Intermediacy of publications

Lovro Šubelj[a,1], **Ludo Waltman**[b], **Vincent Traag**[b], **and Nees Jan van Eck**[b]

[a]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia; [b]Leiden University, Centre for Science and Technology Studies, P.O. Box 905, 2300 AX Leiden, The Netherlands

This manuscript was compiled on December 19, 2018

Citation networks of scientific publications offer fundamental insights into the structure and development of scientific knowledge. We propose a new measure, called intermediacy, for tracing the historical development of scientific knowledge. Given two publications, an older and a more recent one, intermediacy identifies publications that seem to play a major role in the historical development from the older to the more recent publication. The identified publications are important in connecting the older and the more recent publication in the citation network. After providing a formal definition of intermediacy, we study its mathematical properties. We then present two empirical case studies, one tracing historical developments at the interface between the community detection and the scientometric literature and one examining the development of the literature on peer review. We show both mathematically and empirically how intermediacy differs from main path analysis, which is the most popular approach for tracing historical developments in citation networks. Main path analysis tends to favor longer paths over shorter ones, whereas intermediacy has the opposite tendency. Compared to main path analysis, we conclude that intermediacy offers a more principled approach for tracing the historical development of scientific knowledge.

intermediacy | publication | citation network | main path analysis

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). Garfield also developed a software tool called HistCite that visualizes citation networks of scientific publications. This tool supports users in tracing historical developments in science, a process sometimes referred to as *algorithmic historiography* by Garfield (2–4). More recently, a software tool called CitNetExplorer (5) was developed that has similar functionality but offers more flexibility in analyzing large-scale citation networks. Other software tools, most notably CiteSpace (6) and CRExplorer (7, 8), provide alternative approaches for tracing scientific developments based on citation data.

Main path analysis, originally proposed by Hummon and Doreian (9), is a widely used technique for tracing historical developments in science. Given a citation network, main path analysis identifies one or more paths in the network that are considered to represent the most important scientific developments. Many variants and extensions of main path analysis have been proposed (10–16), not only for citation networks of scientific publications but also for patent citation networks (17–21).

In this paper, we 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. These are publications that, based on citation links, are important in connecting the older and the more recent publication.

Like main path analysis, intermediacy can be used to identify one or more citation paths between two publications. However, as we will make clear, there are fundamental differences between intermediacy and main path analysis. Most significantly, we will show that main path analysis tends to favor longer citation paths over shorter ones, whereas intermediacy has the opposite tendency. For the purpose of tracing historical developments in science, we argue that intermediacy yields better results than main path analysis.

Intermediacy

Consider a directed acyclic graph $G = (V, E)$, where V denotes the set of nodes of G and E denotes the set of edges of G. The edges are directed. We are interested in the connectivity between a source $s \in V$ and a target $t \in V$. Only nodes that are located on a path from source s to target t are of relevance. We refer to such a path as a source-target path. We assume that each node $v \in V$ is located on a source-target path.

Definition 1. Given a source s and a target t, a path from s to t is called a *source-target path*.

In this paper, our focus is on citation networks of scientific publications. In this context, nodes are publications and

Significance Statement

Researchers spend a lot of time keeping track of the literature in their field. Computational methods can be used to increase the efficiency with which researchers study the literature. We propose a method called intermediacy that enables tracing the historical development of scientific knowledge. Based on citation relations, intermediacy aims to identify publications that play a major role in the historical development from an older publication to a more recent one. Main path analysis currently is the most commonly used approach for addressing this problem. We show the advantages of intermediacy over main path analysis. When implemented in interactive search interfaces, intermediacy may help to significantly increase the efficiency with which researchers study the literature in their field.

Author contributions: L.Š., L.W., V.T., and N.J.E. designed research; L.Š., L.W., V.T., and N.J.E. performed research; L.Š., V.T., and N.J.E. analyzed data; and L.W. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: lovro.subelj@fri.uni-lj.si.

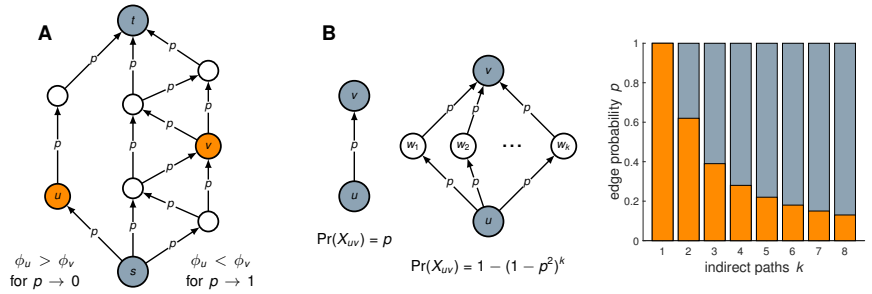

Fig. 1. (**A**) Illustration of the limit behavior of intermediacy. For $p \to 0$, intermediacy favors nodes located on shorter paths and therefore node u has a higher intermediacy than node v. For $p \to 1$, intermediacy favors nodes located on a larger number of edge independent paths and therefore node v has a higher intermediacy than node u. (**B**) Illustration of the choice of the parameter p. Nodes u and v are connected by a single direct path in the left graph and by k indirect paths of length 2 in the right graph. For different values of k, the bar chart shows the values of p for which the probability that there is an active path from node u to node v is higher (in orange) or lower (in gray) in the left graph than in the right graph.

edges are citations. We choose edges to be directed from a citing publication to a cited publication. Hence, edges point backward in time. This means that the source is a more recent publication and the target an older one.

Informally, the more important the role of a node $v \in V$ in connecting source s to target t, the higher the intermediacy of v. To formally define intermediacy, we assume that each edge $e \in E$ is active with a certain probability p. We assume that the probability of being active is the same for all edges $e \in E$. Based on the idea of active and inactive edges, we introduce the following definitions.

Definition 2. If all edges on a path are active, the path is called *active*. Otherwise the path is called *inactive*. If a node $v \in V$ is located on an active source-target path, the node is called *active*. Otherwise the node is called *inactive*.

For two nodes $u, v \in V$, we use X_{uv} to indicate whether there is an active path (or multiple active paths) from node u to node v ($X_{uv} = 1$) or not ($X_{uv} = 0$). The probability that there is an active path from node u to node v is denoted by $\Pr(X_{uv} = 1)$. We use $X_{st}(v)$ to indicate whether there is an active source-target path that goes through node v ($X_{st}(v) = 1$) or not ($X_{st}(v) = 0$). The probability that there is an active source-target path that goes through node v is denoted by $\Pr(X_{st}(v) = 1) = \Pr(X_{sv} = 1)\Pr(X_{vt} = 1)$. This probability equals the probability that node v is active.

