intermediacy of publications

"identifikacija ključnih publikacij za razvoj znanstvenega področja"

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science

Ludo Waltman
Leiden University
Centre for Science and Technology Studies

Vincent Traag
Leiden University
Centre for Science and Technology Studies

Nees Jan van Eck
Leiden University
Centre for Science and Technology Studies

NetSlo '19

intermediacy of publications

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.

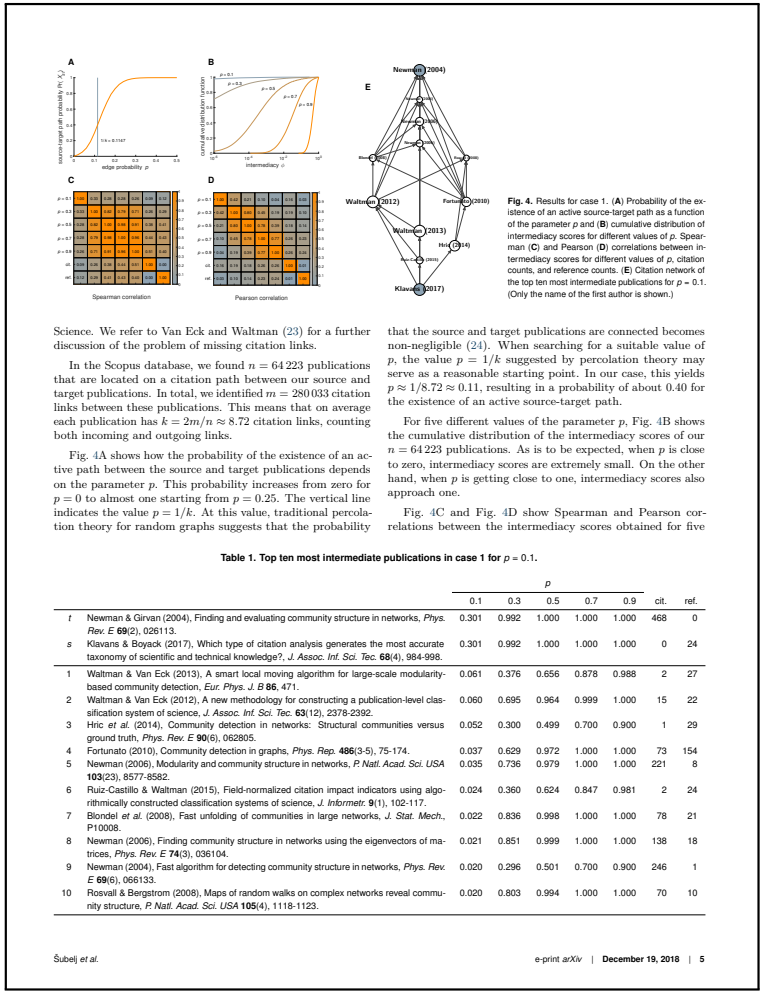


Fig. 4. Results for case 1. (**A**) Probability of the existence of an active source-target path as a function of the parameter p and (**B**) cumulative distribution of intermediacy scores for different values of p. Spearman (**C**) and Pearson (**D**) correlations between intermediacy scores for different 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 $p = 0.1$.

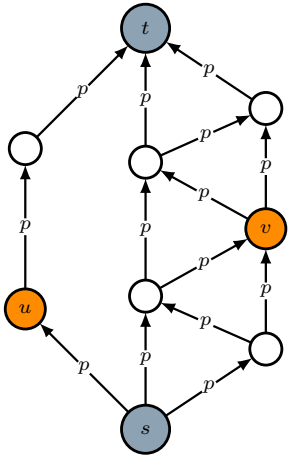
		p						
		0.1	0.3	0.5	0.7	0.9	cit.	ref.
t	Newman & Girvan (2004), Finding and evaluating community structure in networks, *Phys. Rev. E* **69**(2), 026113.	0.301	0.992	1.000	1.000	1.000	468	0
s	Klavans & Boyack (2017), Which type of citation analysis generates the most accurate taxonomy of scientific 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

(**yesterday**) paper **rejected** without review in **PNAS**

problem & motivation

algorithmic historiography for evolution of field (**Garfield, 1964–**)

relying on **citations** between scientific **publications** from **WoS** & **Scopus**



existing approaches include **main paths** (**Hummon & Doreian, 1989**)

