Di Bona et al. *EPJ Data Science* (2023) 12:42
https://doi.org/10.1140/epjds/s13688-023-00418-1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

The concept of decentralization through time and disciplines: a quantitative exploration

Gabriele Di Bona[1,2,3†], Alberto Bracci[4†], Nicola Perra[1], Vito Latora[1,5,6] and Andrea Baronchelli[4,7,8*]

*Correspondence:
abaronchelli@turing.ac.uk
[4]Department of Mathematics, City, University of London, Northampton Square, EC1V 0HB London, United Kingdom
[7]The Alan Turing Institute, British Library, 96 Euston Road, NW1 2DB London, United Kingdom
Full list of author information is available at the end of the article
[†]Equal contributors

Abstract

Decentralization is a pervasive concept found across disciplines, including Economics, Political Science, and Computer Science, where it is used in distinct yet interrelated ways. Here, we develop and publicly release a general pipeline to investigate the scholarly history of the term, analysing 425,144 academic publications that refer to *(de)centralization*. We find that the fraction of papers on the topic has been exponentially increasing since the 1950s. In 2021, 1 author in 154 mentioned *(de)centralization* in the title or abstract of an article. Using both semantic information and citation patterns, we cluster papers in fields and characterize the knowledge flows between them. Our analysis reveals that the topic has independently emerged in the different fields, with small cross-disciplinary contamination. Moreover, we show how Blockchain has become the most influential field about 10 years ago, while Governance dominated before the 1990s. In summary, our findings provide a quantitative assessment of the evolution of a key yet elusive concept, which has undergone cycles of rise and fall within different fields. Our pipeline offers a powerful tool to analyze the evolution of any scholarly term in the academic literature, providing insights into the interplay between collective and independent discoveries in science.

Keywords: Decentralisation; Science of science; Interdisciplinary; Knowledge flows; Complex networks

1 Introduction

"For students of recent domestic affairs it is becoming increasingly evident that '*decentralization*' is a magic word". With these words in 1975 Herbert London opens his article "The meaning of decentralization" [1]. Almost 50 years later, Schneider states that *decentralization* "is called for far more than it is theorized or consistently defined" [2]. *(De)centralization* (i.e., either *Decentralization* or its counterpart *Centralization*) has indeed become almost a buzzword, permeating not only the academic literature, but also the public discussion. The debate between centralized and decentralized contact tracing at the beginning of the COVID-19 pandemic is a clear example [3]. However, one of the

Springer

major drivers of its growth has certainly been the rise of blockchain based technologies such as cryptocurrencies, NFTs and the metaverse. [4–7].

However, *(de)centralization* is not a new concept, and has different connotations across fields. In political science, it usually refers to the delegation of power to local communities with respect to a central government [8]. The concept has similar connotations when referring to educational [9], fiscal [10] and more generally governance systems. Other domains where the term is widely used include public health [11], internet protocols [12], robot swarms [13] and social network analysis [14] among others, with the last one providing one of the few quantitative definitions available thanks to Freeman in 1978. Given this background, some questions naturally arise: have these different disciplines independently developed the concept of *(de)centralization*, maybe even with different meanings (i.e., a case of polysemy)? Have they influenced each other? Which fields have been most influential to the evolution of this concept?

Here, we address these questions by studying a corpus of scientific literature indexed by the Semantic Scholar open database [15]. First, we observe an exponentially growing interest in the topic, with an author in 154 contributing to a paper mentioning *(de)centralization* in its title or abstract in 2021. Then, we map the literature on *(de)centralization* by focusing on the subset of relevant articles and clustering them according to their semantic and citation information. This way, we discover that different academic fields have separately contributed to this topic. We hence study how the different clusters have influenced each other, showing how much more transfer of knowledge between different academic areas is happening in recent years. Interestingly, our analysis reveals that STEM and social sciences did not influence each other. Finally, we focus on two paradigmatic examples: Governance, interpreted generally as "the way that organizations or countries are managed at the highest level, and the systems for doing this" [16], and Blockchain, including all blockchain based technologies, from cryptocurrencies to NFTs and the metaverse. We show how Governance is the first cluster to extensively make use of the term *(de)centralization*, containing the most or second most number of papers each year since its appearance in the 1950s, and playing a leading role in the transfer of knowledge to other fields until the 1990s. Blockchain instead has become both the most influential cluster and the most productive cluster in the past 10 years, showing three different phases in its recent history characterized by different interactions with other fields. Overall, our results shed light on the history and evolution of the more and more important concept of *(de)centralization*. Furthermore, we publicly release the code of the pipeline developed in this study, so that it may be used to study and understand the evolution of other concepts through the lenses of the academic literature.

