THE COMMUNITY STRUCTURE OF COLLABORATION NETWORKS IN COMPUTER SCIENCE AND ITS IMPACT ON SCIENTIFIC PRODUCTION AND CONSUMPTION

A Preprint

Ana Maria Jaramillo
BioComplex Laboratory
Department of Computer Science
University of Exeter, UK
aj499@exeter.ac.uk

Hywel T.P. Williams
SEDA Lab
Department of Computer Science
University of Exeter, UK

Nicola Perra
School of Mathematical Sciences
Queen Mary University of London, UK

Ronaldo Menezes
BioComplex Laboratory
Department of Computer Science
University of Exeter, UK

July 21, 2022

ABSTRACT

Collaboration networks, where nodes represent authors and edges coauthorships among them, are key to understand the consumption, production, and diffusion of knowledge. Due to social mechanisms, biases, and constraints at play, these networks are organized in tight communities with different levels of segregation. Here, we aim to quantify the extent, features, and impact of segregation in collaboration networks. We study the field of Computer Science via the Semantic Scholar Open Research Corpus. We measure segregation of communities using the Spectral Segregation Index (SSI) and find three categories: non-segregated, moderately segregated, and highly segregated communities. We focus our attention on non-segregated and highly segregated communities, quantifying and comparing their structural topology and core location. When we consider communities of both categories in the same size range, our results show no differences in density and clustering, but evident variability in their core position. As community size increases, communities are more likely to occupy a core closer to the network nucleus. However, controlling for size, highly segregated communities tend to be located closer to the network periphery than non-segregated communities. Finally, we analyse differences in citations gained by researchers depending on their community segregation level. Interestingly, researchers in highly segregated communities gain more citations per publication when located in the periphery. They have a higher chance of receiving citations from members of their same community in all cores. Researchers in non-segregated communities accrue more citations per publication in intermediary and central cores. To our knowledge, our work is the first to characterise segregated communities in scientific collaboration networks and to investigate the relationship between segregation and impact measured in terms of citations. Our results help detect and describe highly segregated communities of scientific collaborators and could pave the way to intervention strategies aimed at reinforcing the growth of science.

Keywords science of science · coreness · collaboration networks · segregated communities
1 Introduction

Understanding the social structures behind scientific production may have profound implications in promoting the growth of knowledge, the well-being of our societies and the evolution of research [9]. Indeed, many studies have shown how socially influenced behaviours impact different aspects of the scientific enterprise. Examples include the selection of collaborators, citations, and the review processes which are biased by author attributes, such as prestige [18], gender [34], and country of affiliation [32, 26].

Collaboration networks, where nodes describe researchers and links collaboration patterns among them, have been identified as key to understand and map the production of Science [39, 28, 29]. Particular attention has been devoted to their structural properties. These networks are organized in communities formed by groups of highly collaborative researchers with relatively low external interactions [23]. By looking at the evolution of these networks in time, one might see these communities going from being disconnected components to join the giant component. When comparing the proportion of nodes in the giant component relative to the total number of nodes, there are critical transition points, which have been shown to represent the constitution of new disciplines and the growth of science [3].

As in any activity driven by human interactions, the biases mentioned above influence the formation of communities and their connection/disconnection with other parts of the network. On the one side, previous literature has shown how the lack of exposure to individuals outside their own circle can create segregated groups [35]. In other contexts, such as discussions on social media, this "structural segregation" [15] can increase polarization [31, 30], and reinforce similar opinions [7]. High segregation levels —found in social networks with very fragmented groups— hamper the development of social capital and the emergence of cooperative behaviour, which are both detrimental to innovation, social learning, and problem solving [13], all key elements of scientific practices. In particular, studies on collaboration networks of computer scientists have shown that researchers immersed in segregated groups have disadvantaged positions in accessing information [14].

However, on the other side, segregation in tight communities could increase the exploitation of innovative ideas and interdisciplinary work. For example, groups of researchers organized in efficient structures, characterised for being more interconnected and less clustered, proved to outperform others in solving complex problems [20]. Furthermore, researchers from evolutionary medicine located on the network’s periphery showed to produce better and lasting ideas [27]. There is a tension between consolidating and diversifying collaborations as both might affect the growth of scientific knowledge and research impact. Our understanding of when and how collaborations across communities can help expanding research methods and questions [25], as well as promote the spreading of scientific results [32, 33] is still limited.

In this context, we tackle three specific research questions: i) How to define segregated communities in collaboration networks? ii) Are there differences in the topological structure and core position of communities with different segregation levels? iii) Does the segregation level affect their impact measured by citations?

To answer these questions, we study collaboration networks made by pairwise links between researchers in Computer Science coauthoring a publication. We assume that communities of researchers with very high internal versus low external connectivity can be considered highly segregated. Interestingly we found three categories of segregation and a relation between the size, segregation category, and core position of the communities in which non-segregated communities tend to be in higher cores of the network. In addition, we found that researchers in highly segregated communities gain more citations when on the periphery of the network. In comparison, researchers in non-segregated communities gain more citations in the nucleus (with some variations depending on the productivity of the researchers).

Also, regardless of the core or productivity, highly segregated communities gain a higher proportion of their citations from their own communities.

The paper is organised as follows: Section 2 describes the dataset and network properties used in this study. Section 3 details the procedure and characterisation of community partition. Section 4 defines the structural segregation metric used in this study and categorises communities into non-segregated, moderately segregated, and highly segregated. To characterise the non-segregated and highly segregated categories, in Section 5, we calculate four metrics related to the topology and core position of the communities and compare them using distributions and z-scores. To analyse the implications for researchers in communities with different segregation categories, in Section 6, we compare the number of citations per publication and the proportion of citations received by members of the same community. Finally, Section 7 discusses our main contributions, limitations and final remarks.
Table 1: Characteristics of the collaboration network in 2011. The detailed growth of these metrics per year is in the SM in section S2

| Metric                        | Value          |
|-------------------------------|----------------|
| Number of nodes               | 812,464        |
| Number of edges               | 2,460,213      |
| Density                       | 7.45e-06       |
| Average clustering coefficient| 0.71           |
| Average binary degree         | 6.06           |
| Average weighted degree       | 7.35           |
| Average strength degree       | 1.76           |
| Number of connected components| 101,859        |
| Largest connected component   | 436,435        |

2 Data and networks

We analyse the emergence of segregated communities in the collaboration network of Computer Scientists. To this end, we consider the Semantic Scholar Open Research Corpus [17]. Our analyses correspond to 45 years from 1975 to 2020, and we built a collaboration network for each year. To simplify the manuscript, we display some of the main results of our analysis for one particular year (2011), and we compare the results of another two years in the Supplementary Material (2004 and 2014). The three years have similar results regarding the structure of the communities but differ in some of the citation analyses. Particular events did not inform the selection of years. The three snapshots serve just as an example. We leave the longitudinal analysis across all years for future work.

We built the coauthorship network from approximately 630,000 publications available in the dataset under the discipline of Computer Science during 2011. The characteristics of the resulting scientific network from the coauthorships are in Table 1. The temporal behaviour of these metrics is detailed in the Supplementary Material (SM) in Section S2. A node represents a researcher, and a link is formed if two researchers coauthored at least one scientific publication in the year. The links are weighted using the strength of the pairwise collaboration, defined as the sum of common publications dividing each publication by the number of coauthors [21, 5]. For comparison, Table 1 shows the average binary degree (links with weight 1 if researchers coauthor at least one publication), average weighted degree (links weighted with the number of coauthorships) and average strength degree (links weighted with the common publications over the number of coauthors). Further analyses of the degree distributions for the three ways of giving weight to the links are in the SM in section S1.

3 Community detection and description

To identify the community partition of the entire collaboration network, we test eight commonly used community detection algorithms divided into three categories: modularity optimisation, dynamical processes, and statistical inference [10]. To select which algorithm represents a better community detection, we must consider that all the coauthors of one publication form a clique [22], resulting in high clustering coefficients of our collaboration networks (Table 1). Following the methodology proposed in Ref. [10], we select the results from the Label-propagation algorithm because it finds communities that are less confounded by fully connected cliques and have higher average embeddedness of their nodes. The embeddedness of a node is its internal (inside the community) strength degree over its total strength degree [16]. The results of each algorithm are displayed in the SM in section S3.

