Jaramillo et al. *EPJ Data Science* ( 2023) 12:47
https://doi.org/10.1140/epjds/s13688-023-00411-8

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

# The structure of segregation in co-authorship networks and its impact on scientific production

Ana Maria Jaramillo[1,2*], Hywel T.P. Williams[3], Nicola Perra[4] and Ronaldo Menezes[1,5]

*Correspondence:
ajaramillo@biocomplex.org
[1]BioComplex Laboratory, Department of Computer Science, University of Exeter, Exeter, UK
[2]Complexity Science Hub, Vienna, Austria
Full list of author information is available at the end of the article

## Abstract

Co-authorship networks, where nodes represent authors and edges represent co-authorship relations, are key to understanding the production and diffusion of knowledge in academia. Social constructs, biases (implicit and explicit), and constraints (e.g. spatial, temporal) affect who works with whom and cause co-authorship networks to organise into tight communities with different levels of segregation. We aim to examine aspects of the co-authorship network structure that lead to segregation and its impact on scientific production. We measure segregation using the Spectral Segregation Index (SSI) and find four ordered categories: completely segregated, highly segregated, moderately segregated and non-segregated communities. We direct our attention to the non-segregated and highly segregated communities, quantifying and comparing their structural topologies and *k*-core positions. When considering communities of both categories (controlling for size), our results show no differences in density and clustering but substantial variability in the core position. Larger non-segregated communities are more likely to occupy cores near the network nucleus, while the highly segregated ones tend to be closer to the network periphery. Finally, we analyse differences in citations gained by researchers within communities of different segregation categories. Researchers in highly segregated communities get more citations from their community members in middle cores and gain more citations per publication in middle/periphery cores. Those in non-segregated communities get more citations per publication in the nucleus. To our knowledge, this work is the first to characterise community segregation in co-authorship networks and investigate the relationship between community segregation and author citations. Our results help study highly segregated communities of scientific co-authors and can pave the way for intervention strategies to improve the growth and dissemination of scientific knowledge.

**Keywords:** Co-authorship networks; Science of science; k-core decomposition; Segregation analysis

## 1 Introduction

The social structures behind scientific production may profoundly affect the growth and dissemination of knowledge, the well-being of our societies, and the evolution of academic



Springer

research [1]. Many studies have shown how socially influenced behaviours impact different aspects of the scientific enterprise. Examples include the selection of co-authors, citation rates, and peer review processes, with authors' attributes biases such as prestige [2], gender [3], and country of affiliation [4, 5].

Co-authorship networks, where nodes represent researchers and links represent co-authorship relations between them, have been shown as key to the understanding and mapping of scientific production [6–8]. Particular attention has been devoted to their structural properties. These networks are organised in communities formed by groups of highly collaborative researchers with relatively low external interactions [9]. Looking at the evolution of these networks in time, one might see these communities going from being disconnected components to joining the giant component, as the co-authorship network coalesces. When comparing the proportion of nodes in the giant component relative to the total number of nodes, critical transition points represent the constitution of new disciplines and the growth of science [10].

As in most activities driven by human interactions, the biases mentioned above influence the processes of community formation and their connection/disconnection with other parts of the network. On one side, the previous literature has shown how the lack of exposure to individuals outside their circle can create segregated groups [11]. In different contexts of scientific production, such as discussions on social media, this "structural segregation" [12] can increase polarization [13, 14] and reinforce similar opinions [15]. High segregation levels—found in social networks with very fragmented groups—hamper the development of social capital and the emergence of cooperative behaviour, to the detriment of innovation, social learning, and problem solving [16]. In particular, computer scientists immersed in gender-segregated groups (low female-male connectivity) have disadvantaged positions in accessing information [17]. On the other side, researchers grouped into segregated communities could increase the exploitation of innovative ideas with in-depth work. For example, groups of researchers organised in efficient structures, characterised for being more interconnected and less clustered, proved to outperform others in solving complex problems [18], and researchers from evolutionary medicine produce better and longer-lasting ideas when located on the network's periphery [19].

There is 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 expand research methods and questions [20], as well as promote the spreading of scientific results [4, 21], is still limited.

In this context, we tackle three specific research questions: *(i)* How to identify highly segregated communities in co-authorship networks? *(ii)* Are there differences in communities' topological structure and core position with different segregation categories? *(iii)* Does the segregation category affect success in science as measured by citations?

To answer these questions, we study co-authorship networks using a dataset of publications in Computer Science. We assume that communities of researchers with very high internal connectivity versus low external connectivity can be considered highly segregated. We use four ordered segregation categories and show a relationship between community size, segregation category, and core position. Our main findings are that highly segregated communities tend to be near the network's periphery, and researchers in those communities gain more citations when positioned in the middle or periphery cores, with a higher proportion of those citations from their own communities in middle cores. In compar-

ison, non-segregated communities tend to be near the nucleus. And in the nucleus, the non-segregated researchers do both: gain more citations in total, and a higher proportion of those citations come from their own communities.

The paper is organised as follows: Sect. 2 describes the dataset and network properties used in this study. Section 3 details the procedure and characterisation of the community partition. Section 4 defines the structural segregation metric used in this study and how communities are categorised as completely segregated, highly segregated, moderately segregated and non-segregated. Our analyses focus on understanding non-segregated and highly segregated communities. Section 5 shows four metrics related to these communities' topology and core position, and we compare them using distributions and Z-Scores. In Sect. 6, we compare the number of citations per publication, and the proportion of citations received by members of the same community, to analyse the implications for researchers in communities with different segregation categories. Finally, Sect. 7 summarises our main contributions, limitations and final remarks.

