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 $k \in \{1, \ldots, w\}$. 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 $k \in \{1, \ldots, w\}$. 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 $k \in \{1, \ldots, w\}$, 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

$$\beta = C C^T.$$

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

Note: The dataset provided for this question has been altered in such a way that is is unrecognizable from the original.

The original dataset contains 1,4565,786 observations, each at the `offender_id-event_id-crime_code` level. This is, each observation is a unique person, arrested at a given crime event, and charged with a given crime. Note that at any given crime event, the offender can be charged several different crimes. Within, the number of unique offenders is 538,851, and the number of different crime events are 1,163,423, which occurred between 1997 and 2004, according to Table 1.

In terms of crimes, there is a total of 301 different crime codes. The 5 crimes that occurred the the most are shown in Table 2. Together, they comprise a total of 516,876 perpetrators, and correspond to, approximately, 35% of the sample. In terms of crimes in general, there is high variability. Moreover, a great number of crimes have low occurrences, as can be seen in Figure 1a. To gain more insight, the sample is restricted to crimes with less or equal than 5,000 offenders. For reference, the average number of offenders per crime is, approximately, 4,869. With this, of the 294 different codes, 254 (86.3%) remain. The result can be seen in Figure 1b. The plot changes slightly. Most of the codes have little offenders and, among the rest, there is high variability.

In terms of location, these five most notorious crimes occur mainly in county 682. This could be due to 682 being the most populated county in the database. The details can be found in Table 3. The table shows that, for crime 131, this county accounts for 84,952 offenders. This corresponds to, approximately, 78.63% of all offenders arrested for this crime. The rest of the crimes present a share in county 684 that is large, but much smaller than for crime 131, round between 20% and 29%. To further see the relevance of this county, Table 4 shows, for each crime, in how many different counties it was detected. Additionally, the table shows the contribution each county should have if crimes occurred evenly in each one. This contribution is computed as follows:

$$\text{Contribution}_{\text{crime}} = \frac{100}{\#\text{ Different counties}_{\text{crime}}}$$

This table shows how concentrated is the distribution of crimes in county 682. In particular, it can be seen that crime 131 is present in almost a half of counties in which the rest of the crimes are. This can explain the abrupt jump in contribution in Table 3.

Now, to create the co-offending network, the library `networkx` is used. This Python library is useful to compute metrics in an efficient way. Before using this module, the first step is to create the adjacency matrix, using `scipy`'s sparse matrix class. First, the complete network, including solo offenders is created. A solo offender is a person whose only neighbor in the network is himself. The details of the network are presented in Table 5. By construction, the original network (without removing solo offenders) has the same number of self-edges than of nodes (because offenders were arrested with themselves). After removing solo offenders, the total number of self-edges reduces to 121,119, which implies that the total number of solo offenders was 417,732 (= 585,851 - 121,119). From here onwards, the self-edges of the network were removed, to avoid computational and interpretation issues. Consequently, the network will have 538,851 nodes and 178,372 edges. In particular, some nodes will not be connected to any other node.

The degree distribution of this network is particular. In Figure 2 is presented the degree distribution in log-log scale. In Figure 2a is presented as a histogram and in Figure 2b as a smoothed discrete pdf. Although the figures do not show a full power law distribution, it is possible to observe a linear behavior up to, approximately, degrees 15 or 20. Afterwards, the plots becomes erratic, as fewer nodes have degrees over these caps.

