intermediacy of publications

uncovering important publications for the development of a field

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

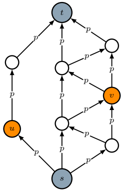
COSTNET '20

work resulting from my STSM at Leiden University funded by COSTNET

problem & motivation

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

relying on **citations between publications** from **WoS/Scopus**



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

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

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

"...citations are valid and valuable means of creating accurate historical descriptions of scientific fields."

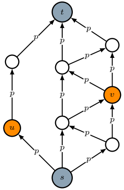
measure of intermediacy

(**setting**) select **source** & **target** publications **s** & **t**

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

(**result**) importance of **publication u** as **intermediacy** $\phi_{\mathbf{u}}$

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



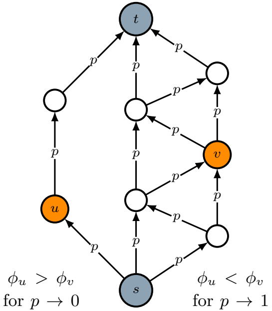
$\mathbf{X_{st}}$ – exists path **from s to t** & $\mathbf{X_{st}^u}$ – exists such path **through u**

$\phi_u = 2\phi_v \not\equiv$ publication u is "twice" as important as publication v

limit case $p \to 0$

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

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

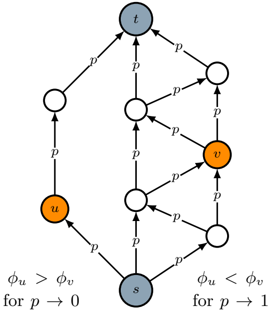


ℓ_u – **length** of **shortest paths** from s to t through u

limit case $p \to 1$

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

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

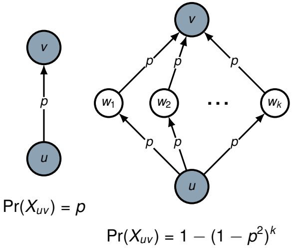


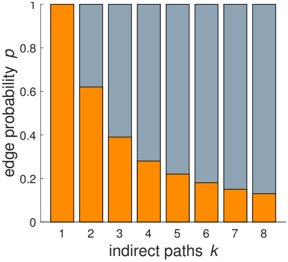
σ_u – **number** of **edge-disjoint paths** from s to t through u

intuition for parameter p

for what p is **direct citation** $\equiv$ k **indirect citations**

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





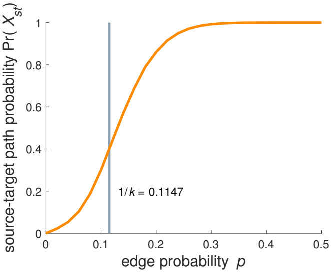
k – **number** of **indirect paths** from u to v

$p = 0.22 \equiv k = 5$ & $p = 0.11 \equiv k = 10$

choice of parameter p

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

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



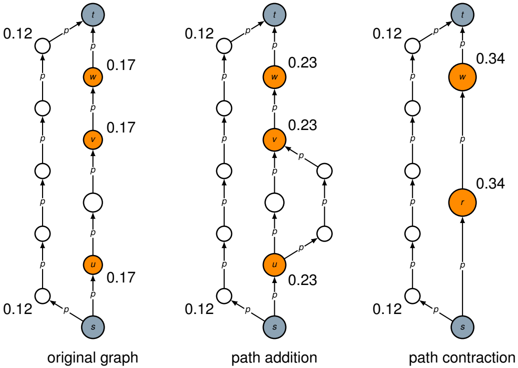
k – **average** number of **citations** & **references**

---

percolation theory suggests that for $k > 1$ probability $\Pr(X_{st})$ is non-negligible

properties of intermediacy

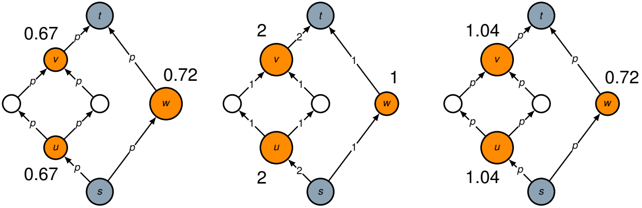
path addition & **contraction** increase intermediacy (**proof**)