## 2 Data and networks

We analyse the emergence of segregated communities in the scientific co-authorship network, focusing on the field of Computer Science. The choice of Computer Science here is pragmatic (manageable size) but also because we can study co-authorships in this field since its early stages; it consolidated as a discipline relatively recently (the late 60s) with the appearance of associations, undergraduate and PhD programmes, and specialised funding agencies [22]. We obtained data from the Semantic Scholar Open Research Corpus [23]. Our analyses correspond to 45 years from 1975 to 2020, for which we have sufficient data. To simplify the manuscript, we display some of the main results of our analysis using one particular year (2010) as an example. The choice of example year is somewhat arbitrary and was driven solely by the idea that approximately 10 years of work after that year should provide enough information about citation trends. For generality, we study other 2 example years (2006 and 2014) with results given in the Additional file 1. Henceforth, all references to results in the Additional file 1 have a prefix "S" (e.g. Section S1, Figure S1, Table S1). All three example years have similar results regarding the structure of the communities but differ in some of the citation analyses. We leave a complete longitudinal analysis across all years for future work, noting that citation comparisons cannot be fairly performed for recent years as works have yet to accrue citations.

For each year of analysis, we build a co-authorship network. Each node represents a researcher. A link is created when two researchers co-author at least one scientific publication in the year of study. For the analyses in this paper, we selected the Largest Connected Component (LCC) of each co-authorship network. The characteristics of the LCC co-authorship networks for the three years studied are shown in Table 1. Values in parentheses represent the proportion of the metric in the LCC compared with the entire co-authorship network. For example, for building the co-authorship network in 2010, we used all of the 615,737 available publications but then just analysed 294,181 publications in the LCC (0.48 of the available publications).

There are different ways to measure the value of the links between 2 researchers. For the current analyses, we use the strength of the link between two researchers $i$ and $j$ as proposed by Newman [25, 26]. The strength captures the idea that two researchers that are the sole co-authors of a paper know each other better than two researchers that co-authored a paper with many other co-authors, hence giving more importance to those

**Table 1** Characteristics of the Largest Connected Component (LCC) co-authorship network in 2006, 2010, and 2014. The values in parentheses correspond to the proportion of each quantity falling within the LCC as a fraction of the entire co-authorship network (e.g. for 2010, there were 294,181 papers forming the LCC, which is 0.48 of all Computer Science papers available in that year). The communities were detected with the *Label-propagation* algorithm [24]. Information about the growth of these metrics per year is given in Section S1

| Metric per year | 2006 | 2010 | 2014 |
|---|---|---|---|
| Number of papers | 194,114 (0.43) | 294,181 (0.48) | 369,304 (0.52) |
| Number of nodes | 249,797 (0.47) | 407,532 (0.54) | 566,835 (0.57) |
| Number of edges | 292,336 (0.22) | 1,453,217 (0.29) | 1,042,623 (0.32) |
| Density | 9.37e–06 | 1.75e-05 | 6.49e-06 |
| Clustering coefficient | 0.78 | 0.99 | 0.89 |
| Mean degree | 4.97 | 13.12 | 6.48 |
| Mean weighted degree | 5.99 | 14.33 | 9.44 |
| Mean strength degree | 1.73 | 1.78 | 1.8 |
| Number of communities ($\geq$3 researchers) | 24,470 | 39,998 | 54,655 |
| Number of researchers in communities ($\geq$3 researchers) | 249,797 | 407,532 | 566,835 |
| Number of internal papers (all the authors within the same community) | 86,354 | 128,415 | 189,072 |

papers with fewer co-authors. The strength is calculated as $w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}$, where $\delta_i^k$ takes the value of 1 if the researcher $i$ co-authored the paper $k$ and $n_k$ refers to the number of authors of the paper $k$. To sum the strength of the links of $i$ leads to the strength degree, which differs from the two well-known options of giving a value of 1 to each link (leading to the degree) or using the number of co-authorships as the weight of the link [27] (leading to the weighted degree). In Table 1, we compare the mean value of the three degrees (degree, weighted degree and strength degree) computed for the LCC. In Section S2, we give toy examples showing how the three degrees are calculated and compare their distributions over the years.

## 3 Community detection and description

To compute the community partition of the entire co-authorship network, we tested six commonly used community detection algorithms divided into two categories: modularity optimisation (Leading-eigenvector, Multilevel, Fast-greedy) and dynamical processes (Infomap, Walktrap, Label-propagation) [28]. To select which algorithm represents a better community detection, we must consider that all the co-authors of one publication form a clique [29], resulting in high clustering coefficients for co-authorship networks (Table 1). Following the methodology proposed by Fortunato and Hric [28], we select the results from the *Label-propagation* algorithm [24] 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 [30]. The results of each algorithm are included in Section S3. In addition, we analysed if our results depend on the community partition in Section S9, where we repeat some of the analyses for communities computed with *Infomap*.

Figure 1 shows the evolution of community structure from 1975 to 2020 based on outputs from the *Label-propagation* algorithm. The number of communities has grown above 50,000 by the end of the study period (Fig. 1A). The distributions of community sizes (number of nodes in each community) for each year are shown in Fig. 1B; the community size frequencies for 2010 are presented in the inset. Interestingly, 90% of the studied communities have fewer than 20 researchers, a constant tendency each year. The maxi-

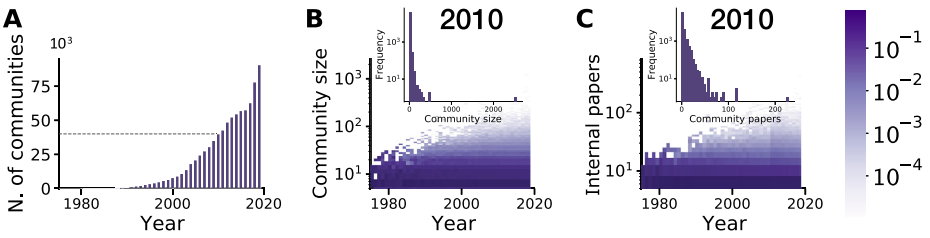


**Figure 1** Computer Science co-authorship community structure from 1975-2020. Plots show community metrics based on the *Label-propagation* algorithm for the Largest Connected Component (see text). (A) Number (in thousands) of communities per year. The dashed line highlights the 39,998 communities with size $\geq$3 in 2010. (B) Community size distribution (i.e., number of researchers per community). (C) Distribution of the number of internal papers (all authors within the same community) per community. The colour bar represents the proportion of communities with a given size and a number of internal papers. The inset panels of (B) and (C) show the frequencies of size and number of internal papers for the example year (2010)

mum community size in the last five years of data (2015-2020) is more than 2000. Finally, analysing the number of internal papers (i.e. papers with all the authors within the same community) written by each community, we found that 94% of communities publish less than ten papers, with an upper limit slightly above 200 papers (Fig. 1C). On average, for all the years, there are 0.38 internal papers per researcher (average number of internal papers over the number of researchers). The last results indicate that most researchers in Computer Science work in medium size groups, with the majority working on a few papers, differing from other disciplines with solo authors or large working groups [31].

