Problem Set 4
Statistics, Computation and Applications

Felipe del Canto

November, 2021

Problem 4.1: Suggesting Similar Papers

For part (a), let N be the number of papers, C be adjacency matrix for the citation network, and let γ be the adjacency matrix for the co-citation network. Observe that papers i and j are co-cited with weight w if and only if there are w papers such that $C^T(i,k) \cdot C(k,j) = 1$ for each paper $k \in \{1,\dots,w\}$. Indeed, if i and j are co-cited with weight w, then there exist w papers such that $C(k,i) = C(k,j) = 1$ for each $k \in \{1,\dots,w\}$. This implies that $C^T(i,k) \cdot C(k,j) = 1$ for each one. Conversely, if there are w papers with $C^T(i,k) \cdot C(k,j) = 1$ for each $k \in \{1,\dots,w\}$, then $C(k,i) = C(k,j)$ and, consequently, papers i and j are co-cited with weight w. The previous claim shows that a first attempt at obtaining the co-citation adjacency matrix is

$$\gamma = C^T C.$$

Since C^T codes the directed edges from the cited papers from the one that cites them, then the product $C^T C$ counts the 2-step directed edges between two papers that are cited by the same third paper. Since a paper is co-cited with itself the number of times another paper cite it, this product also captures it. Consequently, the previous equality is the desired computation.

For part (b), the process is similar to part (a), but focused on the papers that cite, instead of those cited. Now, we are interested in the two-step directed edges between to papers that cite the same paper. Consequently, if β is the bibliographic coupling adjacency matrix, then the first approach should be

$$\beta = CC^T.$$

This product captures the two-step directed edges that start in a citing paper and finish in another citing paper. Note that this product also captures the bibliography coupling between a paper and itself. Since a paper is bibliography coupled with itself with weight equal to the number of other papers it cites, then there are that same amount of two-step directed edges between a that paper and itself.

For part (c), consider two papers belonging to two different disciplines which described two different techniques. In principle, this papers are very different. Suppose, additionally, that these techniques are widely used in a third discipline and, consequently, their respective papers are usually cited by many articles belonging to this latter field. Under these assumptions, both of these papers will have a high co-citation weight, despite not dealing with a related subject. In contrast, their bibliographic coupling should be close (or equal) to zero. Evidently, a scientist writing a paper in the third discipline, will benefit much from knowing these two papers are used together in other articles related to the ones she is writing. However, this does not imply that the two papers are related. Instead, it shows that the research made by this scientist is related to the other work in her discipline. In contrast, consider two survey papers in some particular field, but one of them was published a few years before the other. Since these are surveys, their bibliographic coupling might be large, but since newer papers would tend to cite the most recent survey, their co-cite weight will be close to zero.

To answer which measure is more appropriate to indicate similarities, the first step would be to address in which sense we are considering two articles to be similar. If we are interested in determining the boundaries of a field, it might me more useful to rely on the bibliographic coupling. Since an author interested in inserting her work into a given field would likely cite many of its relevant articles, it is likely that

papers belonging to a given field have many of their references in common. Observe that a similar argument could be done for the co-cite similarity. If two papers belong to a certain field, it is more likely they appear together in other bibliographies. However, the caveat is that the co-cite similarity is very sensitive to the year in which each paper is published. Two recent papers have not had the chance to be included in many bibliographies, making this similarity very sparse for recent articles. What is more, the co-cite similarity is in fact monotone non-decreasing with respect to time, whereas the bibliographic coupling is independent of time.

Now consider that, instead, a researcher is interested in determining which articles should she review before writing about a certain topic. In particular, she is following a certain paper and would like to know which other papers she should review. In this case, looking at similar papers using the bibliographic coupling could be a bad measure for two main reasons. First, each article's bibliography is an endogenous product. The author(s) of the main paper she is following chose carefully which citations they wanted to use, and other papers that share them may not be as useful for this new task. Second, bibliographies are often extensive and only contain limited amounts of information of each of their references. Consequently, the list of similar articles using the bibliographic coupling may be sparse. If, instead, the researcher could access which papers where co-cited with her main reference, then the work may be easier. A higher co-cite similarity implies that many independent authors linked her meaningful reference with other works she may use. This process is exogenous and typically does not depend on the author of the initial paper.

Problem 4.2: Co-offending Network

For part (a),
For part (b),
For part (c),
For part (d),
For part (e),
For part (f),
For part (g),
For part (h),
For part (i),
For part (j),
For part (k),
For part (l),
For part (m),