PNAS
www.pnas.org

Supplementary Information for

Shorter Distances between Papers over Time are Due to More Cross-Field References and Increased Citation Rate to Higher Impact Papers

Attila Varga
Email: attilavarga@email.arizona.edu

**This PDF file includes:**





www.pnas.org/cgi/doi/10.1073/pnas.1905819116

## S1 Generating random networks

The random networks were created with an edge swap algorithm. This approach preserves the two degree distributions of the source and target papers. The process starts by first randomly choosing two edges in the network. In the following step the target papers of the edges swap their links to the two source papers. As an example, let's say that the randomly chosen two edges are X → A and Y → B. By swapping the two edges between the nodes we get X → B and Y → A. No edge pairs are allowed to be swapped if it would result in the creation of multiple edges with the same source and target papers. This procedure is repeated 100 * *E* times for each network, where *E* is the number of edges in the network.

## S2 Edge clustering coefficient

### *S2.1. Computation*

The edge clustering coefficient *C* used in this study is similar to the edge clustering coefficient proposed by Radicchi and his collaborators (1) for simple networks (one-mode networks). In their formulation the edge clustering coefficient is the ratio of the number of triangles formed by the focal edge, and the maximum possible number of triangles given the size of the neighborhood of the two nodes forming the focal edge. There are two differences between this coefficient and *C*. The numerator of *C* is not the number of triangles formed by the focal edge, but rather the number of quadrilaterals. Bipartite networks do not contain triangles, or 3-cycles. It is a common practice in studies of complex networks to measure the cliquishness of bipartite networks with quadrilaterals, or 4-cycles (2-4), which is the largest maximal subgraph (nodes forming all possible links) containing more than two nodes in such graphs.

The second difference is that the denominator is not the maximum number of cycles, but rather it is the randomly expected number of cycles determined by the degree distributions in the neighborhoods. The number of cycles formed by the edge is limited by the size of its neighborhoods, as well as by the degrees of nodes in its neighborhood. The higher the degrees, the more likely it is that the neighboring nodes form an edge. Utilizing this approach to normalize the observed number of edges - instead of using the maximum number of possible edges - helps to make the results across years comparable. This is an important consideration, because the average degrees in the network and the degree distributions change substantially over time. Figure S1 describes in detail how *C* is calculated.

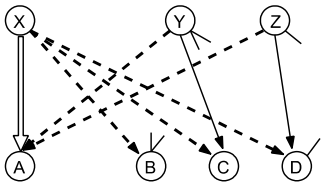


**Figure S1.** The computation of the edge clustering coefficient. The focal edge, or citation, is X ⟶ A. X is the source node or paper, and A is the target node. X cites B, C, and D. A is cited by Y and Z. These two sets of nodes B, C, D and Y, Z are the neighborhoods of X and A respectively. The clustering coefficient is the log ratio of observed (*obs*) and expected (*exp*) edge frequencies between the neighborhoods: $C = \log(\frac{obs+1}{exp})$. The calculation of the expected frequency is based on the assumptions of the configuration model: the expected number of edges in the neighborhood in a random attachment process is the function of the nodal degrees. One is added to the numerator to be able to calculate the coefficient even if there is no edge between the neighborhoods. In the example the number of observed edges is 2, Y ⟶ C and Z ⟶ D. The expected edge frequencies are calculated according to the excess degree distributions of the two neighborhoods. The figure indicates the edges that are outside of the neighborhood with stubs. The neighbors of X have the excess degree distribution (not counting the dashed lines to the focal nodes) 2, 1, 2 and A's neighborhood has the distribution 3, 2. The expected frequency is the product of each degree in the two sets divided by the number of edges (*E*) in the network: $exp = \sum_{i,j} k_i k_j / E$, where $k_i$ and $k_j$ are degrees in the two neighborhoods. In the example, where $E = 100$, $exp = 0.2$, and $C = \log\left(\frac{3}{0.25}\right) = 2.48$.