(**longest/shortest paths**) many **irrelevant**/miss **relevant** publications

(**intermediacy**) important publications should only be **well-connected**

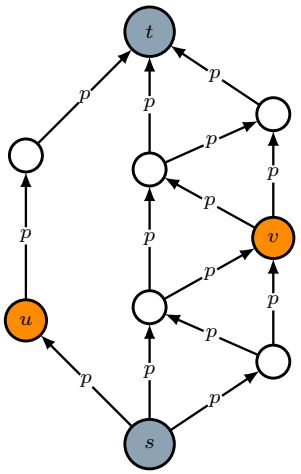
intermediacy measure

(**input**) selected **source** & **target** publications **s** & **t**

(**method**) each citation is relevant/active with **probability p**

(**measure**) importance of **publication u** called **intermediacy** ϕ_u

$$\phi_u = \Pr(X_{st}^u) = \Pr(X_{su}) \Pr(X_{ut})$$

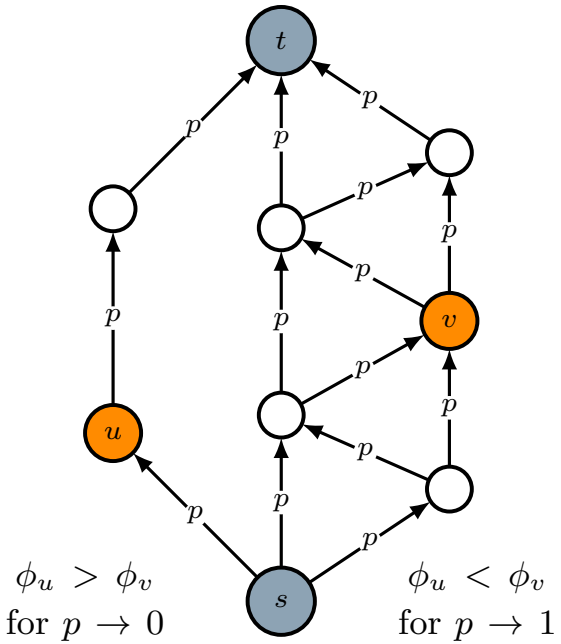


X_{st} exists path **from s to t** & X_{st}^u exists path **through u**

intermediacy for $p \to 0$

for $p \to 0$ intermediacy ϕ **governed by** ℓ (**proof**)

for $p \to 0$ if $\ell_u < \ell_v$ then $\phi_u > \phi_v$

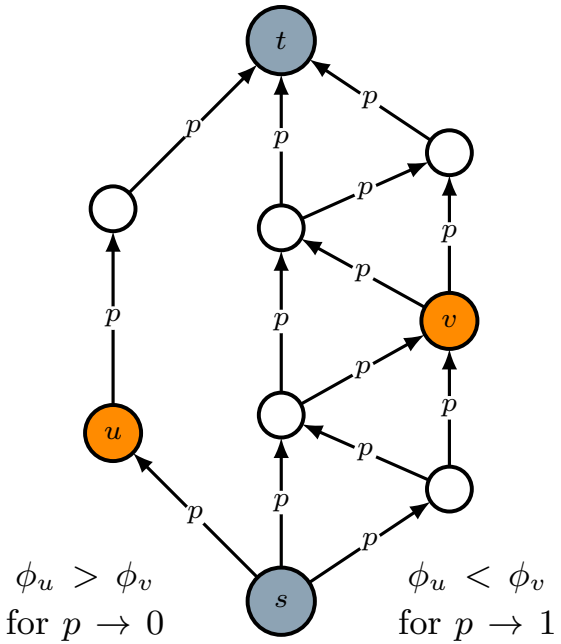


ℓ_u is **length** of **shortest paths** from s to t through u

intermediacy for $p \to 1$

for $p \to 1$ intermediacy ϕ **governed by** σ (**proof**)

