Algebraic structures in the ego article citation network

Ronald Rousseau [1,2,3]

ronald.rousseau@khbo.be
[1]KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium
[2]Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, 2000 Antwerpen, Belgium
[3]K.U.Leuven, Dept. Mathematics, Celestijnenlaan 200B, 3000 Leuven (Heverlee), Belgium

Abstract
New indices characterizing an article in its ego citation network are introduced. Among these we especially mention the outgrow index. Although algebraic aspects are emphasized, a first step towards their interpretation is attempted.

Introduction
Informetrics is not only the study of regularities (the so-called informetric laws) or of citation counting and its consequences, it is also the study of related algebraic structures such as graphs or networks. In its most abstract form it becomes a study of categories (Rousseau, 1992). In this contribution we introduce some new notions related to the structure of citation networks. These notions are described independently from any specific database. We recall that a citation network always contains two points of view: the 'cites' point of view, and the 'is cited by' view. These viewpoints are dual and both will play a role in our contribution. Although most topics discussed in this article apply to any publication, we will study only the article citation network. These articles may be journal articles or articles published in a conference proceedings book, but we assume that no other types of references or citation sources are present in the network under study. This simplifies our treatment. We begin with a purely algebraic, graph-theoretic description, then we consider dynamical aspects, and finally we take a first step in the direction of attaching meaning to our constructs, i.e. we go beyond the purely algebraic construct.

Ego article citation network: the 'cites' relation
We consider an article citation network and focus on one specific target article: the ego, as it is called in network theory (Wasserman & Faust, 1994), see Figure 1. Our contribution follows the lead of Howard D. White who was one of the first to perform ego-centred citation analyses (White, 2000). We refer to this target article as A and consider A's reference list, denoted as Ref(A). Recall that – for simplicity - we assume that this reference list contains only articles. The length of A's reference list, i.e. its number of references, is denoted as $T_{Ref}(A)$. This approach implies that we follow the 'cites' relation (taking A's point of view). Article A and all articles in its reference list form a set, denoted as $ER(A) = Ref(A) \cup \{A\}$, where ER stands for the extended reference list (namely, extended by including A). We will attach a positive number to each element of ER(A). This will allow us to rank ER(A) and we will characterize the relative position of A in this ranked list. Note that references correspond to outlinks in the 'cites' network.
For each element in ER(A) we determine the number of articles by which it is cited. In this contribution, this is the number we are interested in. Yet our basic framework, though not its interpretation, also applies to other numbers one may associate to an article, such as its age or the number of authors. Next we rank all elements in ER(A) according to its number of received citations. Finally the position of A in this list is characterized by its citations-of-references number $CR(A) = 1 - R(A)/(T_{Ref}(A)+1)$, where R(A) denotes the rank of A in this

list. In case of ties we use an average rank. This notion was introduced by us in an earlier article and termed the outgrown index (Rousseau & Hu, 2010). CR(A) is always a number between zero (included) and one (not included). Articles that cite an article from A's reference list form the set of all articles which are bibliographically coupled with A (this set may, or may not, depending on the application, include article A itself). Each of these articles has already a bibliographic coupling strength and a relative bibliographic coupling strength with A (Kessler, 1963). This establishes a relation between the notions of bibliographic coupling and the outgrow or CR-index.
Yet, instead of considering the number of citations received by each element in ER(A) we can also take the number of references in each of these articles. This is yet another number associated to ER(A). In this way we obtain another ranked list and we can determine a reference-reference index RR(A) using RR(A) = 1 – R'(A)/(T_{Ref}(A)+1), where R'(A) is the rank of A in the list determined by the number of references. This approach uses references of references, hence two generations of references, establishing another relation between our new approach and notions studied in the field (Hu et al., 2011a).
Note that the set ER(A) can be studied for its own sake, and this in many different ways: new ways, as introduced here and in (Rousseau & Hu, 2010), but also well-known ways. Indeed, one may determine the average and median number of received citations (for each member); an h-index (Hirsch, 2005) may be determined based on its members' received citations or references given, see (Liang & Rousseau, 2010), while box plots (Egghe & Rousseau, 1990, p.25) can be used to compare ERs for different As.