*S2.2. Observed and expected frequencies*

The median number of connections between the neighboring nodes of the source and target papers, or in other words the median number of 4-cycles formed by a citation, increased from 2 to 5 (Figure S2/A). The similar inverted U-shape trend is observable regarding this quantity as with *C*. It reaches its peak median value 6 between 1995 and 2005, before it falls back to 5. The overall increase can be explained by the fact that the degrees of source and target papers also increased. At the same time the expected number of edges did not increase (Figure S2/B). Although the increased degrees pushed up the chance of observing edges between the neighborhoods of the citation edges, the overall probability of an edge in the network (or the graph density, Table S1) was constantly decreasing. The net effect of these trends is that the expected frequency of edges remained more stable through time.

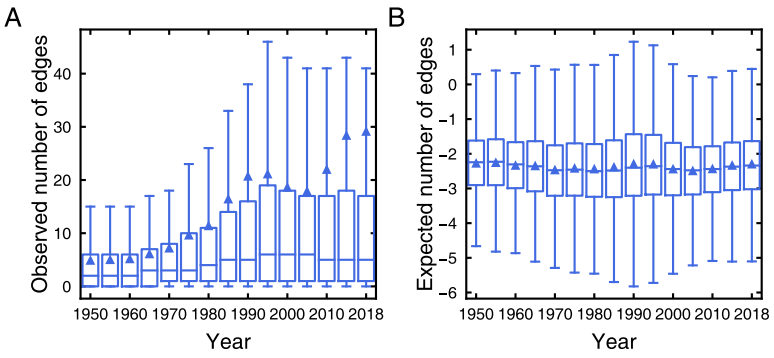


**Figure S2.** The distribution of the (A) observed and (B) expected number of edges in *C*. The figure shows the base 10 logarithm of the expected frequencies.

**Table S1.** Descriptive statistics of the networks. The average citation impact (the degree of the target papers) and the average number of references (the degree of the source papers) were computed for the network only including target papers that have at least one citation. This is a more informative statistic because although papers with only one citation are very frequent, this does not affect the relations between source papers. Note that despite the decreasing graph density, the distances in the random networks are decreasing over time (Figure 1/A). This is somewhat surprising, because lower densities imply larger distances in the network. What reduces the distances substantially in the random networks is the growing importance of high degree target papers. The average distance increases only in 2000, at the same time when the Gini index also falls back slightly (Figure 2/A).

| | # of source papers | # of target papers | # of citations | Target papers with >1 citations | Average citation impact | Average # of references | Giant component | Density |
|---|---|---|---|---|---|---|---|---|
| 1950 | 17,846 | 151,449 | 211,622 | 21.1% | 2.88 | 6.30 | 85.2% | 7.83E-05 |
| 1955 | 22,065 | 201,767 | 289,360 | 22.3% | 2.95 | 7.03 | 88.7% | 6.50E-05 |
| 1960 | 32,971 | 293,458 | 441,695 | 24.4% | 3.07 | 7.56 | 90.9% | 4.57E-05 |
| 1965 | 50,636 | 468,688 | 733,595 | 26.0% | 3.18 | 8.60 | 92.3% | 3.09E-05 |
| 1970 | 92,660 | 788,994 | 1,338,042 | 29.4% | 3.36 | 9.21 | 94.0% | 1.83E-05 |
| 1975 | 133,148 | 1,155,305 | 2,116,564 | 32.2% | 3.59 | 10.70 | 95.6% | 1.38E-05 |
| 1980 | 178,004 | 1,616,074 | 3,074,067 | 33.6% | 3.68 | 11.87 | 96.5% | 1.07E-05 |
| 1985 | 235,568 | 2,226,269 | 4,439,644 | 34.9% | 3.85 | 13.29 | 97.2% | 8.47E-06 |
| 1990 | 274,160 | 2,841,917 | 5,835,565 | 35.3% | 3.99 | 15.13 | 97.7% | 7.49E-06 |
| 1995 | 342,005 | 3,788,492 | 8,200,617 | 36.8% | 4.16 | 17.46 | 98.4% | 6.33E-06 |
| 2000 | 440,153 | 4,856,915 | 10,994,579 | 39.5% | 4.20 | 18.74 | 98.7% | 5.14E-06 |
| 2005 | 528,412 | 6,079,432 | 14,286,191 | 41.5% | 4.26 | 20.62 | 99.1% | 4.45E-06 |
| 2010 | 640,477 | 7,940,220 | 19,526,310 | 43.3% | 4.37 | 23.68 | 99.5% | 3.84E-06 |
| 2015 | 738,550 | 9,809,231 | 24,880,301 | 43.9% | 4.50 | 26.38 | 99.7% | 3.43E-06 |
| 2018 | 759,594 | 10,605,080 | 26,872,152 | 43.9% | 4.49 | 27.66 | 99.8% | 3.34E-06 |