for $p \to 1$ if $\sigma_u < \sigma_v$ then $\phi_u < \phi_v$

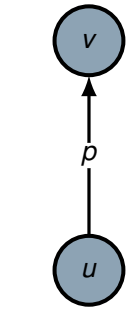


σ_u is **number** of **edge-disjoint paths** from s to t through u

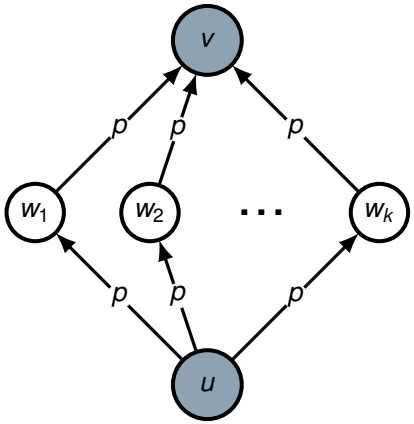
intuition for p

for what p is **direct citation** equivalent to k **indirect citations**

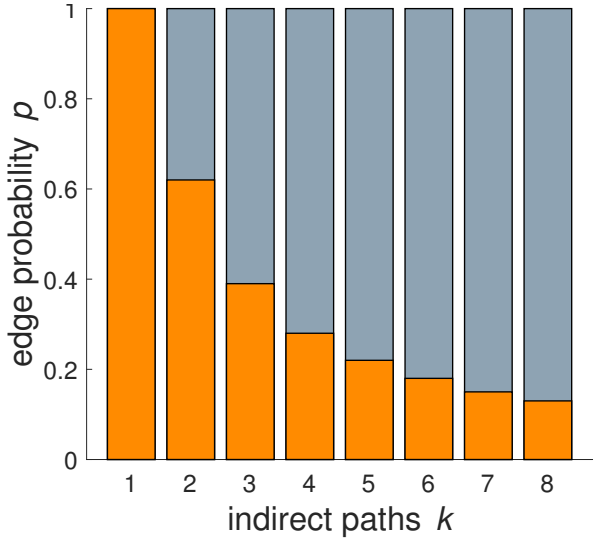
$$\Pr(X_{uv}) = p = 1 - (1 - p^2)^k$$



$\Pr(X_{uv}) = p$



$\Pr(X_{uv}) = 1 - (1 - p^2)^k$

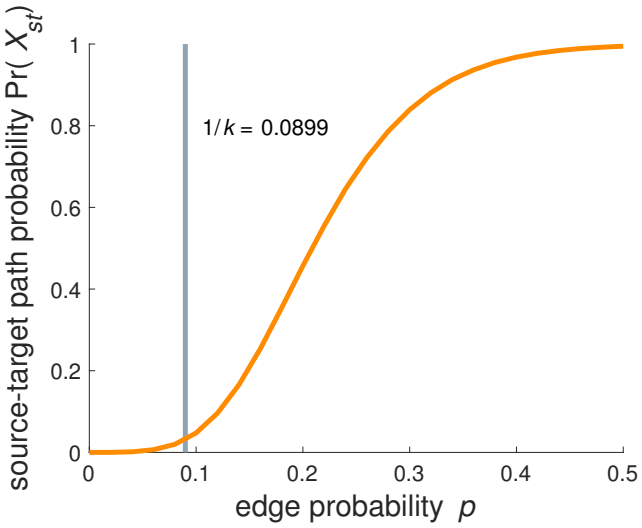


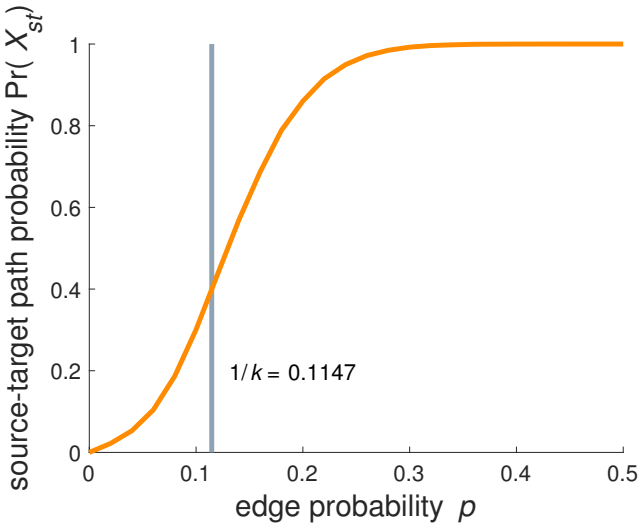
k is **number** of **indirect paths** from u to v

p phase transition

for what p source-target path $\mathbf{Pr(X_{st}) > 0}$ & intermediacy $\exists \mathbf{u} : \phi_{\mathbf{u}} > \mathbf{0}$