2 The pipeline

In this section, we briefly describe the pipeline we have set up and publicly released[1] to select the data and perform the research described in this study. The pipeline is conceptually divided into three steps: (1) data collection, (2) clustering of the dataset using a multilayer hierarchical stochastic block model, and (3) analysis of the influence between clusters over time using knowledge flows.

[1] See https://github.com/alberto-bracci/decentralization.

2.1 Step 1: data collection

The first step consists in collecting the academic publications related to the concept of *(de)centralization*, or potentially to other concepts. To perform a large scale analysis of the academic literature, we exploit the possibility to access the publicly available *Semantic Scholar Academic Graph* (S2AG, pronounced "stag"), which provides monthly snapshots of research papers published in all fields [15]. Launched in 2015 by the Allen Institute for Artificial Intelligence (AI2), Semantic Scholar provides this corpus as an open access database with the specific scope of facilitating scientific analysis of academic publications. It contains about 203.6 million papers (1st Jan. 2022 snapshot), 76.4 million authors, and 2 billion citations. Moreover, this database recently incorporated the Microsoft Academic Graph (MAG) [17], which was shut down at the end of 2021 [18].

From this corpus we extracted the data about papers that contain the root string *"centrali"* in words of the title or abstract, to capture possibly all variations of words related to the concepts of *Centralization* and *Decentralization* (nouns, adjectives, etc.). In this way, we also incidentally captured articles written in different languages, mainly Portuguese, French and Spanish, and also a minority of unrelated articles (e.g., biology articles involving plant species including "centrali" in their name). More information on the frequency of the words containing such root string can be found in Table S1 in the Additional file 1 (SM). The resulting dataset has 425,144 papers characterized by a series of attributes. Among these, of particular interest to us there is the title, the abstract, the authors, all in- and out-citations (respectively citations and references), the year of publication, and the fields of study, which were determined based on machine learning field classifiers leveraging on the existing MAG taxonomy and classification [19]. Notice, however, that some articles miss one or more of these attributes. See Table S2 in the SM for details on how many papers have each of these attributes.

2.2 Step 2: hierarchical clustering

In the Semantic Scholar corpus almost each paper is associated to a list of fields of study. However, these are high-level, as there are in fact only 19 fields of very heterogeneous sizes (see Table S3 in the SM for details on how many papers are classified in each field of study). Moreover, sometimes the fields are not correctly assigned. In the second step of the pipeline, we hence use a multilayer hierarchical stochastic block model (hSBM) [20, 21], developed to find statistically significant clusters at multiple hierarchical levels for the analysis of text data with multiple data types. Here, in fact, we consider two layers. The document layer —where links represent citations between papers— and the text layer —a bipartite network between documents and the words present in their titles. The method naturally produces clusters of documents and topics (word clusters), incorporating the information from both layers in the process. Furthermore, as the name suggests, the model produces a hierarchical clustering, providing a richer structure of both article clusters and topics, which captures both small clusters and topics and how they are related to each other in a higher level structure.

We consider only the papers in our dataset that have a non-empty title and contain at least one citation or reference to another paper in the dataset (42.7% of the initial dataset), as we are interested in how the concept of *(de)centralization* evolved in the academic literature, and citations are the most natural proxy for how knowledge is transferred. We use title texts, instead of abstracts, for various reasons: firstly, because the title is more

frequently available than the abstract (see Table S2 in the SM); secondly, because the title has the advantage of being more distilled compared to abstracts [22]; lastly, because titles contain a significantly smaller number of words than abstracts, allowing us to obtain a text layer similar to the document layer in terms of number of edges by simply cutting out words present in less than 5 documents. It is indeed well known that the hSBM performs optimally when both layers have roughly the same size, otherwise the smaller layer is effectively ignored by the algorithm [21]. The filtered dataset hence consists of 181,605 documents and 15,381 different words, summing up to 590,215 document-to-document citation links and 1,396,830 document-to-word links.

To ensure the robustness of our results, we performed 100 iterations of the algorithm. Notice that the number of clusters and levels of granularity obtained is not fixed, but is automatically suggested by the algorithm. By running the algorithm multiple times, we aimed to capture the inherent variability and uncertainty in the Monte Carlo partitioning process. Subsequently, the consensus partition is calculated by maximizing the overlap with all the partitions from the 100 runs. Such consensus partition serves as a robust representation of the underlying structure within the analyzed data. More statistics comparing the single iterations of the hSBM and the consensus partition are shown in Fig. S2 in the SM.

Afterwards, keywords are assigned to each cluster to roughly represent the content and themes of the articles within them (for more details see Fig. S3, Fig. S4 and Table S4 in the SM, with related section). Keywords are chosen by looking at the most frequent words in the cluster, the most significant topics in the cluster according to the normalized mixture proportion [21], as well as the first 5 papers in the cluster according to different measures (see SM Sect. 2.1 for more details).