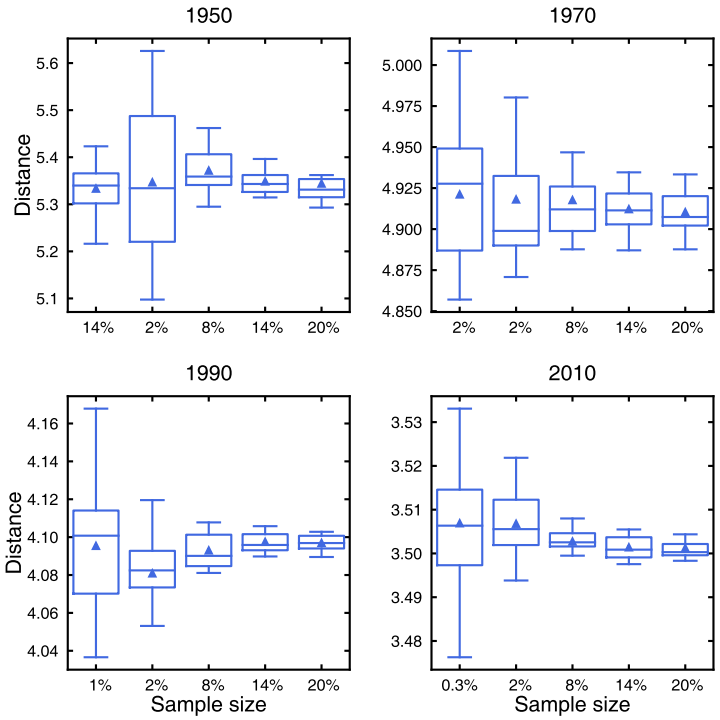


**Figure S3.** Robustness of the sampling procedure. Throughout the paper I use a fixed sample size of 2,000 nodes to measure the distances. However, because the studied networks are of very different sizes, and because degree distributions are heterogeneous, I conducted a robustness test of that sampling approach. The figure shows the results for five different proportional sample sizes in four years. In each year the first box plot represents the average distances of 30 samples containing 2,000 nodes. The proportion of that sample varies by the size of the network, which is shown on the x-axis. The rest of the experiments are similar distributions of average distances in varying sized samples. Each of these proportional samples is repeated 10 times for each year. In each year the average distances, measured for the various sample sizes, cluster together. Furthermore, the studied temporal trend is clearly distinguishable, because the sample distributions do not overlap across the years.

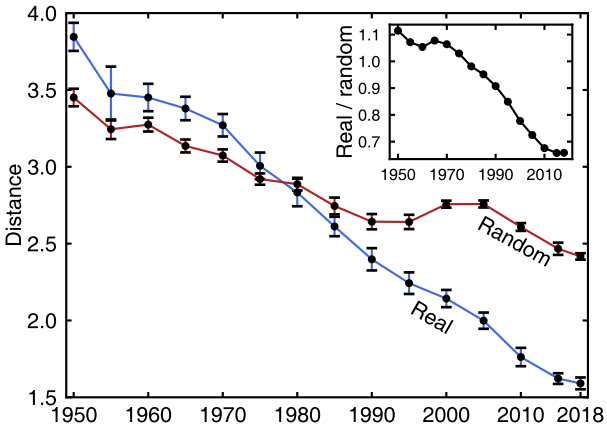


**Figure S4.** Weighted distance. The networks presented here are the projections of the bipartite networks into one-mode networks. The projected nodes in the networks are the source nodes, and their edges are weighted. The weights are the reciprocals of the number of co-cited target papers between the two source papers. For example, if X cites A, B, C, and Y cites B, C, D in the bipartite network, the weighted edge between X and Y in the projected network is 1/2, because X and Y have two overlapping target papers in common (B and C). The smaller the weight the shorter the distance between the two source papers. The figure shows the average distances in those networks and three SDs from the averages. The estimation of the average distances is based on the same repeated sampling procedure described in Figure 1. Each estimation is based on 10 repeated samples. The randomly expected averages are derived from the projections of 10 bipartite random networks.