We display the results of the Label-propagation algorithm over the years in Fig. 1. The number of communities grows to more than 200,000 in the last three years (see Fig. 1 A). The distributions of the size of each community, measured by the number of inner researchers for each year, are shown in Fig. 1 B. The communities have fewer than 20 researchers, a constant tendency each year, and the maximum size of a community rounds 1,000 researchers for the last five years. Finally, analysing the number of papers produced by each community, communities tend to publish less than 40 papers and have an upper limit of around 1,000 papers (Fig. 1 C). The last two metrics show that the detected communities produce more papers than their number of researchers. More information about this analysis is in the SM in section S4.
4 Community segregation

4.1 Spectral segregation index

To measure structural segregation in the detected communities of the collaboration network, we adapt the Spectral Segregation Index (SSI) from Urban Science [8]. The SSI measures segregation at the network, group and individual levels, differing from other segregation metrics and algorithms applied to social networks [4]. The SSI intends to measure the average segregation of the community members by computing the segregation of each community’s nodes based on the segregation of the node’s neighbours inside the community. This metric implies that a node’s segregation is calculated considering the linear combination of the node’s own proportion of internal connectivity and the neighbours’ proportion of internal connectivity (internal refers to links inside the community). Then, the SSI computation implies a reinforcing process in which a node with a high SSI value has neighbours with a high SSI. Following the procedure described by Echenique & Fryer [8], we first select the adjacency matrix and divide each row by the sum of the row (resulting in the transition matrix taking the total strength degree of the node). Then, we choose submatrices $B$ with the columns and rows corresponding to the researchers in each community. Finally, we select the largest eigenvalue $\lambda_1(B)$ of the submatrix $B$ and its corresponding eigenvector $V_1(B)$. The SSI for each community is computed over the submatrix $B$ as in Equation 1.

$$SSI(B) = \frac{\sum_{v\in B} \lambda_1(B) * V_1(B) \cdot v}{\|V_1(B)\|}$$ (1)

Because the eigenvector is computed as the stationary state of a “random walk” process, its values over the submatrix $B$, are shaped by the connectivity patterns within the community. Values of SSI near 0 mean low segregation, while values near 1 mean high segregation. Communities that are disconnected components have a SSI equal to 1, meaning perfect segregation [8]. It is worth mentioning that the SSI metric is different from the Modularity, a metric at the network level that measures the quality of a community partition by computing the difference between the fraction of inter-community edges versus intra-community edges for a given null model [24].

4.2 Defining segregated communities

To define the level of segregation, we computed the SSI considering the sole connections of the year $y$. We normalized SSI values for all communities to range from zero (no segregation) to one (completely segregated). When computing the SSI, we just studied 50,501 (33% from the original 150,972 communities) because the other communities were disconnected components and perfectly segregated (as shown in the SM in section S4). The presence of a large portion of disconnected communities is due to the temporal horizon considered, i.e., one year. We do not study those communities in section 5 because they could bias our results due to not having a core position (see below for details). However, we included completely segregated communities in the analyses of section 6 as a category of completely segregated communities. Most of them correspond to a sole publication implying a complete connected clique (density and clustering equal to one).
In Fig. 2, we show toy networks of highly segregated and non-segregated communities, in panels A and B, respectively. Those toy networks are “ego-networks” of the coloured community. We show in dark grey other neighbouring communities and in light grey links among different communities.

We compute the probability density function (PDF) of the SSI, its mean ($\mu$) and standard deviation ($\sigma$). We select as highly segregated communities those with a relatively high $SSI \leq \mu + \sigma$, and non-segregated communities those with a relatively low $SSI \leq \mu - \sigma$. This approach naturally leads to three categories of segregation: non-segregated, moderately segregated, and highly segregated. In Fig. 2 panel C, we show the PDF of the SSI for the year 2011, the division of segregation categories, and the number of communities in each category. This procedure ends with 7,192 highly segregated and 10,073 non-segregated communities. In grey, there are 32,266 communities with moderate values of SSI, which are not part of our network analyses because we want to study communities on the limits of the segregation spectrum. In section 6, we study the relationship between segregation, we include in the figures in grey color the results of the moderately segregated communities, which show intermediary results between non-segregated and highly segregated communities.

5 Segregated communities characteristics

To check for differences and similarities between non-segregated and moderately/highly segregated communities, we investigate and compare four metrics in total. The first three, regarding the structural properties of the communities, are: size (measured as the number of researchers), density (measured as the proportion of internal links over the set of all possible internal links), and clustering coefficient (measured as the number of triangles over the number of triplets within the community).

The core/periphery position of echo chambers [37] (segregated communities in online social networks) has been shown to influence their ability to spread information during social movements [1]. Therefore, in the context of scientific production, we want to understand if the communities’ positions in the collaboration network also relate to their segregation category. To this end, we compute and compare the core position of each community as the fourth metric. We first create a pruned network in which each community is a node, and links are formed if the members share coauthorships. Then, we apply the k-core decomposition algorithm [2] and assign each community to a correspondent core. These cores range from 1 (periphery) to 12 (nucleus) for the pruned network. More details about this calculation are given in the SM in section S5.

To compare the four metrics (size, density, clustering and core) of the segregated and non-segregated communities, we perform an statistical analysis in section 5.1 and section S8 in the SM.

5.1 Community metric distributions

To understand the overall differences between non-segregated and highly segregated communities, we compare the probability density function (PDF) for each metric in the 2011 network (Fig. 3) (analogous plots for different years are reported in the SM in S10.2). Comparing communities of different sizes can produce misleading results because, for example, a community of 3 nodes has a higher chance of being dense and clustered than a community of 20 researchers. Therefore, we group the communities by size and analyse the z-score of the SSI values in section S8 of the SM. In Fig. S6, we show the distributions for all, non-segregated and highly segregated communities. We found that smaller communities tend to have more significant values of SSI. We perform the following analyses by dividing the non-segregated and highly segregated communities into 10 different size ranges. Further explanations and analyses of the division are in section S7.

We first separate the communities by size, then by segregation category (i.e., non-segregated or highly segregated), and compute the PDFs of the four metrics: size, density, clustering and core location. The larger size of non-segregated communities is up to 37 researchers compared with 761 researchers in the highly segregated ones, showing us a tendency of highly segregated communities to be larger. In Fig. 3, we analyse the behaviour of the metrics comparing both types of communities, those sizes without non-segregated communities are not displayed here. Our results show no differences in the PDF of density and clustering for non-segregated and highly segregated communities inside the same size range, except for those smaller than five researchers, where highly segregated communities are less dense and clustered. As the communities size grows, the density column shows both types of communities decreasing their peak values from 1 to 0.1, and the clustering column shows decrements from 1 to 0.8. In the case of the core position, the differences are high. The core column shows that as both types of communities increase their size, their core positions go from the periphery (1st core) to the nucleus (12th core) but at different paces. The non-segregated communities reach faster the nucleus in sizes smaller than 40 researchers, while the highly segregated ones reach the nucleus when their size is larger. We perform the same PDF analyses without separating by size in the SM in section S9 to highlight
Figure 2: Classifying communities as non-segregated and highly segregated. A, B are toy networks of non-segregated and highly segregated communities, respectively, showing the ego-network of researchers connected to other communities in 2011. C shows the probability density function (PDF) of the spectral segregation index (SSI) for the year 2011. The plot is divided into three categories that denote non-segregated (highly segregated) communities with a value of SSI smaller (larger) than one standard deviation from the mean value of the SSI distribution. In grey are those communities within one standard deviation categorized as moderately segregated communities and are not part of the following analysis.
the importance of considering community size. The results are misleading when we do not divide the communities by size range because the differences are higher in density and clustering, and both types of communities have similar behaviour in their core position.

In conclusion, small communities are denser, more clustered and toward the periphery of the network. When they increase in size, their densities decrease to lower values, their clustering shows moderate decrements, and their core position shifts to the nucleus. There are no large differences in density and clustering between non-segregated and highly segregated communities, with values mainly driven by community size. On the contrary, the most evident difference is their core position. Non-segregated communities reach the nucleus for smaller sizes with respect to the highly segregated ones. We also analyse the z-score of the metrics in section S8 of the SM, and the results remain congruent with statistical differences between non-segregated and highly segregated communities just for the core position.

### 5.2 Comparing size and segregation of communities in different cores

Here, we aim to understand the relationship among the three variables that have shown higher differences between non-segregated and highly segregated communities: size, segregation and core position. To this end, we use kernel density estimators (KDE). Fig. 4 shows the KDE results with smooth 2D curves for segregated and non-segregated communities for all sizes and SSI values divided by core position. We display just 6 of 12 cores for readability, but the complete plot is displayed in the SM in section S9: Fig. S10 and Fig. S12.

At first, when comparing the number of communities per core, there are more small communities in lower cores (i.e., periphery), and few but large communities in higher cores (i.e., nucleus) as we can see in Table 2. This finding is in line with previous results [19], where the shell structure of some empirical and randomized networks showed to be explained by their community structure. When the core position increases the number of researchers also increase with more researchers in the nucleus of the network. The number of non-segregated and highly segregated communities has similar trends inside each core. Still, there are around twice non-segregated communities than highly segregated ones. Contrary, there are more researchers in highly segregated communities than in non-segregated ones. When comparing the size, highly segregated communities increase faster as we move towards the core of the network. Finally, when comparing the SSI, highly segregated communities have a small range of SSI for all cores, while non-segregated communities have a broader range of SSI in higher cores. The last can be explained by the shape of the SSI distribution and by the way we define the three categories.