2.3 Step 3: knowledge flows

In the third step of the pipeline, we want to better understand how the different groups of documents identified by the hSBM have influenced each other throughout history. To do so, we evaluate the knowledge flows between these groups, using article citations as proxy [23]. In particular, we compute the knowledge flow from one cluster a in one year Y_a to another cluster b in a future year Y_b. The computation takes into account the fraction of citations towards papers in a of the year Y_a from papers in b published in the year Y_b with respect to the fraction of citations towards a in Y_a from all papers published in Y_b, as well as the overall fraction of papers of a in Y_a. The citation network suffers indeed from a series of inherent biases: field size, typical number of citations in a field or a journal, typical number of references, age of the fields etc. This method de facto considers the number of citations with respect to a null model, resulting in a link weight which is effectively a z-score.

Mathematically, if a collection of papers is divided in a partition $\mathcal{P}$ of clusters such that different clusters do not overlap and altogether form the collection of papers, then we can define the knowledge flow units $C_{a\to b}(Y_a, Y_b)$ from papers in cluster $a \in \mathcal{P}$ published in the year Y_a to papers in cluster $b \in \mathcal{P}$ in a future year Y_b by counting how many citations have occurred from b to a in the two years, that is,

$$C_{a\to b}(Y_a, Y_b) = \left|\left\{(x, y) : x \in a, Y_x = Y_a, y \in b, Y_y = Y_b \text{ s.t. } y \text{ cites } x\right\}\right|. \tag{1}$$

As said before, we need to normalize this number with respect to a null model, so as to keep into account different sizes of clusters and different norms in citation practices. Hence, the knowledge flow $K_{a\to b}(Y_a, Y_b)$ from a in year Y_a to b in year Y_b can be computed in the following way:

$$K_{a\to b}(Y_a, Y_b) = \begin{cases} 1 & \text{if } \frac{C_{a\to b}(Y_a,Y_b)}{\sum_{c\in\mathcal{P}} C_{c\to b}(Y_a,Y_b)} \Big/ \frac{|x\in a: Y_x = Y_a|}{\sum_{c\in\mathcal{P}} |x\in c: Y_x = Y_a|} \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

After the normalization against the null model, knowledge flows can be indeed treated as z-scores. Hence, in Eq. (2) we consider a knowledge flow as significant (i.e., a binary value of 1) if higher than the threshold 1, and as not significant (i.e., 0) otherwise.

Therefore, we obtain a binary value for each pair of clusters and each pair of years. In other words, the collection of (knowledge flow) links between all pairs of clusters and years generates a temporal network of clusters, which we aggregate in different ways to facilitate the following analysis. In particular, we consider the average knowledge flow $K_{a\to b}(Y_a)$ from a cluster a to another b from a specific year Y_a as the average of the knowledge flows $K_{a\to b}(Y_a, Y_b)$ from cluster a to b from year Y_a to all years $Y_b > Y_a$, taking into account only years Y_b where there is at least one publication in b. Formally, this reads:

$$K_{a\to b}(Y_a) = K_{a\to b}(Y_a, \bullet) = \left\langle K_{a\to b}(Y_a, Y_b) \right\rangle_{\{Y_b > Y_a : \exists x\in b \text{ s.t. } Y_x = Y_b\}}. \tag{3}$$

This value represents, on a scale from 0 to 1, how much publications in cluster a in year Y_a have influenced the future of cluster b. Analogously, we define the average knowledge flow $K_{a\to b}(T)$ from cluster a to cluster b from a period of time T to the future by averaging $K_{a\to b}(Y_a)$ over all years Y_a in T in which there is at least one publication in a, that is,

$$K_{a\to b}(T) = \left\langle K_{a\to b}(Y_a) \right\rangle_{\{Y_a\in T : \exists x\in a \text{ s.t. } Y_x = Y_a\}}. \tag{4}$$

We can also measure the average influence in terms of knowledge flows from a cluster to all other clusters and vice-versa, as well as the average knowledge flow among all clusters, respectively as follows:

$$\begin{aligned} K_{a\to\bullet}(Y) &= \left\langle K_{a\to b}(Y) \right\rangle_b, \\ K_{\bullet\to a}(Y) &= \left\langle K_{b\to a}(Y) \right\rangle_b, \\ K_{\bullet\to\bullet}(Y) &= \left\langle K_{a\to b}(Y) \right\rangle_{a,b}. \end{aligned} \tag{5}$$

Here, $K_{a\to\bullet}(Y)$ refers to the average influence from papers in cluster a published in year Y towards all clusters in the future. On the opposite, $K_{\bullet\to a}(Y)$ refers to the average influence of the papers in all clusters in the year Y towards the future of cluster a. Finally, $K_{\bullet\to\bullet}(Y)$ refers to the average influence (towards the future) of all papers published in year Y.

