two exceptions, the same behavior as for the ER graphs. Figure 4.2 shows the three different characteristics (all results in this figure are for the eigenvector centrality). The middle panel in Figure 4.2 represents the robustness behavior in BA graphs in almost all cases. The robustness is independent of the network size, only the variance decreases with increasing size, whereas the variance is relatively small already. The outer panels show the two exceptional cases. The absence of edges proportional to the degree of edge (left panel) reduces the robustness of the eigenvector centrality with increasing network size. If nodes are missing (right panel), the robustness is, as in most other cases, independent of the network size, but the variance is much larger and declines hardly with increasing size.

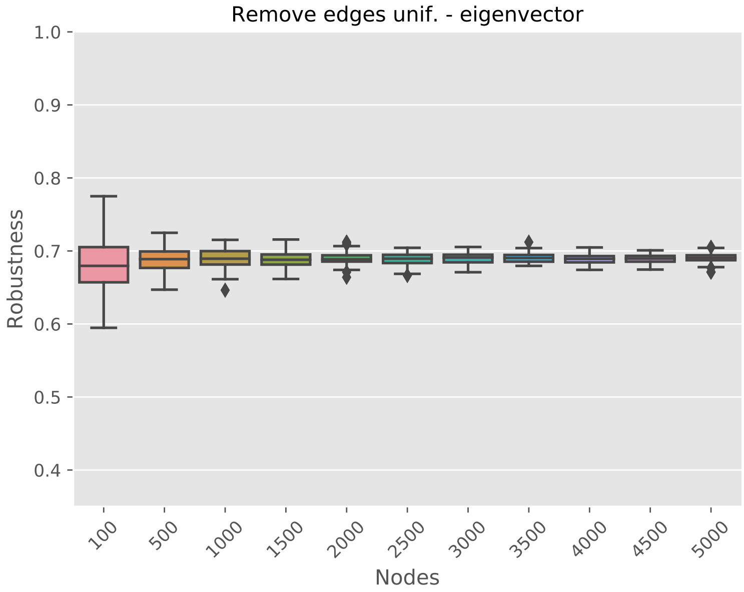


Figure 4.3: The behavior of the robustness in ER graphs with increasing network size (abscissa) for fixed average degree is shown in the figure above. The error level is 0.2.

4.4.2 Experiments with increasing average degree

In this part, we analyze the impact of changes to the average degree on the robustness of centrality measures. We conduct two types of experiments, based on ER and BA graphs, and based on the configuration model. The procedure for the first experiments is similar to the experimental setup described in Section 4.4.1. The difference between these experiments is the generation of the random graphs. For the experiments in this section we fix the network size at $n = 1000$ and select the parameters p and m in such a way that we obtain networks with an average degree between 4 and 100.

With the second type of experiments we examine the effects of changes in the average degree on the robustness of centrality measures in a more realistic setting. We use the degree sequences of the empirical networks from Section 4.3. We “scale” the degree sequences and generate more dense versions of the underlying networks using the configuration model. We use these generated networks as input graph G to analyze the robustness of centrality measures in these types of networks, the remaining procedure is analogous to the experiments already described in this section.

For the results concerning the ER graphs, the pattern consists of two parts, independent of the type of error. The robustness of the eigenvector centrality and the PageRank is, except for the smallest initial increases, constant and thus independent of the increase of the average degree. With degree centrality, on the other hand, robustness decreases with increasing average degree. The decreases occur especially during the initial increases of the average degree (approx. the range between 4 and 25), here the robustness decreases by 0.1. These observations can also be found in the rank correlation between average degree and the robustness (Table 4.5). While degree centrality here always shows strongly negative correlation, in most other cases no or weakly negative correlation can be observed.

Table 4.5: The rank correlations between the average degree and the robustness are listed for the cases of BA and ER graphs under the influence of different error mechanisms with an error level of 0.2.

	BA graphs				ER graphs			
Error mechanism	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-
Centrality								
Degree	0.91	0.92	0.90	0.87	-0.66	-0.64	-0.64	-0.51
Eigenvector	0.82	0.93	0.95	0.69	-0.07	0.02	0.04	0.15
PageRank	0.93	0.94	0.93	0.91	-0.33	-0.25	-0.41	-0.02

The results for the BA graphs show a consistent pattern. In all cases, regardless of centrality measure and error type, a higher average degree is accompanied by a higher robustness. There is a very high, positive rank correlation between the average degree and the associated robustness (see Table 4.5). The increases in robustness associated with the increase in the average degree are particularly strong for initial increases, further increases still have a positive effect on robustness, but the effect of this effect diminishes. The only exception to this is eigenvector centrality, which resembles a linear relationship. The variance is slightly higher for the error type missing nodes than for the other error types. In the case of eigenvector centrality the variance is much higher in this case.

In the previous section, we have observed that for BA graphs, a higher average degree is associated with higher robustness. Although BA graphs have a skewed degree distribution, a property that many empirical networks also have, these networks are nevertheless otherwise rather artificial. Therefore, we now make the previous experiment a little more realistic. For this purpose, we use the degree sequences of empirical networks and manipulate them to increase the average degree and generate networks in order to analyze the robustness of centrality measures in these networks.

The experiment's basic design is similar to that of the experiments in the previous section. The difference lies in the way the networks are created. We take the degree sequences of the empirical networks (Table 4.1) and create several "scaled" versions of them and generate networks based on these degree sequences using the configuration model. To scale a degree sequence of a network G we take a factor s and multiply each entry $\mathrm{ds}(G)_i$ of the degree sequence $\mathrm{ds}(G)$ by this factor. If $\mathrm{ds}(G)_i \cdot s$ is not natural number, we take the integer part of it and add 1 with the probability of the fractional part. For example, if the original degree is 9 and $s = 1.25$, then the scaled degree is $11 + Bern(0.25)$ where $Bern$ is the Bernoulli distribution. We scale the degree with factors between 1 and 5 ($s \in (1.0, 1.25, 1.5, \ldots, 5)$ and repeat the whole procedure 100 times for every combination. We also calculate the robustness of the underlying networks (the networks from which we obtain the degree sequences).

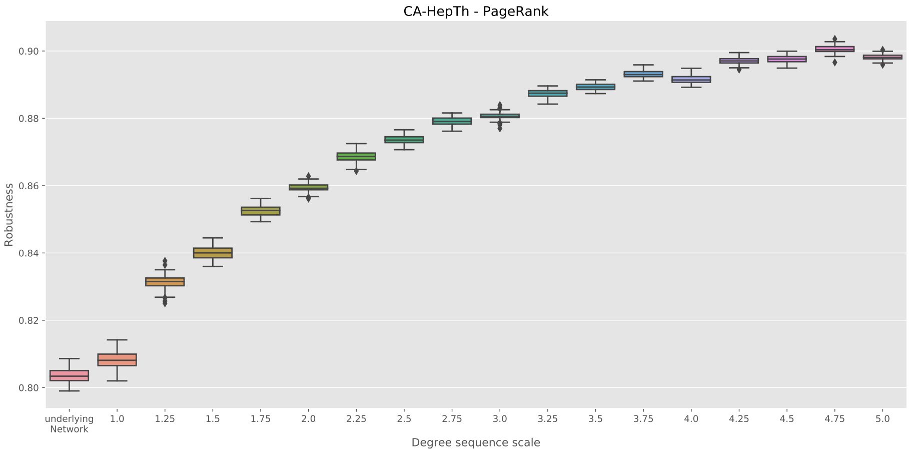


Figure 4.4: Configuration model results for one network (CA-HepTh). The scaling factor is listed on the abscissa. Additionally, the first entry is the robustness of the underlying network.

First, we compare the robustness of centrality measures in "unscaled" random graphs (i.e., $s = 1$) with the robustness of the corresponding empirical network. In the case of the degree centrality, the robustness values of both networks are, in the vast majority of cases and regardless of the type of error, similar. This is, however, not the case the eigenvector centrality and the PageRank. For these measures, the robustness of the random graph and the underlying network are only similar in about 50% of the cases. We observe no clear pattern which could explain this behavior.

Table 4.6: Configuration model: rank correlations between the scaling factor of the degree sequence and the average robustness of centrality measures w.r.t. random graphs with these degree sequences.