$$p \geq n/2m = 1/k$$

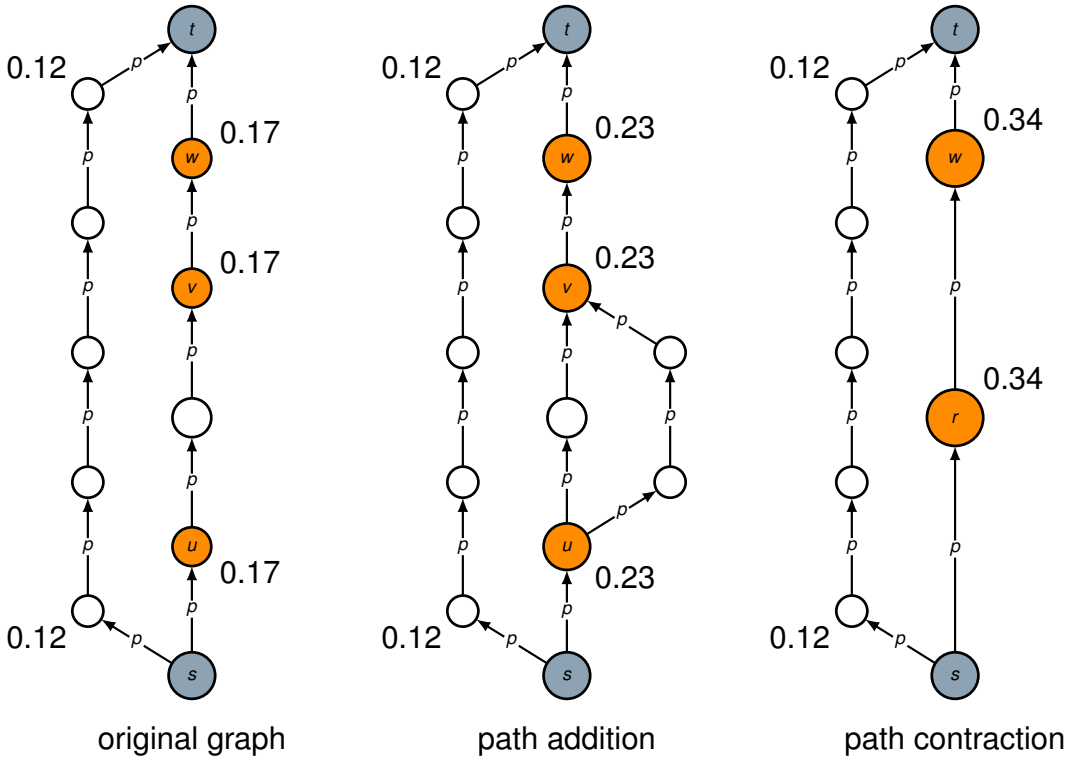




k is **average** number of **citations** & **references**

properties of intermediacy

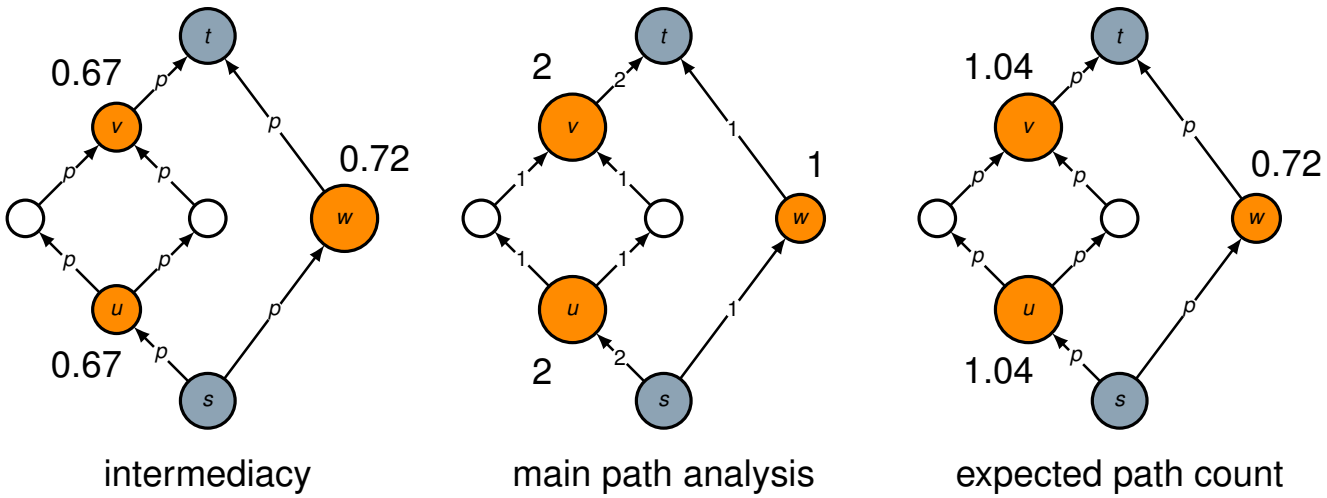
path **addition** & **contraction** increase intermediacy (**proof**)



path from source to target becomes **"easier"** (**intuition**)

alternatives to intermediacy

alternatives are **main paths** & **expected paths** (**state of the art**)