## 4 Community segregation

From this section, all analyses are done considering the researchers, internal papers and communities in the LCC co-authorship network. In addition, because we study the internal connectivity structure of the communities, we analyse communities with at least three researchers. Hence, for 2010, we analysed 128,415 papers authored by 407,532 researchers grouped in 39,998 communities, as shown in the last three rows of Table 1.

### 4.1 Spectral segregation index

We use the Spectral Segregation Index (SSI) proposed by Echenique and Fryer [32] to measure structural segregation in the detected communities of the LCC. The SSI has been proposed to be measured over groups formed by categories of specific attributes, such as gender or race. However, in this current study, we measure segregation over groups based on the network structure and not previously labelled as such using member metadata. Previous research has demonstrated that differences can exist between structural communities and ground-truth communities in collaboration networks [28]. However, these disparities are primarily influenced by the attributes used as ground-truth, as evidenced by a social network study conducted in Indian villages. That study revealed that structural communities aligned well with caste-based[1] ground-truth communities but did not align with other attributes such as gender, age, work, or education [33]. Here, we want to study segregation patterns related to the diffusion capabilities of the researchers, and structural

[1] A fixed social group into which an individual is born within a particular system of social stratification, particularly used in Hinduism.

communities have been shown to help find "invisible colleges" [25] and interdisciplinary groups that are not apparent with just the use of metadata [34].

The SSI measures individual segregation as the linear combination of a node's and its neighbours' fraction of internal connectivity inside the group defined (internal refers to links inside the community in our case). The SSI implies a reinforcing process in which a node with a high SSI value has neighbours with a high SSI. In Section S4, we underscored the non-trivial nature of SSI and demonstrated that it is not merely a proxy for another established metric. There are various segregation metrics, and an interested reader should refer to Bojanowski and Corten [35].

We compute the SSI following the procedure defined by Echenique and Fryer [32]: First, we normalise the LCC 's adjacency matrix $R = [r_{ij}]_{N\times N}$ (which contains the strength of the link between 2 researchers $i$ and $j$). To achieve this, we take the original adjacency matrix and normalise their rows, to sum up to 1 (one). Then, we select a submatrix $B_g$ for each community, $g$, which contains only internal interactions within the community $g$. The value of $SSI_g$ corresponds to the largest eigenvalue $\lambda$ of the submatrix $B_g$ [32].

The eigenvalue $\lambda$ is computed as the stationary state of a "random walk" process. Hence, the connectivity patterns within the community shape the values of $\lambda$, which is, in turn, the average of the individual segregation values within the community. Values of SSI near 0 represent a low segregation categories, while values near 1 represent high segregation. Communities that are disconnected components have an SSI equal to 1, meaning perfect segregation [32], hereafter referred to as completely segregated communities.

### 4.2 Defining segregation categories

We compute the SSI considering only the connections within a calendar year in the co-authorship network, and we use communities with $\geq$3 nodes, considering only the LCC. For 2010, we worked with 39,998 communities (29% from the original 136,967 communities of the entire co-authorship network in 2010). The other communities are completely segregated and do not connect to the LCC. Completely segregated communities: *(i)* can be cliques (i.e., fully connected subgraphs), *(ii)* have few internal papers (1.44 on average), and *(iii)* do not have a core position (computed from the k-core decomposition of the communities network). Their presence is partially due to the time-window considered (i.e., one year); the longer the period considered, the larger the LCC becomes and, consequently, the fewer the isolated components. Then, we do not analyse the structural properties or core positions of completely segregated communities in Sect. 5 because they could skew our results. However, we include a category of completely segregated communities in Sect. 6 when we analyse the relationship between different segregation categories and citations.

The values of SSI are continuous, and no clearly defined categories exist. As we wanted to compare the spectrum's extremes, we searched for specific and ordered categories of segregation: highly, moderately, and non-segregated communities. For this, we compared four different methods of division to identify the three categories over the Probability density function (PDF) of the SSI. In Section S4 we detail the methods and results of dividing the communities by *(i)* standard deviation, *(ii)* Gaussian mixture, *(iii)* $k$-means, and *(iii)* 20th percentile. In the main manuscript, we show the results by the 20th percentile, as this division is the simplest; it results in the evenest number of communities in both categories: highly and non-segregated, and the qualitative results of the structural properties of the communities and the citation comparison do not change.