Summarising, communities tend to have larger sizes towards the nucleus of the network, and the size of highly segregated communities increases faster when the communities go from the periphery to the nucleus. The communities located in the periphery of the network are more numerous and of smaller size, and those highly segregated in the nucleus have considerable larger sizes. This can be explained by the requirement of being a big community in order to be both: highly segregated and located in a high core. We repeat this analysis for the years 2004 and 2014 in the section S10 of the SM with similar results and relations among the four metrics.

### 6 Differences in citations gained by researchers of non-segregated and highly segregated communities

Here, we tackle the third, and final research question, investigating whether the segregation category of a community affects a researcher’s impact measured by citations. We consider both the number of citations, and also the sources of citations, to characterize whether highly segregated communities are more self-referential than non-segregated groups. For each researcher in non-segregated, moderately, and highly segregated communities, we analyse the citations received until 2020 by the publications in 2011 in Computer Science.
Figure 3: **Topological and core position differences among non-segregated and highly segregated communities.** The panels represent the probability density functions (PDF) in each column for the size, density, clustering, and core position of highly segregated (red) and non-segregated (blue) for 8 different size ranges. Each row represents communities with the number of researchers written in the Range label, and the PDFs were computed using just the communities in the sample after separating by size, hence the different y-axis limits.
We compute the cumulative density function (CDF) of four variables for researchers within the specific category of communities: i) Total number of citations, ii) Citations per paper, iii) Proportion of citations received by researchers in the same community, and iv) Proportion of citations from coauthors in the same year. For each variable, we analyse researchers in non-segregated versus researchers in highly segregated communities with different levels of granularity: i) all researchers without grouping them by core position and ii) researchers grouped by the core position of their communities. All curves y-axis start at 0 and end with values of 1, and we show 6 out of 12 cores in the main manuscript to improve visibility but the entire plot is reported in the SM in section S9.

Our results show that the number of citations highly correlates with the number of publications with a Spearman correlation coefficient of 0.44 (p-value < 10^-3). Then, to understand if there are confounding results for the four citation variables, we compare them grouping the researchers by core position and different productivity ranges. We select three ranges for the comparison: 1-5, 6-10, 10< published papers per year. The complete results of grouping the researchers by productivity ranges is in section S9 of the SM.

For comparing each pair of distributions we used two statistical tests: Kolmogorov-Smirnov and Mann-Whitney. The first test was used to compare the distributions shape and the second one to compare differences of medians. Each of the next plots contains in bold letters KS or MW when the p-value of the test is significant following a code of * < 0.1, ** < 0.05, and *** < 0.01. All corresponding values of each comparison are displayed in Table S1 in SM.

To study the first part of the question, we compute the total number of citations and the average number of citations per publication, as shown in Fig. 5 first row, and Fig. 6 first and second rows.

On an aggregated level, our results show no statistically significant differences when researchers are in non-segregated or highly segregated communities (as an artefact of averaging the results). However, the citations gained by researchers in completely segregated communities are considerable less than other communities (darker red in the plot in the first and second panels). When grouping the researchers by the core position of their communities, in Fig. 6, the results can be split into cores: highest (nucleus), middle, and lower (periphery). Researchers in non-segregated communities in the highest cores have significantly more citations than researchers in highly segregated communities (Fig. 6). This trend is stronger for researchers with productivity (>10 papers) (Fig. S10 in SM). When analysing middle cores, the differences are less apparent. However, researchers in non-segregated communities still tend to have more citations and citations per product than researchers in highly segregated communities, with some statistically significant comparisons. Lastly, in lower cores, researchers with high productivity (>10 papers), in highly segregated communities, have significantly more citations than those in non-segregated ones.

The results remain similar when analysing the citations per publication. Also, when comparing the results of 2011 with 2004 and 2014, we have consistent results. For high cores (center of the network), researchers in non-segregated communities get significantly more citations, and in lower cores (periphery of the network), the researchers of highly segregated communities get more citations. In the case of middle cores, both 2011 and 2014 have mixed results that need further exploration (detailed results of 2004 and 2014 in section S10.5 of the SM).

To study the second facet of the question, we compute the proportion of citations coming from members of the same community and the proportion of citations done by all researcher collaborators, as shown in Fig. 5 second row, and
Figure 5: **Citations metrics for all researchers in non-segregated and highly segregated communities** Each panel represents the cumulative density function (CDF) of TC: Total number of citations, CP: Citations per paper, CC: Proportion of citations received by researchers in the same community, and CN: Proportion of citations from coauthors in the same year for researchers that published papers in Computer Science in 2011 and are members of the studied communities. Red graphs correspond to the citations of researchers in highly segregated communities, while the blue ones correspond to researchers in non-segregated communities. The code of colors is: dark red for researchers in completely segregated (CS), gray for moderately segregated (M), light red for highly segregated (S), and blue for non-segregated communities (NS).

Fig. 6 third and fourth rows. For each variable, we analyse the authors citing the publications of Computer Science in 2011. For computing these proportions, we count the number of publications with at least one of the authors in the citing publication satisfying the rule of being in the same community or being a coauthor. Then, we divide these counts over the total number of citations.

On an aggregated level, our results show no statistically significant differences when researchers are in non-segregated or highly segregated communities (as an artefact of averaging the results). But researchers in completely segregated communities are receiving a considerably lower proportion of citations from their own community and their year collaborators than those in other communities (darker red in Fig. 5 in the third and fourth panels). When we group the core position, researchers in highly segregated communities, regardless of the community’s core position (Fig. 6), or the range of productivity (Fig. S10-13 in SM), have a larger proportion of citations from other members of their own communities. Also, researchers in highly segregated communities have a larger proportion of citations done by the coauthors than researchers in non-segregated communities. There is consistency across cores or productivity ranges (Fig. S12-13 in SM) with only few significant results. When comparing these results with 2004 and 2014, they follow similar trends. But in the highest core, researchers in non-segregated communities with the smallest productivity have more internal citations and more citations from coauthors (further analysis in section S9.2 of the SM).

In summary, researchers in highly segregated communities tend to have fewer citations when their communities are located at the network’s core. When they are in the periphery they need high productivity (> 10 papers) to attract more citations. They receive more citations from their own communities in all cores. At the same time, the results of middle cores need more exploration because of the mixed results. When comparing the results of 2011 with 2004 and 2014, the
main difference is that researchers in highly segregated communities in the nucleus of the network with low productivity (< 5 papers) receive a smaller proportion of citations by their community members and their coauthors. For the other cores, results in 2004 and 2014 are similar to those obtained in 2011.

7 Discussion

Due to a range of social mechanisms, processes and biases, collaboration networks are organized in communities [23]. Within-group dynamics might lead to the emergence of segregation and polarization, hampering innovation, social learning, and problem solving [15, 31, 30, 13]. Nevertheless, cohesive groups allow for the development of common narratives and language, offer support and share knowledge. As such, they have been identified as a locus for exploitation (when large in central locations) and exploration (when small in the periphery) of ideas, results, and methods [38, 27].

Still, the understanding of segregated groups in collaboration networks and their possible effects is limited. Here, we tackle this problem by quantifying segregation levels of communities in collaboration networks and characterizing their topological characteristics and position in the network. Furthermore, we study the relation between segregation and impact (measured using as proxy citations). To this end, we analysed the collaboration network of Computer Science in the Semantic Scholar Open Research Corpus. We detect communities with the label-propagation algorithm and compute the SSI as a structural segregation metric considering the links of the community each year separately. We find that communities lay in a spectrum that varies from non-segregated to highly segregated communities. Based on the distribution of the SSI, we identify three main categories and focus on just the two opposite limits: non-segregated and highly segregated communities. We adopt metrics of size, density, and clustering, the core position, and the citations gained by the publications forming the community, to study their topological properties and impact.

Our results show that small, highly segregated communities tend to be more on the periphery, with no large differences in density and clustering with respect to small non-segregated communities. When analysing the total number of
citations, researchers in highly segregated communities tend to receive fewer citations than non-segregated ones. In
addition, when analysing the sources of those citations, for researchers in highly segregated communities, up to 10%
more of those citations come from the same community than in non-segregated communities. Combining both results,
we infer that in terms of spreading ideas and knowledge in the collaboration network, being in highly segregated
communities could have two main effects: i) to attract fewer citations in the center of the network because researchers
get trapped in a chamber that breaks the balance in the exploitation/exploitation of ideas towards exploiting/echoing
scientific research [20]; ii) to gain more citations in the periphery because of the increased diversity of disciplines and
collaborators [40].

Both effects need further analyses because one could expect small and highly segregated communities located in the
periphery to have a smaller impact. Yet, our results show that when comparing communities in the periphery, researchers
in the highly segregated ones gain more citations, with significant differences when they are very productive (>10
papers). Individual success correlates with the exploitation of ideas [20] but also most innovative research (exploration
of new concepts and persistent citations) comes from the periphery of networks [27] and is done by smaller groups
of researchers [38]. Here, our results align with previous evidence showing researchers in the periphery being less
active [37] (i.e., publishing less in our case) but having more impact, where, altogether being a larger population,
become a collective power that can mobilize and spread information [1] (such as scientific theories).