Intermediacy can now be defined as follows.

Definition 3. The *intermediacy* ϕ_v of a node $v \in V$ is the probability that v is active, that is,

$$\phi_v = \Pr(X_{st}(v) = 1) = \Pr(X_{sv} = 1)\Pr(X_{vt} = 1). \quad [1]$$

In the interpretation of intermediacy, we focus on the ranking of nodes relative to each other. We do not consider the absolute values of intermediacy. For instance, suppose the intermediacy of node $v \in V$ is twice as high as the intermediacy of node $u \in V$. We then consider node v to be more important than node u in connecting the source s and the target t. However, we do not consider node v to be twice as important as node u.

We now present an analysis of the mathematical properties of intermediacy. The proofs of the mathematical results provided below can be found in the Materials and Methods section.

Limit behavior. To get a better understanding of intermediacy, we study the behavior of intermediacy in two limit cases, namely the case in which the probability p that an edge is active goes to 0 and the case in which the probability p goes to 1. In each of the two cases, the ranking of the nodes in a graph based on intermediacy turns out to have a natural interpretation. The difference between the two cases is illustrated in Fig. 1A.

Let ℓ_v denote the length of the shortest source-target path going through node $v \in V$. The following theorem states that in the limit as the probability p that an edge is active tends to 0, the ranking of nodes based on intermediacy coincides with the ranking based on ℓ_v. Nodes located on shorter source-target paths are more intermediate than nodes located on longer source-target paths.

Theorem 1. In the limit as the probability p tends to 0, $\ell_u < \ell_v$ implies $\phi_u > \phi_v$.

The intuition underlying this theorem is as follows. When the probability that an edge is active is close to 0, almost all edges are inactive. Consequently, almost all source-target paths are inactive as well. However, from a relative point of view, longer source-target paths are more likely to be inactive than shorter source-target paths. This means that nodes located on shorter source-target paths are more likely to be active than nodes located on longer source-target paths (even though for all nodes the probability of being active is close to 0). Nodes located on shorter source-target paths therefore have a higher intermediacy than nodes located on longer source-target paths.

We now consider the limit case in which the probability p that an edge is active goes to 1. Let σ_v denote the number of edge independent source-target paths going through node $v \in V$. Theorem 2 states that in the limit as p tends to 1, the ranking of nodes based on intermediacy coincides with the ranking based on σ_v. The larger the number of edge independent source-target paths going through a node, the higher the intermediacy of the node.

Theorem 2. In the limit as the probability p tends to 1, $\sigma_u > \sigma_v$ implies $\phi_u > \phi_v$.

Intuitively, this theorem can be understood as follows. When the probability that an edge is active is close to 1, almost all edges are active. Consequently, almost all source-target paths are active as well, and so are almost all nodes. A node is inactive only if all source-target paths going through the node are inactive. If there are σ edge independent source-target paths that go through a node, this means that the node can be inactive only if there are at least σ inactive edges. Consider two nodes $u, v \in V$. Suppose that the number of edge independent source-target paths going through node v is larger than the number of edge independent source-target paths going through node u. In order to be inactive, node v then requires more inactive edges than node u. This means that node v is less likely to be inactive than node u (even

though for both nodes the probability of being inactive is close to 0). Hence, node v has a higher intermediacy than node u. More generally, nodes 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.

Parameter choice. The probability p that an edge is active is a free parameter of intermediacy for which one needs to choose an appropriate value. The results presented above are concerned with the behavior of intermediacy in the limit cases in which the probability p tends to either 0 or 1. Fig. 1B provides some insight into the behavior of intermediacy for values of the probability p that are in between these two extremes. The figure shows two graphs. In the left graph, there is a direct path (i.e., a path of length 1) from node u to node v. There are no indirect paths. In this graph, the probability that there is an active path from u to node v equals p. In the right graph, there is no direct path from node u to node v, but there are k indirect paths of length 2. Each of these paths has a probability of p^2 of being active. Consequently, the probability that there is at least one active path from node u to node v equals $1-(1-p^2)^k$. The bar chart in Fig. 1B shows for different values of k the values of p for which the probability that there is an active path from node u to node v is higher (in orange) or lower (in gray) in the left graph than in the right graph. For instance, suppose that $k=5$. For $p<0.22$, the probability that there is an active path from node u to node v is higher in the left graph than in the right graph. For $p>0.22$, the situation is the other way around. If the probability p that an edge is active is set to 0.22, a direct path between two nodes is considered equally strong as 5 indirect paths of length 2. Based on Fig. 1B, one can set the probability p to a value that one considers appropriate for a particular analysis.

Path addition and contraction. Next, we study two additional properties of intermediacy, the property of path addition and the property of path contraction. We show that both adding paths and contracting paths lead to an increase in intermediacy. Path addition and path contraction are important properties because they reflect the basic intuition underlying the idea of intermediacy.

We start by considering the property of path addition. We define path addition as follows.

Definition 4. Consider a directed acyclic graph $G=(V,E)$ and two nodes $u,v\in V$ such that there does not exist a path from node v to node u. *Path addition* is the operation in which a new path from node u to node v is added. Let ℓ denote the length of the new path. If $\ell=1$, an edge (u,v) is added. If $\ell>1$, nodes $w_1,\ldots,w_{\ell-1}$ and edges $(u,w_1),(w_1,w_2),\ldots,(w_{\ell-2},w_{\ell-1}),(w_{\ell-1},v)$ are added.

This definition includes the condition that there does not exist a path from node v to node u. This condition ensures that the graph G will remain acyclic after adding a path. The following theorem states that adding a path increases intermediacy.

Theorem 3. Consider a directed acyclic graph $G=(V,E)$, a source $s\in V$, and a target $t\in V$. In addition, consider two nodes $u,v\in V$ such that there does not exist a path from node v to node u. Adding a path from node u to node v increases the intermediacy ϕ_w of any node $w\in V$ located on a path from source s to node u or from node v to target t.