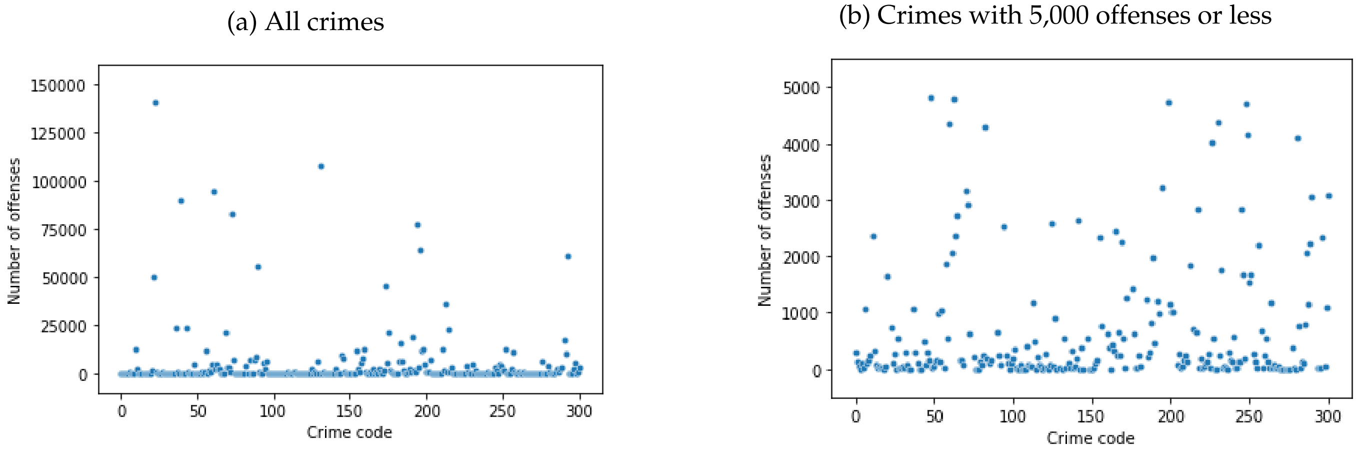
To shed some light about the cumulative distribution of degrees of this network, Table 6 shows the total of nodes with degree at most d, for different values of d. For reference, the total number of nodes with positive degree is 121,119. As can be seen from the table, almost the entire network (in terms of connected nodes) has degree at most 20, with 98.36% of all nodes. This means that the right tail is not so heavy. The cumulative distribution is shown in Figure 3.

Now the attention is drawn to the connected components of the network. This network has 36,086 connected components. The smallest one has 2 nodes, while the largest one comprises 19,921 nodes. The latter is, approximately, a 16.45% of all nodes with positive degree. This largest connected component may include some big criminal organization, and therefore is worth studying separately. Focusing on this largest connected component, the first step is to review the degree distribution. Interestingly, for this sub-network, the degree distribution does not follow a power-law distribution, although there seems to be two linear segments: from degrees 1 to 3, and from degrees 3 to 20. This can be seen in Figure 4 (in log-log scale). As before, in Figure 4a the distribution is presented as a histogram and in Figure 4b as a smoothed discrete pdf.

Following the same exercise as with the whole network, in Table 7 shows the total of nodes with degree at most d, for different values of d. In this case, the total number of nodes with positive degree is equal to the size of the connected component: 19,921. Comparing this table to Table 6, it is possible to observe that the cumulative percentage at all levels is lower. This means that the degree distribution of the largest connected component has a heavy right tail. To see this, in Figure 5 is presented the cumulative distribution of this connected component (in blue), alongside the one for the whole network (in red). The plot shows that both curves are still very much apart at degree 30 and in fact they coincide almost at degree 100.

Finally, another important aspect of this connected component is its shape. Give its size it is not possible to plot, but some descriptive statistics regarding shape can be presented. The diameter of this connected component is 48. This means that the two most separated nodes are at a distance of 48. In the extreme case where this connected component is a line, the diameter would be the number of nodes minus 1, that is 19,920. This is 415 times the diameter obtained. This may imply this network has a more hierarchical structure, which would be expected in an organization. Additionally, the clustering coefficient 0.478. This is large and means that, approximately, half of the triples of connected nodes form a triangle. This could mean that in top of the hierarchical structure, there are groups or "teams" which work together in some crimes and that depend on members with higher relevance. This hypothesis may be further sustained due to the edge density of the connected component. In this case the density is, approximately, 0.00036, which is extremely close to 0. For reference, a full graph (a graph in which al nodes are connected) has a density of 1 and, conversely, a graph with no edges has a density of 0. If the hypothesis of the organizational structure were not true, then much more nodes would have connections, as there is no leader commanding the criminal network. Finally, if the organizational hypothesis is true, then there should be some nodes with high betweenness centrality and low degree (the bosses). Although, it is also possible to have *brokers*, this is, individuals that connect two distinct organizations together. This could, in principle, have higher degree if they are connected with several parts of both sub-organizations. In Table 8 are presented the 10 most important nodes in terms of betweenness centrality, along with their degree. The table shows that some of this nodes have low degree and others have large degree. The former might be important individuals in the organization, while the latter might be brokers, connecting different organizations.