Centrality	Degree				Eigenvector				PageRank			
Error	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-	e+	e-	e-(p)	n-
Network												
zachary	0.60	0.75	0.46	0.81	0.81	0.90	0.96	0.76	0.84	0.87	0.84	0.91
dolphins	0.88	0.79	0.82	0.94	0.96	0.97	0.96	0.96	0.96	0.91	0.90	0.94
jazz	0.66	0.38	0.26	0.96	0.72	0.66	0.50	0.96	0.82	0.78	0.56	0.96
pdzbase	0.63	0.34	-0.31	0.32	0.62	0.96	0.96	0.75	0.66	0.63	0.41	0.65
vidal	0.85	0.79	0.18	0.76	1.00	1.00	1.00	0.97	0.91	0.87	0.78	0.84
facebook	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
powergrid	0.81	0.82	0.85	0.82	1.00	1.00	1.00	1.00	0.78	0.85	0.85	0.85
CA-GrQc	0.87	0.87	0.84	0.85	1.00	1.00	1.00	1.00	0.99	0.97	0.97	0.97
reactome	0.96	0.91	0.91	0.91	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
CA-HepTh	0.87	0.87	0.84	0.87	1.00	1.00	1.00	1.00	0.99	0.97	0.96	0.97
pgp	0.84	0.69	-0.15	0.74	1.00	1.00	1.00	0.97	0.91	0.82	0.56	0.82
CA-HepPh	0.94	0.88	0.82	0.88	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00
CA-AstroPh	0.97	0.93	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CA-CondMat	0.91	0.91	0.90	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
deezer_RO	0.88	0.90	0.91	0.91	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99
deezer_HU	0.93	0.93	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
deezer_HR	0.97	0.96	0.96	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
brightkite	0.85	0.72	-0.25	0.72	1.00	1.00	1.00	0.96	1.00	0.87	0.84	0.88
livemocha	0.97	0.93	0.91	0.91	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.99
petster_cat	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00
douban	0.34	-0.57	-0.69	-0.57	-0.51	0.87	0.62	0.82	0.57	-0.10	-0.69	-0.12
gowalla	0.90	0.82	0.34	0.84	1.00	1.00	0.99	0.81	1.00	0.96	0.90	0.94
dblp	0.87	0.87	0.85	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
petster_dog	0.97	0.96	0.93	0.96	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00

The characteristic behavior regarding the influence of the scaling factor s on the robustness of centrality measures in the random graphs is shown in Figure 4.4. The robustness increases with increasing scaling factor (i.e., for networks with higher average degree). The initial increases of the scaling factor have a larger influence on the robustness than the subsequent increases. The rank correlations between the scaling factor of the degree sequence and the robustness of centrality measures w.r.t. random graphs with these degree sequences for all networks, error mechanisms, and centralities are listed in Table 4.6. For most cases, we observe very high correlations. Multiple realizations of the same experiment yield similar results (i.e., the variance is low). This observation strongly suggests that the robustness is a property of the degree sequence. Notable exceptions are the results for smaller networks (zachary, dolphins, jazz, and pdzbase). In these cases, the correlations are lower and the variances are higher compared to the other networks. We suspect that this instability is related to the size of these networks, they all contain less than 200 nodes.

Among the remaining (larger) networks, it is noticeable that the correlations for the douban network are considerably lower. This might be due to the structure of the douban network: is has a low average degree (4.2) and approx. 2/3 of the nodes have a degree of one. Hence, there is little structure implied by this its degree sequence.

4.5 Analysis and discussion of the degree centrality

In our experiments in Sections 4.3 and 4.4, we observed that, in many cases, not the network size but the average degree is correlated with the robustness of centrality measures. An increasing average degree often leads to a higher robustness, which was observed in the experiments with the empirical networks and also in the experiments with graphs generated by the BA and configuration model. However, the opposite effect was also observed. In the case of ER graphs, the robustness of degree centrality decreases with an increasing average degree. In this section, we take a more detailed look at the two scenarios involving ER and BA graphs.

In the following, we focus on the degree centrality, as it is the most accessible for an analytical perspective (Platig et al., 2013; Tsugawa and Ohsaki, 2015; Murai and Yoshida, 2019). First (Section 4.5.1), we derive an expression for the robustness based on the interpretation of the robustness as a probability. Then we argue that this interpretation is closely related to the robustness measured with Kendall's tau, but more accessible for an analytical perspective. The exact expression for the robustness depends on the type of error. In Section 4.5.2, we derive these expressions for the case of ER graphs and use them in Section 4.5.3 to show that, for ER graphs and sufficiently large network size, the robustness is independent of the network size, as long as the average degree is constant. In Section 4.5.4, we analyze the behavior of the robustness of the degree centrality when the average degree is increasing in ER and BA graphs in more detail.

4.5.1 Analytical approach for the robustness

To analyze the degree centrality in more detail, we use the following terms: G is the unmodified Graph, H is the erroneous graph (a "modified" version of G, i.e., H is on the same node set as G or on a subset of that node set).

We used Kendall's τ to measure the robustness in our experiments to provide comparability to existing studies. Goodman and Kruskal's rank correlation coefficient γ (Goodman and Kruskal, 1954), as explained in Section 4.2.3, allows us to develop an analytical approach. Therefore we will use it in the remainder of this section. Note that both measures differ in their definitions only when there are ties. We can rewrite the robustness of the degree centrality with respect to G and H (as stated in Equation (4.2)) in terms of the probability of concordant (P_c) and discordant (P_d) pairs:

$$\begin{aligned} r_{\gamma,\text{degree}}(G,H) &= \frac{n_c - n_d}{n_c + n_d} & &= \frac{2n_c - (n_c + n_d)}{n_c + n_d} \\ &= 2\frac{n_c}{n_c + n_d} - 1 & &= 2\frac{\frac{n_c}{n}}{\frac{n_c}{n} + \frac{n_d}{n}} - 1 \\ &= 2\frac{P_c}{P_c + P_d} - 1. \end{aligned} \tag{4.3}$$

Now we will derive how these probabilities and thus also the robustness can be calculated.

The error level is denoted by α; from the context, it becomes apparent whether this refers, for example, to the level of deleted nodes or edges. Additionally, consider two nodes v_1, v_2 drawn randomly from $V(H)$. Now let D_i denote the random variable for the degree of node v_i in G and X_i denote the random variable for the degree change of node v_i (i.e., the difference of the degree of node v_i in G and in H), D_1, X_1 and D_2, X_2 are independent and identically distributed (i.i.d). On this basis let $P(D_1 = d_1, X_1 = x_1, D_2 = d_2, X_2 = x_2)$ be the joint probability that specific values for d_1, x_1, d_2, x_2 occur together. We abbreviate this by $P(d_1, x_1, d_2, x_2)$.

Summing $P(d_1, x_1, d_2, x_2)$ over the quadruples that correspond to concordant (discordant) pairs of nodes, we can calculate the probability for (v_1, v_2) to be concordant (discordant). For example, for the case of missing edges, the probability for (v_1, v_2) to be concordant is

$$P_c = \sum_{\substack{d_1<d_2;d_1-x_1<d_2-x_2\\ d_1>d_2;d_1-x_1>d_2-x_2}} P(d_1, x_1, d_2, x_2), \tag{4.4}$$

and the probability for (v_1, v_2) to be discordant is

$$P_d = \sum_{\substack{d_1<d_2;d_1-x_1>d_2-x_2\\ d_1>d_2;d_1-x_1<d_2-x_2}} P(d_1, x_1, d_2, x_2). \tag{4.5}$$

The robustness as defined in Equation (4.3) is thus a function of the probabilities defined in Equations (4.4) and (4.5).

4.5.2 Expressions for the robustness and error types

In the previous section, we showed how the robustness can be expressed in terms of the probability of pairs to be concordant or discordant. These probabilities depend on the type of error and the degree distribution of the graph. In this section, we derive explicit expressions for $P(d_1, x_1, d_2, x_2)$ for the case of missing edges, missing nodes, and additional edges in ER graphs with n nodes and edge probability p.

Since D_1, D_2 and X_1, X_2 are i.i.d., we can express $P(d_1, x_1, d_2, x_2) = P(d_1, x_1) \cdot P(d_2, x_2)$. To derive the actual expression for $P(d_i, x_i)$, we use the fact that $P(d_i, x_i) = P(x_i|d_i) \cdot P(d_i)$, where $P(d_i)$ is the probability that node v_i has a degree of d_i. For ER graphs this is the binomial distribution $D_i \sim Bin(n, p)$. Independently of the specific error mechanism, $P(x_i|d_i)$ describes the effects of that error on node v_i. Hence, to create explicit expressions to that allow us to calculate the actual robustness, we solely need to define $P(x_i|d_i)$ for the corresponding the type of graph and error. In the following, we will provide these for the case of missing edges, missing nodes, and additional edges in ER graphs.

