3 Empirical illustration

We now present two cases that serve as empirical illustrations of the notion of intermediacy. The first case deals with the topic of community detection and its relationship with scientometric research. This case was selected because we are well acquainted with the topic. The second case deals with the topic of peer review. This case is of interest because it was recently examined using main path analysis \cite{Batagelj2017}. Hence, this case enables us to demonstrate some of the differences between intermediacy and main path analysis.

3.1 Case 1: Community detection and scientometrics

In the first case that we consider, 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 and Girvan (2004) to Klavans and Boyack (2017). These are our target and source publications. Newman and Girvan (2004) introduced a new metric for community detection in networks, known as modularity, while Klavans and Boyack (2017) compared different ways in which modularity-based approaches can be used to identify communities in citation networks.

Our analysis relies on data from the Scopus database produced by Elsevier. We also considered the Web of Science database produced by Clarivate Analytics. However, many citation links relevant for our analysis turned out to be missing in the Web of Science database. There are also missing citation links in the Scopus database, but for Scopus the problem seemed less significant than for Web of Science. This is why we work with the Scopus database. We refer to \citet{Vaneck2017} for a further discussion of the problem of missing citation links.

In the Scopus database, we found $n = 64\,223$ publications that are located on a citation path between our source and target publications. In total, we identified $m = 280\,033$ citation links between these publications. This means that on average each publication has $k = 2m / n = 8.72$ citation links, counting both incoming and outgoing links. We used our Monte Carlo algorithm to calculate the intermediacy of each publication for different values of the parameter $p$.

Fig.~\ref{} presents some high-level results. The top-left plot shows how the probability of the existence of an active path between the source and target publications depends on the parameter $p$. This probability increases from a value of zero for $p = 0$ to a value close to one starting from $p = 0.25$. The vertical line indicates the value of $p$ obtained using the rule of thumb from percolation theory, as discussed in Section~\ref{}. This rule of thumb yields $p = 1 / k = 0.11$, resulting in a probability of about $0.4$ for the existence of an active path between the source and target publications.

For five different values of the parameter $p$, the top-right plot shows the cumulative distribution of the intermediacy scores of our $n = 64\,223$ publications. As is to be expected, when $p$ is close to zero, intermediacy scores are extremely small. For instance, for $p = 0.1$, almost all intermediacy scores are below $10^{-6}$. On the other hand, when $p$ is getting close to one, intermediacy scores also approach one. For $p = 0.9$, for instance, there are hardly any intermediacy scores below $0.1$.

The matrices in the bottom of Fig.~\ref{} provide the correlations between the intermediacy scores obtained for five different values of the parameter $p$. Correlations with citation counts and reference counts are provided as well. (We use the term \emph{citation count} to refer to the number of incoming citation links of a publication, while we use the term \emph{reference count} to refer to the number of outgoing citation links of a publication. Only citation links located on a citation path between the source and target publications are counted.) The left matrix reports Spearman correlations, while the right matrix reports Pearson correlations. We consider intermediacy scores to be most useful from an ordinal (as opposed to a cardinal) perspective. From this point of view, Spearman correlations are more relevant than Pearson correlations, but for completeness we report both types of correlations. The Spearman correlations show that values of $0.3$, $0.5$, $0.7$, and $0.9$ for $p$ all yield fairly similar rankings of publications in terms of intermediacy. However, the ranking obtained for $p = 0.1$ is quite different. Pearson correlations tend to be lower than Spearman correlations. Hence, even when different values of $p$ yield similar rankings of publications, there usually does not exist a simple linear relationship between the intermediacy scores. Finally, regardless of the value of $p$, it turns out that intermediacy scores are not very strongly correlated with citation counts and reference counts.

We now explore our results in more detail. Based on our expert knowledge of the topic under study, we found that the most useful results were obtained by setting the parameter $p$ equal to $0.1$. Table~\ref{} lists the ten publications with the highest intermediacy for $p = 0.1$. For each publication, the intermediacy is reported for five different values of $p$. In addition, the table also reports each publication’s citation count and reference count. Fig.~\ref{} shows the citation network of the ten most intermediate publications for $p = 0.1$.