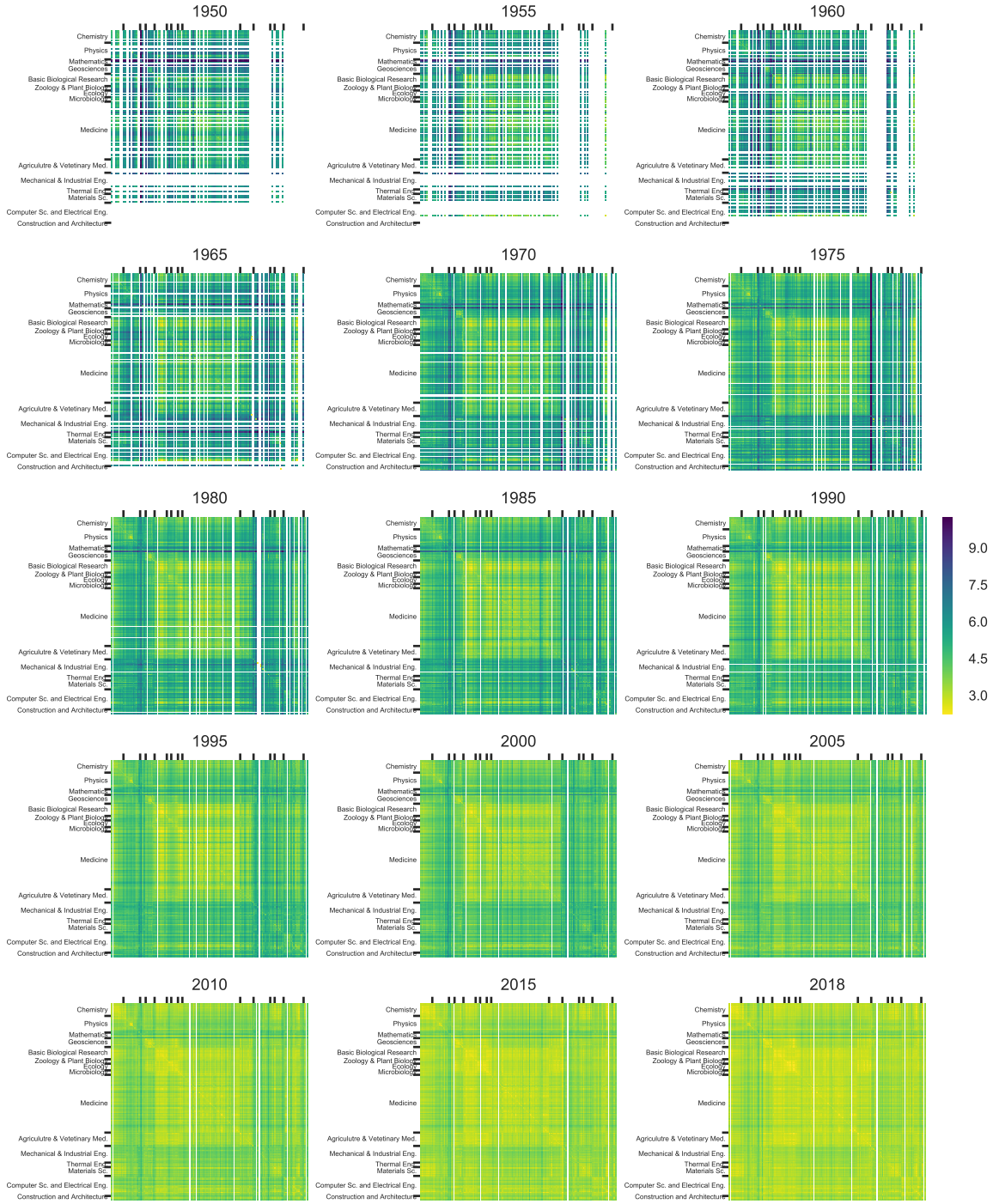


**Figure S5.** Distances between subdisciplines (Subject Categories). Each row and column represents a subdiscipline in the matrix, and the cells show the distances between the pertaining subdisciplines. The estimation of the distances is based on the following sampling procedure. A sample of 10,000 source papers have been selected in each year, and the shortest paths have been calculated to this sample from all source papers. The larger sample size was needed to increase the likelihood of catching smaller subdisciplines with a fewer number of papers in the given year. Subdisciplines that are not represented with a journal in the samples or across the years are blank. Papers that are published in journals with multiple subdisciplines are represented in all their subdisciplines in the matrix. The grouping of subdisciplines is based on the Deutsche Forschungsgemeinschaft's classification system. Within the same branch, the subdisciplines are ordered alphabetically. Note that distance matrices are not completely symmetric. This is because reading the matrix row-wise gives the distance from all papers in a Subject Category, whereas the column-wise reading gives the distances from the sampled papers.