For larger and highly segregated communities, being in central cores decreases their impact. Indeed, we find fewer total
citations in this group and a larger proportion of internal citations. Following previous evidence, we hypothesize those
researchers are exploiting/echoing the same scientific ideas [27]. These results need further exploration because their
central positions in the network’s core increase their chance of outside interactions with non-segregated communities,
which in turn can accelerate the propagation of echoed information (ranging from biased theories to new paradigms)
from local groups to reach the entire network [6]. The inner impact of highly segregated communities and over the
entire network should be measured to intervene, if necessary, and tackle or boost the spread of echoed information to
different groups [14].

7.1 Limitations

Here, we discuss the limitations of the methods and results of this study and how future analyses could address them.
First, our analysis does not generalize for all the years of Computer Science papers available in the Semantic Scholar
database because we studied just three years. We have developed a methodology that could be replicated over several
years, but further analysis needs to be done to understand how the transitions of researchers between different segregation
levels affect their research impact over time.

Second, our analyses do not generalise to all collaboration networks because the publications of Computer Science
in the Semantic Scholar Open Research Corpus represent a vast amount of literature in a discipline prone to working
in small teams [22]. Further analysis of other fields is needed to understand how these patterns apply to different
 collaboration structures.

Third, we did not classify the core-periphery type of our network. Recent work has highlighted the importance of
understanding if the network is prone to be divided into cores as layers (as we did with the k-core decomposition
algorithm) or if a hub/spoke core division is a better descriptor [11]. However, previous results show that authorship
networks are the most prone to have a core-layered typology as we used in the current work [11]. In further analyses,
the definition of segregated communities should also consider the scientific network’s core typology.

Finally, our fourth highlighted limitation relies on the sole use of the collaboration networks and citations to define
segregation levels and impact of communities. A more precise analysis could consider other features and data to provide
a better representation of the consumption and production of scientific knowledge [39]. Future work could consider
publications content, the researchers’ demographic diversity, and the interdisciplinary citations.

7.2 Future research

Future research on this topic could consider: i) the temporal analysis of segregated communities and their relation with
gaining more or less citations over time, ii) the analysis of diversity of the scientific publications inside the communities
using opinion distance [31] and their demographic diversity to understand if the segregated and isolated communities
are not diverse and echoing research to the point of becoming polarised, iii) the definition of lead researchers (using
the hub/spoke core or author position in the publications) and the understanding of their relationship to segregated
communities [12], iv) the measurement of the impact of the segregated communities on the topology of the network
formation and the spreading processes of scientific theories [36].
References

[1] P. Barberá, N. Wang, R. Bonneau, J. T. Jost, J. Nagler, J. Tucker, and S. González-Bailón. The critical periphery in the growth of social protests. *PLoS ONE*, 2015.

[2] V. Batagelj and M. Zaversnik. An O(m) algorithm for cores decomposition of networks. *arXiv:0310049*, 2003.

[3] L. M. Bettencourt, D. I. Kaiser, and J. Kaur. Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210–221, 2009.

[4] M. Bojanowski and R. Corten. Measuring segregation in social networks. *Social Networks*, 2014.

[5] T. J. Cann, I. S. Weaver, and H. T. Williams. *Is it Correct to Project and Detect? Assessing Performance of Community Detection on Unipartite Projections of Bipartite Networks*, volume 812. Springer International Publishing, 2019.

[6] J. T. Davis, N. Perra, Q. Zhang, Y. Moreno, and A. Vespignani. Phase transitions in information spreading on structured populations. *Nature Physics*, 2020.

[7] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, 6:1–12, 2016.

[8] F. Echenique and R. G. Fryer. A measure of segregation based on social interactions. *Quarterly Journal of Economics*, 2007.

[9] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A. L. Barabási. Science of science. *Science*, 359(6379), 2018.

[10] S. Fortunato and D. Hric. Community detection in networks : A user guide. *Physics Reports*, 659:1–44, 2016.

[11] R. J. Gallagher, J. G. Young, and B. F. Welles. A clarified typology of core-periphery structure in networks. *Science Advances*, 2021.

[12] L. Guo, J. A. Rohde, and H. Wu. Who is responsible for Twitter’s echo chamber problem? Evidence from 2016 U.S. election networks. *Information Communication and Society*, 23(2):234–251, 2020.

[13] A. D. Henry, P. Prałat, and C. Q. Zhang. Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8605–8610, 2011.

[14] Z. S. Jalali, W. Wang, M. Kim, H. Raghavan, and S. Soundarajan. On the information unfairness of social networks. *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020*, pages 613–621, 2020.

[15] S. Kim. Directionality of information flow and echoes without chambers. *PLoS ONE*, 14(5):1–22, 2019.

[16] A. Lancichinetti, J. Saramäki, M. Kivelä, and S. Fortunato. Characterizing the community structure of complex networks. *PLoS ONE*, 2010.

[17] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The Semantic Scholar Open Research Corpus. 2020.

[18] F. B. Lynn. Diffusing through disciplines: Insiders, outsiders, and socially influenced citation behavior. *Social Forces*, 93(1):355–382, 2014.

[19] I. Malvestio, A. Cardillo, and N. Masuda. Interplay between $\$k\$-core and community structure in complex networks. *Scientific Reports*, 10(1):14702, 2020.

[20] W. Mason and D. J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3):764–769, 2012.

[21] M. E. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. *Complex Networks*, pages 337–370, 2004.

[22] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 2001.

[23] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[24] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[25] M. W. Nielsen, C. W. Bloch, and L. Schiebinger. Making gender diversity work for scientific discovery and innovation, 2018.
[26] T. Opthof, R. Coronel, and M. J. Janse. The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular Research*, 56(3):339–346, 12 2002.

[27] D. T. Painter, B. C. Daniels, and M. D. Laubichler. Innovations are disproportionately likely in the periphery of a scientific network. *Theory in Biosciences*, 140(4):391–399, 2021.

[28] R. K. Pan, K. Kaski, and S. Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, 2(1):902, 2012.

[29] R. K. Pan, A. M. Petersen, F. Pammolli, and S. Fortunato. The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3):656–678, 2018.

[30] N. Perra and L. L. E. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9(1):1–11, 2019.

[31] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 2021.

[32] M. J. Smith, C. Weinberger, E. M. Bruna, and S. Allesina. The scientific impact of nations: Journal placement and citation performance. *PLoS ONE*, 9(10):1–6, 2014.

[33] D. H. Sonnenwald. Scientific collaboration. *Annual Review of Information Science and Technology*, 2007.

[34] C. R. Sugimoto, V. Lariviere, C. Ni, Y. Gingras, and B. Cronin. Global gender disparities in science. *Nature*, 504:211–213, 2013.

[35] C. R. SUNSTEIN. #Republic. Princeton University Press, ned - new edition edition, oct 2018.

[36] P. Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE*, 2018.

[37] H. T. Williams, J. J. R. McMurray, T. Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[38] L. Wu, D. Wang, and J. A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

[39] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley. The science of science: From the perspective of complex systems, 2017.

[40] C. Zingg, V. Nanumyan, and F. Schweitzer. Citations driven by social connections? A multi-layer representation of coauthorship networks. *Quantitative Science Studies*, 1(4):1493–1509, 2020.
S1  Degree distribution comparison

This section compares the distributions over time for the three possible weights of the links in collaboration networks, as mentioned in Section 2 of the Main Manuscript. In Fig. S1 we can see how the distributions of weighted (in green) and binary (in blue) degrees grow exponentially over the years, while in the case of the strength (in orange), the growth is more linear with some higher values being consistently high over the last 10 years. Because we want to understand which collaboration networks make more segregated communities, we decide to use the strength as the weight of the links. This decision is taken to have stronger collaborations that have fewer coauthors. When considering the strength, coauthors with high values would have a more peer-to-peer interaction for writing a paper, and their values are consistent for more years.

S2  Growth of the collaboration network metrics

This section analyses the behaviour over time for the 7 variables displayed in the Table 1 of the Main Manuscript. The order of the panels from A to G has the same order of variables in the table: “Number of nodes” to “Size of the largest connected component”. We can see in Fig. S2 (A) how the number of nodes grows over time as well as the number of edges (B). However, as expected, the number of real edges growth is not as fast as the number of possible edges ($\frac{(N\times(N-1))}{2}$) because not all newly added researchers (nodes) can collaborate (get connected) with the already existing ones. Hence, the density decreases over the years (C). Collaboration networks have the particular characteristic of forming full connected cliques among all the coauthors of one paper. Then, it is expected that their value of clustering coefficient is relatively high, considering the size of the network. In addition, there are new trends of working in teams over the last decades, the clustering coefficient increases from 1990 (D) together with the increments in the number of papers per community (as seen in Fig. 1 of the Main Manuscript). In the case of the average binary degree the number of coauthors (E) show peaks of increments over time. Finally, the number of connected components always increase (F) and there is a faster increase in the number of nodes in the Largest connected component (G) after during the 2000’s.
S3 Results and comparison of community detection algorithms

We analysed the community structure of collaboration networks to understand the formation of groups that could be closed in nature, as mentioned in Section 3 of the Main Manuscript. To avoid biased results in this analysis, we applied the eight most-used community detection algorithms divided into three categories: optimisation, dynamics, and statistical inference [?].