3 Results

3.1 The decentralized evolution of (de)centralization

We start by analysing the number of papers mentioning *(de)centralization* over the years (see *The pipeline* section for more details), comparing it to the total number of academic

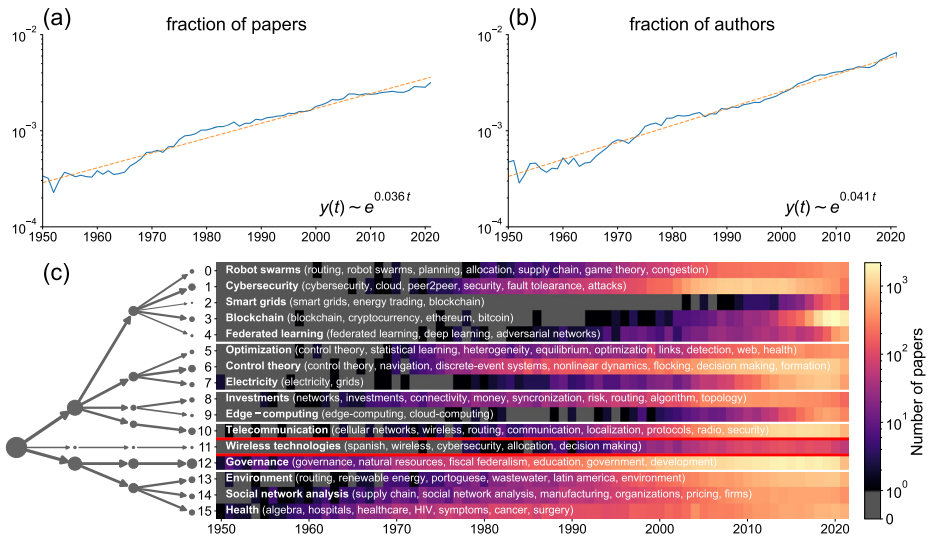

Figure 1 The rising interest of academic literature towards *(de)centralization*. *(a)* Fraction of papers mentioning *(de)centralization* in the Semantic Scholar corpus. *(b)* Fraction of authors mentioning *(de)centralization* in the Semantic Scholar corpus. Both fractions have been steadily increasing since the 1950s, showing growing interest in the topic. *(c)* Number of papers in the clusters at the 3rd level of the hierarchy in each year. Clusters are ordered respecting the hierarchical network on the left, in which node and link sizes are proportional to the total number of papers in the related cluster. In the heatmap, white lines individuate clusters belonging to the same cluster at the 2nd level of the hierarchy, while red lines divide different clusters at the 1st level. The representative keyword of the clusters at level 3 is reported in bold in the respective rows, while all the specific keywords identified at the 4th hierarchical level are shown within brackets. Clusters with less than 500 papers in total are not shown in the figure

outputs (papers, books etc.) produced in time, which is known to be exponentially increasing [24]. As shown in Fig. 1(a), the fraction of papers mentioning *(de)centralization* has been exponentially increasing in time since the 1950s, rising to around one paper every 315 in 2021. The growing interest in this topic is also reflected by the increasing number of authors involved in such academic research. Indeed, as shown in Fig. 1(b), the fraction of authors producing such research has risen exponentially by more than one order of magnitude, with almost one academic every 154 writing a paper mentioning the topic in 2021. This growth is also seen in terms of raw number of publications and authors, as shown in Fig. S1 in the SM, where we compare these numbers for the S2AG corpus and the *(de)centralization* dataset and find a stronger exponential rise for the latter. Notice that for both papers and authors there are some periods with a higher or lower increase in the fraction, showing spikes of interest at particular times. For example, in Fig. 1(a,b) we can see that between the late 1970s and the 1980s the growth rate was faster than the overall exponential fit.

In order to understand what has characterized the origins and evolution of the topic, we set to identify topics and clusters of papers in the dataset by using the hSBM algorithm [20, 21] described in *The pipeline* section. In the following analysis, we focus our attention only to years after 1950. Before this date there are only around 100 papers in our *(de)centralization* dataset. The very first is a political science one from 1851 on local self-governments versus centralized governments [25]. Among the others in this period, apart from around 50 papers that relate to *(de)centralization* in governments, organizations and states, we have detected 30 papers that are actually false positives of the selection process.

Considering also how, in general, digitalization issues may have contributed to the small number of papers before 1950, the reliability and coverage of the first 99 years of the data are unclear, and we opted to exclude them from the analysis.

