Uncovering important intermediate publications

The importance of nodes in a network is of considerable interest. Much research has focused on different types of centralities. We here discuss the case where we want to assess the importance of a node in connecting two other nodes, which we call the intermediacy. This is particularly relevant in citation networks, where we want to uncover the relative importance of publications. In this abstract, we therefore limit ourselves to directed acyclic graphs.

Let G=(V,E) be a directed acyclic graph with nodes V and directed edges E. We are provided with two nodes s and t and we want to determine how important nodes are for getting from node s to node t. We use a probabilistic framework to assess the importance of a node. With probability p each edge is said to be active, and with probability 1 - p each edge is inactive. Intermediacy is then defined as the probability that there is a path of only active edges from s to t that passes through node v.

Intermediacy obviously depends on the probability p that an edge is active. We prove that intermediacy has a quite simple intuitive understanding in the two extremities of p going to 0 and p going to 1. For p going to 0 intermediacy is determined by the path length: the shorter the shortest path from s to t going through v, the higher the intermediacy of v. For p going to 1 intermediacy is determined by the number of edge-independent paths: the larger the number of edge-independent paths from s to t going through v, the higher the intermediacy of v. Intermediate values of p interpolate between these two extremities, and both path length and the multitude of (edge-independent) paths affect intermediacy.

We have devised an exact algorithm based on a decomposition of the probability of the existence of a path. This decomposition yields an algorithm based on contracting and removing edges. Although the algorithm can be relatively efficiently implemented, it runs in exponential time, so that it can only be applied to relatively small graphs. Given the problem itself is NP-hard, even for directed acyclic graphs, it is unlikely that a more efficient algorithm exists. We therefore also developed an efficient Monte Carlo algorithm for calculating approximate intermediacy scores. We do so by repeatedly performing depth-first searches where each edge is considered with probability p, which runs in linear time.

Intermediacy is of particular interest when uncovering relevant intermediate publications in citation networks. Main path analysis is another method that is often used for determining relevant intermediate publications. The two methods provide quite different results: main path analysis tends to favour longer paths, whereas intermediacy tends to favour shorter paths.

Finally, we illustrate our method on some case studies. We find that publications with the highest intermediacy seem to reflect well the intellectual development.