What do we learn from the results presented in Table~\ref{} and Fig.~\ref{}? 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 and Van Eck (2012) proposed the use of modularity-based approaches for constructing classification systems of scientific publications, while Waltman and 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 in Table~\ref{} and Fig.~\ref{} are classical publications on community detection in general and modularity in particular. This applies to the three publications by Newman and also to Rosvall and Bergstrom (2008) and Fortunato (2010). The publications by Newman all deal with modularity-based community detection. Rosvall and Bergstrom (2008) proposed an alternative approach to community detection. They applied their approach to a citation network of scientific journals, which explains the connection with the scientometric literature. 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. Hric et al. (2014) is a more recent publication on community detection. This publication focuses on the challenges of evaluating the results produced by community detection methods. This issue is very relevant in a scientometric context, and therefore the publication was cited by our source publication (Klavans & Boyack, 2017). Finally, there is one more scientometric publication in Table~\ref{} and Fig.~\ref{}. This publication (Ruiz-Castillo & Waltman, 2015) is one of the first studies presenting a scientometric application of classification systems of scientific publications constructed using a modularity-based approach. The publication was also cited by our source publication.

The citation counts reported in Table~\ref{} show that some publications, especially the more recent ones, have a high intermediacy even though they have been cited only a very limited number of times. This makes clear that a ranking of publications based on intermediacy is quite different from a citation-based ranking of publications. The publications in Table~\ref{} that have a high intermediacy and a small number of citations do have a substantial number of references.

Fig. ???. High-level results for case 1. Top left: Probability of the existence of an active path between the source and target publications as a function of the parameter $p$. Top right: Cumulative distribution of intermediacy scores for different values of $p$. Bottom: Spearman and Pearson correlations between intermediacy scores (for different values of $p$), citation counts, and reference counts.

Table ???. Top 10 most intermediate publications (for $p = 0.1$) in case 1.

Fig. ???. Citation network of the top 10 most intermediate publications (for $p = 0.1$) in case 1 (left) and case 2 (right).

3.2 Case 2: Peer review

We now turn to our second case, in which we analyze the literature on peer review. The second case is based on data from the Web of Science database. We make use of the same data that was also used in a recent paper by \citet{Batagelj2017}. The authors of this paper kindly made their data available to us.

We started with a citation network of $45\,965$ publications dealing with peer review. This is the citation network that was labeled CiteAcy by \citet{Batagelj2017}. We selected Cole and Cole (1967) and Garcia et al. (2015) as our target and source publications. The main path analysis carried out by \citet{Batagelj2017} suggests that these are central publications in the literature on peer review. For the purpose of our analysis, only publications located on a citation path between our source and target publications are of relevance. Other publications play no role in the analysis. We therefore restricted the analysis to the $n = 615$ publications located on a citation path from Garcia et al. (2015) to Cole and Cole (1967). These publications turned out to be connected by $m = 3\,420$ citation links, resulting in an average of $k = 2m / n = 11.12$ citation links per publication. Like in the first case discussed above, our Monte Carlo algorithm was used to calculate intermediacy scores for different values of the parameter $p$.

Fig.~\ref{} presents high-level results. These results are quite similar to the corresponding results for our first case, presented in Fig.~\ref{}. As can be seen in the top-left plot, the rule of thumb from percolation theory yields $p = 1 / k = 0.09$, which is close to the value of $0.11$ obtained in our first case. However, the probability of the existence of an active path between the source and target publications equals $$, which is much lower than the probability of $0.4$ in our first case. Furthermore, intermediacy scores tend to be higher in our second case than in our first case. This can be seen by comparing the top-right plot in Fig.~\ref{} to the corresponding plot in Fig.~\ref{}. We note that the former plot has a linear horizontal axis, while the horizontal axis in the latter plot is logarithmic.

Fig. ???. High-level results for case 2. Top left: Probability of the existence of an active path between the source and target publications as a function of the parameter $p$. Top right: Cumulative distribution of intermediacy scores for different values of $p$. Bottom: Spearman and Pearson correlations between intermediacy scores (for different values of $p$), citation counts, and reference counts.

Table ???. Top 10 most intermediate publications (for $p = 0.1$) in case 2.