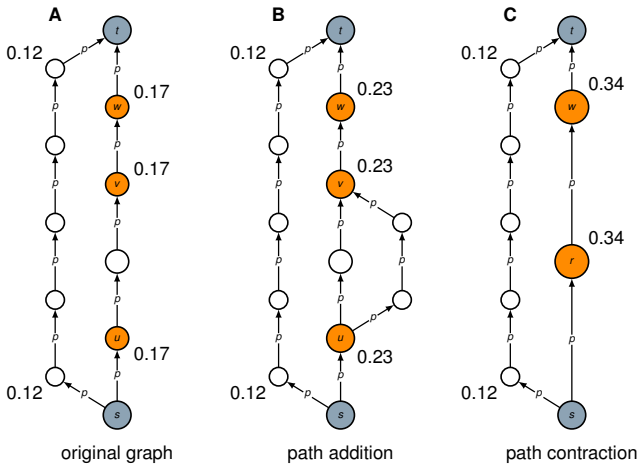

Fig. 2. Illustration of the properties of path addition and path contraction. Comparing (**B**) to (**A**) shows how path addition increases intermediacy. Comparing (**C**) to (**B**) shows how path contraction increases intermediacy. For some nodes in (**A**), (**B**), and (**C**), the intermediacy is reported, calculated using a value of 0.7 for the probability p.

Theorem 3 does not depend on the probability p. Adding a path always increases intermediacy, regardless of the value of p. To illustrate the theorem, consider Fig. 2A and Fig. 2B. The graph in Fig. 2B is identical to the one in Fig. 2A except that a path from node u to node v has been added. As can be seen, adding this path has increased the intermediacy of nodes located between source s and node u or between node v and target t, including nodes u and v themselves. While the intermediacy of other nodes has not changed, the intermediacy of these nodes has increased from 0.17 to 0.23. This reflects the basic intuition that, after a path from node u to node v has been added, going from source s to target t through nodes u and v has become 'easier' than it was before. This means that nodes located between source s and node u or between node v and target t have become more important in connecting the source and the target. Consequently, the intermediacy of these nodes has increased.

We now consider the property of path contraction. We use V_{uv} to denote the set of all nodes located on a path from node u to node v, including nodes u and v themselves. Path contraction is then defined as follows.

Definition 5. Consider a directed acyclic graph $G=(V,E)$ and two nodes $u,v\in V$ such that there exists at least one path from node u to node v. *Path contraction* is the operation in which all nodes in V_{uv} are contracted. This means that the nodes in V_{uv} are replaced by a new node r. Edges pointing from a node $w\notin V_{uv}$ to nodes in V_{uv} are replaced by a single new edge (w,r). Edges pointing from nodes in V_{uv} to a node $w\notin V_{uv}$ are replaced by a single new edge (r,w). Edges between nodes in V_{uv} are removed.

The following theorem states that contracting paths increases intermediacy.

Theorem 4. Consider a directed acyclic graph $G=(V,E)$, a source $s\in V$, and a target $t\in V$. In addition, consider two nodes $u,v\in V$ such that there exists at least one path from

node u to node v and such that nodes in V_{uv} do not have neighbors outside V_{uv} except for incoming neighbors of node u and outgoing neighbors of node v. Contracting paths from node u to node v increases the intermediacy ϕ_w of any node $w \in V$ located on a path from source s to node u or from node v to target t.

Like Theorem 3, Theorem 4 does not depend on the probability p. Theorem 4 is illustrated in Fig. 2B and Fig. 2C. The graph in Fig. 2C is identical to the one in Fig. 2B except that paths from node u to node v have been contracted. As a result, there has been an increase in the intermediacy of nodes located between source s and node u or between node v and target t, including nodes u and v themselves (which have been contracted into a new node r). While the intermediacy of other nodes has not changed, the intermediacy of these nodes has increased from 0.23 to 0.34. This reflects the basic intuition that, after paths from node u to node v have been contracted, going from source s to target t through nodes u and v has become 'easier' than it was before. In other words, nodes located on a path from source s to target t going through nodes u and v have become more important in connecting the source and the target, and hence the intermediacy of these nodes has increased.

Alternative approaches. How does intermediacy differ from alternative approaches? We consider two alternative approaches. One is main path analysis (9). This is the most commonly used approach for tracing the historical development of scientific knowledge in citation networks. The other alternative approach is the expected path count approach. Like intermediacy, the expected path count approach distinguishes between active and inactive edges and focuses on active source-target paths. While intermediacy considers the probability that there is at least one active source-target path going through a node, the expected path count approach considers the expected number of active source-target paths that go through a node.

Consider the graph shown in Fig. 3A. To get from source s to target t, one could take either a path going through nodes u and v or the path going through node w. Based on intermediacy, the latter path represents a stronger connection between the source and the target than the former one. This follows from the path contraction property.

Interestingly, main path analysis gives the opposite result, as can be seen in Fig. 3B. For each edge, the figure shows the search path count, which is the number of source-target paths that go through the edge. There are two source-target paths that go through (s, u) and (v, t), while all other edges are included only in a single source-target path. Because the search path counts of (s, u) and (v, t) are higher than the search path counts of (s, w) and (w, t), main path analysis favors paths going through nodes u and v over the path going through node w. This is exactly opposite to the result obtained using intermediacy. Fig. 3B makes clear that main path analysis yields outcomes that violate the path contraction property. Main path analysis tends to favor longer paths over shorter ones. For the purpose of identifying publications that play an important role in connecting an older and a more recent publication, we consider this behavior to be undesirable. There are various variants of main path analysis, which all show the same type of undesirable behavior.

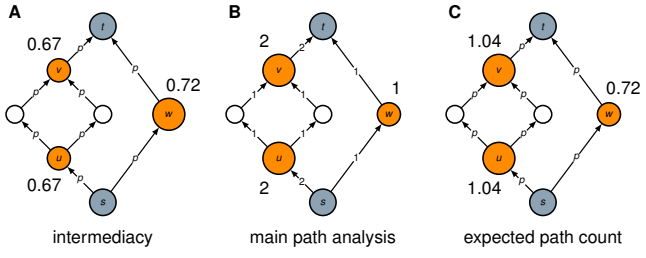

Fig. 3. Comparison of intermediacy (**A**), main path analysis (**B**), and expected path count (**C**). For nodes *u*, *v*, and *w*, the intermediacy (**A**), path count (**B**), and expected path count (**C**) are reported, using a value of 0.85 for the probability *p* in the calculation of intermediacy and expected path count.