- Community detection algorithms based on optimisation: These algorithms perform optimisation techniques related to the modularity measure, which compares the number of edges within a community with its number of edges expected by chance [?].
  - Leading eigenvector: This algorithm expresses modularity in terms of the eigenvectors, and it performs spectral partitioning for community detection on the modularity matrix [?].
  - Multilevel: This algorithm maximises modularity in two phases. First, each node is the sole member of a community; then each node is grouped with each neighbour community, and it computes the modularity again. If the new modularity is larger, the grouped nodes form a new community; otherwise, the new group is discarded. In the case of a local maximum, i.e., no positive gains in modularity, the second phase converts each node in a community with weighted edges having the number of common edges and start the first phase again. Finally, iterate both phases until achieving a maximum modularity [?].
  - Fast greedy: This algorithm maximises the modularity in faster ways. First, assign each node to a sole member community, and compute a matrix with the gaining in modularity between each pair of communities. Second, find and select the maximum gaining in the matrix, merge both communities, and update the gaining in modularity matrix with the new communities. Finally, repeat the second step until one community remains [?].

- Community detection algorithms based on spreading process dynamics: These algorithms perform flow-based approaches in which the state of the nodes changes as a function of the neighbours’ states following spreading process dynamics with information of the links involved in the community detection as paths of flow.
  - Spinglass: In this algorithm a spin state represents the group index of each node in which a quality function, i.e. Hamiltonian function, contains four elements: 1) reward for internal links in the community, 2) reward for non-links between different communities, 3) penalisation for non-links in the community, and 4) penalisation for links between different communities. This algorithm starts by dividing the nodes into two groups, calculate the Hamiltonian function and divide each group again in two until achieving the minimum Hamiltonian function [?].
Supplementary Figure S3: Community detection results for the Semantic Scholar Strength network. The left panel represents the number of communities over time, the central panel represents the number of communities that are not disconnected components, and the right panel is the number of strong communities based on the embeddedness of their nodes. A community is considered strong if all nodes in the community have $k_{iC}/k_i > 0.5$, where $k_{iC}$ is the strength degree of the node $i$ inside the community $C$, and $k_i$ is the total strength degree of the node $i$ for the year.

- **Infomap:** This algorithm applies coding theory to compress streams that represent the probability of paths in the network traversed by a random walker. The entropy of frequencies of each path is computed, and nodes are grouped when they are part of paths with less entropy in the coding compression [7].

- **Walktrap:** This algorithm computes the Euclidean distance of two communities based on a random walker’s probability of being traversed. It should be larger when nodes are in different communities and smaller for nodes in the same community. First, each node is the sole member of a community, and then the pair of communities with the lowest distance of the iteration merge [8].

- **Label propagation:** This algorithm starts with a unique label for each node, and each node turns its label to the most common label in its neighbours. If there are ties, the label is chosen uniformly at random. This algorithm iterates until each node has the most common label in its neighbourhood [9].

- **Community detection algorithms based on statistical inference:** These algorithms perform stochastic models based on the probability of a connection between nodes.

  - **Stochastic Block Models:** This algorithm fits a generative network model on the data with maximum log-likelihood, considering that the probability of two neighbour nodes being in the same community exceeds the probability of two neighbour nodes being in different communities [10].

After computing the community detection algorithms over the 37 years, we can see how the number of communities grows over time with the size of the network. Those communities calculated with algorithms based on dynamics (Label propagation, Walktrap and Infomap) have almost the double number of communities that the algorithms based on optimisation (Leading eigenvector, Multilevel, and Fast greedy) as we can see in Fig. S3 in the left panel. The central panel shows the cause of the large difference in the number of communities: the algorithms based on dynamical processes have a larger number of connected communities, and we can conclude that dynamical algorithms tend to confound fewer communities with disconnected connected components.

To select the appropriate algorithm, we also analysed their internal and external connectivity. For the eight algorithms, we studied the strength of each community and its behaviour over time. For each node in all communities, we calculated its embeddedness as their internal community degree strength $k_{iC}$ over their total degree strength for the year. We labelled strong communities those with all nodes embedded in the community: $k_{iC}/k_i > 0.5$. Our results show that the community detection algorithms based on dynamic processes (Walktrap and Labelpropagation) have more strong communities than optimisation algorithms (Fast Greedy, Multilevel and Eigen Vector), as we can see in Fig. S3 in the left panel. From these results, we can conclude that the detected communities are cohesive, and then the network presents a well-defined community structure over the 55 years timeline. In terms of connected and strong communities and their strength, the Label propagation and Walktrap algorithms presented better results than the other algorithms. For this analysis, we chose the results of communities from the Label propagation algorithm to have a larger number of connected communities.
Supplementary Figure S4: **Comparison of community size and scientific productivity inside the communities.**

The panels represent the number (left) and the proportion (right) of internal scientific publications in all the communities (y-axis) vs the number of researchers inside each community (x-axis).

**S4 Number and proportion of papers produced vs the size of communities**

In this section, we analyse the number and proportion of publications that a community produces compared with the number of researchers that a community has, as we mentioned in Section 3 of the Main Manuscript. Our results show how there is an apparent tendency of communities to produce less papers than the number of researchers that they have, which on average would be less than a paper per person. When we calculate the number of communities above the diagonal line, we find just 0.98% (1,455 from 148,950) of the communities, resulting in a very small result for the scientific production inside the communities. An important disclaimer here is that the spread of communities in the upper diagonal is broader than in the bottom diagonal, which means that the community’s size is generally related to the productivity of a group.

After we divide the communities by highly segregated and non-segregated as described in Section 4.2 in the Main Manuscript, we also computed the size vs productivity of the researchers.

**S5 Network of communities with core location**

This section shows the network of communities with each community located in a core from periphery (1) to nucleus (12) as we can see in Fig. S5. We can see from the periphery to the core of the network how communities grow in size and that non-segregated (blue) and highly segregated (red) are in all cores of the network. However, in periphery cores there are more highly segregated communities, in the middle cores are more non-segregated communities and in the core of the network because the size of highly segregated communities increase it is difficult to see which communities are the most dominant. We analysed the statistically the relation among core position, size, and segregation category in the main manuscript Section 5.

**S6 Distributions of Spectral Segregation Index for different ranges of size**

This section compares the distribution of all, highly segregated, and non-segregated communities by ten different ranges of sizes from 3-5 nodes in the first range until 46-761 nodes in the tenth range, as shown in Fig S6. Highly segregated communities tend to be larger than non-segregated ones as we can see there are no non-segregated communities with more than 37 nodes. Also, when we compute the value of SSI, highly segregated communities have values less expected than the chance due to values smaller (larger) than -1(1) of z-score. In contrast, in the case of non-segregated communities, the values are mainly inside the [-1,1] z-score.
Supplementary Figure S5: Network of communities with the shell layout for highly segregated in red and non-segregated communities in blue. Each number makes reference to k-core in which each community is located.

Supplementary Figure S6: SSI distribution for different range sizes when considering all, highly segregated and non-segregated communities. The dashed lines represent a Z-score of -1 and 1 in each panel, as a visual guide for the communities with the behaviour expected by chance in the size range. Notice that the number of communities in the Segregated and Non-Segregated panels range does not sum up the number of communities in the panel All because there were communities not classified in these two categories.
Supplementary Figure S7: Distribution of topological and core values of highly segregated and non-segregated communities without dividing by small or large communities. Panels A-D represent the probability density functions (PDF) for the size, density, clustering and core position of highly segregated (red) and non-segregated (blue) communities.

S7 Comparing highly segregated and non-segregated communities without differentiating by size

This section shows the PDF’s without separating the communities by size range. From Fig. S6 we can observe that highly segregated communities tend to be larger, less dense, less clustered and there are no differences in the core position of the communities. However, as we see in the Section 5 of the main manuscript, when we divide the communities by size the results change because the density and clustering coefficient are not different between both categories of segregation but the core position is the one with high differences. Here, we argue that the analysis of the main manuscript makes it more reliable.

S8 Community metric comparison with z-scores correcting by size

This study uses network metrics with values highly influenced by the size of the communities, and from the previous analyses, we saw how the metrics change with size. Therefore, this part of the analysis aims to measure how significant are the differences between non and highly segregated communities in the same size range. Because size is the control metric, we solely analyse the density, clustering coefficient, and core position. We perform z-scores comparisons for each metric across segregation categories controlling for size. For example, we take the density of each highly segregated community in the size range: [3,5], and we calculate a z-score with the density of all non-segregated communities in the same size range. Then, we analyse the PDF distributions of all the z-scores as shown in Fig. S8 for 2011. We also compare the PDFs without separating them by size, and the results and analyses are in the SM in Section S8.

