The embedding of strong citations

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

[1] *University of Ljubljana, Faculty of Computer and Information Science*
[2] *Leiden University, Centre for Science and Technology Studies*

1 Introduction

Scientometrics have long studied citations to learn more about the structure and flow of knowledge. However, until recently, it was difficult to discern between different citations. Not all citations are equally relevant to a publication. Some earlier works may be really foundational for a publication, whereas others are only tangentially related. We rely on a large-scale full-text database to study the strength of citations. We define the strength of a citation as the number of in-text references that are made to that citation.

The strength of ties has for long been a subject of discussion and analysis in the social sciences. Most famously, Granovetter [1] suggested the so-called "weak-ties hypothesis". The weak-ties hypothesis captures two elements of information flow within social network:

1. Stronger ties are embedded in a more tightly knit network.

2. Weak ties carry new information.

Especially the latter gained some familiarity in the phrase "the strength of weak ties" (also the title of Granovetter's work). Although the second element is perhaps the most innovative insight, it is more difficult to study. The first element has been frequently analysed in many different settings, and has generally been found to hold in social networks.

We here analyse the strength of citations and the extent to which they are embedded in a densely connected citation network. We find that this hypothesis holds in citation network. However, contrary to social network, we may expect most of the information to flow over the stronger citations. The stronger citations are the ones that seem to be more important to the publication. If the publications that are so strongly cited would not have been published, it would perhaps be questionable whether the new publication could have been written. Hence, the weak citations seem to serve a less clear purpose: they are less clearly embedded in a larger citation network, nor can we expect much new information to flow through them. This flow of new information is something not yet analysed, but we plan to do so in the near future.

2 Methodology

For a number of publishers we dispose of the full-text articles (see [2] for a more detailed description), among which Elsevier, on which we rely for this current study. For a number of selected journals, we analyse all citations made by the full-text articles in that journal to any other article covered in the Web of Science (WoS). We then also retrieve all citations to those articles. Finally, we include all citations between the cited/citing articles of the full-text articles. We hence obtain insight in the number of triads in which a citation of a certain strength is embedded. We analyse three journals: Cell, Journal of Informetrics (JoI) and Physics Letters B (PLB).

There are in total three different types of triads in which a citation strength can be embedded (Fig. 1). Let us assume that the citation for which we are analysing the strength goes from publication i to j. Then citation $i \to j$ can be embedded in different ways in a triad with some

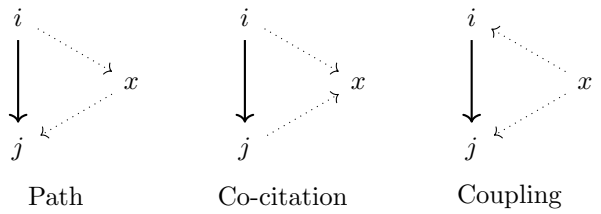


Figure 1: Different possible configurations of triads and their denomination. Note that the "reverse" path generally does not exist in a citation network, because that would generate a cycle. We use "triangles" to simply denote any of the three configuration. The thick edge indicates the citation for which we know the citation strength. The dotted edges indicate citations for which we do not necessarily know the citation strength.

Table 1: Pearson correlation between citation strength and triads.

	Cell	JoI	PLB
Triangles	0.25	0.25	0.23
Paths	0.13	0.12	0.09
Co-citations	0.24	0.19	0.22
Couplings	0.18	0.18	0.16

other publication x. The first possibility is that i cites x and that x also cites j, which we call a *path*. The second possibility is that both i and j cite x, which we call a *co-citation*. The third possibility is that x cites i and j, which we call a *coupling*.