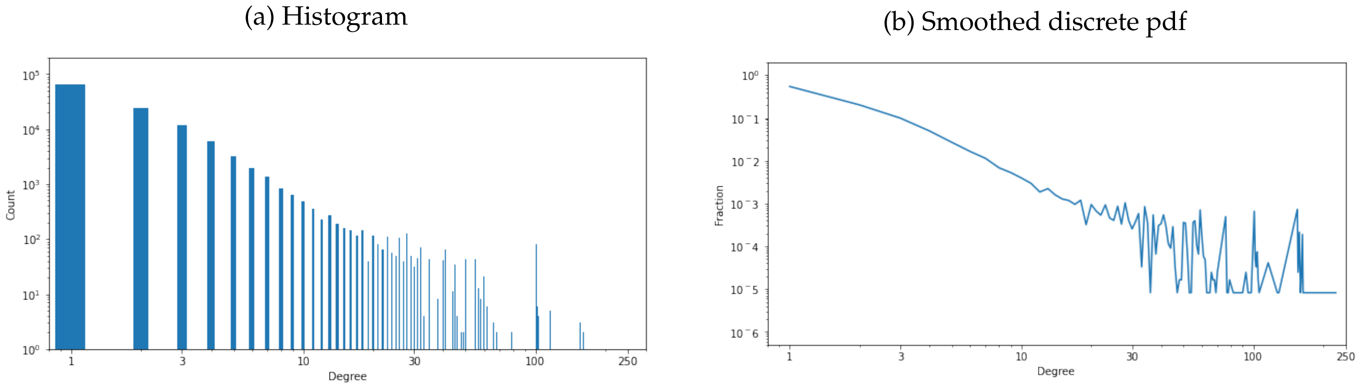
Figure 1: Total offenses per crime code. In panel (a), all crime codes are considered. In panel (b), only crime codes with a number offenses lower or equal to 5,000. For reference, the average number of offenses per crime code is 4,869.



(a) All crimes

(b) Crimes with 5,000 offenses or less

Figure 2: Degree distribution of the co-offending network. In panel (a), the degree distribution as a histogram. In panel (b), as smoothed discrete pdf.



(a) Histogram

(b) Smoothed discrete pdf

Figure 3: Cumulative degree distribution of the co-offending network.

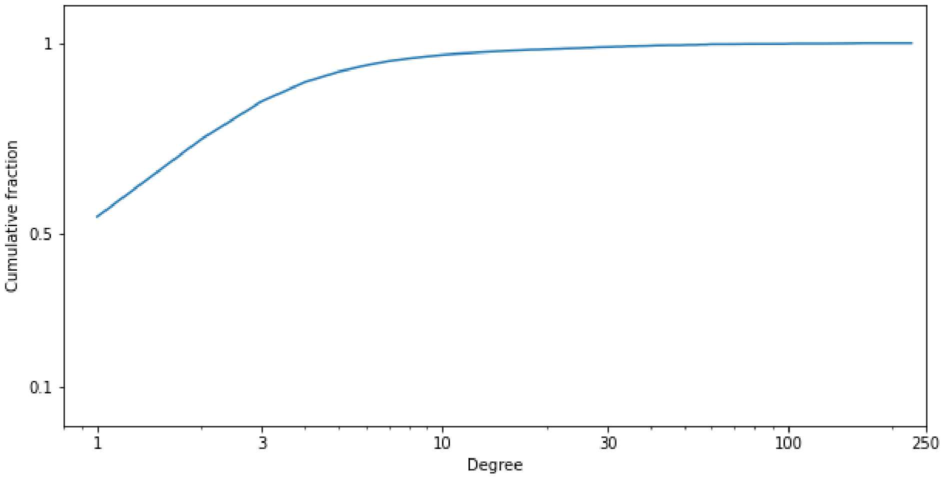
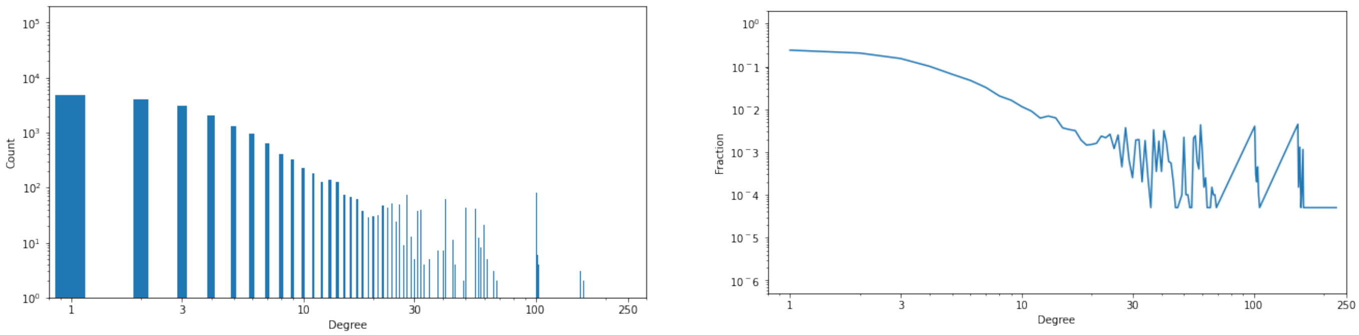


Figure 4: Degree distribution of the largest connected component. In panel (a), the degree distribution as a histogram. In panel (b), as smoothed discrete pdf.