In line with the results of the previous section, the z-scores comparisons show no large differences in the density and clustering coefficient for both types when we control for size. The z-scores have their largest peak at zero for most cases (Fig. S7 first and second columns), meaning that these communities are not significantly different from other communities of the same size with a different segregation category. There is a modest difference in the density and clustering coefficient of small communities of size [3,5] in which highly segregated communities show to be less dense and clustered (Fig. S8 first row, first and second panels). The main difference is consistent with the core position of the communities. When controlling for size, all communities of different sizes have similar patterns. Highly segregated communities are in lower network cores (i.e., towards the periphery) than non-segregated ones in higher network cores (i.e., towards the nucleus) in all range sizes of comparison (Fig. S8 third column).

S8.1 Comparing segregated and non-segregated communities z-scores without differentiating by size

This section shows the results from the null-models of comparing the topological metrics and core position of segregated and non-segregated communities without correcting by size. The comparison is done by computing the z-score of each community with at least 30 communities of the opposite category. For example, we compute the z-score of one segregated community vs other 30 non-segregated communities. From Fig. S9 we can observe that segregated communities tend to less dense, less clustered and there are no differences in the core position of the communities. However, as we see in the Section 5 of the main manuscript, when we divide the communities by size the results change because the density and clustering coefficient are not different between both categories of segregation but the core position is the one with high differences. Here, we argue that the analysis of the main manuscript is more reliable because values of density and clustering are mainly driven by the size of the community.
Supplementary Figure S8: **Comparison of topological and core position of non-segregated and highly-segregated communities corrected by size.** The panels represent the probability density functions (PDF) for the z-score of comparing the density, clustering and core position of highly segregated (red) and non-segregated (blue) communities with opposite communities, i.e. highly segregated compared with non-segregated, of the same range size. The PDFs were computed using just the z-scores of comparisons that had at least 30 communities of the opposite category and the same range size to compare. The dashed line in zero represents no significant difference while being above zero represents a higher value of the variable, and being below zero implies smaller values.
Supplementary Figure S9: Comparison of topological and core position of segregated and non-segregated communities corrected by size without dividing by small or large communities. Panels A-C represent the probability density functions (PDF) for the z-score of comparing the density, clustering and core position of segregated (red) and non-segregated (blue) communities with opposite communities, i.e. segregated compared with non-segregated of the same size. The PDF’s were computed using just the z-scores of comparisons that had at least 30 communities of the opposite category and the same size to compare. The dashed line in zero represents no significant difference, while being above zero represent a higher value of the variable, and being below zero implies smaller values.

S9  Comparing citations of segregated and non-segregated communities

This section shows the results and procedure to compute and compare the citations of researchers in segregated and non-segregated communities for the year 2011. Because the number of citations received by an author is related with their number of publications (Section 5.4 of the main manuscript), we first divide the researchers by their number of publications to fairly compare the citations. We chose 3 categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year. The computed metrics were: i) Number of citations per product, ii) Number of citations, iii) Proportion of citations received by researchers inside the community, and iv) Proportion of citations from coauthors in the same year.

S9.1 Differences in citation numbers

This subsection shows the results for two metrics computed to answer the first component of the question: Does the segregation category of the community that a researcher is in affects their impact measured by citations? In subsection 5.4 of the main manuscript, we show a summarised version of Fig. S10. This component shows the number and number per product of citations gained during the period 2011-2020. For comparing each pair of distributions we used two statistical tests: Kolmogorov-Smirnov and Mann-Whitney. The first test was used to compare the distributions shape and the second one to compare differences of medians. Each of the next plots contains in bold letters KS or MW when the p-value of the test is significant following a code of * < 0.1, ** < 0.05, and *** < 0.01. All corresponding values of each comparison are displayed in Table S1. We can see from both plots that when the publications are small (1-10 papers), there is not a high difference between being in a segregated and non-segregated communities in periphery cores. In contrast, in middle and central cores the researchers in non-segregated communities gain more citations. There are larger differences when we compare the citations and citations per product with values of publications more than 10. In this case, researchers in non-segregated communities tend to have higher values of citations per product when in middle and central cores. Researchers in segregated communities have more total and per product citations when the community is on the periphery, as shown in both Fig. S11 and Fig. S10.

The previous results are significant when the core position of the communities increases towards the nucleus of the network. Specifically, we can see that the CDFs are more different when the productivity range increases. However, the number of researchers with the highest values for each metric is not enough to make the differences in distributions significant. Instead, the differences are more significant when the productivity is at its lowest value, in which the tales have considerable more values in the number of citations and citations per product. Also, the differences in the number of citations per product have more significant results than the entire number of citations.
Supplementary Figure S10: **Number of citations per product** Each panel represents the complementary density function (CDF) of the number of citations per product (2011-2020) received by researchers that published papers in Computer Science in 2011 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2011. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.

Supplementary Figure S11: **Number of citations** Each panel represents the complementary density function (CDF) of the number of citations (2011-2020) received by researchers that published papers in Computer Science in 2011 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2011. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S12: Proportion of citations received by researchers inside the community. Each panel represents the complementary density function (CDF) of the proportion of citations (2011-2020) from researchers that are part of the same community. This metric is calculated for researchers that published papers in Computer Science in 2011 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of publications in 2011. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.

S9.2 Differences in citation sources

This subsection shows the results for two metrics computed to answer the second component of the question: Does the segregation category of the community that a researcher is in affects their impact measured by citations? In subsection 5.4 of the main manuscript, we show a summarised version of Fig. S12. This component shows the proportion of citations received by researchers inside the community and by coauthors gained during the period 2011-2020. For comparing each pair of distributions we used two statistical tests: Kolmogorov-Smirnov and Mann-Whitney. The first test was used to compare the distributions shape and the second one to compare differences of medians. Each of the next plots contains in bold letters KS or MW when the p-value of the test is significant following a code of * < 0.1, ** < 0.05, and *** < 0.01. All corresponding values of each comparison are displayed in Table S1. We can see from both plots that researchers in segregated communities have larger proportion of citations from the same community, as shown in Fig. S12, and from coauthors, as shown in Fig. S13, than researchers in non-segregated ones. We can see how these differences are higher for researchers with values of publications larger than 10. An interesting result is that the proportion of citations gained by community member is slightly larger than the proportion of citations gained by the coauthors which could be explained by researchers from the same laboratory citing each other but not necessarily writing publications with each other.

The previous results are also significant when the core position of the communities increases towards the nucleus of the network. Specifically, we can see that the CDFs are more different when the productivity range increases. In this case, the number of researchers with the highest values for each metric is enough to make significant the differences in distributions. Here the differences are less significant when the productivity is at its lowest value, in which the tales do not have enough values of proportions of citations gained from the same community and coauthors of the same year. In this case, the differences in the proportion of citations gained by researchers of the same community have more significant results than the proportion of citations from coauthors in the same year.
Supplementary Figure S13: **Proportion of citations from coauthors in the same year** Each panel represents the complementary density function (CDF) of the proportion of citations (2011-2020) received from coauthors of the researchers that published papers in Computer Science in 2011 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2011. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.

**S10  Sensitive results to different years**

To understand if our results were sensitive to the year chosen, we analyse other two years before and after 2011: 2004 and 2014. The selection of these three years does not obey any specific rule. Further analysis of the temporal aspect of being part of segregated communities will be done in future work. We built the coauthorship network of these years for computer science with the main characteristics listed in Table S2. The analyses done in this section will be mainly comparative with the results of the year 2011 to check any temporal sensitivity in our analysis. The coauthorship network in 2004 has around half of the nodes and edges than in 2011, and there are 100 thousand more nodes in 2014. The density and clustering do not follow a linear pattern, and the LCC corresponds to 43% of the nodes in 2004 compared to 53% in 2011, and 56% in 2014, as we can see in Table S2.

**S10.1  Defining segregated communities**

For defining the segregated communities in 2004 and 2014 we followed the same procedure than in 2011. We compute the eight most used algorithms of community detection and select the strongest communities based on the embeddedness of their nodes and the less confounded with connected components. As we can see in Fig. S3 the results of the label propagation algorithms remain the best. Then, we compute the SSI to the resulted communities and we divide the communities into the three categories: segregated, mixed, and non-mixed. Comparing Fig. S14 for 2004 and 2014, and Fig. 2 for 2011 in the main manuscript. We can see similar distributions with two peaks near the inflection points. In both cases the largest category is the mixed communities with 65% of the communities for 2011 and 2014, and 75%. The proportion of segregated communities has increased from 5% to 14% since 2004 to 2011 and remained stabled in 14% in 2014.