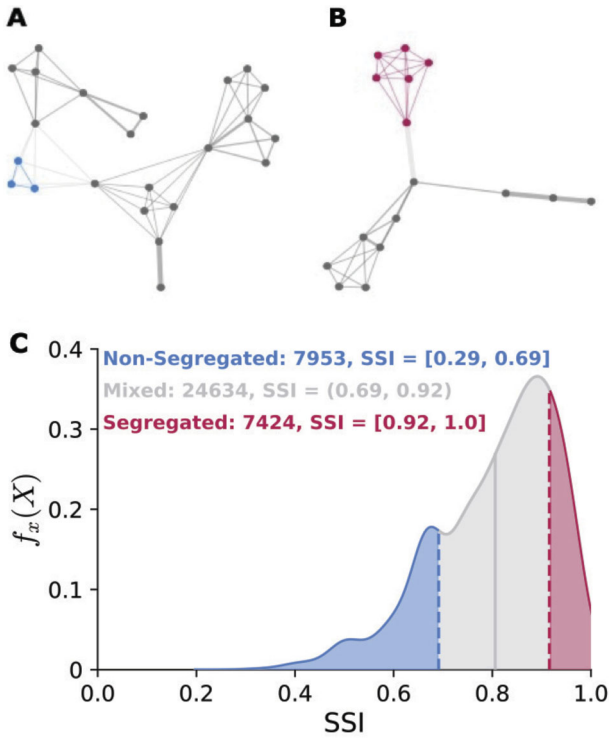


**Figure 2** Classifying communities as non-segregated and highly segregated. (A) and (B) are examples of ego-networks of co-authorships in 2010 of non-segregated and highly segregated communities, in light blue and red, respectively. Ego-networks are sub-graphs induced by the connections between central nodes, i.e., ego (colored nodes belonging to the selected community) and their one-step neighbours, i.e., alters (dark grey nodes belonging to other communities connected to the colored community). Edges inside the communities have the color of the nodes, while links across communities are in light grey. (C) shows the probability density function (PDF) of the spectral segregation index (SSI) for 2010. The plot is divided into three categories that denote non-segregated (light blue), moderately (grey), and highly segregated (light red) communities. The complete procedure is in Section S4. For the distribution, we use a Gaussian kernel density estimation with the "rule of thumb" for the bandwidth selection [36]

In Fig. 2C, we show the PDF of SSI for 2010, the division of segregation categories, and the number of communities in each category. From the 20th percentile division, we ended up with 7953 non-segregated, 24,634 moderately segregated, and 7424 highly segregated communities. We compute the same analysis with the 20th percentile method in Section S4 for 2006 and 2010.

In Fig. 2, we show toy networks of non-segregated and highly segregated communities in panels A and B, respectively. Those toy networks show communities with their members in colour, grey for nodes from other neighbouring communities and in light grey links among different communities.

In the following analyses, we study two categories: non-segregated and highly segregated communities, as we want to study the extremes of the SSI spectrum. However, in the first subsection of Sect. 6, we compare the citation patterns of the four ordered segregation categories: completely segregated, highly segregated, moderately segregated and non-segregated communities.

## 5 Characterisation of communities in different segregation categories

We compare four metrics in total to investigate the characteristics of non-segregated and highly segregated communities. The first three metrics refer to the structural properties of the communities to understand if the segregation categories are related to a community's internal connections. We compute the 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) [37].

The fourth metric refers to the core position of the communities because the core/periphery position of segregated communities in online social networks (i.e. echo chambers) [38] has been shown to influence their ability to spread information during social movements [39]. Therefore, in the context of scientific production, we want to understand if the communities' position in the co-authorship network also relates to their segregation category. We first create a network in which each community is a node, and links between these nodes exist if their members share co-authorships. Then, we apply the k-core decomposition algorithm [40] and assign each community to a correspondent core. The core values range from 1 (periphery) to N (nucleus), where N depends on how many cores we have in a particular year, 11 in the case of 2010. See Section S5 for more details about calculating the core decomposition of the communities networks.

As a previous step, we group the communities by different size ranges (detailed explanation and more analyses in Section S6). For the comparison, we first separate the communities by size range and segregation category (i.e., highly or non-segregated). Then, we perform a statistical analysis to compare the PDF of the four metrics (size, density, clustering, and core position) of the non-segregated and highly segregated communities, with results for dividing the communities by the 20th percentile in the 2010 network in Fig. 3 and analogous plots for the other division methods and different years in Section S7.

Our results show that small communities tend to be denser, more clustered and toward the network's periphery. As expected, their densities and clustering decrease when they increase in size, though less visibly for clustering. There are mild differences in density and clustering between non-segregated and highly segregated communities, with values mainly driven by community size. Moreover, there is a difference in their core position, with more large non-segregated communities in the nucleus and more highly segregated communities in peripheral cores.

We performed five additional analyses, reported in the Supporting Materials: *(i)* the same analyses for the other three division methods with similar results in Section S7.1. *(ii)* the same analyses without separating the communities by size with misleading results in Section S7.2 as there are differences in density and clustering (driven by size) but no differences in the core position. *(iii)*the same analysis for 2006 and 2014 in Section S7.3.1 with similar results. *(iv)* We compare the Z-Score of the metrics and compute kernel density estimators for comparing size, SSI, and core position at the same time in Section S7.3.2. with congruent results. *(v)* We repeat the procedures of this section in Section S9.2 with the results of *Infomap*. We found that both algorithms have similar results.

## 6 The effect of segregation on citations

This work's third and final research question relates to understanding the relationship between segregation and citation levels. Citations are a well-known measure of scientific success, but we also encourage reading our results critically, as citations have been