The consensus partition, obtained by collecting the outcomes of 100 runs of the hSBM algorithm, consists of 7 hierarchical levels. On the left of Fig. 1(c), we draw this hierarchy only until the 3rd level (starting from the common root at level 0) for visualization clarity. More information on the number of clusters in each level can be found in Sec. S2.1 in the SM, where we also conduct a comparative analysis of the consensus partition with the individual runs of the hSBM algorithm.

On the right of Fig. 1(c), we also show the heatmap of the number of papers in time for each cluster at the 3rd level, for a total of 16 different clusters after excluding other 5 clusters with less than 500 papers not included in this analysis. The keywords shown in the heatmap have been manually assigned to roughly characterize each cluster. In particular, the keywords between parentheses have been chosen amongst the most frequent and most significant in the clusters at the 4th level, while the most representative keyword at the 3rd level has been chosen and printed in bold. For details on how they were assigned see *The pipeline* section and Sec. S2.2 in the SM.

In the following, we refer to a cluster at the 3rd level by its representative keyword (capitalized). As shown by the hierarchy and by the horizontal red lines in the heatmap, clusters are divided in three main branches. Looking at the two biggest branches, we can see a clear division between more STEM oriented documents (top branch) and those in Political sciences, Social sciences, as well as Medicine (bottom branch). Notably, a third smaller branch appears isolated, including papers at the intersection of the other two, mostly about Wireless technologies and their applications. Going into more details, in the STEM branch we notice how Cybersecurity, Control theory and Telecommunication (clusters 1, 6, and 10 respectively) are the ones producing most publications, with Blockchain (cluster 3) becoming the most relevant in the last 5 years in terms of number of papers published per year. On the other branch, Governance (cluster 12), including works in Political science, Education and Fiscal federalism, is the most relevant cluster, while Environment, Social network analysis and Health clusters (respectively clusters 13, 14, and 15) have produced a smaller number of papers. Furthermore, see Fig. S5 and Fig. S6 in the SM respectively for a similar plot done at the 4th level and for a bipartite hierarchical network showing how clusters are represented in the various topics.

In Fig. 1(c) we have shown how the number of papers in each of these clusters has evolved over time. Looking more into details of the early history of *(de)centralization*, the first papers adopting the term have all been in the Social sciences branch, most importantly the Governance cluster, followed by Social network analysis and Health. In the 1950s, indeed, there are 58 papers in Governance, which represents the first cluster to adopt the term and use it extensively. Some of these articles refer, among other things, to democracy as a form of centralized decision-making system [26]. Other clusters with more than 10 papers refer to Social network analysis and Health, as seen for example in Kaufman's "Toward an interactional conception of community" [27]. Here, *centralization* is depicted as a force gradually destroying the concept of community as a social unit. Notably, most of these papers have no citations from other articles in the full corpus, with only some citations within the cluster of governance. In the 1960s, the largest growth is found again in the Governance and Health clusters, both reaching around 150 papers each in the decade.

An important example of the former is that of Bachrach et al. [28], where they highlight how different disciplines (i.e., social and political sciences) reach completely opposite conclusions about the *(de)centralization* of power. In the same decade *(de)centralization* also appears in other relevant clusters, namely Social network analysis (50 papers) and Investments (29 papers), with a significant number of citations in both directions between them. The term is picked up from the STEM branch only later in the 1970s, especially through works in Control theory and Optimization [29], coming significantly to a popular domain as Cybersecurity only in the 1980s.

We have seen how different domains have picked up the concept of *(de)centralization* at different times. It is therefore natural to ask whether they developed such uses separately, or they influenced each other in some way. The hierarchical clustering partially answers this question, as it gives a degree of separation between domains based on citation and semantic information. However, significant information is still present in the citations between papers of different clusters. We exploit this by computing knowledge flows [23], whose aim is to quantify the transfer of knowledge given by the citations between groups of papers through a comparison with a null model (for more details see *The pipeline* Section). We thus study the average knowledge flow $K_{a\to b}(T)$ from papers in a cluster a, at level 3 of the hierarchy, in a period of time T to future publications in another cluster b, represented by a number between 0 and 1 showing how significant this influence has been.