We also compute the k-core decomposition of the network of communities for the years 2004 and 2014. The three years show more segregated communities in the peripheric cores while the non-segregated communities are more prominent in central cores, as we can see in Fig. S5 for 2011 and Fig. S15 for 2004 and 2014. Also, there are more cores in more recent years, and the central cores in 2011 and 2014 have more segregated and non-segregated nodes than in 2004, in which the mixed communities are not depicted and then appear empty in the figure.
Supplementary Table S1: P-values of kolmogorov-Smirnov (KS) and Mann-whitney (MS) statistical tests for difference of distributions of citations metrics between researchers in segregated and non-segregated communities differentiated by core position and productivity range.

| Core | Range   | Number of citations | Number of citations per product | Prop. of citations from coauthors in the same year | Prop. of citations received by researchers inside the com. |
|------|---------|---------------------|-------------------------------|-----------------------------------------------|-----------------------------------------------|
|      |         | KS 1.081         | MW 0.175                      | KS 0.571                                      | MW 0.220                                      |
| 1    | [1, 8]  | 0.336            | 0.168                         | 0.832                                         | 0.409                                         |
| 1    | [9, 80] | 0.175            | 0.109                         | 0.832                                         | 0.420                                         |
| 1    | [81, 91]| **0.004***       | **0.001***                    | **0.004***                                    | **0.011***                                    |
| 1    | [92, 109]| 0.034**          | 0.181                         | 0.336                                         | 0.377                                         |
| 2    | [1, 8]  | 0.832            | 0.228                         | 0.832                                         | 0.271                                         |
| 2    | [9, 80] | 0.081*           | 0.124                         | 1.000                                         | 0.462                                         |
| 2    | [81, 91]| **0.034**        | **0.057**                     | 0.832                                         | 0.430                                         |
| 2    | [92, 109]| 0.832            | 0.399                         | 0.983                                         | 0.405                                         |
| 3    | [1, 8]  | 0.832            | 0.280                         | 0.983                                         | 0.347                                         |
| 3    | [9, 80] | 0.983            | 0.347                         | 0.832                                         | 0.299                                         |
| 3    | [81, 91]| 0.571            | 0.473                         | 0.832                                         | 0.362                                         |
| 3    | [92, 109]| 0.832            | 0.462                         | 0.983                                         | 0.462                                         |
| 4    | [1, 8]  | **0.004***       | 0.033                         | **0.012**                                    | **0.045**                                    |
| 4    | [9, 80] | 0.832            | 0.368                         | 0.571                                         | 0.308                                         |
| 4    | [81, 91]| 0.336            | 0.182                         | 0.832                                         | 0.364                                         |
| 4    | [92, 109]| 0.175            | 0.166                         | 0.175                                         | 0.190                                         |
| 5    | [1, 8]  | 0.571            | 0.420                         | 0.832                                         | 0.495                                         |
| 5    | [9, 80] | 0.983            | 0.399                         | 0.336                                         | 0.262                                         |
| 5    | [81, 91]| **0.012**        | 0.036                         | 0.571                                         | 0.325                                         |
| 5    | [92, 109]| 0.175            | 0.166                         | 0.571                                         | 0.222                                         |
| 6    | [1, 8]  | 0.175            | 0.125                         | 0.175                                         | 0.113                                         |
| 6    | [9, 80] | 0.175            | 0.168                         | **0.012**                                    | **0.052**                                    |
| 6    | [81, 91]| 1.000            | 0.399                         | 0.175                                         | 0.159                                         |
| 6    | [92, 109]| 0.983            | 0.333                         | 0.175                                         | 0.054*                                        |
| 7    | [1, 8]  | **0.034**        | **0.044**                     | 0.336                                         | 0.196                                         |
| 7    | [9, 80] | 0.175            | 0.154                         | 0.175                                         | 0.167                                         |
| 7    | [81, 91]| 0.571            | 0.375                         | 0.081                                         | 0.227                                         |
| 7    | [92, 109]| 0.571            | 0.285                         | 0.832                                         | 0.473                                         |
| 8    | [1, 8]  | **0.004***       | **0.017**                     | **0.004***                                    | 0.016                                         |
| 8    | [9, 80] | 0.832            | 0.299                         | 0.832                                         | 0.420                                         |
| 8    | [81, 91]| 0.571            | 0.407                         | 1.000                                         | 0.440                                         |
| 8    | [92, 109]| 0.175            | 0.139                         | 0.336                                         | 0.051                                         |
| 9    | [1, 8]  | 0.336            | 0.094                         | 0.175                                         | 0.076                                         |
| 9    | [9, 80] | 0.336            | 0.182                         | 0.571                                         | 0.149                                         |
| 9    | [81, 91]| 0.832            | 0.458                         | 1.000                                         | 0.494                                         |
| 9    | [92, 109]| 0.571            | 0.254                         | 0.336                                         | 0.110                                         |
| 10   | [1, 8]  | **0.004***       | **0.033**                     | **0.081**                                    | **0.093**                                    |
| 10   | [9, 80] | **0.000***       | **0.012**                     | **0.000***                                    | **0.003**                                    |
| 10   | [81, 91]| 0.983            | 0.331                         | **0.081**                                    | **0.065**                                    |
| 10   | [92, 109]| 0.336            | 0.282                         | **0.081**                                    | 0.134                                         |
| 11   | [1, 8]  | 0.571            | 0.322                         | 0.983                                         | 0.473                                         |
| 11   | [9, 80] | 0.571            | 0.462                         | **0.034**                                    | **0.066**                                    |
| 11   | [81, 91]| 0.983            | 0.288                         | 0.983                                         | 0.465                                         |
| 11   | [92, 109]| 0.571            | 0.281                         | 0.336                                         | 0.222                                         |
| 12   | [1, 8]  | 0.571            | 0.308                         | 0.832                                         | 0.332                                         |
| 12   | [9, 80] | 0.571            | 0.328                         | 0.983                                         | 0.495                                         |
| 12   | [81, 91]| 0.571            | 0.397                         | 0.832                                         | 0.494                                         |
| 12   | [92, 109]| 0.175            | 0.127                         | 0.832                                         | 0.386                                         |
Supplementary Table S2: Characteristics of the collaboration network in 2004 and 2014. The detailed growth of these metrics per year is in the SM in Section S2.

| Metric                              | Value 2004 | Value 2014 |
|-------------------------------------|------------|------------|
| Number of nodes                     | 424,658    | 998,211    |
| Number of edges                     | 1,602,254  | 3,232,835  |
| Density                             | 1.77e-05   | 6.48e-06   |
| Average clustering coefficient      | 0.98       | 0.89       |
| Average binary degree               | 7.54       | 6.47       |
| Average weighted degree             | 7.35       | 7.35       |
| Average strength degree             | 1.76       | 1.76       |
| Number of connected components      | 64,880     | 118,074    |
| Largest connected component         | 183,635    | 567,037    |

Supplementary Figure S14: **Classifying communities as segregated and non-segregated for 2004 and 2014** Probability density function (PDF) of the spectral segregation index (SSI) for the years 2004 and 2014. The plot is divided into three categories that denote non-segregated (segregated) communities with a value of SSI smaller(larger) than one standard deviation from the mean value of the SSI distribution. In grey are those communities within one standard deviation categorized as mixed communities and are not part of the following analysis.

Supplementary Figure S15: **Network of communities with the shell layout** for segregated in red and non-segregated communities in blue. Each node is located in the core obtained after computing the k-core decomposition.
S10.2 Communities metrics distributions comparison

When comparing the probability density functions (PDFs) of size, density, clustering, and core position for the years 2004, 2011 and 2014, we can see similar results. For communities with 6 to 30 nodes, the results of size, density and clustering of the PDFs are not highly different among segregated and non-segregated communities. When comparing the smallest communities, we can see in Fig. S16 compared with Fig. S17 and Fig. 3 of the main manuscript that for both years, the segregated communities are larger (skewed to 5 nodes), less dense and less clustered. In the case of the core position, the behaviour remains similar: segregated communities are in lower cores, and where communities are bigger, both groups move to the core of the network, but non-segregated communities move faster and tend to be in higher cores. The main difference between both years is for the largest group of communities in which segregated communities seem denser and more clustered than non-segregated ones in the year 2004, which is not the case for 2011 either for 2014.

Suppose we do not divide the communities by size. In that case, the PDFs for all the metrics follow the same shape in the three years resulting in misleading interpretation as we can see in Fig. S6 and Fig. S18: Non-segregated communities seem to be smaller, denser, clustered and without differences in the core position when comparing with segregated communities. Contradictory results to the ones obtained dividing the communities by size.
Supplementary Figure S16: **Topological and core position differences among segregated and non-segregated communities for 2004** The panels represent the probability density functions (PDF) in each column for the size, density, clustering, and core position of segregated(red) and non-segregated(blue) for 7 different size ranges. Each row represents communities with the number of researchers written in the Range label, and the PDF’s were computed using just the communities in the sample after separating by size, hence the different y-axis limits.
Supplementary Figure S17: **Topological and core position differences among segregated and non-segregated communities for 2014.** The panels represent the probability density functions (PDF) in each column for the size, density, clustering, and core position of segregated (red) and non-segregated (blue) for 9 different size ranges. Each row represents communities with the number of researchers written in the Range label, and the PDF’s were computed using just the communities in the sample after separating by size, hence the different y-axis limits.
Supplementary Figure S18: Distribution of topological and core values of segregated and non-segregated communities without dividing by small or large communities for 2004 and 2014 Panels A-D represent the probability density functions (PDF) for the size, density, clustering and core position of segregated (red) and non-segregated (blue) communities.