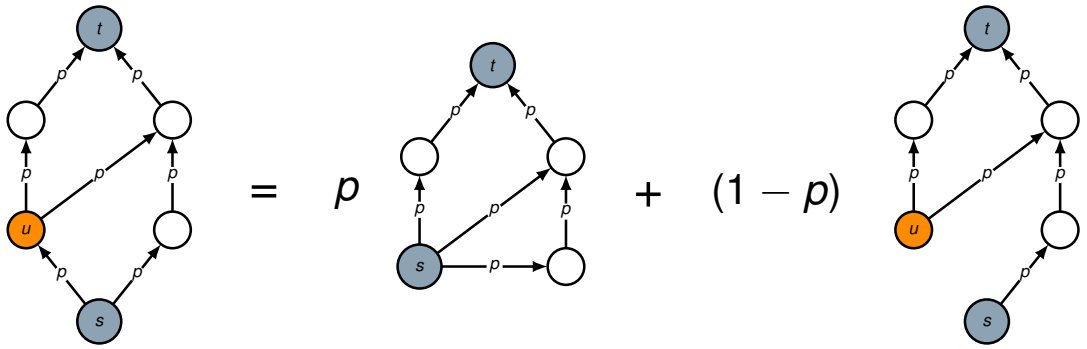


alternatives **violate** path contraction property (**example**)

exact algorithm

decomposition algorithm by edge **contraction** & **removal** (**Ball, 1979**)

$$\Pr(X_{st} \mid G) = p \Pr(X_{st} \mid G/(s,u)) + (1-p) \Pr(X_{st} \mid G - (s,u))$$



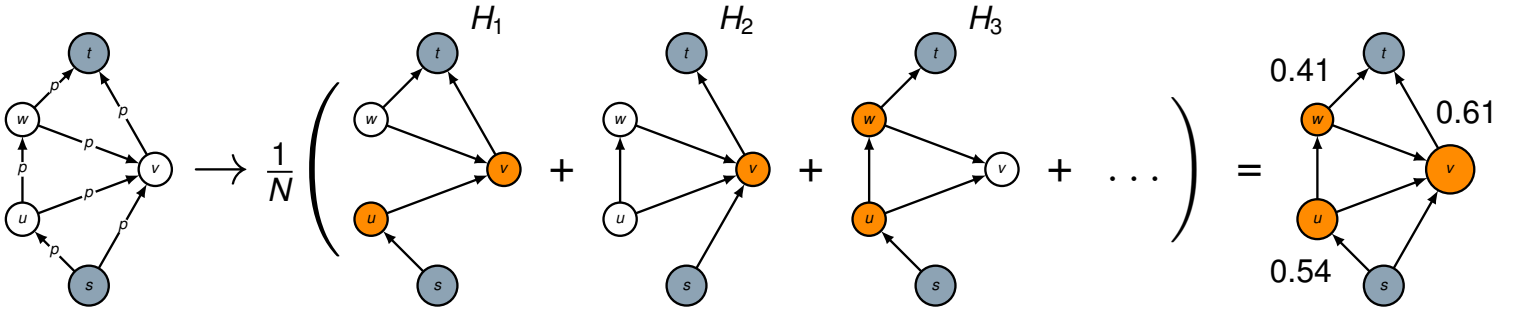
graph G contraction $G/(s,u)$ removal $G - (s,u)$

runs in **exponential time** since NP-hard even in DAG (**Johnson, 1984**)

approximate algorithm

simple **Monte Carlo** simulation algorithm by edge **sampling**

$$\phi_u = \Pr(X_{st}^u \mid G) = \frac{1}{N} \sum_{k=1}^{N} \mathrm{I}(X_{st}^u \mid H_k)$$



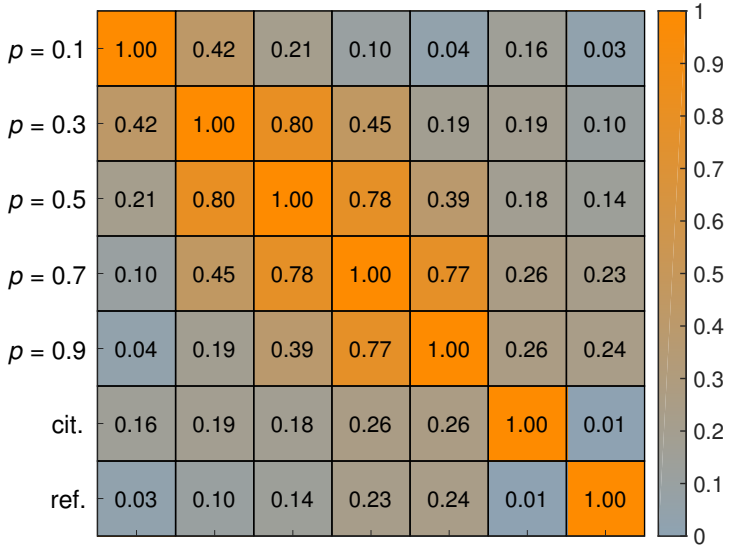
graph G — N samples — intermediacy ϕ

runs in **linear time** using probabilistic DFS over say **10^6 samples**

intermediacy $\neq$ centrality

correlation coefficient between **intermediacies** ϕ & **citations**/**references**