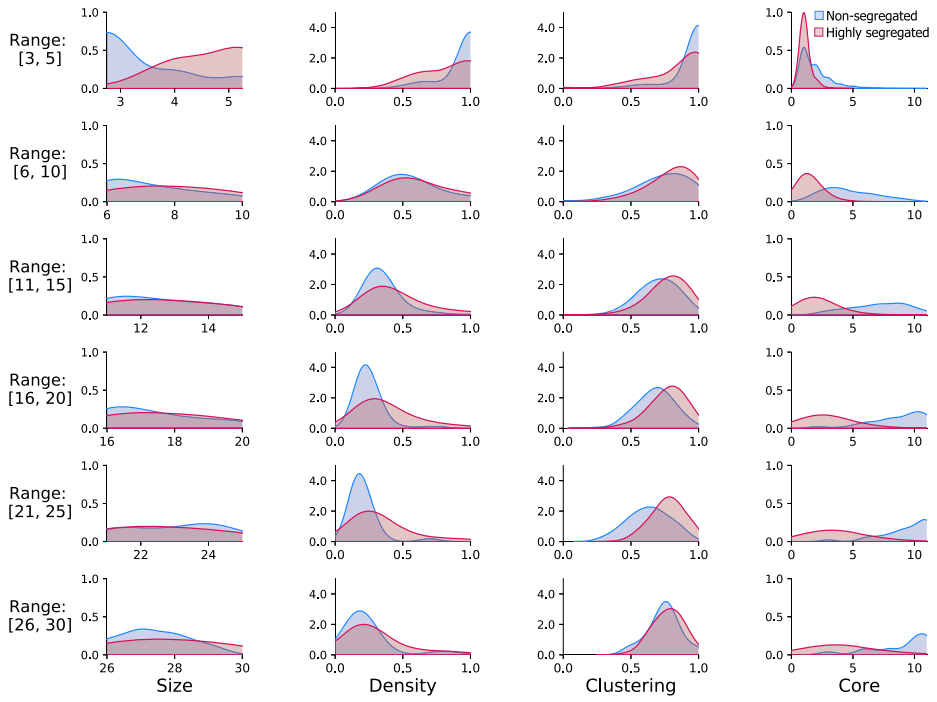


**Figure 3** Topological and core position differences among non-segregated and highly segregated communities by 20th percentile. The panels represent the probability density functions (PDF) in each column for the size, density, clustering, and core position of highly segregated (light red) and non-segregated (light blue) communities for different size ranges per row. When plotting the curves, we use a Gaussian kernel density estimation with the "rule of thumb" for the bandwidth selection [36]

related to selection biases, mainly affecting underrepresented communities, e.g. women and non-western researchers publishing non-English content [41]. Here, we consider both the number of citations and the origin of the citations (in terms of the community partition) to characterise whether highly segregated communities have more self-citations than non-segregated ones. For each researcher in non-segregated and highly segregated communities, we analyse the citations received until 2020 by the publications of 2010.

First, we investigate whether the number of internal papers correlates with *(i)* the total number of citations and *(ii)* the average number of citations per paper. We find low correlations of 0.29 ($p$-value $< 10^{-3}$) and 0.10 ($p$-value $< 10^{-3}$), respectively. We use the Spearman correlation in both cases because the number of papers has a non-linear relationship with citations and citations per paper (Figure S13).

Second, 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 from the same community, and *(iv)* proportion of all citations from the same year's co-authors (2010 for the main manuscript).

For each variable, we analyse researchers at two levels of granularity. *(i)* All researchers without grouping them by core position for the four categories: completely segregated, highly segregated, moderately segregated and non-segregated in Fig. 4, and *(ii)* researchers grouped by the core position of their communities for two categories: non-segregated and highly segregated in Fig. 5. We did not analyse our results by different ranges of internal papers due to the low correlation with the citation variables.

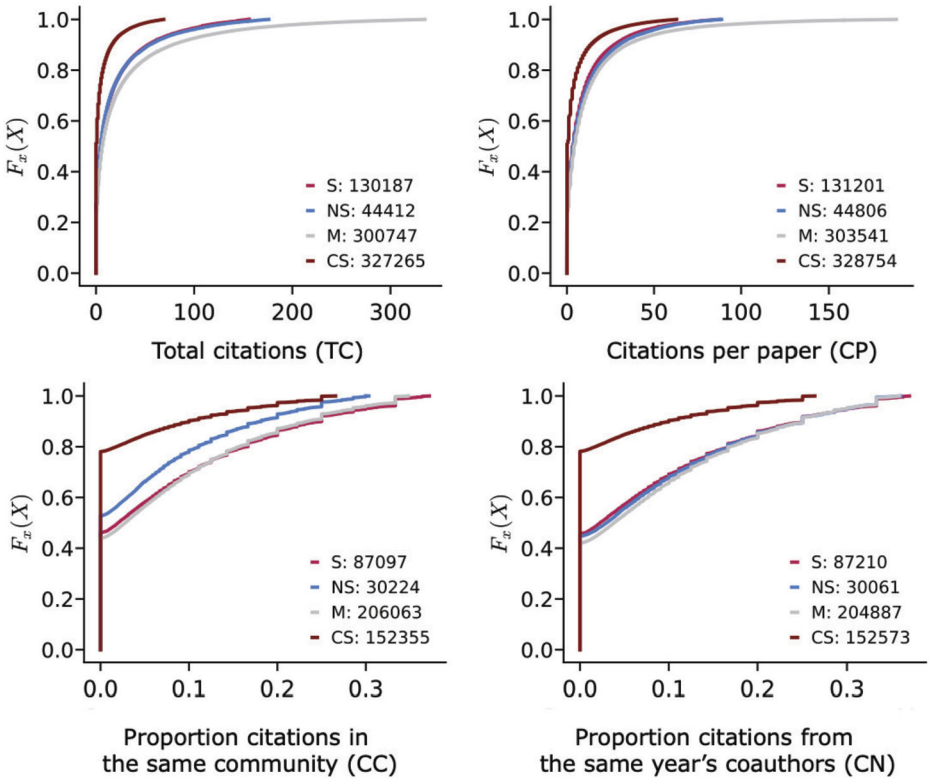


**Figure 4** Citation metrics for all researchers in communities of different segregation categories. Each panel represents the cumulative density function (CDF) for the total citations (TC), the citations per paper (CP), the proportion of citations from the same community (CC), and the proportion of citations from the same year's co-authors (CN). The code of colours is: dark red for researchers in completely segregated (CS), grey for moderately segregated (M), light red for highly segregated (S), and blue for non-segregated communities (NS). Letters **KS** or **MW** appear when there are significant *p*-values for Kolmogorov-Smirnov (different distribution shapes) and Mann-Whitney (different distribution medians) for the CDFs of non-segregated and highly segregated communities. Significance levels are denoted as follows: * < 0.1, ** < 0.05, and *** < 0.01

We use two statistical tests to compare the CDFs of non-segregated and highly segregated communities: Kolmogorov-Smirnov (KS) and Mann-Whitney (MW). The first test compares the shape of the distributions, and the second compares the differences between medians.

We first analyse the CDFs for the *(i)* Total citations (TC) and *(ii)* Citations per paper (CP). On an aggregated level, in Fig. 4 top row, our results indicate that there are no differences between highly and non-segregated researchers in terms of TC nor CP, we see that completely segregated researchers (darker red in the plot) have smaller values than other researchers, with no significant differences. However, the previous results hide some information because they are averaging over all network cores. Then, in Fig. 5, we group the researchers by the core position of their communities, and we split the results into the nucleus, middle, and periphery. In middle and periphery cores, highly segregated researchers have more TC than non-segregated ones, with opposite results in the nucleus (top row). For the CP (second row), there are no differences in the middle or periphery cores, but non-segregated researchers have more CP in the nucleus.