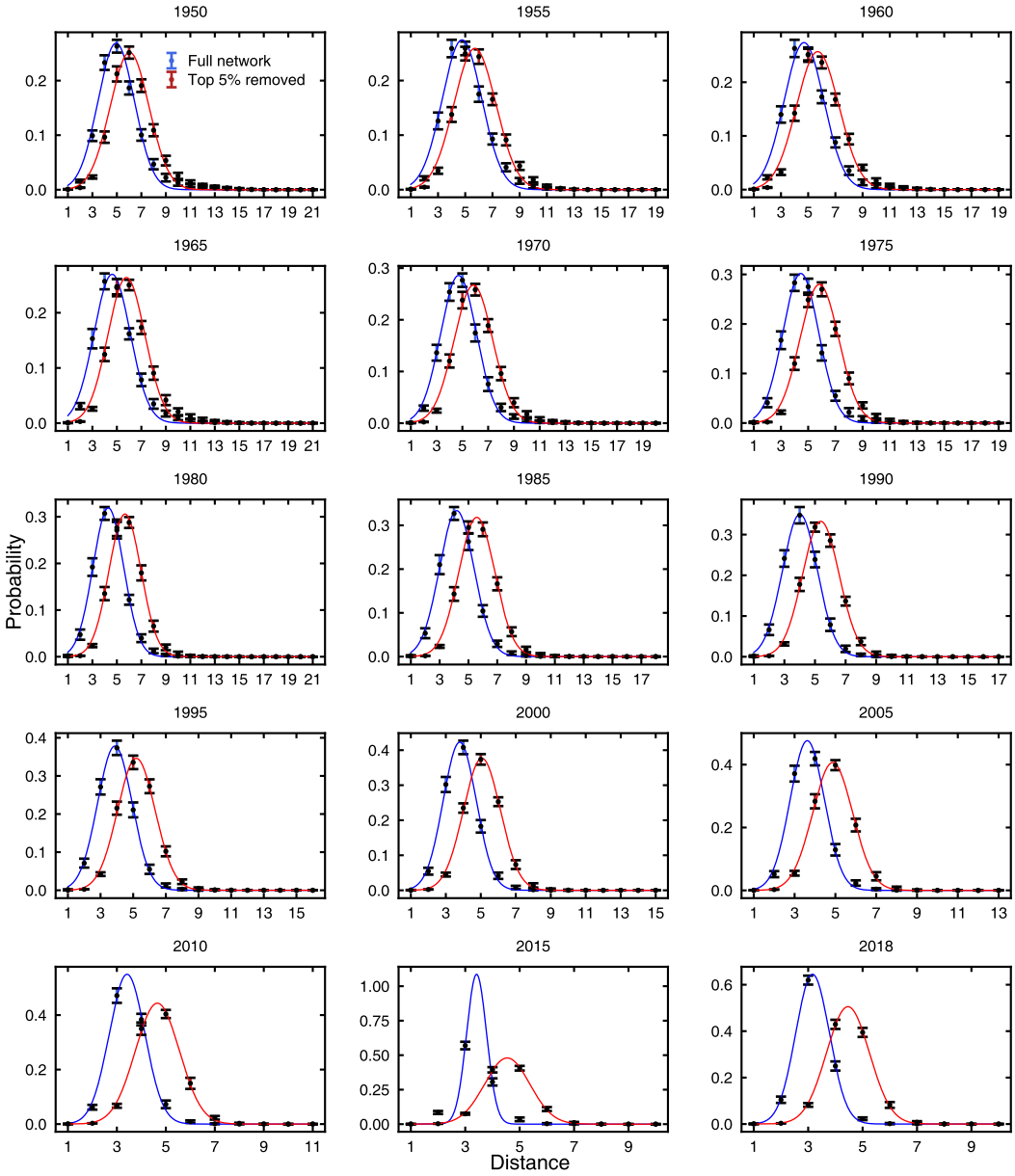


**Figure S6.** Distribution of distances (blue), and the resulting distribution of distances after removing the top 5% highly cited papers (red). Each marker on the figure represents the average probability of a given distance in the thirty repeated samples. Error bars are three SDs. Normal curves are fitted to the distributions.