Missing edges For the case of missing edges, $P(x_i|d_i)$ is the probability that x_i edges are removed from a node with degree d_i. This is binomial distributed with $P(X_i = x_i|D_i = d_i) \sim Bin(d_i, \alpha)$. The fraction of missing edges is denoted by α. The restrictions for the quadruples which are used for the calculation of robustness are the same as in Equation (4.4) and (4.5).

- The marginal distribution is as follows:

$$P(D_i = d_i, X_i = x_i) =: P(d_i, x_i) =$$
$$\binom{d_i}{x_i}\alpha^{x_i}(1-\alpha)^{d_i-x_i}\binom{n}{d_i}p^{d_i}(1-p)^{n-d_i}, \text{ for } i \in \{1,2\}. \quad (4.6)$$

- As we assume that the edges are deleted independently of each other, $P(d_1, x_1, d_2, x_2) = P(d_1, x_1) \cdot P(d_2, x_2)$, and thus calculate the robustness of the degree centrality $r_{\gamma,\text{degree}}(G, H)$ with

$$P_c = \sum\nolimits_{\substack{d_1<d_2;d_1-x_1<d_2-x_2\\ d_1>d_2;d_1-x_1>d_2-x_2}} P(d_1, x_1) \cdot P(d_2, x_2)$$

and

$$P_d = \sum\nolimits_{\substack{d_1<d_2;d_1-x_1>d_2-x_2\\ d_1>d_2;d_1-x_1<d_2-x_2}} P(d_1, x_1) \cdot P(d_2, x_2).$$

Missing nodes In the case of missing nodes, the restrictions for the quadruples, which are used for the calculation of robustness and the degree distribution, are the same as above. The error level α is the fraction of nodes that are missing in H. For the conditional distribution of the degree decrease $P(x_i|d_i)$ we note that: $n\alpha$ nodes are deleted, $n(1-\alpha)$ nodes are not deleted, d_i is the degree of node v_i — the number of neighbors “drawn” from the set of n nodes (Actually $n-1$, but for large n, the difference becomes negligible.).

- With this, we can specify the distribution of $P(x_i|d_i)$ as a $HGeom(n\alpha, n(1-\alpha), d_i)$:

$$P(x_i|d_i) = \binom{n\alpha}{x_i}\frac{\binom{n(1-\alpha)}{d_i-x_i}}{\binom{n}{d_i}}, \text{ for } i \in \{1,2\}. \quad (4.7)$$

- Hence, with $n' = n\alpha$:

$$P(d_i, x_i) = P(x_i|d_i)P(d_i) = \binom{n'}{x_i}\frac{\binom{n-n'}{d_i-x_i}}{\binom{n}{d_i}}\binom{n}{d_i}p^{d_i}(1-p)^{n-d_i}, \quad (4.8)$$

for $i \in \{1,2\}$.

- Similar to the case of missing edges, we can use the fact that $P(d_1, x_1, d_2, x_2) = P(d_1, x_1) \cdot P(d_2, x_2)$, and thus calculate the robustness of the degree centrality $r_{\gamma,\text{degree}}(G, H)$ with

$$P_c = \sum\nolimits_{\substack{d_1<d_2;d_1-x_1<d_2-x_2\\ d_1>d_2;d_1-x_1>d_2-x_2}} P(d_1, x_1) \cdot P(d_2, x_2)$$

and

$$P_d = \sum_{\substack{d_1<d_2;d_1-x_1>d_2-x_2\\ d_1>d_2;d_1-x_1<d_2-x_2}} P(d_1,x_1)\cdot P(d_2,x_2).$$

Additional edges While, in the case of additional edges, the degree distribution is still the same as above, x_i now refers to the degree increase of node v_i and α is the fraction of edges added to the graph.

- The conditional distribution for the degree increase is $Bin(n-d_i, \alpha\frac{p}{1-p})$ and hence,

$$P(d_i,x_i) = \binom{n}{d_i} p^i (1-p)^{(n-d_i)} \binom{n-d_i}{x_i} (\alpha\frac{p}{1-p})^{x_i} \\ (1-\alpha\frac{p}{1-p})^{(n-d_i-x_i)}, \text{ for } i \in \{1,2\}. \quad (4.9)$$

- Again, we can use the fact that $P(d_1,x_1,d_2,x_2) = P(d_1,x_1)\cdot P(d_2,x_2)$, and thus calculate the robustness of the degree centrality $r_{\gamma,\text{degree}}(G,H)$. It is important to note, however, that the conditions for the summations change in the case of additional edges:

$$P_c = \sum_{\substack{d_1<d_2;d_1+x_1<d_2+x_2\\ d_1>d_2;d_1+x_1>d_2+x_2}} P(d_1,x_1)\cdot P(d_2,x_2), \quad (4.10)$$

and

$$P_d = \sum_{\substack{d_1<d_2;d_1+x_1>d_2+x_2\\ d_1>d_2;d_1+x_1<d_2+x_2}} P(d_1,x_1)\cdot P(d_2,x_2). \quad (4.11)$$

4.5.3 The case of constant average degree

In the following, we use the expressions developed in the previous section to study the impact of increasing network size (while the average degree is constant) on the degree centrality in more detail. For the case of missing edges and additional edges, we prove that the robustness of the degree centrality is independent of the network size.

Missing edges For the case of missing edges, we derived the expression for $P(d_i,x_i)$ in Equation (4.6). The degree distribution can be approximated by an exponential distribution with $\lambda = np$, hence $P(d_i) = \frac{e^{-\lambda}\lambda^{d_i}}{d_i!}$. If we replace the corresponding term in Equation (4.6), then:

$$P(d_i,x_i) = \binom{d_i}{x_i} \alpha^{x_i} (1-\alpha)^{d_i-x_i} \frac{e^{-\lambda}\lambda^{d_i}}{d_i!}, \text{ for } i \in \{1,2\}. \quad (4.12)$$

In Equation (4.12), the probability and hence robustness, Equation (4.3), does not directly depend on n, as long as $\lambda = np$ is constant (which implies that the average degree stays constant), the robustness does no change if the network size increases. This shows that the robustness of degree centrality in ER graphs for this case does not directly depend on the network size. For arbitrary network size, there exist robust and non-robust networks. This also explains the observations from the experiments conducted in Section 4.4.1.

Additional edges For the case of additional edges, we derived the expression for $P(d_i, x_i)$ in Equation (4.9). With $q = \alpha\frac{p}{1-p}$, we can rewrite Equation (4.9) as:

$$P(d_i, x_i) = Bin(n, p; d_i) \cdot Bin(n, q; x_i) \cdot \text{correction term, for } i \in \{1, 2\}, \tag{4.13}$$

where the correction term is

$$\frac{(n-d_i)(n-d_i-1)\cdots(n-d_i-x_i+1)}{n(n-1)\cdots(n-x_i+1)}(1-q)^{-d_i} \tag{4.14}$$

and converges to 1.

Finally, we can apply the Poisson approximation ($\lambda = np$) to Equation (4.13), so the probability no longer depends on the network size:

$$P(d_i, x_i) \approx \text{Pois}(\lambda; d_i) \cdot \text{Pois}(\alpha\lambda; x_i), \text{ for } i \in \{1, 2\}. \tag{4.15}$$

This result shows that, also for the case of additional edges, the robustness of degree centrality in ER graphs does not depend on the network size for large networks.

4.5.4 The case of increasing average degree

In the following, we take a closer look at the observations of Section 4.4.2, the behavior of the robustness of degree centrality with increasing average degree. In Section 4.4.2, we observed that the robustness of the degree centrality in ER graphs decreases with increasing average degree. For BA graphs, in contrast, the robustness increased with increasing average degree. To further explore this observation, recall that a pair of nodes is discordant, if the degree change induced by the error is larger than the degree difference between those nodes in the error-free graph. We therefore suspect that the ratio of the variance of degree values to the variance of degree changes could explain the robustness.