(a) Histogram

(b) Smoothed discrete pdf

Figure 5: Cumulative degree distribution of the co-offending network.

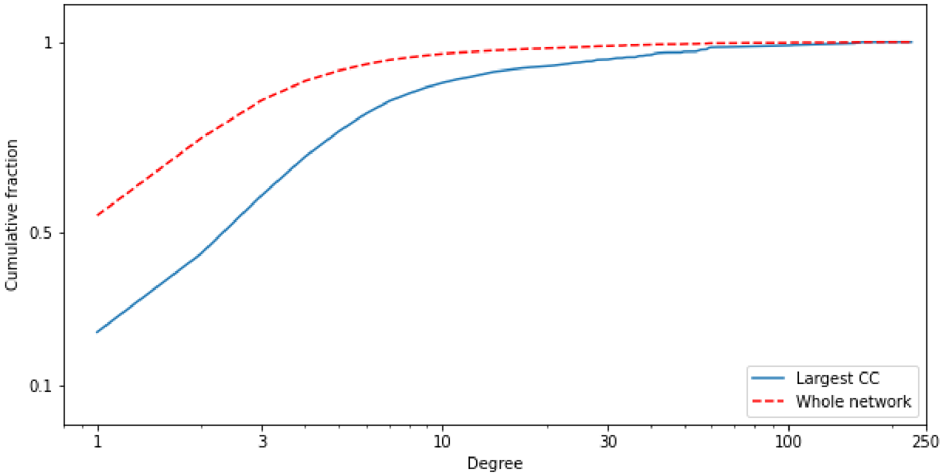


Table 1: Number of arrests per year

Year	Number of arrests
1997	22,404
1998	142,582
1999	163,346
2000	220,197
2001	234,495
2002	246,397
2003	251,298
2004	185,067

Table 2: Crime codes with most offenses. Only the top five are shown. The last row shows the total offenses for these five crimes.

Crime code	Number of offenses
22	140,901
131	108,031
61	94,655
39	90,179
73	83,110
Total	**516,876**

Table 3: Most notorious crimes and their occurrence in county 682. Column "Total offenses" is taken from Table 2. Column "Contribution" is the percentage of offenses that occur in county 682.

Crime code	Offenses in county 682	Total offenses	Contribution
131	84,952	108,031	78.63%
22	28,247	140,901	20.05%
61	24,101	94,655	25.46%
73	23,743	83,110	28.57%
39	21,698	90,179	24.06%

Table 4: Most notorious crimes and number of counties they occur in. Column "Uniform contribution" represents the contribution of each county if all occurrences where evenly distributed among counties.

Crime code	Number of different counties	Uniform contribution
131	652	0.15%
22	1,228	0.08%
61	1,084	0.09%
73	1,099	0.09%
39	1,212	0.08%

Table 5: Network descriptive statistics

	With solo offenders	Without solo offenders
Total nodes	538,851	538,851
Self edges	538,851	121,119
Non-self edges	178,372	178,372
Total edges	717,223	299,491

Table 6: Cumulative degree distribution of the co-offending network. Each row shows the number of nodes (and the percentage) that have at most degree *d*, for different values of *d*.

d	Number of nodes	Percentage
3	102,689	84.78%
5	112,042	92.51%
10	117,373	96.91%
15	118,576	97.90%
20	119,134	98.36%
25	119,498	98.66%
30	119,849	98.95%

Table 7: Cumulative degree distribution of the largest connected component. Each row shows the number of nodes (and the percentage) that have at most degree *d*, for different values of *d*.

d	Number of nodes	Percentage
3	11,937	59.92%
5	15,253	76.57%
10	17,793	89.32%
15	18,433	92.53%
20	18,660	93.67%
25	18,958	94.66%
30	19,009	95.42%

Table 8: Betweenness centrality of the ten most important nodes. Their degrees are also provided.

Node	Betweenness centrality	Degree
3	0.1218	15
5	0.1112	9
10	0.0966	9
15	0.0793	16
20	0.0690	14
25	0.0625	22
25	0.0553	8
25	0.0541	14
25	0.0522	10
30	0.0496	21