Then, we analyse the CDFs for *(iii)* the proportion of Citations from the same community (CC) and *(iv)* Proportion of citations from the same year's co-authors (CN). For

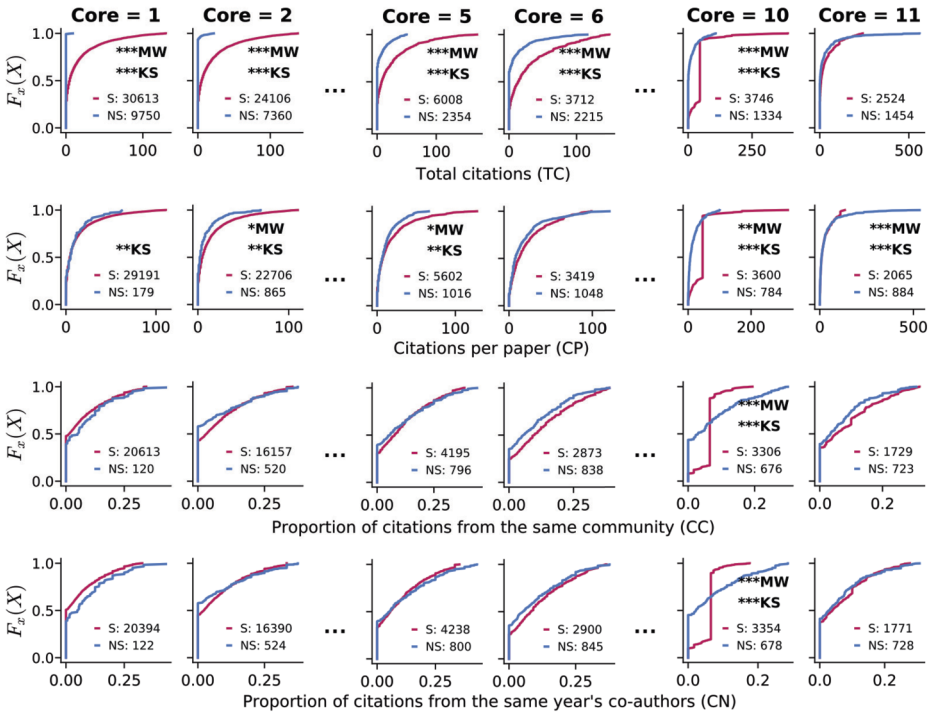


**Figure 5** Citation metrics for researchers in communities of different segregation categories and core positions. Each row represents the cumulative density function (CDF) for the total citations TC, the citations per paper (CP), the proportion of citations from the same community (CC), and the proportion of citations from the same year's co-authors (CN). The code of colours is: light red for highly segregated (S) and blue for non-segregated communities (NS). Letters **KS** or **MW** appear when there are significant *p*-values for Kolmogorov-Smirnov (different distribution shapes) and Mann-Whitney (different distribution medians) for the CDFs of non-segregated and highly segregated communities. Significance levels are denoted as follows: * < 0.1, ** < 0.05, and *** < 0.01. Here, we show 7 out of 11 cores to guide the reader, but Figure S15 shows results for the 11 cores of 2010

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 (for CC) or co-author (for CN, regardless of the community). Then, we divide these counts by the total number of citations.

On an aggregated level (Fig. 4 second row), our results show that highly segregated researchers have more CC than non-segregated ones while there is no difference for CN. In addition, completely segregated researchers (darker red) receive lower CC and CN than others. There are no differences in the periphery when we group by the core position (Fig. 5 third and fourth rows). However, in middle cores, highly segregated researchers have more CC and CN; in the nucleus, non-segregated researchers have larger values.

We compare the results of 2010 with those in 2006 and 2014 in Section S8. For TC, highly segregated researchers outperform non-segregated in the periphery and middle cores, but there are no significant differences for CP. In the nucleus, non-segregated researchers do better for both TC and CP. There are no differences in CC and CN for non-segregated and highly segregated researchers, but for 2014 the trends are similar to those in 2010.

In summary, highly segregated researchers tend to have more citations per paper when they locate in peripheral cores and more citations from their communities in middle cores.

At the same time, non-segregated researchers show higher values for the four metrics when they are in cores near the nucleus.

## 7 Discussion

Due to a range of social mechanisms, processes, and biases, co-authorship networks are organised in communities [9]. Within-group dynamics might lead to the emergence of segregation and polarisation, hampering innovation, social learning, and problem-solving [12–14, 16]. 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 [19, 42]. Still, understanding segregated groups in co-authorship networks and their possible effects is limited. Here, we tackle this problem by quantifying segregation categories of communities in co-authorship networks and characterising their topological properties and position in the network.

For our case study, we analyse the co-authorship network of Computer Science in the Semantic Scholar Open Research Corpus [23]. We detect communities with the *Label-propagation* algorithm and compute a structural segregation metric considering the community's links: the Spectral Segregation Index (SSI). Based on the distribution of the SSI, we identify three main categories and focus on the two opposite limits: non-segregated and highly segregated communities. Then, we compare the communities' size, density, clustering, and core position between categories. Furthermore, we study the relationship between segregation and impact using citations from the community's publications.

Our results indicate that highly segregated communities tend to be more on the periphery, with some differences in density and clustering with non-segregated communities. When we analyse the total number of citations, researchers in highly segregated communities receive more citations than non-segregated ones in middle and peripheral cores. In addition, when we analyse the sources of those citations, for researchers in highly segregated communities, up to 5% more of those citations come from the same community than non-segregated communities in middle cores. Combining both results and based on previous literature, we speculate that in terms of spreading ideas and knowledge in the co-authorship network: *(i)* researchers in highly segregated communities attract more citations in the periphery of the network because most cited papers are not the internal ones but rather those across communities with diverse disciplines and co-authors [43]. And *(ii)* researchers in non-segregated communities in the nucleus are citing themselves more and are exploiting/echoing scientific research [18].