Instead of focusing on the probability of the existence of at least one active source-target path, as is done by intermediacy, one could also focus on the expected number of active source-target paths going through a node. This alternative approach, which we refer to as the expected path count approach, is illustrated in Fig. 3C. As can be seen in the figure, nodes u and v have a higher expected path count than node w. Paths going through nodes u and v may therefore be favored over the path going through node w. Fig. 3C shows that, unlike intermediacy, the expected path count approach does not have the path contraction property. Depending on the probability p, contracting paths may cause expected path counts to decrease rather than increase. Because the expected path count approach does not have the path contraction property, we do not consider this approach to be a suitable alternative to intermediacy.

Empirical analysis

We now present two case studies that serve as empirical illustrations of the use of intermediacy. Case 1 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. Case 2 deals with the topic of peer review. This case is of interest because it was recently examined using main path analysis (22). Hence, it enables us to demonstrate the key differences between intermediacy and main path analysis. In both case studies, the intermediacy of publications was calculated using the Monte Carlo algorithm presented in the Materials and Methods section.

Case 1: Community detection and scientometrics. 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 measure 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 are missing in Web of Science. There are also missing citation links in Scopus, but for Scopus the problem is less significant than for Web of

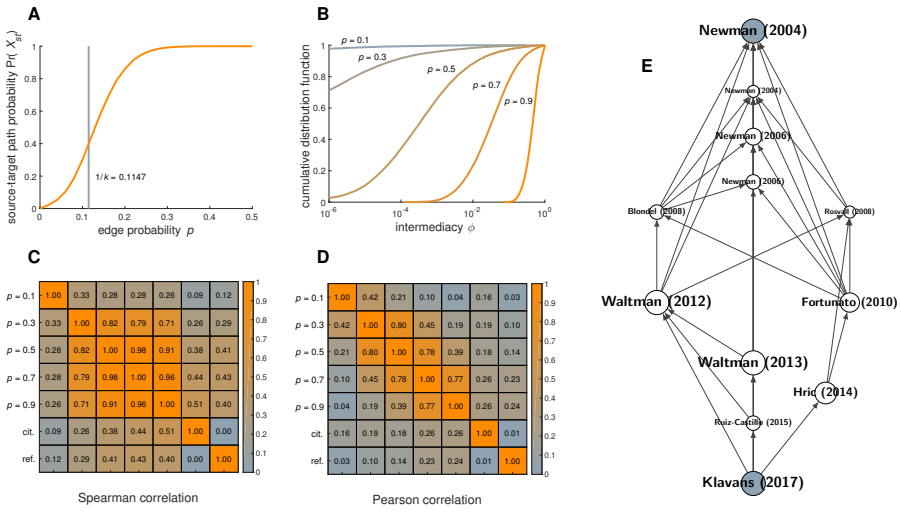

Fig. 4. Results for case 1. (**A**) Probability of the existence of an active source-target path as a function of the parameter p and (**B**) cumulative distribution of intermediacy scores for different values of p. Spearman (**C**) and Pearson (**D**) correlations between intermediacy scores for different values of p, citation counts, and reference counts. (**E**) Citation network of the top ten most intermediate publications for $p = 0.1$. (Only the name of the first author is shown.)

Science. We refer to Van Eck and Waltman (23) 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 $k = 2m/n \approx 8.72$ citation links, counting both incoming and outgoing links.

Fig. 4A 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 zero for $p = 0$ to almost one starting from $p = 0.25$. The vertical line indicates the value $p = 1/k$. At this value, traditional percolation theory for random graphs suggests that the probability that the source and target publications are connected becomes non-negligible (24). When searching for a suitable value of p, the value $p = 1/k$ suggested by percolation theory may serve as a reasonable starting point. In our case, this yields $p \approx 1/8.72 \approx 0.11$, resulting in a probability of about 0.40 for the existence of an active source-target path.

For five different values of the parameter p, Fig. 4B 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. On the other hand, when p is getting close to one, intermediacy scores also approach one.

Fig. 4C and Fig. 4D show Spearman and Pearson correlations between the intermediacy scores obtained for five

Table 1. Top ten most intermediate publications in case 1 for $p = 0.1$.

		p						
		0.1	0.3	0.5	0.7	0.9	cit.	ref.
t	Newman & Girvan (2004), Finding and evaluating community structure in networks, *Phys. Rev. E* **69**(2), 026113.	0.301	0.992	1.000	1.000	1.000	468	0
s	Klavans & Boyack (2017), Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?, *J. Assoc. Inf. Sci. Tec.* **68**(4), 984-998.	0.301	0.992	1.000	1.000	1.000	0	24
1	Waltman & Van Eck (2013), A smart local moving algorithm for large-scale modularity-based community detection, *Eur. Phys. J. B* **86**, 471.	0.061	0.376	0.656	0.878	0.988	2	27
2	Waltman & Van Eck (2012), A new methodology for constructing a publication-level classification system of science, *J. Assoc. Inf. Sci. Tec.* **63**(12), 2378-2392.	0.060	0.695	0.964	0.999	1.000	15	22
3	Hric *et al.* (2014), Community detection in networks: Structural communities versus ground truth, *Phys. Rev. E* **90**(6), 062805.	0.052	0.300	0.499	0.700	0.900	1	29
4	Fortunato (2010), Community detection in graphs, *Phys. Rep.* **486**(3-5), 75-174.	0.037	0.629	0.972	1.000	1.000	73	154
5	Newman (2006), Modularity and community structure in networks, *P. Natl. Acad. Sci. USA* **103**(23), 8577-8582.	0.035	0.736	0.979	1.000	1.000	221	8
6	Ruiz-Castillo & Waltman (2015), Field-normalized citation impact indicators using algorithmically constructed classification systems of science, *J. Informetr.* **9**(1), 102-117.	0.024	0.360	0.624	0.847	0.981	2	24
7	Blondel *et al.* (2008), Fast unfolding of communities in large networks, *J. Stat. Mech.*, P10008.	0.022	0.836	0.998	1.000	1.000	78	21
8	Newman (2006), Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev. E* **74**(3), 036104.	0.021	0.851	0.999	1.000	1.000	138	18
9	Newman (2004), Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**(6), 066133.	0.020	0.296	0.501	0.700	0.900	246	1
10	Rosvall & Bergstrom (2008), Maps of random walks on complex networks reveal community structure, *P. Natl. Acad. Sci. USA* **105**(4), 1118-1123.	0.020	0.803	0.994	1.000	1.000	70	10

different values of the parameter p. We consider intermediacy scores to be most useful from an ordinal 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 substantially 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 clear linear relationship between the intermediacy scores.