In order to investigate this, we have repeated the experiments regarding the degree centrality from Section 4.4.2 and recorded the changes of the degree values. Figure 4.5 shows the robustness of degree centrality dependent on the parameter that controls the average degree of the ER and BA graphs (red curve). In addition, the ratio between variance of degree and variance of degree change (blue curve) is also shown. For both types of random graphs and all four error mechanisms, these results show a consistent pattern. If this ratio is increasing, the robustness also rises, if it falls, the robustness also falls.

Next, we consider three cases, missing nodes, missing edges, and additional edges, for which we argue why the ratio of the variances for the case of ER graphs behaves like this. Recall, $D \sim Bin(n, p)$ is the degree distribution for the ER graph, and X is the random variable for the degree change.

Missing edges For the case of missing edges, α is the probability for edge deletion and $X \sim Bin(n, p\alpha)$. Let $f(p) = Var(D)/Var(X)$ be the ratio of variance of the degree to variance of the degree change, then:

$$f(p) = \frac{np(1-p)}{np\alpha(1-p\alpha)}, \text{ hence} \tag{4.16}$$

$$f'(p) = \frac{\alpha - 1}{\alpha} \cdot \frac{1}{(1-p\alpha)^2}. \tag{4.17}$$

Since the first part in Equation (4.17) is < 0 and its second part is > 0, the quotient of $Var(D)/Var(X)$ is (strictly) monotonically decreasing.

Additional edges For the case of additional edges, the total number of edges added is $\alpha p\binom{n}{2}$, the total number of edges that could be added to the graph is $\binom{n}{2}(1-p)$. Hence, the probability for a nonexistent edge to be added is $\frac{\alpha p\binom{n}{2}}{\binom{n}{2}(1-p)} = \alpha\frac{p}{1-p}$. Therefore, X, the degree change of a node, is distributed as follows: $X \sim Bin(n - D, \alpha\frac{p}{1-p})$. Now, let $D' \sim Bin(n, 1-p)$, then $X \sim Bin(D', \alpha\frac{p}{1-p})$ and thus $X \sim Bin(n, p\alpha)$. Which is the same distribution as for the case of missing edges. Consequently, the same results as derived above for the case of missing edges also holds for the case of additional edges.

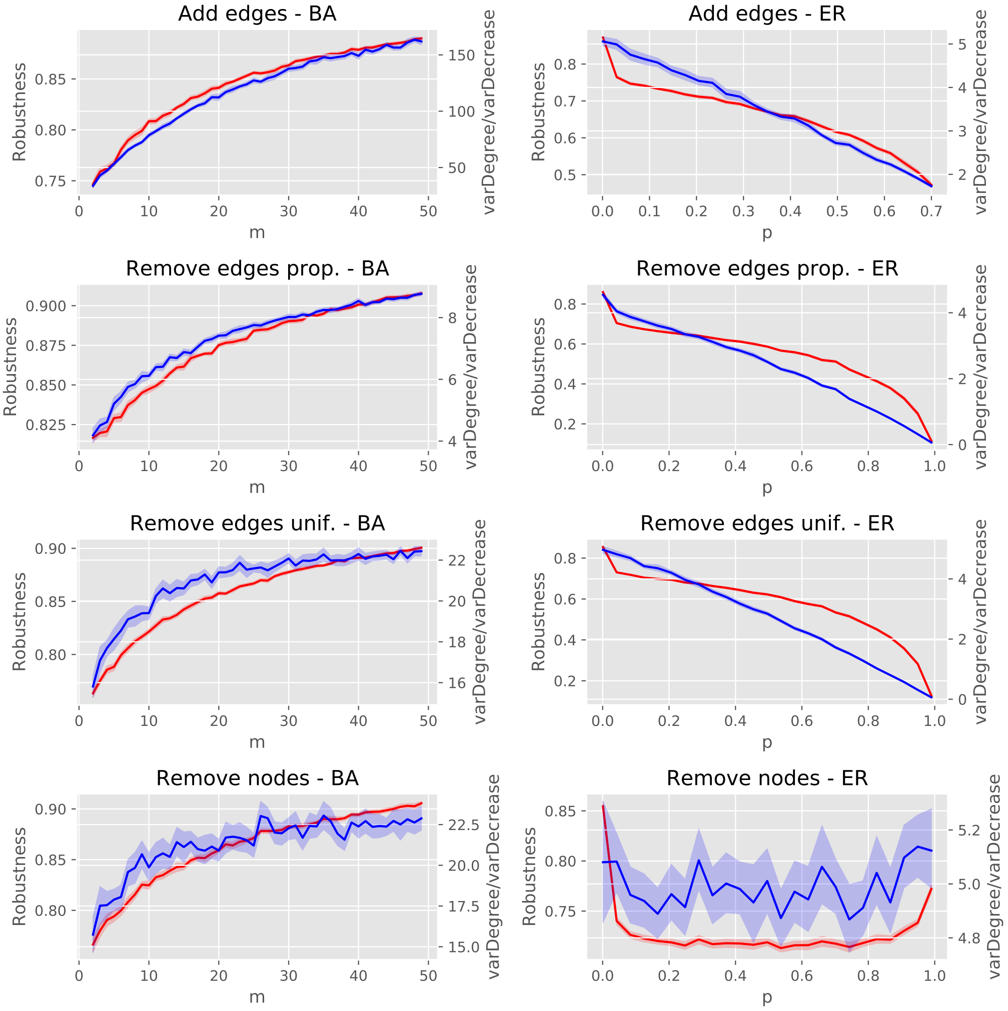


Figure 4.5: Robustness of the degree centrality (red curve, axis labels on the left side) and ratio between the variance of the degree and the variance of the degree change (blue curve, axis labels on the right side).

Missing nodes For the case of missing nodes, the degree change X is distributed as follows: $X \sim Bin(\alpha n, p)$. The ratio of variance of the degree to the variance of the degree change for this case is as follows:

$$\frac{Var(D)}{Var(X)} = \frac{np(1-p)}{\alpha np(1-p)} = \frac{1}{\alpha}, \tag{4.18}$$

which is a constant function with respect to p, in contrast to Equation (4.16).

Equations (4.17) and (4.18) demonstrate the difference between the edge error mechanisms (missing and additional edges) and the case of missing nodes: in the latter the quotient $Var(D)/Var(X)$ is a constant which explains the behavior of the degree centrality if nodes are missing in ER graphs which we observe in our experiments, see bottom right panel in Figure 4.5. In this case, the robustness does not change with increasing p (and thus increasing average degree). For the case of missing edges and additional edges, it is plausible that the robustness is related to the quotient $Var(D)/Var(X)$. If the variance of the degree change increases more strongly than the variance of the degree, the probability that concordant pairs become discordant pairs also increases. This illustrates the behavior shown in the first and third panel on the right-hand side of Figure 4.5.

4.6 Conclusions and final remarks

Network data is often erroneous, which compromises the reliability of centrality measures and the conclusions of subsequent analyses. We investigated the robustness behavior of the degree, eigenvector centrality, and the PageRank in empirical networks of different size and structure, as well as in random graphs, under the influence of missing and additional edges and missing nodes. We were primarily interested in the relationship between the robustness of these centrality measures and the network size respectively the average degree. We observed that a higher average degree was frequently associated with higher robustness for cases where nodes or edges are missing. Additionally, the degree was always the most robust measure and the variance of the eigenvector centrality was, in most cases, substantially higher than the variance of the degree centrality or the PageRank. Moreover, we observed that there exist small networks that are robust and larger networks that are not robust w.r.t. centrality measures. These results also demonstrate that the study of the robustness of centrality measures in the context of larger networks is highly relevant.

For further insight, we conducted experiments on random graphs. In the first type of experiment that we performed on ER and BA graphs, the average degree was constant, but the network size was increasing. The increasing network size did, however, not affect the robustness. In the second type of experiment, the network size was fixed, but the average degree increased. In the case of ER graphs, the robustness was either not affected by the change of the average degree (eigenvector centrality and PageRank), or even decreased (degree centrality). In the case of BA graphs, the robustness increased — for all centrality measures and errors that we considered. This was also the case for random graphs generated by the configuration model which generates more realistic networks since we used the degree distributions of the empirical networks of the first part of our study. These results suggest that centrality measures are more robust the higher the average degree in the network, as long as the networks have a skewed degree distribution.