Here, in Fig. 2 we consider three different periods of time T: 1970-1989, 1990-2007 and 2008-2020. Similarly to what we will do in the next figures, we have excluded the year 2021 as a source of knowledge flow, because our dataset ends at the end of 2021, thus meaning that we cannot evaluate knowledge flows from papers of 2021 to future years. In the figure, all clusters are ordered as in Fig. 1(c), with the representative keywords shown in the legend below. For each period T, the color of the cell of row a and column b of the

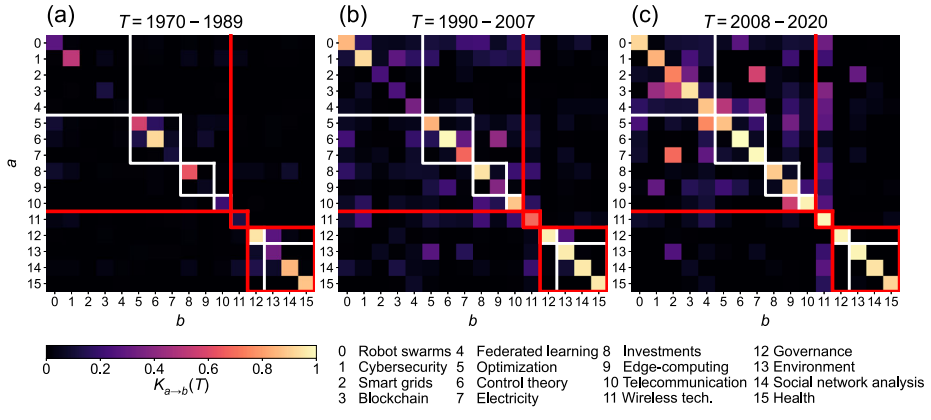

Figure 2 Temporal evolution of the influence between clusters. Average knowledge flows $K_{a\to b}(T)$ from each cluster a to each cluster b at level 3 in the period T = 1970-1989 *(a)*, T = 1990-2007 *(b)* and T = 2008-2020 *(c)*, represented by the different colors according to the colorbar in the bottom left part of the figure. A representative keyword for each cluster is reported on the bottom right part of the figure. White lines denote clusters belonging to the same 2nd level cluster, whereas red lines mark different branches at the 1st level. In the first period, little to no communication is happening between different clusters. In recent years, more communication happens inside the same 2nd level cluster, and towards the middle branch (cluster 11). However, little communication happens between the two other branches, roughly representing the STEM and social sciences communities respectively

heatmap refers to the average knowledge flow $K_{a\to b}(T)$ from papers in cluster a in that period of time to future papers in cluster b, according to the colormap shown below.

Starting from the first period in Fig. 2(a), between 1970 and 1989 clusters have little to no influence on the future of the other ones. We have previously seen how in these years the use of *(de)centralization* started to rise across some domains, mostly being Governance, Control theory, Social network analysis, Health, Cybersecurity, and Investments. However, apart from Governance and Control theory (clusters 12 and 6), these clusters have low knowledge flow even to themselves, meaning that the use of *(de)centralization* was only relegated to sporadic and not so influential papers in the literature. This also confirms that the topic has appeared independently at this early stage.

Then, looking at the period from 1990 and 2007 in Fig. 2(b), we can see how much more transfer of knowledge has occurred between clusters. As shown in Fig. 2(c), this trend is even more pronounced in the last and more recent period, whose start coincides with the creation and rise of blockchain technologies. Interestingly, these transfers reflect the structure of the hierarchy and denote significant differences between the high-level domains. The STEM branch (made of the clusters 0 to 10) shows clear communication between clusters belonging to the same group both at the 2nd and 1st level (respectively within white and red lines), whereas the right bottom branch shows almost no communication with the other domains, especially after 2008. The only significant knowledge flow from this branch in the middle period goes from Environment (cluster 13) to Optimization (cluster 5), while in the last period this is only relegated between Environment and Smart grids (cluster 2). The middle branch instead shows clear influence from the other two, and little influence towards them, especially in the last period.

Finally, notice how the highest knowledge flows between different clusters in Fig. 2(b) are from those that, in the first period, were starting to be more influential. Similarly in the last period the clusters with highest influence on other ones are mostly within the STEM branch. Overall, the heatmaps show a clear decentralized birth of the concept of *(de)centralization*, appearing in different fields and domains with little to no communication between each other. Instead, in recent years, we find a more coordinated evolution, even though still sectorial in some cases, and mainly lead by STEM related clusters.

3.2 The case of governance and blockchain

Having analyzed the concept of *(de)centralization* in the general academic landscape, we now focus on two of the most important clusters in the history of this topic: Governance and Blockchain. As shown in Fig. 1(c) and in Fig. S7 in the SM, these two clusters are among the biggest across time in terms of number of papers. The Governance cluster has always been first or second with respect to the other clusters at the 3rd level, while Blockchain was barely present before 2008, the year of the bitcoin white paper [4]. After that, Blockchain gradually increased in size and had an exponential explosion after 2015, coincidentally with the increasing hype around the technology and its applications, in particular bitcoin and ethereum [30–32]. Finally, it has become the most productive cluster since 2019, surpassing governance.