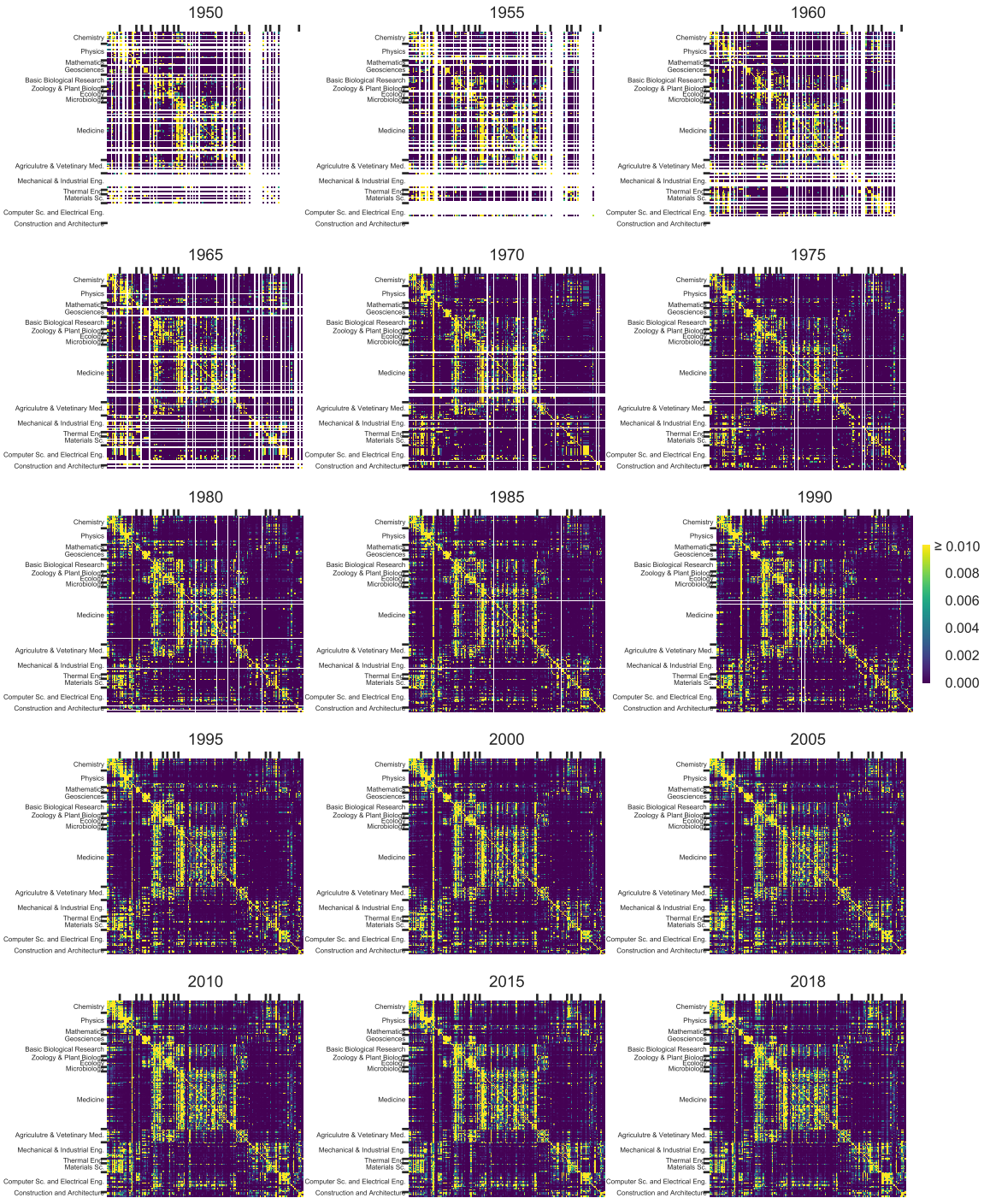


**Figure S7.** Subdiscipline to subdiscipline citation percentage matrices. Each cell in the matrices shows the percentage of citations from the row to the column subdiscipline. The maximum brightness of the cells is set to indicate cells with ≥ 1% frequency. The intention behind this coloring is to highlight the dispersion of citations across the years. Papers that are published in journals with multiple subdisciplines are represented in all their subdisciplines in the matrix. The grouping of subdisciplines is based on the Deutsche Forschungsgemeinschaft's classification system. Within the same branch, the subdisciplines are ordered alphabetically.