In the third part of our study, we introduced an analytical approach for the robustness in terms of a rank correlation. Focusing on Goodman and Kruskal's rank correlation, we derived explicit expressions for the robustness of the degree centrality in ER graphs for the case of missing nodes, missing edges, and additional edges. We showed that the robustness for these type of networks is independent of their size, as long as the average degree is constant. For arbitrary network size, there exist robust and non-robust networks w.r.t. all centrality measures used in this study. Moreover, we argued that the quotient of the variance of the degree and the variance of the degree change may explains the robustness behavior at least to some extent. In addition, we studied the behavior of this quotient analytically for ER graphs.

These findings contribute to a better understanding of the robustness of centrality measures. Researchers should, therefore, pay particular attention to error-free data collection if it is known that the particular network is sparse. When centrality measures are used on this type of network, the results may be interpreted with caution.

This study also provides a basis for further research. The findings may be incorporated into procedures for the treatment of erroneous network data. Further investigations about sophisticated error mechanisms (e.g., extending the missing edges proportional error mechanism or mixtures of error mechanisms (Platig et al., 2013)) would be interesting. Another direction would be the use of random graph models in which other properties apart from the average degree can be controlled separately. Moreover, further investigation of the relationship between the robustness and the quotient of variances would be interesting.

Conflicts of interest

The authors have nothing to disclose.

4.7 References

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(July):378–382.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(October):509–512.

Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P., and Weinstein, H. (2005). PDZBase: A Protein–Protein Interaction Database For PDZ-Domains. *Bioinformatics*, 21(6):827–828.

Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., and Arenas, A. (2004). Models of Social Networks based on Social Distance Attachment. *Phys. Rev. E*, 70(5):056122.

Bolland, J. M. (1988). Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182.

Borgatti, S. P., Carley, K. M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136.

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, 25(2):163–177.

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and Mobility: User Movement in Location-based Social Networks. In *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, pages 1082–1090.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307.

De Las Rivas, J. and Fontanillo, C. (2010). Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807.

Dünker, D. and Kunegis, J. (2015). Social Networking by Proxy: Analysis of Dogster, Catster and Hamsterster. *Proc. Int. Conf. on World Wide Web Companion*, pages 361–362.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Erman, N. and Todorovski, L. (2015). The effects of measurement error in case of scientific network analysis. *Scientometrics*, 104(2):453–473.

Frantz, T. L., Cataldo, M., and Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328.

Ghoshal, G. and Barabási, A.-L. (2011). Ranking stability and super-stable nodes in complex networks. *Nature communications*, 2:394.

Gleiser, P. M. and Danon, L. (2003). Community Structure in Jazz. *Advances in Complex Systems*, 6(4):565–573.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15.

Holzmann, H., Anand, A., and Khosla, M. (2019). Delusive PageRank in Incomplete Graphs. In Aiello, L. M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., and Rocha, L. M., editors, *Complex Networks and Their Applications VII*, pages 104–117, Cham. Springer International Publishing.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: A Knowledgebase of Biological Pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432.

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251.

Kim, P. J. and Jeong, H. (2007). Reliability of rank order in sampled networks. *European Physical Journal B*, 55(1):109–114.

Koschützki, D., Lehmann, K., and Peeters, L. (2005). Centrality Indices. In Brandes, U. and Erlebach, T., editors, *Network Analysis: Methodological Foundations*, pages 16–61. Springer Berlin Heidelberg.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.

Kunegis, J. (2013). KONECT - The koblenz network collection. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web.*

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph Evolution: Densification and Shrinking Diameters. *ACM Trans. Knowl. Discov. Data*, 1(1).

Leskovec, J. and Mcauley, J. J. (2012). Learning to Discover Social Circles in Ego Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 539–547. Curran Associates, Inc.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations. *Behavioral Ecology and Sociobiology*, 54:396–405.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16(1):435–463.

Martin, C. and Niemeyer, P. (2019). Influence of measurement errors on networks: Estimating the robustness of centrality measures. *Network Science*, 7(2):180–195.

Murai, S. and Yoshida, Y. (2019). Sensitivity analysis of centralities on unweighted networks. In *The World Wide Web Conference*, WWW '19, pages 1332–1342, New York, NY, USA. ACM.

Newman, M. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256.

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64.

Niu, Q., Zeng, A., Fan, Y., and Di, Z. (2015). Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131.

Platig, J., Ott, E., and Girvan, M. (2013). Robustness of network measures to link errors. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6).

Rozemberczki, B., Davies, R., Sarkar, R., and Sutton, C. (2019). GEMSEC: Graph embedding with self clustering. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*, pages 65–72. ACM.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., and Ayivi-Guedehoussou, N. (2005). Towards a Proteome-scale Map of the Human Protein–Protein Interaction Network. *Nature*, (7062):1173–1178.

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, 107(3):1283–1298.

Smith, J. A. and Moody, J. (2013). Structural Effects of Network Sampling Coverage I: Nodes Missing at Random. *Social Networks*, 35(4).

Smith, J. A., Moody, J., and Morgan, J. H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48:78–99.

Tsugawa, S. and Ohsaki, H. (2015). Analysis of the Robustness of Degree Centrality against Random Errors in Graphs. In *Studies in Computational Intelligence*, volume 597, pages 25–36.

Wang, C., Butts, C. T., Hipp, J. R., Jose, R., and Lakon, C. M. (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to Add Health. *Social Networks*, 45:89–98.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409.

Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics of ‘Small-world’ Networks. *Nature*, 393(1):440–442.

Yang, J. and Leskovec, J. (2012). Defining and Evaluating Network Communities based on Ground-truth. In *Proc. ACM SIGKDD Workshop on Mining Data Semantics*, page 3. ACM.

Zachary, W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *J. of Anthropological Research*, 33:452–473.

Zafarani, R. and Liu, H. (2009). Social Computing Data Repository at ASU.

A process for the evaluation of node embedding methods in the context of node classification

Christoph Martin and Meike Riebeling

Martin, C., & Riebeling, M. (2020). A process for the evaluation of node embedding methods in the context of node classification. ArXiv E-Prints, arXiv:2005.14683.

The layout has been revised.

Abstract

Node embedding methods find latent lower-dimensional representations which are used as features in machine learning models. In the last few years, these methods have become extremely popular as a replacement for manual feature engineering.

Since authors use various approaches for the evaluation of node embedding methods, existing studies can rarely be efficiently and accurately compared. We address this issue by developing a process for a fair and objective evaluation of node embedding procedures w.r.t. node classification. This process supports researchers and practitioners to compare new and existing methods in a reproducible way.

We apply this process to four popular node embedding methods and make valuable observations. With an appropriate combination of hyperparameters, good performance can be achieved even with embeddings of lower dimensions, which is positive for the run times of the downstream machine learning task and the embedding algorithm. Multiple hyperparameter combinations yield similar performance. Thus, no extensive, time-consuming search is required to achieve reasonable performance in most cases.

5.1 Introduction

Networks are used to model phenomenons in various domains such as social relations, molecular graphs, biological structures, or recommender systems. Networks represent the relations (edges) between different entities (nodes). Social networks contain information about individuals or communities and the dynamics among them. This information can, for example, be used for segmentation or recommendation tasks. Networks capture not only social relationships, but also citations, biological information, or knowledge relations (Newman, 2003). Developing and experimenting with methods that leverage the information captured by these networks are important endeavors in business and research communities (Yang et al., 2015; Zhang et al., 2018).

In various fields, network data is to be used as input for machine learning models. This poses the challenge that network data must first be transformed in order to serve as features. Traditionally, handcrafted features have been created to represent the nodes. This type of feature engineering, however, has considerable weaknesses. It is very time-consuming on the one hand and, on the other hand, the handcrafted features can often not be reused (Hamilton et al., 2017). Node embeddings map the nodes of a graph to a lower-dimensional vector which can subsequently be used as input for other machine learning techniques. However, due to the particular data structure of a network, the quality of network embeddings depends on preserving the structural properties of a graph while incorporating node attributes. This can be difficult as the structural similarity of nodes can either be portrayed as nodes close to each other or as nodes with similar roles in the network, node embeddings have to respect local and global node similarities together (Cai et al., 2018; Zhang et al., 2018; Goyal and Ferrara, 2018b).

Node embedding methods have enormous potential, thus this area continues to be a highly active field of research. In recent years, several surveys have been published, which summarize the progress made in this area and address the comparison and categorization of node embedding methods (Hamilton et al., 2017; Goyal and Ferrara, 2018b; Zhang et al., 2018; Cai et al., 2018). Due to the popularity of embedding methods, a unified way to compare them has become increasingly important. Methods proposed by existing studies can rarely be compared to each other since authors use different approaches to evaluate node embeddings.

