Resampling-Based Framework for Estimating Node Centrality of Large Social Network

Kouzou Ohara[1], Kazumi Saito[2], Masahiro Kimura[3], and Hiroshi Motoda[4]

[1] Department of Integrated Information Technology, Aoyama Gakuin University, Japan
ohara@it.aoyama.ac.jp
[2] School of Administration and Informatics, University of Shizuoka, Japan
k-saito@u-shizuoka-ken.ac.jp
[3] Department of Electronics and Informatics, Ryukoku University, Japan
kimura@rins.ryukoku.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University, Japan
School of Computing and Information Systems, University of Tasmania, Australia
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We address a problem of efficiently estimating value of a centrality measure for a node in a large social network only using a partial network generated by sampling nodes from the entire network. To this end, we propose a resampling-based framework to estimate the approximation error defined as the difference between the true and the estimated values of the centrality. We experimentally evaluate the fundamental performance of the proposed framework using the closeness and betweenness centralities on three real world networks, and show that it allows us to estimate the approximation error more tightly and more precisely with the confidence level of 95% even for a small partial network compared with the standard error traditionally used, and that we could potentially identify top nodes and possibly rank them in a given centrality measure with high confidence level only from a small partial network.

Keywords: Error estimation, resampling, node centrality, social network analysis.

1 Introduction

Recently, Social Media such as Facebook, Digg, Twitter, Weblog, Wiki, etc. becomes increasingly popular on a worldwide scale, and allows us to construct large-scale social networks in cyberspace. An article that is posted on social media can rapidly and widely spread through such networks and can be shared by a large number of people. Since such information can substantially affect our thought and decision making, a large number of studies have been made by researchers in many different disciplines such as sociology, psychology, economy, and computer science [8,4] to analyze various aspects of social networks and information diffusion on them.

In the domain of social network analysis, several measures called centrality have been proposed so far [7,5,1,3,13]. They characterize nodes in a network based on its structure, and give an insight into network performance. For example, a centrality provides us with the information of how important each node is through node ranking

S. Džeroski et al. (Eds.): DS 2014, LNAI 8777, pp. 228–239, 2014.

derived directly from the centrality. It also provides us with topological features of a network. For example, scale free property is derived from the degree distribution. As a social network in World Wide Web easily grows in size, it is becoming pressingly important that we are able to efficiently compute values of a centrality to analyze such a large social network. However, if a centrality measure is based not only on local structure around a target node, e.g. its neighboring nodes, but also on global structure of a network, e.g. paths between arbitrary node pairs, its computation becomes harder as the size of the network increases. Thus, it is crucial to reduce the computational cost of such centralities for large social networks. Typical examples are the closeness and the betweenness centralities which we consider in this paper (explained later).

It is worth noting that such a centrality is usually defined as a summarized value of more primitive ones that are derived from node pairs in a network. For example, the closeness centrality is defined as the average of the shortest path lengths from a target node to each of the remaining nodes in a network. Considering this fact, it is inevitable to employ a sampling-based approach as a possible solution of this kind of problem on scalability. It is obvious that using only a limited number of nodes randomly sampled from a large social network can reduce the computational cost. However, the resulting value is an approximation of its true value, and thus it becomes important to accurately estimate the approximation error. It is well known from the statistical view point that the margin of error (difference between sample mean and population mean) is $\pm 2 \times \sigma / \sqrt{N}$ with the confidence level of 95%, where σ and N are the standard deviation of a population and the number of samples, respectively. However, this traditional boundary does not necessarily give us a tight approximation error.

In this paper, we propose a framework that provides us with a tighter error estimate of how close the approximation is to the true value. The basic idea is that we consider all possible partial networks of a fixed size that are generated by resampling nodes according to a given coverage ratio, and then estimate the approximation error, referred to as *resampling error*, using centrality values derived from those partial networks. We test our framework using two well-known centrality measures, the closeness and the betweenness centralities, both of which require to use the global structure of a network for computing the value of each node. Extensive experiments were performed on three real world social networks varying the sample coverage for each centrality measure. We empirically confirmed that the proposed framework is more promising than the traditional error bound in that it enables us to give a tighter approximation error with a higher confidence level than the traditional one under a given sampling ratio. The framework we proposed is not specific to computation of node centralities for social network analysis. It is very generic and is applicable to any other estimation problems that require aggregation of many (but a finite number of) primitive computations.

The paper is organized as follows. Section 2 gives the formal definitions of both the resampling-based framework that we propose and the traditional bound of approximation error. Section 3 explains the closeness and the betweenness centralities we used to evaluate our framework and presents how to estimate their approximation error. Section 4 reports experimental results for these centralities on three real world networks. Section 5 concludes this paper and addresses the future work.