Pearson correlation



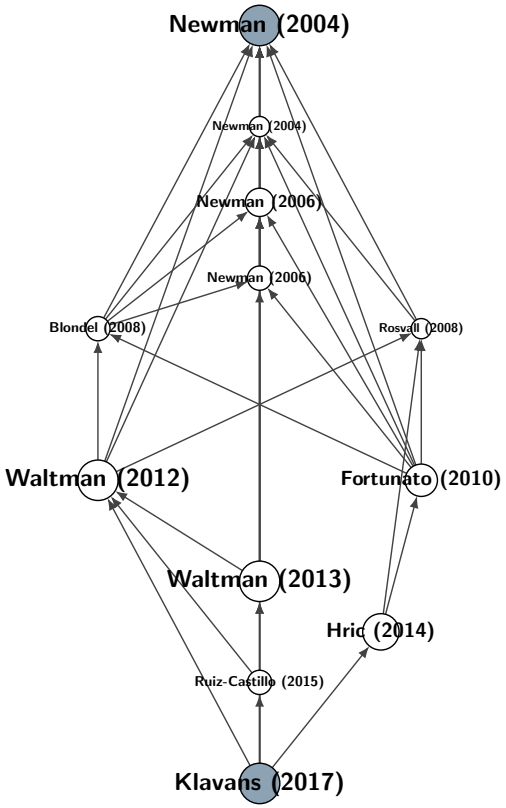
Pearson correlation

intermediacy ϕ **uncorrelated** with standard **centrality measures**

modularity example

(**target**) Newman & Girvan (2004), **Finding and evaluating community...**, *Phys. Rev. E* **69**(2), 026113.

(**source**) Klavans & Boyack (2017), **Which type of citation analysis generates...**, *JASIST* **68**(4), 984-998.



1 Waltman & Van Eck (2013), A smart local moving algorithm for large-scale modularity-based community detection, *EPJB* **86**, 471.

2 Waltman & Van Eck (2012), A new methodology for constructing a publication-level classification system..., *JASIST* **63**(12), 2378-2392.

3 Hric et al. (2014), Community detection in networks: Structural communities versus ground truth, *Phys. Rev. E* **90**(6), 062805.

4 Fortunato (2010), Community detection in graphs, *Phys. Rep.* **486**(3-5), 75-174.

5 Newman (2006), Modularity and community structure in networks, *PNAS* **103**(23), 8577-8582.

6 Ruiz-Castillo & Waltman (2015), Field-normalized citation impact indicators using algorithmically..., *J. Informetr.* **9**(1), 102-117.

7 Blondel et al. (2008), Fast unfolding of communities in large networks, *J. Stat. Mech.*, P10008.

8 Newman (2006), Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev. E* **74**(3), 036104.

9 Newman (2004), Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**(6), 066133.

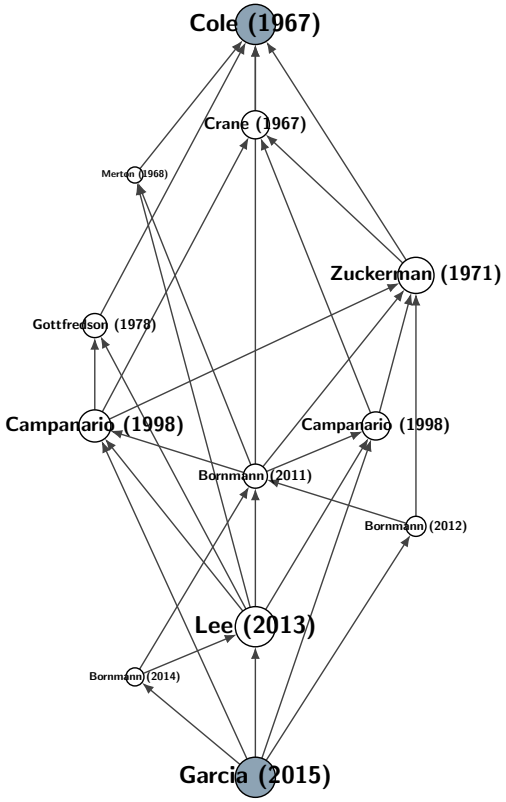
10 Rosvall & Bergstrom (2008), Maps of random walks on complex networks reveal community structure, *PNAS* **105**(4), 1118-1123.

in-house version of Scopus database at **CWTS**

peer review example

(**target**) Cole & Cole (1967), **Scientific output and recognition**, *Am. Sociol. Rev.* **32**(3), 377-390.