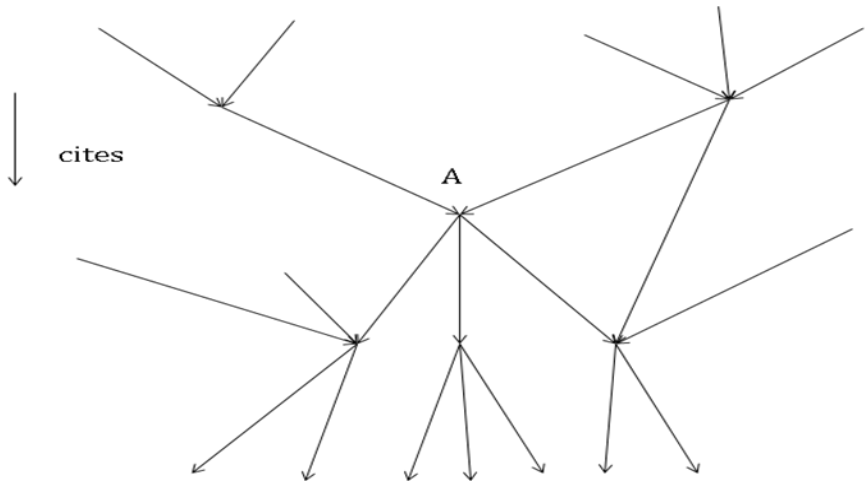

Figure 1. Article A's (the ego) citation network.

Ego article citation network: the 'is cited by' relation

Now we consider all articles that cite A and denote this set as Cit(A) (A is cited by all elements in Cit(A)). The number of elements in Cit(A) is denoted as T_{Cit}(A). This means that, taking A's point of view, we now follow the 'is cited by' relation. Article A and all citing articles form a set, denoted as EC(A) = Cit(A) ∪ {A}. Again we will attach a positive number to each element of C(A), leading to a ranking of EC(A). As was the case for the 'cites' relation, a number between zero and one will be used to characterize the relative position of A in this ranked list.
First we determine for each element in EC(A) the number of articles by which it is cited. Next we rank all elements in EC(A) according to its number of received citations. Finally the position of A in this list is characterized by its citation-citation number CC(A) = 1 – R"(A)/(T_{Cit}(A)+1), where R"(A) denotes the rank of A in this new list. Note that as for

ER(A), also the set EC(A) can be studied for its own sake, and this again in many different ways. Note that this approach uses citing articles of citing articles, hence two citation generations as studied e.g. in (Hu et al. 2011a).
Finally, instead of considering the number of citations received by each element in EC(A) we can also take the number of references in each of these articles. In this way we obtain another ranked list and determine a reference-citation index RC(A) using the relation RC(A) = 1 – R'''(A)/(T_{Cit}(A)+1), where R'''(A) is the rank of A in the list determined by the number of references. Articles that are cited by an article citing A form the set of all articles which are co-cited with A. Each of these articles has a co-citation strength and a relative co-citation strength with A (for a definition of these notions, we refer to (Small, 1973)).
Note that we have taken full advantage of the existing duality (Rousseau, 2010) in a citation network. Indeed, with every 'cites' relation there is a 'is cited by' relation. In this way the notions of bibliographic coupling and co-citation are dual notions. Returning to the 'cites' and 'is cited by' arrows originating from the target article, we point out that the number of references and the number of received citations are just the in- and out-degree of the target article in the citation network. We also note that the CR- or outgrow index is determined by two sets: BC(A), the set of all articles that are bibliographically coupled with A, and Cit(A), the set of all articles that cite A. Similarly, the RC-index is determined by Ref(A) and CoCiT(A), the set of all articles co-cited with A.

Dynamical aspects

We first point out which notions are static, i.e. do not change once the target article is introduced in the citation network, and which are dynamic, i.e. change or can change over time.
The set of references of article A, and ER(A) are static and hence also A's outdegree in the 'cites' network. Consequently also the references of references index, RR(A) is a static indicator. Once another article, say B, is published the bibliographic coupling coefficient between A and B is fixed. Yet, most notions mentioned above are dynamic, namely all those involving received citations. The sets of articles that are bibliographically coupled, BC(A), or co-cited, CoCit(A), with A also change over time.
At the moment an article is published (introduced in the network) its RR-index is fixed, while its CR-index is usually zero as each reference item is cited at least once, namely by A, while A is usually not cited. An exception might be the case that A is cited in the same journal issue as the one in which A is published. Such cases are from now on excluded. CC(A) and RC(A) are zero when A is published as at that moment EC(A) = {A}. The three indices CR(A), RC(A) and CC(A) are dynamic ones. Indeed, the relative position of A in ER(A) and EC(A) may change over time. When A receives its first citation (say by just one other article) then CR(A), the outgrow index) may or may not increase. RC(A) stays 0 or becomes 0.5 and CC(A) becomes 0.5. Even if A stays always the most-cited article in EC(A) its CC(A) still increases. This follows from the definition of the CC-index. Dynamical aspects of the outgrow index were studied in (Hu et al., 2011b).