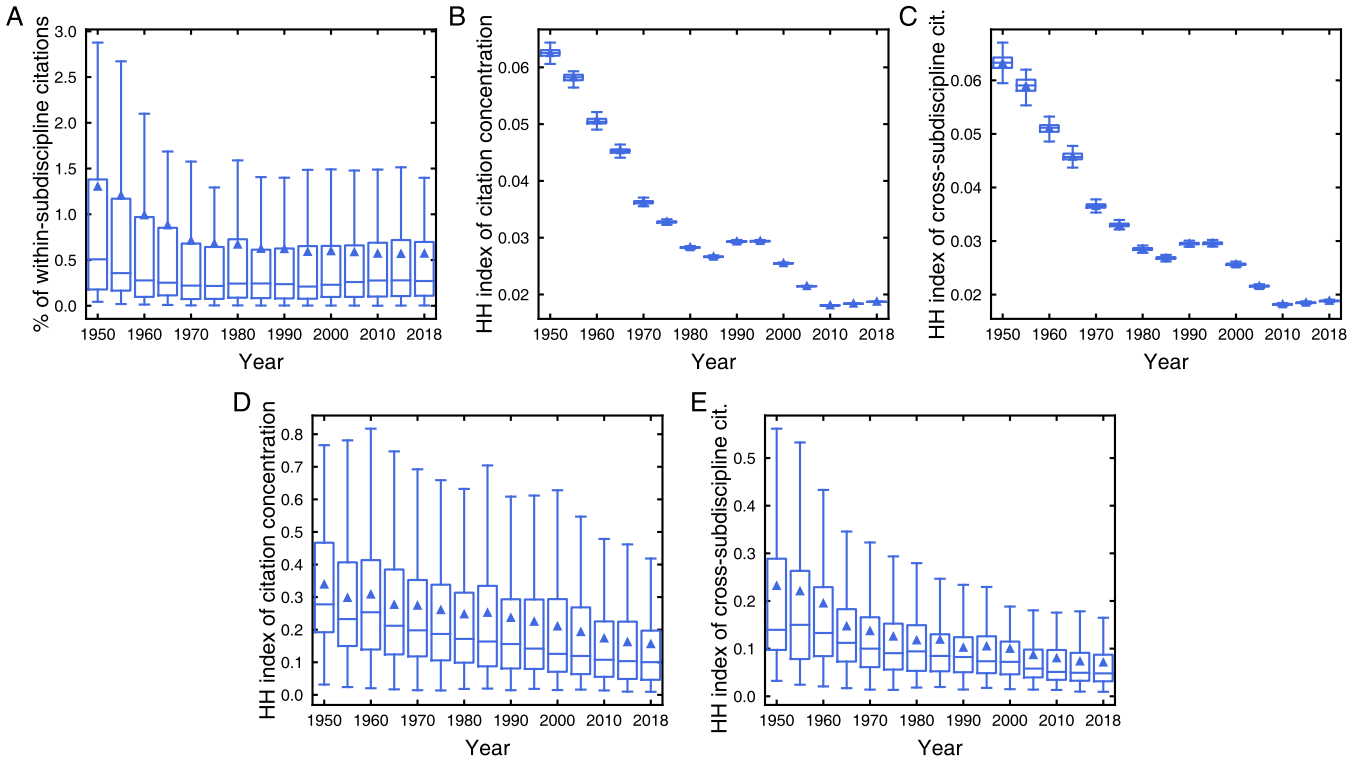


**Figure S8.** Randomly expected distribution of citations between subdisciplines and alternative HH index calcuation. The first three subplots are essentially the same as Figure 3, but this figure shows a single random network. Note that the size and the number of each subdiscipline are the same as in the real data, and the edge swap algorithm does not alter these quantities. (A) The marked decrease of within-subdiscipline citations from 1950 to 1975 is due to the increasing number of subdisciplines, and the later stabilization of that number. Because there are more categories by time, the citations disperse more. The number of subdisciplines in the dataset increased by 95% from 1950 to 1975, and from 1975 to 2018 it increased by 19%. (B-C) The concentration index of randomly distributed edges in the categories reflects the skewness of the size distribution of categories. Recall that the row-wise calculation of the HH index shows the concentration of target papers cited by the source papers in the row categories. In the randomized network, these row-wise distributions of target paper categories simply follow the overall target paper degree distribution, which is the column marginal distribution of the matrix. (D-E) The purpose of calculating the concentration index here is to remove the possible distorting effect of the marginal distributions. Because the HH index of these distributions shows a decreasing temporal trend as seen above, this may affect the trend of the real HH index distributions in Figure 3/B-C. The index here is based on the column-wise percentages. In the original version of the HH index the input vectors are row-wise percentages: the vectors that are used for the calculation are the percentage distributions of target papers across the subdiscipline categories for each row. In this version a vector represents the column