We address this issue by developing a process (Section 5.3) for a fair and objective evaluation of node embedding procedures w.r.t. node classification. Building on and extending existing work (Goyal and Ferrara, 2018b; Zhang et al., 2018; Khosla et al., 2019; Goyal et al., 2019), we explicitly address the choice of the hyperparameters in the process presented here, under consideration of the downstream machine learning task, in this case node classification. This process supports researchers to compare new and existing methods in a reproducible way. Furthermore, end users can use this process to find the optimal method for the particular use case.

In the case study in Section 5.4, we apply the process to four popular node embedding methods and make valuable observations, especially for practitioners. The default hyperparameters for node embedding procedures are generally not a good choice. With an appropriate combination of hyperparameters, good performance can be achieved even with embeddings of lower dimensions, which is positive for the run times of the downstream machine learning task. Multiple hyperparameter combinations yield similar performance; hence usually there is no extensive, time-consuming search required to achieve reasonable performance.

5.2 Node embeddings

Let G be a graph on N nodes with vertex set $V(G) = \{v_1, v_2, \ldots, v_N\}$. Node embeddings are d-dimensional representations of the nodes in G; usually, these are lower-dimensional (i.e., $d \ll N$). These embeddings are commonly used as input for machine learning algorithms. Node embedding methods have the objective to find such a mapping $f : V(G) \rightarrow \mathbb{R}^d$, where nodes which are "similar" to each other in the graph also "similar" to each other in the vector space. The definition of similarity differs between methods. In the literature, the terms graph embedding or network relational learning are also used for this purpose (Perozzi et al., 2014; Zhang et al., 2018; Hamilton et al., 2017).

In our case study (Section 5.4), we use the following four frequently cited and widely used node embedding methods: node2vec (Grover and Leskovec, 2016), GraRep (Cao et al., 2015), Deep Network Graph Representation (DNGR) (Cao et al., 2016), and Large-scale Information Network Embedding (LINE) (Tang et al., 2015). We use the implementations provided by (Goyal and Ferrara, 2018a; Natural Language Processing Lab at Tsinghua University, 2019).

5.3 A process for the comparison of node embedding methods

In this section, we develop the evaluation process for node embedding methods. This process enables researchers and practitioners to perform a fair and objective evaluation of node embedding procedures. We present this process for two main reasons. The first is to compare new and existing methods in a reproducible way. Furthermore, it helps end users to find the optimal method for the particular use case. We start by arguing why the procedure for selecting hyperparameters cannot easily be transferred from previous machine learning methods to node embedding learning. Then we propose an approach and integrate it with the process.

The evaluation of algorithms and methods is an essential part of machine learning and network analysis research (Caruana and Niculescu-Mizil, 2006; Daelemans and Hoste, 2002). Particularly, algorithm selection is a widely discussed topic and an essential part of the application of machine learning algorithms in practice. This is due to the fact that there is not one single method optimal for all problem settings (Kou et al., 2012; Wolpert and Macready, 1997).

Essential components of evaluation experiments in machine learning are the data set, feature selection, feature representation, and hyperparameter settings (Daelemans and Hoste, 2002). The components of an evaluation process for node embeddings are slightly different. The data set and the hyperparameter settings can be transferred to node embeddings as essential components of the evaluation (Zhang et al., 2018). However, the feature selection process and the data representation have to be altered. Node embedding methods naturally take a network and the contained information as feature input, essentially making the step of feature selection unnecessary. The necessary representation of the network might differ between algorithms, hence the data representation is implied by choice of the embedding method.

Node embedding methods constitute an unsupervised problem setting traditionally; semi-supervised methods also exist (e.g., Kipf and Welling (2017), Shchur et al. (2018)), but these are not addressed in this paper. An application task is, therefore, necessary to evaluate the quality of node embeddings and is thus an essential component of the process. In summary, the core components of the process are the network data, the application task, the evaluation metric, and the hyperparameter configuration.

Network data The choice of the network data depends on the setting in which the process is applied and the node embedding methods considered. Practitioners who are looking for the best method for their particular application should use data that is close to the production data. For the comparative evaluation of new and existing embedding methods, in the interests of reproducibility, we recommend using publicly available networks of different size and structure. These may be, for example, the data sets used in the case study in Section 5.4.

Application task and evaluation metric The most popular application task is node classification, which is often applied when presenting a new embedding method. Classification aims at finding class labels for each node. The vector representation serves as feature input for a classifier (Cai et al., 2018; Goyal and Ferrara, 2018b). Training a classifier requires training data, which means that labels have to be available at least for a part of the network. Common evaluation metrics in this context include F_1-score, precision, recall, or accuracy. We propose to use the F_1-score since it takes precision and recall into account, $\mathrm{F}_1\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. In the case of multi-class and multi-label classification problems, we use the macro and micro variant of the F_1-score. Here the classes, respectively, the individual observations, are weighted equally (Powers, 2011).

For the classification task, we propose to use two popular and often used algorithms in machine learning: a logistic regression model (one-vs-rest classification for the multi-label model) and a random forest model. The regression model because of the frequent usage in the evaluation of embedding methods. The random forest is a widely used model in practical

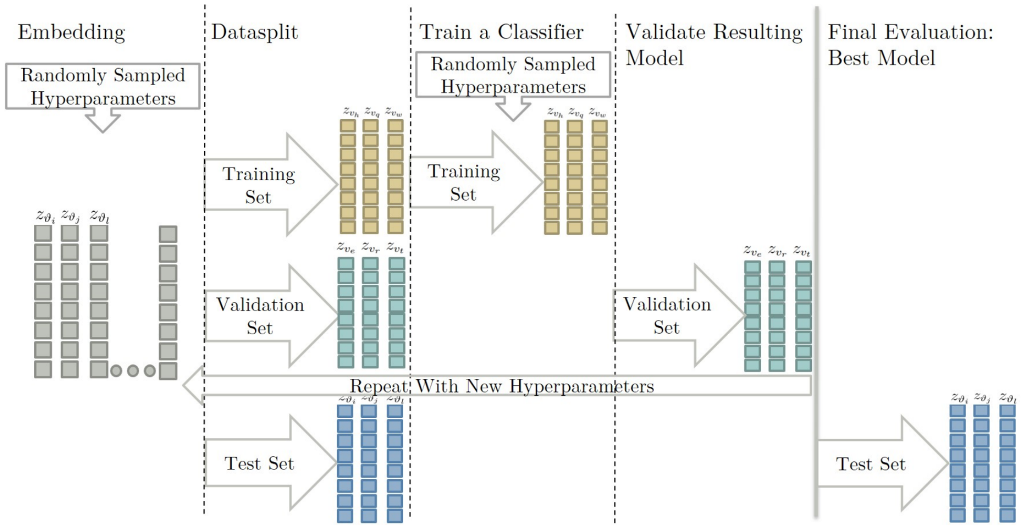

Embedding
Datasplit
Train a Classifier
Validate Resulting Model
Final Evaluation: Best Model
Randomly Sampled Hyperparameters
Randomly Sampled Hyperparameters
z_{ϑ_i} z_{ϑ_j} z_{ϑ_l}
Training Set
Validation Set
Test Set
z_{v_h} z_{v_q} z_{v_w}
z_{v_e} z_{v_r} z_{v_t}
Training Set
Validation Set
Repeat With New Hyperparameters
Test Set

Figure 5.1: **Setup for the hyperparameter selection** The embedding algorithm is applied to the whole network resulting in a vector representation marked by gray squares on the left side which is subsequently divided into three splits: training set (yellow), validation set (turquoise) and test set (blue). The training set is used to train the classifier, the validation set is used in that last part of the hyperparameter selection to validate the performance of the combined model. The test set is not used during the hyperparameter selection. After repeating this process several times, the best hyperparameter combination is selected for the final model, which is evaluated on the test set (final step on the right separated by the bold line).

machine learning applications. Nevertheless, it is usually not applied in node embedding research. Therefore we suggest to use it in this context because it is very flexible and leads to good results on different data sets (Fernández-Delgado et al., 2014).