(**source**) Garcia et al. (2015), **The author-editor game**, *Scientometrics* **104**(1), 361-380.



1 Lee et al. (2013), Bias in peer review, *JASIST* **64**(1), 2-17.

2 Zuckerman & Merton (1971), Patterns of evaluation in science: Institutionalisation, structure and functions. . . , *Minerva* **9**(1), 66-100.

3 Campanario (1998), Peer review for journals as it stands today: Part 1, *Sci. Commun.* **19**(3), 181-211.

4 Crane (1967), The gatekeepers of science: Some factors affecting the selection of articles for scientific journals, *Am. Sociol.* **2**(4), 195-201.

5 Campanario (1998), Peer review for journals as it stands today: Part 2, *Sci. Commun.* **19**(4), 277-306.

6 Gottfredson (1978), Evaluating psychological research reports: Dimensions, reliability, and correlates. . . , *Am. Psychol.* **33**(10), 920-934.

7 Bornmann (2011), Scientific peer review, *Annu. Rev. Inform. Sci.* **45**(1), 197-245.

8 Bornmann (2012), The Hawthorne effect in journal peer review, *Scientometrics* **91**(3), 857-862.

9 Bornmann (2014), Do we still need peer review? An argument for change, *JASIST* **65**(1), 209-213.

10 Merton (1968), The Matthew effect in science, *Science* **159**(3810), 56-63.

snapshot of WoS collected by (**Batagelj et al., 2017**)

small-world example

(**target**) Watts & Strogatz (1998), **Collective dynamics of 'small-world' networks**, *Nature* **393**(6684), 440-442.

(**source**) Backstrom et al. (2012), **Four degrees of separation**, In: *Proceedings of the WebSci '12*, pp. 45-54.

1. Newman (2003), The structure and function of complex networks, *SIAM Rev.* **45**(2), 167-256.
2. Albert & Barabási (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**(1), 47-97.
3. Li et al. (2005), Towards a theory of scale-free graphs: Definition, properties, and implications, *Internet Math.* **2**(4), 431-523.
4. Leskovec et al. (2007), Graph evolution: Densification and shrinking diameters, *ACM Trans. Knowl. Discov. Data* **1**(1), 1-41.
5. Liben-Nowell et al. (2005), Geographic routing in social networks, *P. Natl. Acad. Sci. USA* **102**(33), 11623-11628.
6. Strogatz (2001), Exploring complex networks, *Nature* **410**(6825), 268-276.
7. Boldi et al. (2011), Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks, In: *Proceedings of the WWW '11*, pp. 587-596.
8. Dorogovtsev (2002), Evolution of networks, *Adv. Phys.* **51**(4), 1079-1187.
9. Ye et al. (2010), Distance distribution and average shortest path length estimation in real-world networks, In: *Proceedings of the ADMA '10*, pp. 322-333.
10. Lattanzi et al. (2011), Milgram-routing in social networks, In: *Proceedings of the WWW '11*, pp. 725-734.

in-house version of Scopus database at **CWTS**

scale-free example

(**target**) Barabási & Albert (1999), **Emergence of scaling in random networks**, *Science* **286**(5439), 509-512.

(**source**) Liu et al. (2011), **Controllability of complex networks**, *Nature* **473**(7346), 167-173.

1. Albert & Barabási (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**(1), 47-97.
2. Strogatz (2001), Exploring complex networks, *Nature* **410**(6825), 268-276.
3. Boguñá et al. (2004), Cut-offs and finite size effects in scale-free networks, *Eur. Phys. J. B* **38**(2), 205-209.
4. Nishikawa et al. (2003), Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize?, *Phys. Rev. Lett.* **91**(1), 014101.
5. Kim & Motter (2009), Slave nodes and the controllability of metabolic networks, *New J. Phys.* **11**, 113047.
6. Newman (2003), The structure and function of complex networks, *SIAM Rev.* **45**(2), 167-256.
7. Sorrentino et al. (2007), Controllability of complex networks via pinning, *Phys. Rev. E* **75**(4), 046103.
8. Dorogovtsev (2002), Evolution of networks, *Adv. Phys.* **51**(4), 1079-1187.
9. Pastor-Satorras et al. (2001), Dynamical and correlation properties of the Internet, *Phys. Rev. Lett.* **87**(25), 258701.
10. Yu et al. (2009), On pinning synchronization of complex dynamical networks, *Automatica* **45**(2), 429-435.

in-house version of Scopus database at **CWTS**

deep learning example

(**target**) LeCun et al. (2015), **Deep learning**, *Nature* **521**(7553), 436-444.