Fig. 4C and Fig. 4D also show correlations of intermediacy scores with citation counts and reference counts. The term *citation count* refers to the number of incoming citation links of a publication, while the term *reference count* refers 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. Regardless of the value of p, intermediacy scores are not very strongly correlated with citation counts or reference counts.

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 1 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. 4E shows the citation network of the ten most intermediate publications for $p = 0.1$.

Using our expert knowledge to interpret the results presented in Table 1 and Fig. 4E, we are able to trace how a method for community detection ended up in the scientometric literature. 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 1 and Fig. 4E are classical publications on community detection in general and modularity in particular. 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 1 and Fig. 4E. 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 1 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 1 that have a high intermediacy and a small number of citations do have a substantial number of references.

Case 2: Peer review. We now turn to case 2, in which we analyze the literature on peer review. The analysis 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 Batagelj *et al.* (22).

We started with a citation network of 45 965 publications dealing with peer review. This is the citation network that was labeled CiteAcy by Batagelj *et al.* (22). We selected Cole and Cole (1967) and Garcia *et al.* (2015) as our target and source publications. The main path analysis carried out by Batagelj *et al.* (22) 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 are connected by $m = 3\,420$ citation links, resulting in an average of $k = 2m/n \approx 11.12$ citation links per publication.

As can be seen in Fig. 5A, percolation theory suggests a value of $1/k \approx 1/11.12 \approx 0.09$ for the parameter p. This is close to the value of 0.11 obtained in case 1. However, the probability of the existence of an active path between the source and target publications equals 0.03, which is much lower than the probability of 0.40 in case 1. Intermediacy scores tend to be higher in case 2 than in case 1. This can be seen by comparing Fig. 5B to Fig. 4B. We note that the former figure has a linear horizontal axis, while the horizontal axis in the latter figure is logarithmic. The Spearman and Pearson correlations are somewhat higher in case 2 (Fig. 5C and Fig. 5D) than in case 1 (Fig. 4C and Fig. 4D).

Table 2 lists the ten publications with the highest intermediacy, where we use a value of 0.1 for the parameter p, like in Table 1. Fig. 5E shows the citation network of the ten most intermediate publications. There are numerous paths in this citation network going from our source publication (Garcia *et al.*, 2015) to our target publication (Cole & Cole, 1967). We regard these paths as the core paths between the source and target publications.

The core paths shown in Fig. 5E can be compared to the results obtained by Batagelj *et al.* (22) using main path analysis. Different variants of main path analysis were used by Batagelj *et al.* (22). Both using the original version of main path analysis (9) and using a more recent variant (12), the paths that were identified were rather lengthy, as can be seen in Figs. 9 and 10 in Batagelj *et al.* (22). The shortest main paths included about 20 publications. This confirms the fundamental difference between intermediacy and main path analysis. Main path analysis tends to favor longer paths over

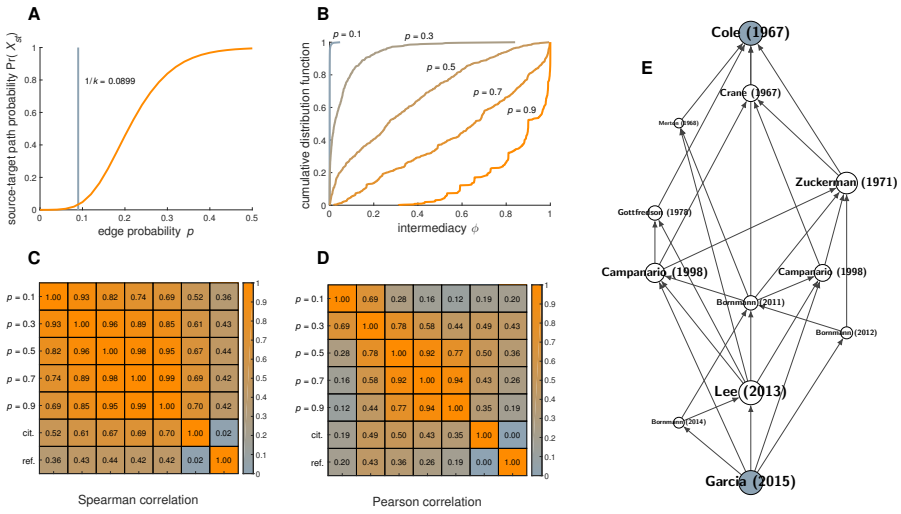

Fig. 5. Results for case 2. (**A**) Probability of the existence of an active source-target path as a function of the parameter p and (**B**) cumulative distribution of intermediacy scores for different values of p. Spearman (**C**) and Pearson (**D**) correlations between intermediacy scores for different values of p, citation counts, and reference counts. (**E**) Citation network of the top ten most intermediate publications for $p = 0.1$. (Only the name of the first author is shown.)

shorter ones, whereas intermediacy has the opposite tendency.

Using the results presented in Table 2 and Fig. 5E, experts on the topic of peer review could discuss the historical development of the literature on this topic. Since our own expertise on the topic of peer review is limited, we refrain from providing an interpretation of the results.

Discussion

Citation networks provide valuable information for tracing the historical development of scientific knowledge. For this purpose, citation networks are usually analyzed using main path analysis (9). However, the idea of a main path is relatively poorly understood. The algorithmic definition of a main path is clear, but the underlying conceptual motivation remains somewhat obscure. As we have shown in this paper, main path analysis has the tendency to favor longer paths over shorter ones. We consider this to be a counterintuitive property that lacks a convincing justification.

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. The older publication and the more recent one are referred to as the target and the source, respectively. Publications with a high intermediacy are important in connecting the source and the target publication in a citation network. As we have shown, intermediacy has two intuitively desirable properties, referred to as path addition and path contraction. Because of the path contraction property, intermediacy tends to favor shorter paths over longer ones. This is a fundamental difference with main path analysis. Intermediacy also has a free parameter that can be used to fine-tune its behavior. This parameter enables interpolation between two extremes. In one extreme, intermediacy identifies publications located on a

Table 2. Top ten most intermediate publications in case 2 for $p = 0.1$.