Hyperparameter configuration The selection of the best hyperparameters is a debated topic in research. The impact of different tuning parameters on each other and how they affect the performance is only poorly understood (Li et al., 2016). In practice, a widely used method to find a set of hyperparameters is random search, where the search space of hyperparameters is randomly explored and evaluated. Bergstra and Bengio (2012) showed that this type of search leads to equally good or even superior models, compared to grid search, while only a fraction of the time is needed.

In addition to the way the hyperparameter selection is performed, the data utilized for tuning is an important topic. Usually, in machine learning data is split into a test, training, and validation set, in which the test set is only used once for the final validation. The training of the algorithm is performed on the training set with a subsequent evaluation of the performance using the validation set (James et al., 2013). For network embedding procedures, this is not possible. Splitting the network data into different sub-graphs would significantly alter the results of the embedding methods as they rely on representing the whole graph mirroring the structural context information of a node and its position in the whole network. Only using part of the network for an embedding would lead to a completely different representation with important context information missing. The proposed solution for the described challenges is a combined tuning of hyperparameters of the embedding and the subsequent application algorithm. The application task serves as the basis for the performance evaluation governing the hyperparameter selection. As shown in Figure 5.1, the representation for the whole graph is learned, whereas only part of this data is used in the application task (for example, classification) to evaluate the hyperparameter selection. For both algorithms — the embedding algorithm and the classification algorithm — hyperparameters are selected randomly. This process is repeated several times. Finally, the best model combination using the best hyperparameters for both algorithms is picked and evaluated on the test set.

5.4 Case study for the comparison process

In this section, we use the process to compare four frequently cited and widely used node embedding methods: node2vec, GraRep, LINE, and DNGR. Especially, we are interested in the impact of the number of dimensions and the amount of training data used on the performance in the domain of node classification.

We use data sets with varying characteristics (i.e., directed and undirected as well as binary, multi-class, and multi-label classification) to get an understanding of how embedding procedures behave under different conditions. Table 5.1 lists basic statistics about these networks. For training and model selection, we use 50% for the training set and 25% for the validation set and test set. For the second part of the experiment, where we analyze the impact of varying amounts of training data, we use $10\%, 20\%, \ldots, 100\%$ of the training data. All of these values refer to the node embedding vectors.

Table 5.1: Summary of networks used in the case study.

Network	Nodes	Edges	Directionality	# Labels	Source
Moreno Blogs	1,224	19,025	directed	2 (binary)	Adamic and Glance (2005)
CiteSeer	3,312	4,660	directed	6 (multi-class)	Getoor (2005)
Facebook	4,039	88,234	undirected	4 (multi-class)	Leskovec and Mcauley (2012)
BlogCatalog	10,312	333,983	undirected	39 (multi-label)	Tang and Liu (2009)

Overall results The performance of the embedding methods w.r.t. the different classifiers and measures are listed in Table 5.2. The scores for the logistic regression scenarios reveal that most of the tested algorithms perform similar across the networks. The highest score for the BlogCatalog network is 0.35, which was reached by node2vec. LINE and GraRep reach equal scores of 0.34 on that network. For Facebook, the scores are even closer together, the values vary between 0.45 and 0.52. The same trend can be found in the results of the Moreno network. For the Moreno network, the score of LINE, GraRep, DNGR, and node2vec are the same with 0.95. The best scores for CiteSeer range from 0.53 to 0.57. Only the deep learning-based method yield worse results, DNGR does not work well with a score of 0.25. Overall, the results indicate that very similar scores can be reached across different methods. The observed performance of node2vec, LINE, and GraRep on the BlogCatalog data set are in line with the results reported in the literature. For GraRep and node2vec, evaluation experiments were also conducted using a one-vs-rest logistic regression (Cao et al., 2015; Grover and Leskovec, 2016). Moreover, in Cao et al. (2015), LINE was included as a baseline. For all three networks, the performance was around 0.4; the slightly lower performance observed in this paper might be explained by the use of only 50% of the networks for training, due to the data split in training, validation and test set explained above.

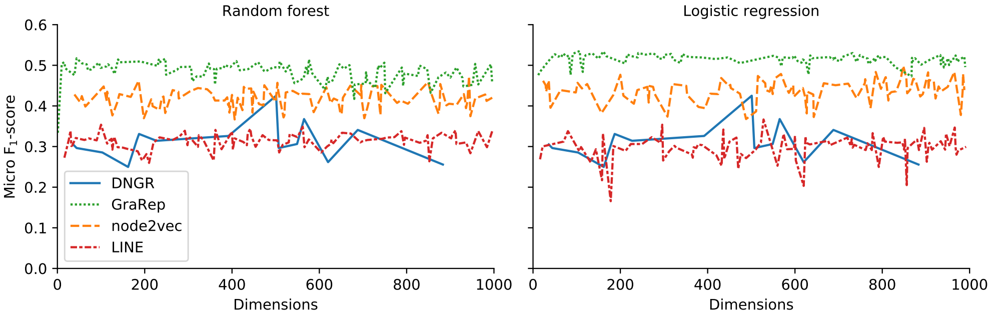
Analysis of the number of dimensions The dimensionality of the embedding is the only hyperparameter shared by all node embedding methods. The performance of embedding algorithms should, intuitively, increase with an increasing number of dimension until reaching a plateau where no substantial improvement of performance happens with increasing dimensionality. Grover and Leskovec (2016) observed this behavior for the node2vec algorithm. Experimenting with the number of dimensions resulted in a saturation of performance improvement at a dimension of around 100. Similar results are reported by Wang et al. (2016). They noticed a decline in performance after saturation at about 100. For some algorithms like GraRep, little influence of the dimensionality on performance was observed.

Table 5.2: Results for the experiments in the case study.

Score	Classifier	Network Embedding	Blog.	CiteSeer	Facebook	Moreno
Macro F_1	Random forest	DNGR	0.020	0.180	0.434	0.941
		GraRep	0.111	0.555	0.489	0.944
		LINE	0.020	0.240	0.458	0.954
		node2vec	0.032	0.505	0.456	0.951
	Log. regression	DNGR	0.068	0.153	0.485	0.954
		GraRep	0.181	0.514	0.505	0.951
		LINE	0.195	0.469	0.427	0.948
		node2vec	0.212	0.493	0.412	0.954
Micro F_1	Random forest	DNGR	0.052	0.266	0.450	0.941
		GraRep	0.244	0.607	0.505	0.944
		LINE	0.054	0.273	0.507	0.954
		node2vec	0.088	0.563	0.514	0.951
	Log. regression	DNGR	0.182	0.248	0.507	0.954
		GraRep	0.342	0.574	0.525	0.951
		LINE	0.339	0.536	0.450	0.948
		node2vec	0.354	0.534	0.496	0.954
Most frequent label			0.090	0.212	0.336	0.520

The reported relation between dimension and performance is almost steady, with a slight decrease after 64 dimensions (Cao et al., 2015). In Figure 5.2, the performance depending on the dimension for the case of the Facebook network is shown. These results indicate that higher dimensions do not necessarily lead to better performance. This behavior also occurs for the other networks. However, analyzing the performance with different dimensions lead to high variances. The reason might lie in the high amount of different hyperparameter combinations since the performance is not only dependent on the dimension but on the combination of parameters picked. Nonetheless, the findings suggest that in combination with the right hyperparameters, small dimensions are sufficient to reach scores that are comparable to the performance with higher dimensions. The results highlight the influence of all hyperparameters on each other. Therefore, the optimal performance of an embedding method depends on all hyperparameters, the network, and the application task. Moreover, the results suggest that equally well results can be reached with many different hyperparameter combinations, indicating that a reasonable performance can be reached without an extensive hyperparameter search. This may also explain the difference between our results to the above. We consider the performance of the final application task (node classification) when finding embeddings.

Figure 5.2: Impact of the dimensionality on the performance for the Facebook network.



Hyperparameter for node2vec A more detailed analysis of the hyperparameter for node2vec is listed in Table 5.3. The results lead to the conclusion that the best hyperparameter combination depends on the network and the application task. In the case of the BlogCatalog data set, there are also apparent differences between the two classification algorithms: The hyperparameter search leads the algorithm towards different learning strategies. The values for the sampling parameters p and q are 2 and 0.25 in the random forest case and 2 and 1 for the logistic regression. Thus for the Blog Catalog network, the random forest benefits from a depth-first sampling strategy preferring nodes further away from the source node, whereas the sampling strategy for the logistic regression is not biased towards one sampling strategy. The parameter p is 2 for both classification cases. Hence, the likelihood of revisiting a node is low.