(**source**) Silver et al. (2017), **Mastering the game of Go without human knowledge**, *Nature* **550**(7676), 354-359.

1 Silver et al. (2016), Mastering the game of Go with deep neural networks and tree search, *Nature* **529**(7587), 484-489.

2 Jouppi et al. (2017), In-datacenter performance analysis of a tensor processing unit, In: *Proceedings of the ISCA '17*, pp. 1-12.

3 Reagen et al. (2016), Minerva: Enabling low-power, highly-accurate deep neural network accelerators, In: *Proceedings of the ISCA '16*, pp. 267-278.

4 Chen et al. (2016), DianNao family: Energy-efficient hardware accelerators for machine learning, *Commun. ACM* **59**(11), 105-112.

5 Chen et al. (2016), Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks, In: *Proceedings of the ISCA '16*, pp. 127-138.

6 Moravčík et al. (2017), DeepStack: Expert-level artificial intelligence in heads-up no-limit poker, *Science* **356**(6337), 508-513.

7 Albericio et al. (2016), Cnvlutin: Ineffectual-neuron-free deep neural network computing, In: *Proceedings of the ISCA '16*, pp. 1-13.

8 Han et al. (2016), EIE: Efficient inference engine on compressed deep neural network, In: *Proceedings of the ISCA '16*, pp. 243-254.

9 Shafiee et al. (2016), ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars, In: *Proceedings of the ISCA '16*, pp. 14-26.

10 Adolf et al. (2016), Fathom: Reference workloads for modern deep learning methods, In: *Proceedings of the IISWC '16*, pp. 1-10.

in-house version of Scopus database at **CWTS**

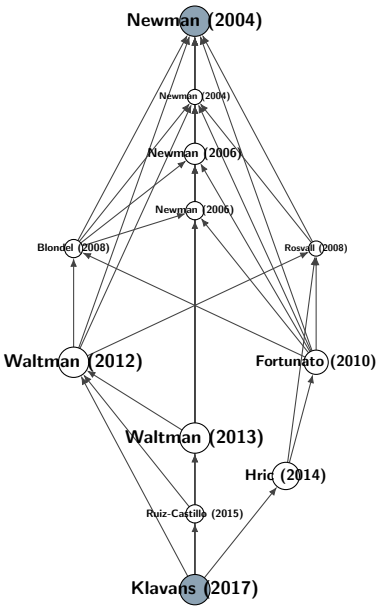
conclusions & future work

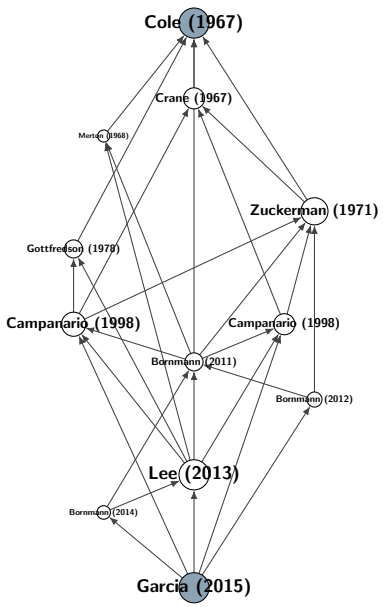
(**proposal**) measure of importance of publications called **intermediacy**

(**theory**) **conceptually clear** & provable behavior in **extreme cases**

(**practice**) intermediacy shows promising results in **case studies**

(**future**) **research app**! applicability to **other networks**?





(**paper**) arxiv.org/abs/**1812.08259**

(**code**) github.com/lovre/**intermediacy**

Lovro Šubelj
University of Ljubljana
lovro.subelj@fri.uni-lj.si
http://lovro.lpt.fri.uni-lj.si

Ludo Waltman
Leiden University
waltmanlr@cwts.leidenuniv.nl
http://www.ludowaltman.nl

Vincent Traag
Leiden University
v.a.traag@cwts.leidenuniv.nl
http://www.traag.net

Nees Jan van Eck
Leiden University
ecknjpvan@cwts.leidenuniv.nl
http://www.neesjanvaneck.nl