To better understand their role in the evolution of the literature on *(de)centralization*, we consider the average knowledge flows between clusters for each year, that is looking at $K_{a\to\bullet}(Y)$, $K_{\bullet\to a}(Y)$, and $K_{\bullet\to\bullet}(Y)$, defined in Eq. (5). Therefore, in Fig. 3 we rank clusters year by year using $K_{a\to\bullet}(Y)$ in (a) and $K_{\bullet\to a}(Y)$ in (b), i.e., looking at how much the papers

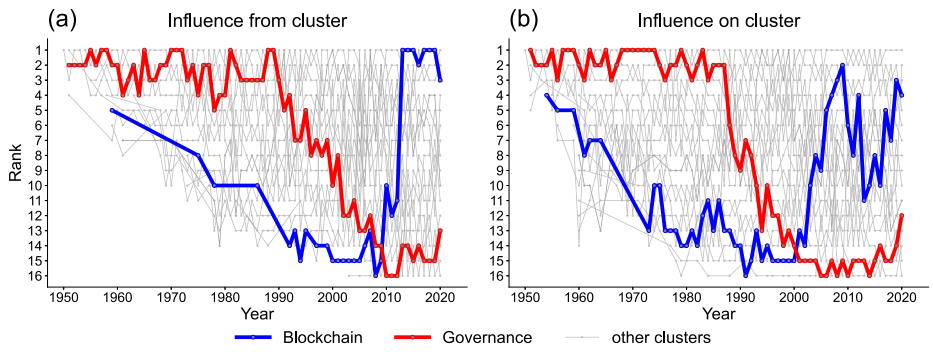

Figure 3 Ranking the influence of Blockchain and Governance in the *(de)centralization* literature. *(a)* Ranking in time of the influence coming from a cluster in the 3rd hierarchical level, computed on the average knowledge flow $K_{a\to\bullet}(Y)$ from papers in cluster a published in the year Y towards all other future papers. *(b)* Ranking in time of the influence to a cluster in the 3rd hierarchical level, computed on the average knowledge flow $K_{\bullet\to a}(Y)$ from papers published in the year Y towards future papers in cluster a. The blockchain cluster, highlighted in blue, has become a central actor in the recent literature on *(de)centralization*, supplanting the governance cluster, highlighted in red

of a cluster a in a year Y have influenced, on average, the future of all other clusters (a), or, vice versa, how much all clusters have influenced the future of a (b). From these plots we can see how, on the one hand, Governance has been in the top ranks until the late 1980s, both as a source and target of knowledge flows. However, in the early 1990s it started to decrease in importance, reaching the bottom ranks in the 2000s, despite being the first cluster in terms of number of papers each of these years. On the other hand, in Fig. 3(a) we notice that the rise of Blockchain started only in 2010, being almost always outside of the *(de)centralization* literature discussion until this point. Then, very sharply, Blockchain becomes the first cluster in terms of influence towards other clusters in 2013, maintaining its position in the following years. Hence, the literature on Blockchain has been key in the development of the *(de)centralization* discussion in the most recent years. Moreover, looking at Blockchain in Fig. 3(b), papers of other clusters before early 2000s have had almost no impact on the scientific future of Blockchain. Interestingly, it has received a lot of influence from publications between 2006 and 2012, that is about when the blockchain and bitcoin originated [4], as well as after 2017, mostly due to the increasing amount of applications using blockchain in the most diverse contexts in recent years. Finally, notice the loss of influence on Blockchain from papers between 2013 and 2016.

These results are corroborated by the time evolution of the average knowledge flow compared to the overall average $K_{\bullet\to\bullet}(Y)$. Indeed, in Fig. S8 in the SM we show how Governance has been increasingly important in influencing other clusters until the 1980s, while since the 1990s it has had a lower average knowledge flow than the average among all clusters. Similarly to what shown by the ranks, after 2013 Blockchain starts to have a much higher influence towards the other clusters compared to the average. Moreover, in Fig S9 in the SM we compare the average knowledge flow within the same cluster and towards other clusters, isolating the clusters Governance and Blockchain from the rest of the clusters. We find that on the one hand Governance has maintained a very high knowledge flow to future papers in the same cluster throughout the years. On the other hand, starting from 1990s, the exchange in knowledge flow towards and from other clusters has decreased. Blockchain, instead, has received a higher than average knowledge flow from other clus-

ters in the 2000s, starting to provide maximum knowledge flow to itself and more than average to other clusters in the next and most recent decade.