wise percentages: each value in the vector is the percentage of target papers in the given subdiscipline that fall into the given row-wise source paper's subdiscipline. The input vectors still represent the importance of cited subdisciplines for a given subdiscipline, but are not affected by the column-wise marginal distribution. This vector does not add up to 100%, or one, and therefore I normalized it to conform to the calculation of the HH index. These transformations ensure that the index is not affected by the marginal distributions.

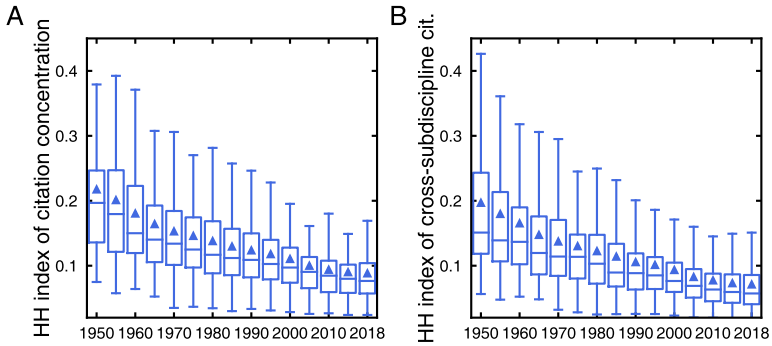


**Figure S9.** Normalized HH index. The formula for the normalization is (HH-1/N)/(1-1/N), where N is the number of subdisciplines in the given year. This normalization ensures that the number of subdisciplines does not affect the index. (A) The index is taking into account all the subdisciplines. (B) The index is calculated for citations where the source and target subdiscipline differ.

**Software**

The data analysis was conducted in Python. The Python package igraph (5) was used for measuring shortest paths. Pajek (6) was used for cross-validation purposes.

**References**


1. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci USA 101: 2658–2663.
2. Robins G, Alexander M (2004) Small worlds among interlocking directors: Network structure and distance in bipartite graphs. Computational & Mathematical Organization Theory 10(1): 69-94.
3. Lind PG, Gonzalez MC, Herrmann HJ (2005) Cycles and clustering in bipartite networks. Physical Review E 72(5): 056127.
4. Latapy M, Magnien C, Del Vecchio N (2008) Basic notions for the analysis of large two-mode networks. Social Networks 30(1): 31-48.
5. Csárdi G, Nepusz T (2006) The igraph software package for complex network research, Inter. J. Complex Sys. 1695
6. V. Batagelj, A. Mrvar (1998) Pajek - Program for Large Network Analysis. Connections 21(2): 47-57.