;

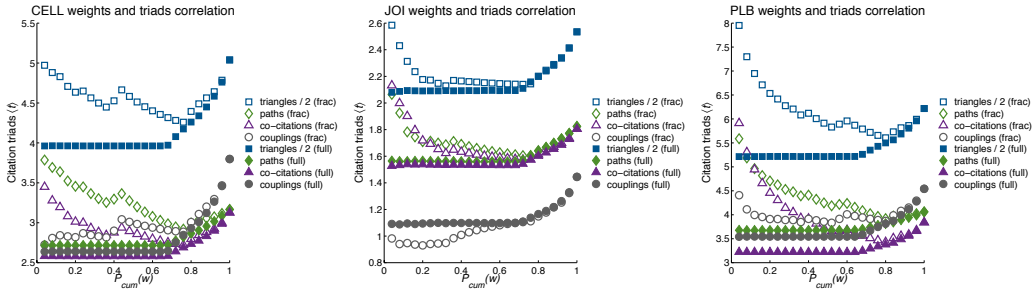
3 Results

Similar to [3] we analyse the average number of triads a citations is involved in. In particular, we analyse how the average number of triads (of different configuration) depends on the cumulative citation strength (i.e. we calculate the average number of triads for citations that a strength higher than a certain percentile than the rest of the citations). We find that there is a clear correlation between all types of triads and the citation strength. However, this correlation is relatively weak (around 0.25 at most). There is a clear ordering of which configuration correlates higher with citation strength across all three different journals. The highest correlation is obtained by the "triangles" that include any possibly type of triad. The second highest correlation is obtained by the "co-citation" type, closely followed by the "coupling" type. The "path" type finally achieves clearly the lowest correlation.

We look in more detail at the dependency of the number of triads on the citation strength (Fig. 2). This clearly shows that the average number of triads is mostly constant up until a relatively high percentile of the weight. This is a results of the many number of citations that only have a weight of 1 (i.e. they are only references once in the full-text). The strength of the correlations are more difficult to estimate from the figures. It does make clear that on average, Cell and PLB tend to have citations that are embedded in many more triangles than JoI.

4 Discussion

Our analysis clearly shows that stronger citations are more strongly embedded in a citation network. This effect is most clearly pronounced for the co-citation triad, in which the cited paper j



(a) Cell (b) Journal of Informetrics (c) Physical Letters B

Figure 2: Dependency of triads on citation strength.

also cites a third paper x that is also cited by the citing paper i. This seems to point in the direction that the stronger citations point to more recent work that covers a certain topic or research question. For example, there is some recent advance made by paper j on which the new paper i wishes to improve, which both work on a problem introduced by paper x. This is also in line with [2] who observe that citations that are more frequently mentioned in the full-text tend to be more recent papers.

Stronger citations are also quite clearly embedded in coupling triads. The explanation of that phenomenon is probably quite different from the co-citation triad. The coupling triad suggest that a publication that is strongly cited both tend to be co-cited by many other papers[1]. For example, if some problem was introduced by paper j which paper i solved, they may often be co-cited by many papers x.

The explanation of the path triad is in a sense the inverse of the co-citation. Indeed a co-citation triad will also appear in our analysis as a path triad (with j and x inverted). As we just explained, papers presumably tend to be founded on the more recent relevant papers in some topic, rather than on older papers. From this perspective, it actually makes sense that the path triad correlates clearly lower with citation strength than co-citation.

In conclusion, we find that strong citations are clearly embedded in a citation network, congruent with the weak tie hypothesis in social networks. To what extent new information flows over the weak citations or the strong citations is still something that is open to discussion. However, we would a priori expect that stronger citations would have more new information flowing over them than weaker citations. This is something that should be further investigated, and we plan to do so in the (near) future.

Acknowledgements

This work has been supported in part by the European Union COST Action number CA15109.

References

[1] Mark S. Granovetter. The Strength of Weak Ties. *Am. J. Sociol.*, 78(6):1360, 1973. ISSN 0002-9602. doi: 10.1086/225469.

[2] Kevin W. Boyack, Nees Jan van Eck, Giovanni Colavizza, and Ludo Waltman. Characterizing in-text citations in scientific articles: A large-scale analysis. oct 2017. URL http://arxiv.org/abs/1710.03094.

[3] JP Onnela, J Saramäki, J Hyvönen, G Szabó, D Lazer, K Kaski, J Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.*

[1] This suggest we should rename the denominations of the triads.

S. A., 104(18):7332–6, may 2007. ISSN 0027-8424. doi: 10.1073/pnas.0610245104. URL http://www.pnas.org/content/104/18/7332.shorthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1863470{&}tool=pmcentrez{&}rendertype=abstract.