Both effects need further analysis because, as expected, highly segregated communities located on the periphery have a larger impact. Individual success correlates with the exploitation of ideas [18]. Still, also the most innovative research (exploration of new concepts and persistent citations) comes from the periphery of networks [19], and it is done by smaller groups of researchers [42]. Here, our results align with previous evidence showing nodes in the periphery being less active [38] (i.e. publishing less in our case) but having more impact. In addition, researchers in those communities are a large population that could become a collective power that can mobilise and spread information [39] (such as scientific theories).

Researchers in larger and non-segregated communities in the nucleus also increase their impact. These results need further exploration because their central positions in the net-

work's nucleus increase their chance of outside interactions with highly segregated communities, which can accelerate the propagation of echoed information (ranging from biased theories to new paradigms) from local groups to reach the entire network [44]. The inner impact of highly segregated communities and their impact on the whole network should be measured to intervene, if necessary, and tackle or boost the spread of echoed information to different groups [17].

### 7.1 Limitations

First, our analysis does not generalise for all the years of Computer Science papers available in the Semantic Scholar database because we study just three years. We have developed a repeatable methodology and replicated our findings over several years. Still, further analysis is needed to understand how the transitions of researchers between different segregation categories affect their research impact over time.

Second, our analyses only generalise to some co-authorship 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 [29]. Further analysis of other fields is needed to understand how these patterns apply to different co-authorship 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 [45]. However, their results show that authorship networks are the most prone to have a core-layered typology, as we used in the current work. In further analyses, the definition of segregated communities should also consider the co-authorship network's core typology.

Finally, our fourth limitation relies on using the extreme values of the SSI 's PDF from the co-authorship networks to define segregation categories of communities. A more precise analysis could consider continuous values of the SSI, other features and data to represent better the consumption and production of scientific knowledge [6]. Future work could consider a continuous comparison of the metrics used in this analysis, publications' content, researchers' demographic diversity, and interdisciplinary citations.

### 7.2 Future research

Future research on this topic could consider: *(i)* the temporal analysis of segregated communities and their relation to gaining more or fewer citations over time, *(ii)* the analysis of the diversity of the scientific publications inside the communities using opinion distance [13] 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 [46], iv) the measurement of the impact of segregated communities on the topology of the network formation and the spreading processes of scientific theories [47].

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00411-8.

**Additional file 1.** (PDF 29.8 MB)

**Acknowledgements**
The authors would like to thank the US Army Research Office for the partial support provided to RM under grant number W911NF-18-1-0421. AMJ is funded by a PhD studentship from the UK Engineering and Physical Sciences Research Council. No funding bodies had any influence over the content of this report.

**Abbreviations**
SSI, Spectral Segregation Index; LCC, Largest Connected Component; PDF, Probability density function; CDF, Cumulative density function; TC, Total citations; CP, Citations per paper; CC, Citations from the same community; CN, Proportion of citations from the same year's co-authors.

**Availability of data and materials**
The datasets generated and analysed during the current study are available in the Semantic Scholar repository, https://www.semanticscholar.org/product/api

## Declarations

**Competing interests**
The authors declare that they have no competing interests.

**Author contributions**
All authors conceived and designed the research. AMJ acquired the data. AMJ, HTPW, NP and RM analysed the data. All authors discussed the research and wrote and approved the final version of the manuscript.

**Author details**
[1]BioComplex Laboratory, Department of Computer Science, University of Exeter, Exeter, UK. [2]Complexity Science Hub, Vienna, Austria. [3]SEDA Lab, Department of Computer Science, University of Exeter, Exeter, UK. [4]School of Mathematical Sciences, Queen Mary University of London, London, UK. [5]Department of Computer Science, Federal University of Ceará, Fortaleza, Brazil.

Received: 8 November 2022 Accepted: 7 August 2023 Published online: 09 October 2023

**References**

1. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási AL (2018) Science of science Science 359(6379). https://doi.org/10.1126/science.aao0185
2. Lynn FB (2014) Diffusing through disciplines: insiders, outsiders, and socially influenced citation behavior. Soc Forces 93(1):355–382. https://doi.org/10.1093/sf/sou069
3. Sugimoto CR, Lariviere V, Ni C, Gingras Y, Cronin B (2013) Global gender disparities in science. Nature 504:211–213
4. Smith MJ, Weinberger C, Bruna EM, Allesina S (2014) The scientific impact of nations: journal placement and citation performance. PLoS ONE 9(10):1–6. https://doi.org/10.1371/journal.pone.0109195
5. Opthof T, Coronel R, Janse MJ (2002) 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. Cardiovasc Res 56(3):339–346. https://doi.org/10.1016/S0008-6363(02)00712-5
6. Zeng A, Shen Z, Zhou J, Wu J, Fan Y, Wang Y, Stanley HE (2017) The science of science: From the perspective of complex systems. https://doi.org/10.1016/j.physrep.2017.10.001
7. Pan RK, Kaski K, Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. Sci Rep 2(1):902. https://doi.org/10.1038/srep00902
8. Pan RK, Petersen AM, Pammolli F, Fortunato S (2018) The memory of science: inflation, myopia, and the knowledge network. J Informetr 12(3):656–678. https://doi.org/10.1016/j.joi.2018.06.005. arXiv:1607.05606
9. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104. https://doi.org/10.1103/PhysRevE.74.036104
10. Bettencourt LMA, Kaiser DI, Kaur J (2009) Scientific discovery and topological transitions in collaboration networks. J Informetr 3(3):210–221. https://doi.org/10.1016/j.joi.2009.03.001
11. Sunstein CR (2018) #Republic: Divided Democracy in the Age of Social Media, Ned - new edition edn. Princeton University Press, Princeton, pp 59–97. https://doi.org/10.2307/j.ctv8xnhtd
12. Kim S (2019) Directionality of information flow and echoes without chambers. PLoS ONE 14(5):1–22. https://doi.org/10.1371/journal.pone.0215949
13. Sasahara K, Chen W, Peng H, Ciampaglia GL, Flammini A, Menczer F (2021) Social influence and unfollowing accelerate the emergence of echo chambers. J Comput Soc Sci. https://doi.org/10.1007/s42001-020-00084-7. arXiv:1905.03919
14. Perra N, Rocha LEC (2019) Modelling opinion dynamics in the age of algorithmic personalisation. Sci Rep 9(1):1–11. https://doi.org/10.1038/s41598-019-43830-2. arXiv:1811.03341
15. Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W (2016) Echo chambers: emotional contagion and group polarization on Facebook. Sci Rep 6:1–12. https://doi.org/10.1038/srep37825. arXiv:1607.01032
16. Henry AD, Prałat P, Zhang CQ (2011) Emergence of segregation in evolving social networks. Proc Natl Acad Sci USA 108(21):8605–8610. https://doi.org/10.1073/pnas.1014486108
17. Jalali ZS, Wang W, Kim M, Raghavan H, Soundarajan S (2020) On the information unfairness of social networks. In: Proceedings of the 2020 Siam international conference on data mining, SDM 2020, pp 613–621. https://doi.org/10.1137/1.9781611976236.69
18. Mason W, Watts DJ (2012) Collaborative learning in networks. Proc Natl Acad Sci USA 109(3):764–769. https://doi.org/10.1073/pnas.1110069108

