Estimating a Variant of Random Walk Centrality on Graphs

Sung-ho Justin Oh (Capstone Abstract)

Determining the important members of a network is an important area of study with many practical applications. With such knowledge, one can effectively inform communities of emergencies or advertise events and products. We focus specifically on random walk centrality as a measure of importance of webpages. For graph $G = (V, E)$ the random walk centrality of a node $v \in V$ is the sum over all $u \neq v$ of the expected time for a random walk starting from node u to first reach v. One can use this measure of centrality to model the average time it takes for a web surfer to reach a webpage via hyperlinks, and to determine which webpages are most valuable in placing information such as advertisements. Moreover this measure is easily computed given the adjacency matrix of the graph and with some linear algebra. However, there are discrepancies between our definition of random walk centrality, and the web surfing model we intend it for. One major issue is that a random walk may repeatedly traverse the same nodes, but the average human web surfer will rarely click the same link (a "purple" link) twice.

We thus define a variant of random walk centrality in which we instead calculate the expected time for a *simple* random walk to first reach a node v. In the simple random walk model, the walk progresses by selecting an *unvisited* node uniformly at random from the neighbors of the current node. This new variant is no longer easily computable with linear algebra, and we conjecture that computing it exactly is hard. We first present a study of the relationship between the original definition of random walk centrality and our new variant. We then discuss how to randomly sample random walks in order to estimate this measure of centrality, and provide approaches that can estimate the centrality of every node in a graph to within additive error ϵ with probability $1 - \delta$ in $O\left(\frac{|V|^3}{\epsilon^2} \ln \frac{1}{\delta}\right)$ samples. Finally, we discuss how to extend our definition of centrality of single nodes to a definition of centrality for subsets of k nodes. We show that finding the subset of k nodes of optimal centrality is NP-hard, and provide techniques for finding approximate solutions to this NP-hard problem. The discussion ends with a suggestion of further variants of random walk centrality that may be worth studying.