		p						
		0.1	0.3	0.5	0.7	0.9	cit.	ref.
t	Cole & Cole (1967), Scientific output and recognition: A study in the operation of the reward system in science, *Am. Sociol. Rev.* **32**(3), 377-390.	0.048	0.841	0.995	1.000	1.000	14	0
s	Garcia *et al.* (2015), The author-editor game, *Scientometrics* **104**(1), 361-380.	0.048	0.841	0.995	1.000	1.000	0	8
1	Lee *et al.* (2013), Bias in peer review, *J. Assoc. Inf. Sci. Tec.* **64**(1), 2-17.	0.018	0.510	0.865	0.986	1.000	5	71
2	Zuckerman & Merton (1971), Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system, *Minerva* **9**(1), 66-100.	0.016	0.336	0.622	0.847	0.981	73	2
3	Campanario (1998), Peer review for journals as it stands today: Part 1, *Sci. Commun.* **19**(3), 181-211.	0.013	0.592	0.967	0.999	1.000	23	35
4	Crane (1967), The gatekeepers of science: Some factors affecting the selection of articles for scientific journals, *Am. Sociol.* **2**(4), 195-201.	0.009	0.270	0.498	0.700	0.900	34	1
5	Campanario (1998), Peer review for journals as it stands today: Part 2, *Sci. Commun.* **19**(4), 277-306.	0.009	0.517	0.952	0.999	1.000	15	30
6	Gottfredson (1978), Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments, *Am. Psychol.* **33**(10), 920-934.	0.008	0.320	0.622	0.847	0.981	26	2
7	Bornmann (2011), Scientific peer review, *Annu. Rev. Inform. Sci.* **45**(1), 197-245.	0.008	0.333	0.776	0.975	1.000	6	71
8	Bornmann (2012), The Hawthorne effect in journal peer review, *Scientometrics* **91**(3), 857-862.	0.007	0.259	0.500	0.700	0.900	1	20
9	Bornmann (2014), Do we still need peer review? An argument for change, *J. Assoc. Inf. Sci. Tec.* **65**(1), 209-213.	0.007	0.275	0.500	0.700	0.900	1	17
10	Merton (1968), The Matthew effect in science, *Science* **159**(3810), 56-63.	0.005	0.243	0.497	0.701	0.901	29	1

shortest path between the source and the target publication. In the other extreme, it identifies publications located on the largest number of edge independent source-target paths.

We have also examined intermediacy in two case studies. In the first case study, intermediacy was used to trace historical developments at the interface between the community detection and the scientometric literature. This case study has shown that intermediacy yields results that appear sensible from the point of view of a domain expert. The second case study, in which intermediacy was applied to the literature on peer review, has provided an empirical illustration of the differences between intermediacy and main path analysis.

There are various directions for further research. First of all, a more extensive mathematical analysis of intermediacy can be carried out, possibly resulting in an axiomatic foundation for intermediacy. Intermediacy can also be generalized to weighted graphs. In a citation network, a citation link may for instance be weighed inversely proportional to the total number of incoming or outgoing citation links of a publication. Another way to generalize intermediacy is to allow for multiple sources and targets. The ideas underlying intermediacy may also be used to develop other types of indicators for graphs, such as an indicator of the connectedness of two nodes in a graph. In empirical analyses, intermediacy can be applied not only in citation networks of scientific publications, but for instance also in patent citation networks or in completely different types of networks, such as human mobility and migration networks, world trade networks, transportation networks, and passing networks in sports.

Materials and Methods

Proofs. Below we provide the proofs of the theorems presented in the main text. We first need to introduce some additional notation. We use $\Pr(X_{uv})$ as a shorthand for $\Pr(X_{uv} = 1)$. To make explicit that this probability depends on a graph G, we write $\Pr(X_{uv} \mid G)$. Furthermore, we use A_e to indicate whether an edge e is active. Hence, $A_e = 1$ if edge e is active and $A_e = 0$ if edge e is not active.

Proof of Theorem 1 . Let $m = |E|$ denote the number of edges in the graph G. Suppose that the m edges are split into two sets, one set of M edges and another set of $m - M$ edges. The probability that the edges in the former set are all active while the edges in the latter set are all inactive equals

$$P_M = p^M (1-p)^{m-M}. \quad [2]$$

Consider a node $v \in V$. The shortest source-target path that goes through node v has a length of ℓ_v. This means that at least ℓ_v edges need to be active in order to obtain an active source-target path that goes through node v. Hence, the probability that there is an active source-target path that goes through node v can be written as

$$\phi_v = \sum_{i=\ell_v}^{m} n_{vi} P_i, \quad [3]$$

where $n_{vi} > 0$ for all $i = \ell_v, \ldots, m$. Note that this probability equals the intermediacy of node v. Now consider two nodes $u, v \in V$ with $\ell_u < \ell_v$. In the limit as p tends to 0, ϕ_u and ϕ_v both tend to 0. However, they do so at different rates. More specifically, in the limit as p tends to 0, we have

$$\begin{aligned}
\lim_{p\to 0} \phi_v/\phi_u &= \lim_{p\to 0} \frac{\sum_{i=\ell_v}^{m} n_{vi} P_i}{\sum_{i=\ell_u}^{m} n_{ui} P_i} \quad [4]\\
&= \lim_{p\to 0} \frac{\sum_{i=\ell_v}^{m} n_{vi} P_i / P_{\ell_u}}{\sum_{i=\ell_u}^{m} n_{ui} P_i / P_{\ell_u}} \\
&= \lim_{p\to 0} \frac{\sum_{i=\ell_v}^{m} n_{vi} p^{i-\ell_u} (1-p)^{\ell_u - i}}{\sum_{i=\ell_u}^{m} n_{ui} p^{i-\ell_u} (1-p)^{\ell_u - i}} \\
&= 0/n_{u\ell_u} \\
&= 0.
\end{aligned}$$

Hence, in the limit as p tends to 0, $\phi_u > \phi_v$. □

Proof of Theorem 2 . Let $m = |E|$ denote the number of edges in the graph G, and let q denote the probability that an edge is inactive, that is, $q = 1 - p$. Suppose that the m edges are split into two sets, one set of M edges and another set of $m - M$ edges. The probability that the edges in the former set are all inactive while the edges in the latter set are all active equals

$$Q_M = q^M (1-q)^{m-M}. \quad [5]$$

Consider a node $v \in V$. There are σ_v edge independent source-target paths that go through node v. This means that at least σ_v edges need to be inactive in order for there to be no active source-target path that goes through node v. Hence, the probability that there is no active source-target path that goes through node v can be written as