19. Painter DT, Daniels BC, Laubichler MD (2021) Innovations are disproportionately likely in the periphery of a scientific network. Theory Biosci 140(4):391–399. https://doi.org/10.1007/s12064-021-00359-1
20. Nielsen MW, Bloch CW, Schiebinger L (2018) Making gender diversity work for scientific discovery and innovation. Nat Hum Behav 2(10):726–734. https://doi.org/10.1038/s41562-018-0433-1
21. Sonnenwald DH (2008) Scientific collaboration. Annu Rev Inf Sci Technol 41(1):643–681
22. Tedre M (2017) In: The science of computing: shaping a discipline, CRC Press, Boca Raton
23. Lo K, Wang LL, Neumann M, Kinney R, Weld D (2020) S2ORC: the semantic scholar open research corpus. https://doi.org/10.18653/v1/2020.acl-main.447. arXiv:1911.02782
24. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3). https://doi.org/10.1103/physreve.76.036106
25. Newman MEJ (2004) Who is the best connected scientist?a study of scientific coauthorship networks. J Complex Netw, 337–370. https://doi.org/10.1007/978-3-540-44485-5_16
26. Cann TJB, Weaver IS, Williams HTP (2018) Is it correct to project and detect? Assessing performance of community detection on unipartite projections of bipartite networks. In: Complex networks and their applications VII. Springer, Cham, pp 267–279. https://doi.org/10.1007/978-3-030-05411-3_22
27. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci 101(11):3747–3752. https://doi.org/10.1073/pnas.0400087101
28. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44. https://doi.org/10.1016/j.physrep.2016.09.002
29. Newman MEJ (2001) The structure of scientific collaboration networks. In: PNAS
30. Lancichinetti A, Saramäki J, Kivelä M, Fortunato S (2010) Characterizing the community structure of complex networks. PLoS ONE arXiv:1005.4376. https://doi.org/10.1371/journal.pone.0011976
31. Fanelli D, Larivière V (2016) Researchers' individual publication rate has not increased in a century. PLoS ONE 11(3):1–12. https://doi.org/10.1371/journal.pone.0149504
32. Echenique F, Fryer RG (2007) A measure of segregation based on social interactions. Q J Econ. https://doi.org/10.1162/qjec.122.2.441
33. Montes F, Jimenez RC, Onnela J-P (2017) Connected but segregated: social networks in rural villages. J Complex Netw 6(5):693–705. https://doi.org/10.1093/comnet/cnx054. https://academic.oup.com/comnet/article-pdf/6/5/693/26058916/cnx054.pdf
34. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
35. Bojanowski M, Corten R (2014) Measuring segregation in social networks. Soc Netw. https://doi.org/10.1016/j.socnet.2014.04.001
36. Scott DW (1992) Multivariate density estimation. Wiley, Huston. https://doi.org/10.1002/9780470316849
37. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. https://doi.org/10.1103/PhysRevE.69.026113
38. Williams HTP, McMurray JRJR, Kurz T, Hugo Lambert F (2015) Network analysis reveals open forums and echo chambers in social media discussions of climate change. Glob Environ Change 32:126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006
39. Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, Tucker J, González-Bailón S (2015) The critical periphery in the growth of social protests. PLoS ONE. https://doi.org/10.1371/journal.pone.0143611
40. Batagelj V, Zaversnik M (2003) An O(m) algorithm for cores decomposition of networks. arXiv:cs/0310049
41. Cronin B, Sugimoto CR (2015) Scholarly metrics under the microscope: from citation analysis to academic auditing. ASIST monograph series, Medford, NJ. https://doi.org/10.5596/c15-025
42. Wu L, Wang D, Evans JA (2019) Large teams develop and small teams disrupt science and technology. Nature 566(7744):378–382. https://doi.org/10.1038/s41586-019-0941-9
43. Zingg C, Nanumyan V, Schweitzer F (2020) Citations driven by social connections? A multi-layer representation of coauthorship networks. Quant. Sci. Stud. 1(4):1493–1509. https://doi.org/10.1162/qss_a_00092. arXiv:1909.13507
44. Davis JT, Perra N, Zhang Q, Moreno Y, Vespignani A (2020) Phase transitions in information spreading on structured populations. Nat Phys. https://doi.org/10.1038/s41567-020-0810-3
45. Gallagher RJ, Young JG, Welles BF (2021) A clarified typology of core-periphery structure in networks. Sci Adv. https://doi.org/10.1126/sciadv.abc9800. arXiv:2005.10191
46. Guo L, Rohde JA, Wu HD (2020) Who is responsible for Twitter's echo chamber problem? Evidence from 2016 U.S. election networks. Inf Commun Soc 23(2):234–251. https://doi.org/10.1080/1369118X.2018.1499793
47. Törnberg P (2018) Echo chambers and viral misinformation: Modeling fake news as complex contagion. PLoS ONE. https://doi.org/10.1371/journal.pone.0203958


## Publisher's Note