S10.3 Communities metrics comparison with z-scores correcting by size

We can also see similar results when comparing the probability density functions (PDFs) of z-scores correcting by size for the density, clustering, and core position for the years 2004, 2011 and 2014. In both years, there is a high difference in the core position of segregated and non-segregated communities for communities with more than six nodes, the former being in lower cores while the latter in higher cores. When comparing the smallest communities, we can see in Fig. S19, Fig. S20, and Fig. 4 of the main manuscript that for the three years, the non-segregated communities are denser and more clustered, but they do not differ in core position. When the communities have more than 26 nodes in 2004, 31 nodes in 2011, there are some differences in density and clustering: segregated communities become slightly denser and more clustered. Both years have very similar results, but the differences in density and clustering are more pronounced in 2004 than in 2011, while there are not high differences for 2014.

Suppose we do not divide the communities by size. In that case, the PDFs of the z-scores for the metrics follow the same shape in the three years resulting in misleading interpretation as we can see in Fig. S9 and Fig. S21: Non-segregated communities are denser, clustered and do not have differences in the core position when comparing with segregated communities. Contradictory results to the ones obtained dividing the communities by size.

S10.4 Comparing the size and segregation of communities in different cores

When we compare the three variables with more pronounced differences: SSI, size, and core position, the results remain similar among the three years. Perhaps there are six cores in 2004, twelve cores in 2011, and thirteen in 2014, the behaviour of segregated communities being larger and with narrower values of SSI remains in the three years. Also, when we go towards the nucleus of the network, increasing the core position, the size of the communities gets higher, and this behaviour happens more for segregated communities as we can see in Fig. S22 for 2004 and 2014, and in Fig. 5 of the main manuscript for 2011.
Supplementary Figure S19: **Comparison of topological and core position of segregated and non-segregated communities corrected by size for 2004** The panels represent the probability density functions (PDF) for the z-score of comparing the density, clustering and core position of segregated (red) and non-segregated (blue) communities with opposite communities, i.e. segregated compared with non-segregated, of the same range size. The PDF’s were computed using just the z-scores of comparisons that had at least 30 communities of the opposite category and the same range size to compare. The dashed line in zero represents no significant difference while being above zero represents a higher value of the variable, and being below zero implies smaller values.
Supplementary Figure S20: Comparison of topological and core position of segregated and non-segregated communities corrected by size for 2014 The panels represent the probability density functions (PDF) for the z-score of comparing the density, clustering and core position of segregated(red) and non-segregated(blue) communities with opposite communities, i.e. segregated compared with non-segregated, of the same range size. The PDF’s were computed using just the z-scores of comparisons that had at least 30 communities of the opposite category and the same range size to compare. The dashed line in zero represents no significant difference while being above zero represents a higher value of the variable, and being below zero implies smaller values.
Supplementary Figure S21: Comparison of topological and core position of segregated and non-segregated communities corrected by size without dividing by small or large communities for 2004 and 2014. Panels A-C represent the probability density functions (PDF) for the z-score of comparing the density, clustering and core position of segregated (red) and non-segregated (blue) communities with opposite communities, i.e. segregated compared with non-segregated of the same size. The PDF’s were computed using just the z-scores of comparisons that had at least 30 communities of the opposite category and the same size to compare. The dashed line in zero represents no significant difference, while being above zero represents a higher value of the variable, and being below zero implies smaller values.

S10.5 Differences in citations gained by researchers of segregated and non-segregated communities

Because the number of citations received by an author is related with their number of publications (Section 5.4 of the main manuscript), we first divide the researchers by their number of publications to fairly compare the citations. We chose 3 categories of productivity 1-5 papers, 6-10 papers, and more than 10 papers published per year. When comparing the number of citations per product, the results are more significant for 2004 and 2014 than 2011. As shown in Fig. S24 in the first core, the CDF distributions are significantly different when the segregated communities have more citations per product than the non-segregated ones. When the core position increases, the results get opposite, and the non-segregated communities in the nucleus of the network are those with a significant difference and higher number of citations per product. The previous results are opposite to the ones in 2011, in which the segregated communities in the nucleus are the ones gaining more citations. However, when considering the results in 2014 in most of the cases the non-segregated communities have more citations per product with some exceptions, Fig. S27. In addition, in 2014 when the number of papers is higher than 88 the citations per product and total number of citations for researchers in segregated communities is higher. As we can see in Fig. S26, the previous comparison remains similar when comparing the total amount of citations.
Supplementary Figure S22: Relation between communities size, segregation and core position in the network for 2004 and 2014 Each panel represents the Kernel Density Estimation (KDE) for size in the y-axis, SSI in the x-axis and the core position of the communities. In red are those segregated, while non-segregated communities are in blues. Darker colours show a higher proportion of communities, while lighter colours represent fewer ones. Each panel shows the number of communities of each type used to compute the corresponding plot. The first cores show communities in the periphery, while the 6th and 13th cores show the communities in the network’s nucleus.
Supplementary Figure S23: **Citations metrics for all researchers in non-segregated and highly segregated communities** Each panel represents the cumulative density function (CDF) of TC: Total number of citations, CP: Citations per paper, CC: Proportion of citations received by researchers in the same community, and CN: Proportion of citations from coauthors in the same year for researchers that published papers in Computer Science in 2004 (top) and 2014 (bottom) and are members of the studied communities. Red graphs correspond to the citations of researchers in highly segregated communities, while the blue ones correspond to researchers in non-segregated communities. The code of colors is: dark red for researchers in completely segregated (CS), gray for moderately segregated (M), light red for highly segregated (S), and blue for non-segregated communities (NS).

Supplementary Figure S24: **Number of citations per product for 2004** Each panel represents the complementary density function (CDF) of the number of citations per product (2004-2020) received by researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2004. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S25: Number of citations per product for 2014 Each panel represents the complementary density function (CDF) of the number of citations per product (2014-2020) received by researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2014. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S26: Number of citations for 2004. Each panel represents the complementary density function (CDF) of the number of citations (2004-2020) received by researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2004. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S27: Number of citations for 2004 Each panel represents the complementary density function (CDF) of the number of citations (2014-2020) received by researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2014. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S28: Proportion of citations received by researchers inside the community for 2004. Each panel represents the complementary density function (CDF) of the proportion of citations (2004-2020) from researchers that are part of the same community. This metric is calculated for researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of publications in 2004. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.

When analysing the source of the citations with the proportion of citations received from members of the same community and coauthors of the same year the results for 2004 can be different to those obtained in 2011 and 2014. As we can see in Fig. S28, the proportion of citations from members of the same community is larger for those researchers in segregated communities in most of the cases for 2004 with significant results. However, differing from the results in 2011 in Fig. 7 of the main manuscript and Fig. S29 for 2014, there are some significant differences in the fifth core of different sizes. For 2004, in the smallest communities in the cores 4 to 6, the researchers in non-segregated communities have a higher proportion of citations from their same community. For 2014, just the researchers in the smallest non-segregated communities in the nucleus of the network (core 13) have significant more citations from their same community contrary to the panels with researchers in segregated communities having more internal citations.
Supplementary Figure S29: Proportion of citations received by researchers inside the community for 2014. Each panel represents the complementary density function (CDF) of the proportion of citations (2014-2020) from researchers that are part of the same community. This metric is calculated for researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of publications in 2014. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.
Supplementary Figure S30: Proportion of citations from coauthors in the same year for 2004 Each panel represents the complementary density function (CDF) of the proportion of citations (2004-2020) received from coauthors of the researchers that published papers in Computer Science in 2004 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2004. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.

When comparing the values of citations from coauthors of the same year, there is also a high difference in the results of 2004 and 2011-2014. Here it is interesting how for small communities in higher cores in 2004, the researchers in non-segregated communities get a larger proportion of citations from their coauthors than researchers in segregated communities, as we can see in Fig. S30. Also, in 2014 there are less significant differences between both categories and researchers in non-segregated communities tend to have more significant citations from their coauthors. These differences need further analyses in future work to understand the temporal mechanisms of gaining citations due to the segregation of the communities in which researchers are.
Supplementary Figure S31: Proportion of citations from coauthors in the same year for 2014. Each panel represents the complementary density function (CDF) of the proportion of citations (2014-2020) received from coauthors of the researchers that published papers in Computer Science in 2014 and are members of the studied communities. Each column corresponds to the core position of the community, and each row is a different range in the number of papers published in 2014. Red graphs correspond to the citations of researchers in segregated communities, while the blue ones correspond to researchers in non-segregated communities. Values for the division of rows show the chosen three categories of productivity: 1-5 papers, 6-10 papers, and more than 10 papers published per year.