Intermediacy of publications

L. Šubelj*, L. Waltman†, V.A. Traag†, N.J. van Eck†
* Faculty of Computer and Information Science, University of Ljubljana, Slovenia
† Centre for Science and Technology Studies, Leiden University, the Netherlands

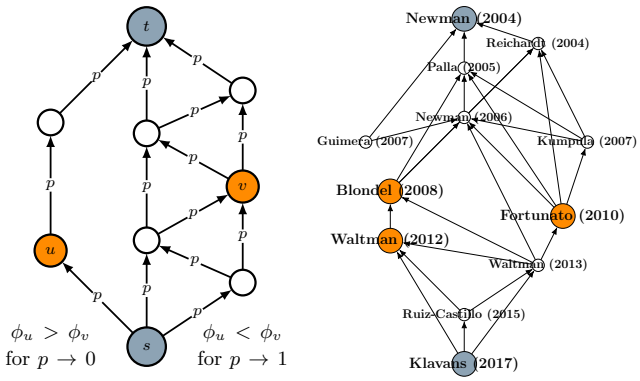
THE importance of nodes in a network is of considerable interest. Much research has focused on different types of centralities. We here discuss the case where we want to assess the importance of a node in connecting two other nodes, which we call the *intermediacy*. This is particularly relevant in citation 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 ϕ_v 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 $p \to 0$ and $p \to 1$, which is illustrated in Fig. 1(a). For $p \to 0$ 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 $p \to 1$ 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 [1], 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 for uncovering intermediate publications in citation networks. Main path analysis is another method that is often used for determining relevant intermediate publications [2]. The two methods provide quite different results: main path analysis tends to favour longer paths, whereas intermediacy tends to favour shorter paths.



(a) Illustration (b) Example

Fig. 1: Illustration of extremities and top ten most intermediate publications based for $p = 0.25$ between Newman (2004) and Klavans & Boyack (2017).

We illustrate our method on the development in the scientific literature of using community detection for publication classification. In particular, we are interested in the intermediate steps of how modularity in 2004 [3] leads to a study in publication classification in citation networks in 2017 [4] (see Fig. 1(b)). In total there are more than 63 000 publications in Web of Science in between the former and the latter publication. We find that the top ten publications with the highest intermediacy for $p = 0.25$ indeed reflect well the intellectual development. The publication with the highest intermediacy is a well-known review article of community detection, the second most intermediate publication introduced modularity for publication classification and the third most intermediate publication presented the Louvain algorithm. Overall, intermediacy does not simply reflect the number of citations received by a publication: the publication with the sixth highest intermediacy is cited only two times by the other intermediate publications.

References

[1] R. Johnson, "Network reliability and acyclic orientations," *Networks*, vol. 14, no. 4, pp. 489–505, 1984.
[2] N. P. Hummon and P. Doreian, "Connectivity in a citation network: The development of DNA theory," *Soc. Networks*, vol. 11, no. 1, pp. 39–63, 1989.
[3] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, pp. 026 113+, feb 2004.
[4] R. Klavans and K. W. Boyack, "Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?" *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 984–998, apr 2017.