$$\Phi_v = \sum_{i=\sigma_v}^{m} n_{vi} Q_i, \quad [6]$$

where $n_{vi} > 0$ for all $i = \sigma_v, \ldots, m$. Note that the intermediacy of node v equals 1 minus this probability, that is, $\phi_v = 1 - \Phi_v$. Now consider two nodes $u, v \in V$ with $\sigma_u > \sigma_v$. In the limit as p tends to 1, Φ_u and Φ_v both tend to 0. However, they do so at different rates. More specifically, in the limit as p tends to 1, we have

$$\begin{aligned}
\lim_{p\to 1} \Phi_u/\Phi_v &= \lim_{p\to 1} \frac{\sum_{i=\sigma_u}^{m} n_{ui} Q_i}{\sum_{i=\sigma_v}^{m} n_{vi} Q_i} \quad [7]\\
&= \lim_{p\to 1} \frac{\sum_{i=\sigma_u}^{m} n_{ui} Q_i / Q_{\sigma_v}}{\sum_{i=\sigma_v}^{m} n_{vi} Q_i / Q_{\sigma_v}} \\
&= \lim_{p\to 1} \frac{\sum_{i=\sigma_u}^{m} n_{ui} q^{i-\sigma_v} (1-q)^{\sigma_v - i}}{\sum_{i=\sigma_v}^{m} n_{vi} q^{i-\sigma_v} (1-q)^{\sigma_v - i}} \\
&= 0/n_{v\sigma_v} \\
&= 0.
\end{aligned}$$

Hence, in the limit as p tends to 1, $\Phi_u < \Phi_v$, which implies that $\phi_u > \phi_v$. □

Proof of Theorem 3 . Suppose that node w is located on a path from source s to node u. Let H denote the graph obtained after the path from node u to node v has been added, and let E_{uv} denote the set of newly added edges. The intermediacy of node w in graph G can be factorized as $\phi_w(G) = \Pr(X_{sw} \mid G) \Pr(X_{wt} \mid G)$. Similarly, for graph H, we have $\phi_w(H) = \Pr(X_{sw} \mid H) \Pr(X_{wt} \mid H)$. Clearly, $\Pr(X_{sw} \mid G) = \Pr(X_{sw} \mid H)$, since the paths from node s to node w are identical in graphs G and H. Furthermore, $\Pr(X_{wt} \mid G) = \Pr(X_{wt} \mid H \text{ and } \forall e \in E_{uv} : A_e = 0)$. Since $\Pr(X_{wt} \mid H \text{ and } \forall e \in E_{uv} : A_e = 0) \leq \Pr(X_{wt} \mid H)$, it follows that $\Pr(X_{wt} \mid G) \leq \Pr(X_{wt} \mid H)$. This means that $\phi_w(G) \leq \phi_w(H)$.

An analogous proof can be given if node w is located on a path from node v to target t. □

Proof of Theorem 4 . Suppose that node w is located on a path from source s to node u. Let H denote the graph obtained after paths from node u to node v have been contracted, and let E_{uv} denote the set of all edges between nodes in V_{uv}. The intermediacy

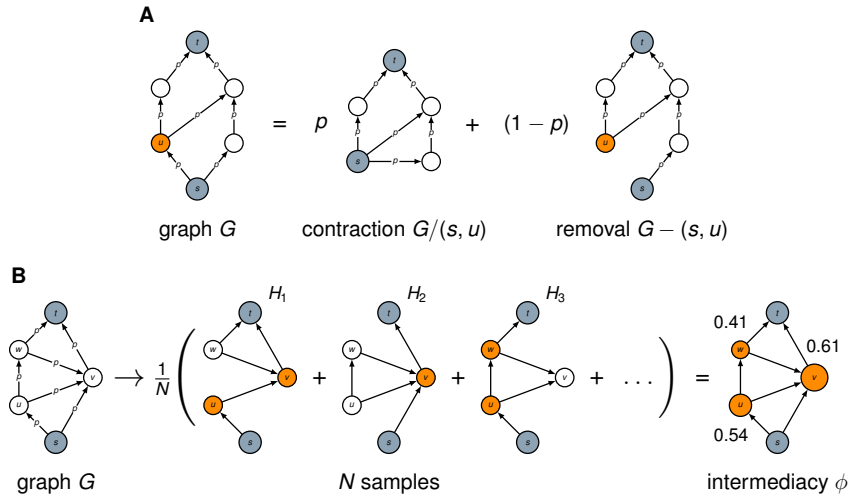

Fig. 6. Illustration of the calculation of intermediacy using the exact algorithm (**A**) and using the Monte Carlo algorithm for $p = 0.7$ (**B**).

of node w in graph G can be factorized as $\phi_w(G) = \Pr(X_{sw} \mid G) \Pr(X_{wt} \mid G)$. Similarly, for graph H, we have $\phi_w(H) = \Pr(X_{sw} \mid H) \Pr(X_{wt} \mid H)$. Clearly, $\Pr(X_{sw} \mid G) = \Pr(X_{sw} \mid H)$, since the paths from node s to node w are identical in graphs G and H. Furthermore, because nodes in V_{uv}, except for nodes u and v, do not have neighbors outside V_{uv}, we have $\Pr(X_{wt} \mid H) = \Pr(X_{wt} \mid G$ and $\forall e \in E_{uv} : A_e = 1)$. Since $\Pr(X_{wt} \mid G$ and $\forall e \in E_{uv} : A_e = 1) \geq \Pr(X_{wt} \mid G)$, it follows that $\Pr(X_{wt} \mid H) \geq \Pr(X_{wt} \mid G)$. This means that $\phi_w(H) \geq \phi_w(G)$.

An analogous proof can be given if node w is located on a path from node v to target t. □

Algorithms. Intermediacy depends on the probability that there exists a path between two nodes in a graph. Determining this probability is known as the problem of network reliability. This problem is NP-hard (25). Below we provide an outline of an exact algorithm for calculating intermediacy. Because of its exponential runtime, the exact algorithm can be used only in relatively small graphs. We therefore also propose a Monte Carlo algorithm that approximates intermediacy.

Exact algorithm. The exact algorithm, illustrated in Fig. 6A, is based on contraction and deletion of edges (26). Suppose we have a graph $G = (V, E)$. The probability that there exists a path between two nodes $u, v \in V$ can be written as

$$\Pr(X_{uv} \mid G) = p \Pr(X_{uv} \mid G/e) + (1 - p) \Pr(X_{uv} \mid G - e), \quad [8]$$

where G/e denotes the contraction of an edge $e \in E$ and $G - e$ denotes the deletion of an edge $e \in E$. Edge contraction must respect reachability (27). Eq. 8 yields a recursive algorithm for calculating $\Pr(X_{uv})$. For a node $v \in V$, this algorithm can be used to calculate $\Pr(X_{sv})$ and $\Pr(X_{vt})$. The intermediacy ϕ_v of node v is then given by Eq. 1. We are usually interested in calculating the intermediacy of all nodes in a graph G, not just of one specific node. This can be performed efficiently by calculating $\Pr(X_{sv})$ and $\Pr(X_{vt})$ for all nodes $v \in V$ in a single recursion.

