

Analysis of the Robustness of Degree Centrality against Random Errors in Graphs

Sho Tsugawa¹ and Hiroyuki Ohsaki²

¹ University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
s-tugawa@cs.tsukuba.ac.jp

² Kwansei Gakuin University, Sanda, Hyogo 669-1337, Japan
ohsaki@kwansei.ac.jp

Abstract. Research on network analysis, which is used to analyze large-scale and complex networks such as social networks, protein networks, and brain function networks, has been actively pursued. Typically, the networks used for network analyses will contain multiple errors because it is not easy to accurately and completely identify the nodes to be analyzed and the appropriate relationships among them. In this paper, we analyze the robustness of centrality measure, which is widely used in network analyses, against missing nodes, missing links, and false links. We focus on the stability of node rankings based on degree centrality, and derive Top_m and Overlap_m , which evaluate the robustness of node rankings. Through extensive simulations, we show the validity of our analysis, and suggest that our model can be used to analyze the robustness of not only degree centrality but also other types of centrality measures. Moreover, by using our analytical models, we examine the robustness of degree centrality against random errors in graphs.

1 Introduction

Research on network analysis, which is used to analyze large-scale and complex networks such as social networks, protein networks, and brain function networks, has been actively pursued [1, 6, 8, 20–22]. In network analysis, relationships among entities in the real world are represented by a graph. In social network analysis (SNA), individuals are represented as nodes in a graph, and the social ties among them, such as similarities, social relations, interactions, and flows, are represented as links [6, 22]. In brain function network analysis, brain regions are represented as nodes, and temporal correlations in activity among them are represented as links [20].

Among various indices proposed for network analysis, centrality measures (e.g., degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality) [4, 11] have been widely used in actual analyses [3, 5, 25]. Centrality measures are indices that express the influence of one node on others, and such measures have been used for various purposes, such as discovering which person plays a central role in a community [3, 5] and inferring which brain regions are important for the task of interest [25].

Typically, the graphs used for network analyses will contain multiple errors because it is not easy to accurately and completely identify the entities to be analyzed and the appropriate relationships among them [7, 9, 14–16, 19]. For instance, graphs used in SNA

can contain several errors of different types, such as *missing nodes*, *missing links*, and *false links*. In traditional SNA, graphs are generated from the results of questionnaires, and so non-responses and inaccurate answers will cause such errors [24]. Even in recent SNA used for analyzing online social networks, such errors can be present due to sampling bias and restrictions on social network data, which is typically accessed by means of application programming interfaces. In biological network analyses, such as analyses of protein interaction networks and gene regulatory networks, graphs often contain errors such as missing links and false links as a result of measurement errors [19, 23].

Several analyses on the robustness of centrality measures used for network analyses against errors in the graphs (simulated as noise created by random addition and deletion of nodes and links) have been performed [7, 9, 12–17, 19]. In [7, 16], how centrality measures of nodes in networks are affected by the random addition and deletion of nodes and links is experimentally investigated. Robustness of centrality measures against link weight noises has also been experimentally investigated, such as in [13, 17].

Most existing studies use an experimental approach to understand the robustness of centrality measures, but some recent studies adopt a theoretical approach. Ghoshal *et al.* [12] analyze node-ranking stability based on the PageRank algorithm against random rewiring of links. Platig *et al.* [19] develop an analytical model to quantify the robustness of degree centrality against link errors (i.e., missing links and false links). They derive correlation coefficients r between the degree measures of the ground-truth graph and those of graphs with errors.

Our study builds on prior work and contributes to developing an analytical model that can be used to quantify the robustness of centrality measures. Since one of the most typical errors in network analysis is missing nodes [7, 9, 24], we extend the model of [19] to include these, and analyze the robustness of degree centrality against missing nodes as well as against missing links and false links. As discussed in the previous works [7, 19], centrality measures are used mainly for node ranking. We therefore focus on the stability of node ranking and derive Top_m and Overlap_m , which evaluate the robustness of node rankings [7, 16, 17, 19]. Through extensive simulations, we show the validity of our analysis. Moreover, by using our analytical models, we examine the robustness of centrality measures against random errors in graphs.

The remainder of this paper is organized as follows. Section 2 introduces related work. In Section 3, we analyze the robustness of degree centrality against three types of errors (i.e., missing nodes, missing links, and false links). Section 4 examines the validity of our analysis through comparison between numerical examples of our analysis and results of simulations, and also discusses the robustness of centrality measures against random errors in graphs. Finally, Section 5 contains our conclusions and a discussion of future work.

2 Related Work

Most existing studies use a simulation to understand the robustness of centrality measures by adding errors to a ground-truth graph and investigating the relation between the centrality measures of the ground-truth graph and those of the graphs with errors [7, 9, 14–16]. In contrast, some recent studies use a theoretical approach [12, 19].