Table 5.3: Results for the hyperparameter analysis of node2vec.

	Random forest				Logistic regression			
	Blog.	Face.	Cite.	Moreno	Blog.	Face.	Cite.	Moreno
Micro F$_1$-score	0.04	0.47	0.27	0.95	0.22	0.5	0.35	0.95
Dimension	74	943	103	733	197	848	245	600
Return parameter: p	2	0.5	0.25	1	2	2	0.75	2
In-out parameter: q	0.25	0.5	0.5	4	1	0.5	0.25	0.5
Number of walks: l	25	42	40	20	33	5	48	39
Walk length: k	70	56	11	45	46	28	11	24

In the paper introducing node2vec, experiments were also conducted on the BlogCatalog network. The authors described an increase in performance with small values for p and q. The results of the presented experiments suggest higher values for p and smaller values for q. Differences in the findings for the return parameter p are probably due to the variants in the remaining hyperparameters. Grover and Leskovec (2016) used default values for all remaining hyperparameters and only experimented with the values for p and q. The findings of the random search suggest a strong effect of the interaction between the parameters. Even though a lower p is optimal in the case of default parameters, a higher p — leading to node sequences containing samples further away from the source — leads to better results, when combined with more and longer walks. These observations highlight the importance of tuning the hyperparameters of node embeddings based on the application task instead of simply using the default parameters.

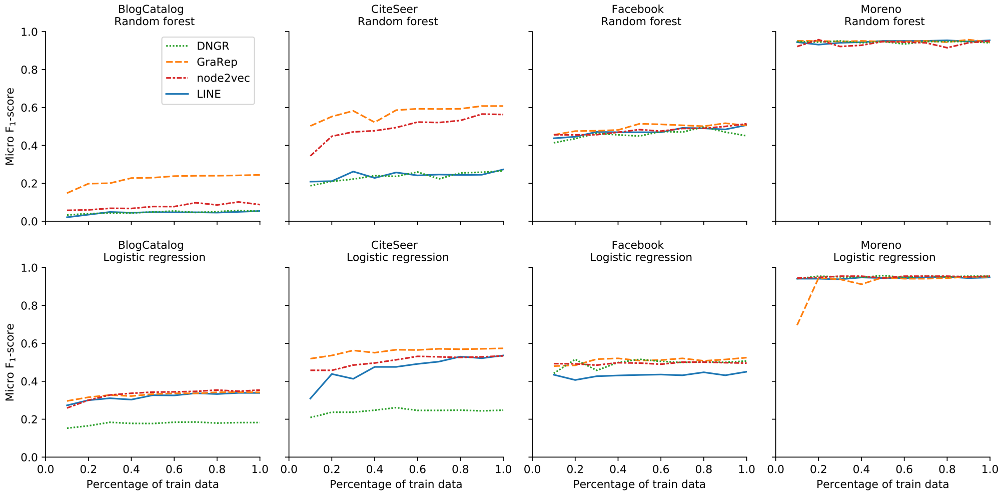


Figure 5.3: Results for the impact of an increasing amount of training data on the classification performance.

Impact of % training data on the performance Figure 5.3 shows the impact of increasing the amount of training data on the performance, the overall impact is small. The behavior of the curves, however, shows that with small ratios, an increase in the amount of training data has a high impact on the performance. At some point, an increase in the data only leads to a small improvement. As an example, for both classification methods, the performance for the Moreno network reaches a peak in performance increase at around 20% of the training data. After that point, the impact on performance is relatively low. Similarly, the score for the embedding methods on the BlogCatalog network are increasing until a ratio of 0.2 to 0.3, the performance score of node2vec in the logistic regression scenario is 0.26 with 10% of the data. However, with 30% of the data, the score is already 0.33. There is only little improvement thereafter as the best score is 0.35. This is consistent with previous studies that showed that the performance of node2vec shows large improvements until 30% after that, the increase in performance is small (Goyal and Ferrara, 2018b). For CiteSeer differences between the random forest and the logistic regression scenario can be observed. In the case of the random forest combined with GraRep and node2vec there is a substantial increase in performance. The starting value of node2vec, for example, is 0.34, whereas the best performance is 0.56. However, in the logistic regression scenario, the difference is only 0.05, which is consistent with a similar experiment conducted by Zhang et al. (2018), who compared the results using 5% and 50% of the whole network data and found an increase of 0.08 points for CiteSeer. The reason for these differences in the two application scenarios is not apparent. However, it might be because the random forest needs more labeled observations to separate them efficiently. The CiteSeer network has many labels with only a few observations. Therefore, a small amount of data might lead to an underrepresentation of training data for some labels.

5.5 Conclusions

Recently, node embeddings became popular as an alternative to handcrafted feature engineering (Hamilton et al., 2017). In this paper, we proposed a process for the comparison of node embedding methods w.r.t. node classification. This process enables researchers and practitioners to perform a fair and objective evaluation of node embedding procedures and helps end users to find the optimal method for the particular use case.

Moreover, in a case study, we applied this process to four popular node embedding methods. These experiments showed that the introduced process provides a foundation for a standardized evaluation of node embedding methods. Additionally, we made valuable observations, especially for practitioners: The default parameters for node embedding procedures are generally not a good choice. We analyzed this in detail for node2vec. Analyzing the impact of the dimensionality of the embeddings, we noticed that the appropriate combination of hyperparameters yields good performance with a lower number of dimensions, which is positive for the run times of the downstream machine learning task and the embedding algorithm. We also observed that multiple hyperparameter combinations yield similar performance. Hence there no extensive, time-consuming search required to achieve reasonable performance.

Although the proposed process provides a robust foundation for the comparison of node embedding methods, there are some aspects which should be addressed by future research. For example, the application task link prediction. It would be particularly interesting to understand how the procedure has to be adjusted differently for missing and future link prediction. A comprehensive comparison of semi-supervised methods would also be of interest.

5.6 References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43. ACM.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Cai, H. Y., Zheng, V. W., and Chang, K. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Cao, S., Lu, W., and Xu, Q. (2015). GraRep: Learning Graph Representations with Global Structural Information. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pages 891–900.

Cao, S., Lu, W., and Xu, Q. (2016). Deep neural networks for learning graph representations. In *AAAI*, pages 1145–1152.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168. ACM.

Daelemans, W. and Hoste, V. (2002). Evaluation of machine learning methods for natural language processing tasks. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*. European Language Resources Association (ELRA).

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.

Getoor, L. (2005). Link-based classification. In *Advanced Methods for Knowledge Discovery from Complex Data*, pages 189–207. Springer.

Goyal, P. and Ferrara, E. (2018a). GEM: A Python package for graph embedding methods. *Journal of Open Source Software*, 3(29).

Goyal, P. and Ferrara, E. (2018b). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Goyal, P., Huang, D., Goswami, A., Chhetri, S. R., Canedo, A., and Ferrara, E. (2019). Benchmarks for graph embedding evaluation. *CoRR*, abs/1908.06543.

Grover, A. and Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 855–864.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Representation Learning on Graphs: Methods and Applications. *IEEE Data Eng. Bull.*, 40(3):52–74.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *Introduction to Statistical Learning*, volume 112. Springer.

Khosla, M., Anand, A., and Setty, V. (2019). A comprehensive comparison of unsupervised network representation learning methods. *CoRR*, abs/1903.07902.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kou, G., Lu, Y., Peng, Y., and Shi, Y. (2012). Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, 11(01):197–225.

Leskovec, J. and Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, pages 539–547.

Li, L., Jamieson, K. G., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2016). Efficient Hyperparameter Optimization and Infinitely Many Armed Bandits. *CoRR*, abs/1603.06560.

Natural Language Processing Lab at Tsinghua University (2019). OpenNE: An open source toolkit for Network Embedding. https://github.com/thunlp/OpenNE. [accessed on 20-May-2019].

Newman, M. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). DeepWalk. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, pages 701–710, New York, New York, USA. ACM Press.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. (2018). Pitfalls of graph neural network evaluation. *CoRR*, abs/1811.05868.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee.

Tang, L. and Liu, H. (2009). Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 817–826. ACM.

Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234. ACM.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. Y. (2015). Network Representation Learning with Rich Text Information. In *IJCAI*, pages 2111–2117.

Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2018). Network representation learning: A survey. *IEEE Transactions on Big Data*.