Interpretation

For one target article a dynamic indicator seems more interesting than a static one. Yet, static ones may also be of interest when comparing these indices for different As.
What is the meaning of the RR-index? This index characterizes the length of A's reference list with respect to the reference lists of its own references. Assuming that an article's reference list represents the knowledge on which it is built, we may observe that on the one hand, an article with a short reference list may still, be it indirectly, be built on a rich amount of knowledge. In this case A's RR-index will be low. On the other hand an article with a long

reference list may be built on many articles each studying a narrow topic (hence each with a short reference list). In that case A's RR index will be quite high. Given the fact that reference lists tend to increase (Persson et al., 2004; Althouse et al., 2010), one may expect the RR-index to be rather high than low, but this is just an observation based on averages, independent of its behaviour in concrete cases.

The CR-index has already been studied (Rousseau & Hu, 2010) and because of its practical meaning been termed the outgrow index. It is, indeed an index that characterizes to which extent an article outgrows (in terms of citations) the referenced items on which it is based. It is an indicator of the relation between the visibility of the target article and its references. One may expect that this index increases over time.

If an article is highly cited and has a high CC-index then it is a leading article in its environment, perhaps describing a really innovative idea. If after some years an article has a rather low CC-index then it probably is not really important or only used in a minor role.

If an article is a review article then it probably has a high RC-index. If it is a normal article then a low RC-index may show that the article is mainly mentioned in review articles, while a high RC-value may indicate that it has been used mainly in highly focused articles. Generally one may expect for normal articles that this index decreases over time.

Conclusions and directions for applications

This new look on a well-known network may provide new ways of studying small subfields. Could these indices and subsets reflect changes in interest and intellectual patterns? Can they point to co-occurrences of ideas? Are they correlated?

One step further can be taken by not determining the number of citations received or references given for the elements in ER(A) and EC(A), but the number of scientific fields that give or receive citations. In this way the rankings (now based on fields used or reached) and the resulting indices connect our approach to diffusion theory (Liu & Rousseau, 2010), the study of interdisciplinarity (Rafols & Meyer, 2010) and the general study of knowledge integration and diffusion (Liu, Rafols & Rousseau, 2011).

In which type of graphs can these ideas be applied? We see possible applications in other citation graphs, e.g. patent citation graphs (Singh, 2005; Wang et al., 2010), but also in totally different networks such as food webs (ecology) and Hasse diagrams such as the partial order graph derived from a Lorenz curve (Nijssen et al., 1998).

Acknowledgments

The author thanks Xiaojun Hu (Zhejiang University) and Raf Guns (Antwerp University) for stimulating discussions related to the outgrow index and its possible applications.

References

Althouse, B.M., West, J.D., Bergstrom, T. & Bergstrom, C.T. (2009). Differences in impact factor across fields and over time. *Journal of the American society for Information Science and Technology*, 60(1), 27-34.

Egghe, L. & Rousseau, R. (1990). *Introduction to Informetrics. Quantitative methods in library, documentation and information science.* Amsterdam: Elsevier.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA*, 102(16), 16569-16572.

Hu, XJ., Rousseau, R. & Chen J. (2011a). On the definition of forward and backward citation generations. *Journal of Informetrics*, 5 (1), 27-36.

Hu, XJ., Rousseau, R. & Chen J. (2011b). Time series of outgrow indices. *Journal of Informetrics, 5 (to appear)*.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.

Liang, LM. & Rousseau, R. (2010). Reference analysis: a view in the mirror of citation analysis. *Geomatics and Information Science of Wuhan University*, 35 (spec), 6-9.
Liu, YX., Rafols, I. & Rousseau, R. (2011). A framework for knowledge integration and diffusion. *Journal of Documentation* (accepted for publication).
Liu, YX. & Rousseau, R. (2010). Knowledge diffusion through publications and citations: a case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2), 340-351.
Nijssen, D., Rousseau, R. & Van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses*, 13(1), 33-38.
Persson, O., Glänzel, W. & Danell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics,* 60(3), 421-432.
Rafols, I. & Meyer, M. (2010), Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience, *Scientometrics,* 82(2), 263-287.
Rousseau, R. (1992). Category theory and informetrics: information production processes. *Scientometrics*, 25(1), 77-87.
Rousseau, R. (2010). Bibliographic coupling and co-citation as dual notions. In B. Larsen (Ed.), *The Janus Faced Scholar. A Festschrift in Honour of Peter Ingwersen* (pp. 173-183). ISSI e-zine (*special* volume).
Rousseau, R. & Hu, XJ. (2010). An outgrow index. *Annals of Library and Information Studies*, 57(3), 287-290.
Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5), 756-770.
Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.
Wang, JC., Chiang, CH. & Lin, SW. (2010). Network structure of innovation: can brokerage or closure predict patent quality? *Scientometrics*, 84(3), 735-748.
Wasserman, S. & Faust, K. (1994). *Social Network Analysis*. Cambridge: Cambridge University Press.
White, H.D. (2000). Toward ego-centered citation analysis. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield* (pp. 475-496). Medford (NJ): Information Today.