5^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2019, University of Amsterdam, The Netherlands

Intermediacy of publications

L. Šubelj[1], L. Waltman[2], V.A. Traag[2], and N.J. van Eck[2]

[1]Faculty of Computer and Information Science
University of Ljubljana, Slovenia
[2]Centre for Science and Technology Studies (CWTS)
Leiden University, the Netherlands

citation network, main path analysis, intermediate publications

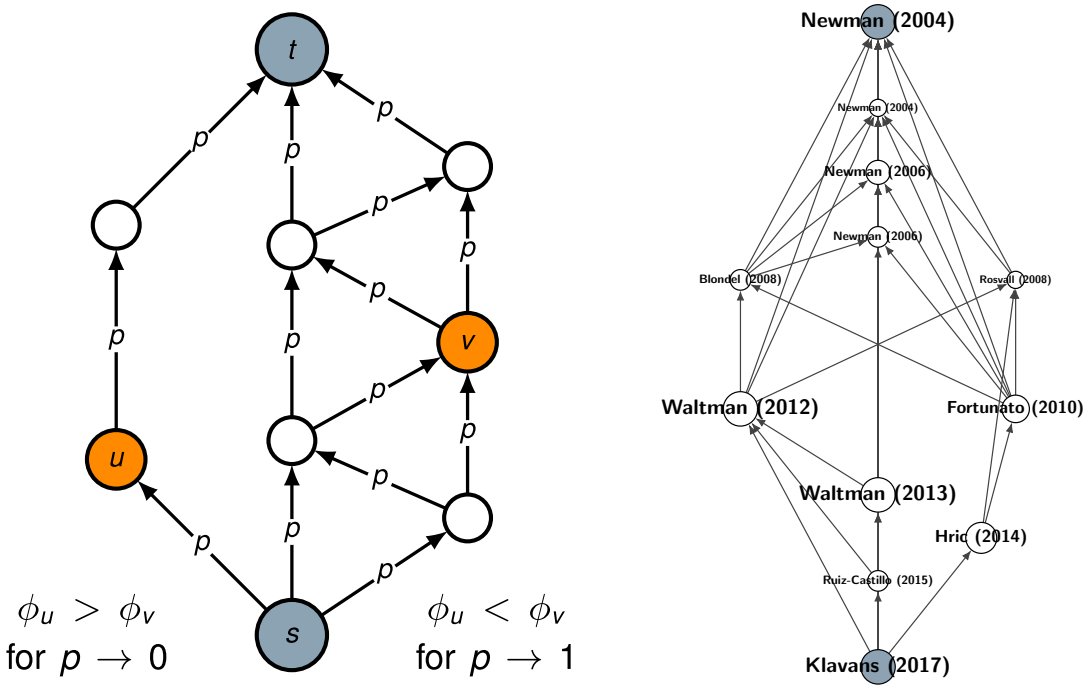
Citation networks provide invaluable information for tracing historical developments in science. The idea of tracing scientific developments based on citation data goes back to Eugene Garfield, the founder of the Science Citation Index. In a report published more than 50 years ago, Garfield and his co-workers concluded that citation analysis is "a valid and valuable means of creating accurate historical descriptions of scientific fields" [1]. Main path analysis, originally proposed by Hummon and Doreian [2], is a widely used technique for tracing historical developments in science.

We here introduce a new approach for tracing historical developments in science based on citation networks. We propose a measure called intermediacy. Given two publications dealing with a specific research topic, an older publication and a more recent one, intermediacy can be used to identify publications that appear to play a major role in the historical development from the older to the more recent publication. There are fundamental differences between intermediacy and main path analysis. Whereas main path analysis favors longer citation paths over shorter ones, intermediacy has the opposite tendency.

Consider a directed acyclic graph $G = (V,E)$, with nodes V and edges E. We are interested in the connectivity between a source node $s \in V$ and a target node $t \in V$. We assume that each edge $e \in E$ is active with the same probability p. We define *intermediacy* ϕ_v of a node $v \in V$ as the probability that there is a path from s to t consisting of only active edges. Determining this probability is equivalent to the problem of network reliability, which is known to be NP-hard [3]. We have constructed a Monte Carlo algorithm to approximate this probability.

In the limit of $p \to 0$, nodes that are located on shorter source-target paths have a higher intermediacy than nodes located on longer source-target paths (Fig. 1a). In the limit of $p \to 1$, nodes that are located on a larger number of edge independent source-target paths have a higher intermediacy than nodes located on a smaller number of edge independent source-target paths. Values of p between 0 and 1 interpolate in some way between the two extremes. For lower values of p the path lengths are relatively more important, whereas for higher values of p the number of edge independent paths are more important. Intermediacy also has two important properties. The addition of new paths will always increase intermediacy. Similarly, contracting existing paths will always increase intermediacy.

This differs clearly from main path analysis [2], which favors longer paths over shorter ones, and hence violates the path contraction property. We consider this behavior of main path analysis to be undesirable. Instead of focusing on the probability of the existence of at least



(a) Limit behavior

(b) Case study

Figure 1: Intermediacy in theory and in practice.

one active source-target path, as is done by intermediacy, one could also focus on the expected number of active source-target paths. That alternative approach suffers from the same problem as main path analysis, and also violates the path contraction property.

To illustrate the use of intermediacy empirically, we performed a case study (Fig. 1b). We analyze how a method for community detection in networks ended up being used in the field of scientometrics to construct classification systems of scientific publications. In particular, we are interested in the development from Newman & Girvan (2004) to Klavans & Boyack (2017). We use our expert knowledge to interpret the results. The two publications with the highest intermediacy (Waltman & Van Eck, 2012, 2013) played a key role in introducing modularity-based approaches in the scientometric community. Waltman & Van Eck (2012) proposed the use of modularity-based approaches for constructing classification systems of scientific publications, while Waltman & Van Eck (2013) introduced an algorithm for implementing these modularity-based approaches. This algorithm can be seen as an improvement of the so-called Louvain algorithm introduced by Blondel et al. (2008), which is also among the ten most intermediate publications. Most of the other publications are classical publications on community detection in general and modularity in particular. Fortunato (2010) is a review of the literature on community detection. The intermediacy of this publication is probably strongly influenced by its large number of references. We note that a ranking of publications based on intermediacy is quite different from a ranking based on the number of citations.

In conclusion, intermediacy, introduced in this paper, offers an alternative to main path analysis. It provides a principled approach for identifying publications that appear to play a major role in the historical development from an older to a more recent publication. It has two intuitively desirable properties, referred to as path addition and path contraction, and favors shorter paths over longer ones. This is a fundamental difference with main path analysis.

References

[1] E. Garfield, I. Sher, and R. Torpie, *The use of citation data in writing the history of science*, Tech. Rep. F49(638)-1256 (The Institute for Scientific Information, 1964).

[2] N. Hummon and P. Doreian, Social Networks **11**, 39 (1989).

[3] M. Ball, Networks **10**, 153 (1980).