The runtime of the exact algorithm is exponential in the number of edges m. The algorithm has a complexity of $\mathcal{O}(2^m)$. In the special case of a so-called series-parallel graph, the runtime of the algorithm can be reduced from exponential to polynomial (28).

Monte Carlo algorithm. The Monte Carlo algorithm, illustrated in Fig. 6B, is quite straightforward. Suppose we have a graph $G = (V, E)$ and we are interested in the intermediacy ϕ_v of a node $v \in V$. A subgraph H can be obtained by sampling the edges in the graph G, where each edge $e \in E$ is sampled with probability p. Given a subgraph H, it can be determined whether in this subgraph node v is located on a path from source s to target t. We sample N subgraphs $H_1, \ldots, H_N$. We then approximate the intermediacy of node v by $\phi_v \approx \frac{1}{N} \sum_{i=1}^{N} I_{st}(v \mid H_i)$, where $I_{st}(v \mid H_i)$ equals 1 if there exists a path from source s to target t going through node v in graph H_i and 0 otherwise.

The Monte Carlo algorithm can be implemented efficiently by simultaneously sampling subgraphs and checking path existence. To do so, we perform a probabilistic depth first search. We maintain a stack of nodes that still need to be visited. We start by pushing source s to the stack. We then keep popping nodes from the stack until the stack is empty. When a node v has been popped from the stack, we determine for each of its outgoing edges whether the edge is active. An edge is active with probability p. If an edge (v, u) is active and if node u is not yet on the stack, then node u is pushed to the stack. At some point, target t may be reached, resulting in the identification of nodes that are located on a path from source s to target t. This implementation of the Monte Carlo algorithm is especially fast for smaller values of the probability p. The runtime of the Monte Carlo algorithm is linear in the number of edges m.

Source code. In this paper, we use a Java implementation of the Monte Carlo algorithm. The source code is available at https://github.com/lovre/intermediacy (29).

ACKNOWLEDGMENTS. We would like to thank Vladimir Batagelj for sharing the data used to study the literature on peer review (22). This work has been supported in part by the Slovenian Research Agency under the programs P2-0359 and P5-0168 and by the European Union COST Action number CA15109.

1. Garfield E, Sher I, Torpie R (1964) The use of citation data in writing the history of science, (The Institute for Scientific Information), Technical Report F49(638)-1256.
2. Garfield E, Pudovkin A, Istomin V (2003) Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology* 54(5):400–412.
3. Garfield E, Pudovkin A, Istomin V (2003) Mapping the output of topical searches in the Web of Knowledge and the case of Watson-Crick. *Information Technology and Libraries* 22(4):183–187.
4. Garfield E (2004) Historiographic mapping of knowledge domains literature. *Journal of Information Science* 30(2):119–145.
5. van Eck N, Waltman L (2014) CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics* 8(4):802–823.
6. Chen C (2006) Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57(3):359–377.
7. Marx W, Bornmann L, Barth A, Leydesdorff L (2014) Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology* 65(4):751–764.
8. Thor A, Marx W, Leydesdorff L, Bornmannd L (2016) Introducing CitedReferencesExplorer (CREXplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics* 10(2):503–515.
9. Hummon N, Doreian P (1989) Connectivity in a citation network: The development of DNA theory. *Social Networks* 11(1):39–63.
10. Batagelj V (2003) Efficient algorithms for citation network analysis. *e-print arXiv:cs/0309023v1* pp. 1–27.
11. Lucio-Arias D, Leydesdorff L (2008) Main-path analysis and path-dependent transitions in HistCite™-based historiograms. *Journal of the American Society for Information Science and Technology* 59(12):1948–1962.
12. Liu J, Lu L (2012) An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology* 63(3):528–542.

13. Batagelj V, Doreian P, Ferligoj A, Kejžar N (2014) *Understanding Large Temporal Networks and Spatial Networks*. (Wiley, Chichester).
14. Yeo W, Kim S, Lee JM, Kang J (2014) Aggregative and stochastic model of main path identification: A case study on graphene. *Scientometrics* 98(1):633–655.
15. Liu J, Kuan CH (2016) A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology* 67(2):465–476.
16. Tu YN, Hsu SL (2016) Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology* 67(8):2016–2031.
17. Verspagen B (2007) Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems* 10(1):93–115.
18. Park H, Magee C (2017) Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PLoS ONE* 12(1):e0170895.
19. Gwak J, Sohn S (2018) A novel approach to explore patent development paths for subfield technologies. *Journal of the Association for Information Science and Technology* 69(3):410–419.
20. Kim J, Shin J (2018) Mapping extended technological trajectories: Integration of main path, derivative paths, and technology junctures. *Scientometrics* 116(3):1439–1459.
21. Kuan CH, Huang MH, Chen DZ (2018) Missing links: Timing characteristics and their implications for capturing contemporaneous technological developments. *Journal of Informetrics* 12(1):259–270.
22. Batagelj V, Ferligoj A, Squazzoni F (2017) The emergence of a field: A network analysis of research on peer review. *Scientometrics* 113(1):503–532.
23. van Eck N, Waltman L (2017) Accuracy of citation data in Web of Science and Scopus in *Proceedings of the 16th International Conference on Scientometrics & Informetrics ISSI '17*. (Wuhan, China), pp. 1087–1092.
24. Newman M (2018) *Networks*. (Oxford University Press, Oxford), 2nd edition.
25. Ball M (1980) Complexity of network reliability computations. *Networks* 10(2):153–165.
26. Moskowitz F (1958) The analysis of redundancy networks. *Transactions of the American Institute of Electrical Engineers* 77(5):627–632.
27. Page L, Perry J (1989) Reliability of directed networks using the factoring theorem. *IEEE Transactions on Reliability* 38(5):556–562.
28. Misra K (1970) An algorithm for the reliability evaluation of redundant networks. *IEEE Transactions on Reliability* R-19(4):146–151.
29. Šubelj L (2018) Intermediacy of publications (http://dx.doi.org/10.5281/zenodo.1424365).