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

alternatives to intermediacy

alternatives include **main paths** & **resistance** (**state of the art**)



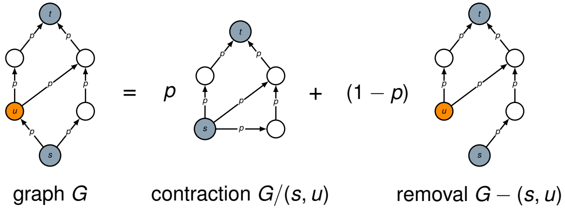
intermediacy | main path analysis | expected path count

alternatives **violate** path addition/contraction property (**examples**)

exact algorithm

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

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



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

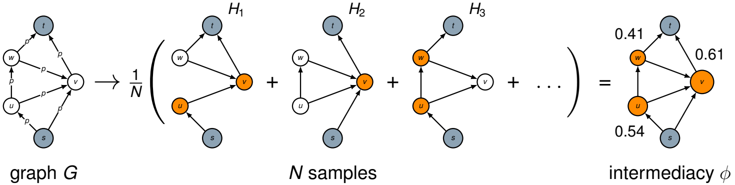
---

decomposition algorithm is cute & elegant, but useless in practice :(

approximate algorithm

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

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



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

$\ll$ 30 min for network with 9 145 771 nodes and 81 771 723 edges :)

intermediacy $\neq$ centrality

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

p = 0.1	1.00	0.33	0.28	0.28	0.26	0.09	0.12
p = 0.3	0.33	1.00	0.82	0.79	0.71	0.26	0.29
p = 0.5	0.28	0.82	1.00	0.98	0.91	0.38	0.41
p = 0.7	0.28	0.79	0.98	1.00	0.96	0.44	0.43
p = 0.9	0.26	0.71	0.91	0.96	1.00	0.51	0.40
cit.	0.09	0.26	0.38	0.44	0.51	1.00	0.00
ref.	0.12	0.29	0.41	0.43	0.40	0.00	1.00

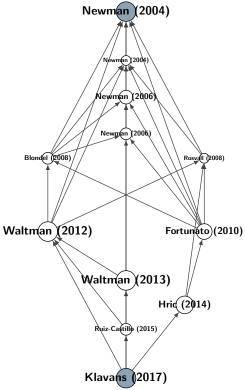
intermediacy **not correlated** with standard **centrality measures**

intermediacy most useful from ordinal perspective $\equiv$ Pearson $<$ Spearman correlation

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.



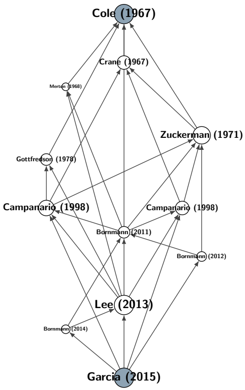
1. Waltman & Van Eck (2013), A smart local moving algorithm for large-scale modularity-based community detection, *EPJB* **86**, 471.
2. Waltman & Van Eck (2012), A new methodology for constructing a publication-level classification system..., *JASIST* **63**(12), 2378-2392.
3. Hric et al. (2014), Community 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.

we set $p = 0.1$ & use 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.

we set $p = 0.1$ & use 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.

we set $p = 0.1$ & use 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.

---

we set $p = 0.1$ & use in-house version of **Scopus** database at **CWTS**

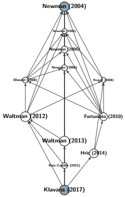
conclusions & future

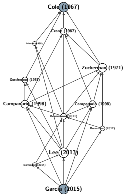
(**proposal**) measure of **importance of publications** called intermediacy

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

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

(**extensions**) multiple sources & targets, weighted networks





(**future**) **online app**! other networks, axiomatic foundation etc.

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

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

Šubelj, Waltman, Traag & Van Eck (2020) Intermediacy of publications, *Royal Society Open Science*, **7**(1), 190207.

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