So far we have seen how Governance has been influential in the early literature about *(de)centralization*, and how Blockchain has risen in recent years as the most important influential cluster, contributing in terms of knowledge flow towards other branches of literature. It is therefore a natural next step to investigate with higher granularity which clusters in particular have influenced or have been influenced by Governance first and Blockchain then, and see how these interactions have changed over time. We start this analysis from the more recent case of Blockchain. This cluster started to appear only around 2008 with the bitcoin white paper [4]. Moreover, we notice a decrease in the influence on this cluster in mid 2010s. We therefore divide the 2008–2020 time span in three parts, following the blockchain history: 2008–2014, representing the origin of blockchain applications before the advent of ethereum; 2015–2018, when the field got more recognition thanks to ethereum and bitcoin; and the final 2019–2020 period, in which we have seen the explosion of academic literature production and the widespread success of multiple applications such as DeFi, NFTs and the metaverse.

In Fig. 4 we plot, in a decreasing order, which clusters have been most influenced by (a) and have most influenced (b) Blockchain during the three periods. To this end, we use a Sankey diagram, showing how the overall picture has changed in the three different phases. The plot is done using the average knowledge flows $K_{a\to b}(T)$, where T is the selected period, while a and b are fixed to Blockchain in Fig. 4(a) and Fig. 4(b) respectively.

Firstly, in Fig. 4(a) we analyze the influence of Blockchain on the these other clusters. The early literature of Blockchain has had a strong impact on most of the clusters. As a matter of fact, there are only a few cases where the average knowledge flow from Blockchain to another cluster is zero, shown by a circle in the respective node and a lighter color in

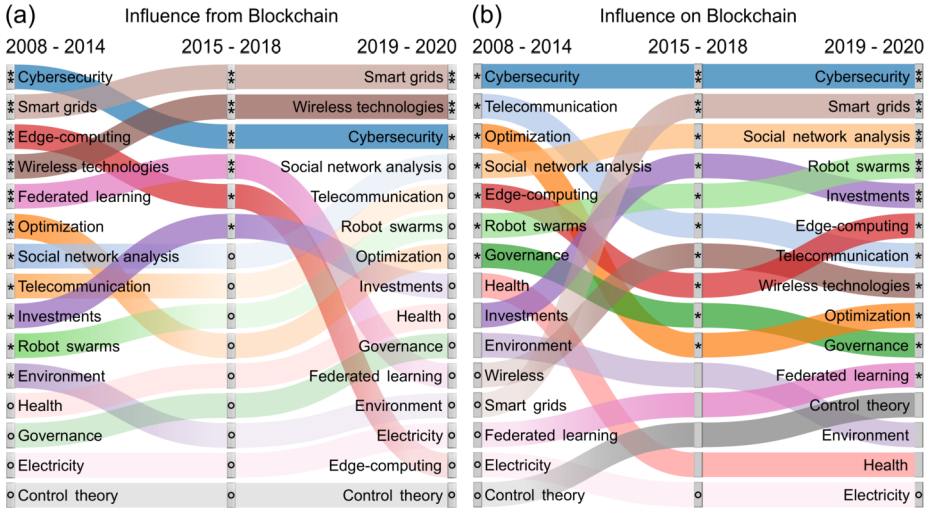

Figure 4 Influences between Blockchain and the other clusters on *(de)centralization*. *(a)* Change in the ranking of the most influenced clusters by Blockchain between its early period (2008-2014), its middle period (2015-2018), and its late period (2019-2020), calculated using the average knowledge flows $K_{a\to b}(T)$, where T is the selected period, and a is fixed to be Blockchain. *(b)* Change in the ranking of the clusters having most influenced the Blockchain literature (same periods as in the previous panel). In both cases, if $K_{a\to b}(T) = 0$, we print a circle in the corresponding gray node and use a lighter color in the respective link. Moreover, we print a star when $0.01 < K_{a\to b}(T) \leq 0.1$, and two stars when $K_{a\to b}(T) > 0.1$

the corresponding link. We also notice that Cybersecurity, Smart grids, Edge-computing, Wireless technologies, and Federated learning have a very significant average knowledge flow from Blockchain, i.e., $K_{a\to b}(T) > 0.1$, shown by the double stars, while other clusters with $0.01 < K_{a\to b}(T) \leq 0.1$ are represented with only one stars. Notice how Blockchain has continued to have a big impact on these mentioned clusters, in particular Smart grids and Wireless technologies, as well as Cybersecurity to a lesser extent. Looking altogether at the three periods, notice how Cybersecurity and Edge-computing have lost influence from Blockchain over time. Moreover, in the last period there is no significant knowledge flow to all other clusters, which is peculiar if we consider that, for example, Federated learning and Edge-computing received a very significant knowledge flow in the previous years. We argue that the recent decrease in knowledge flow is mostly due to the time needed for a paper to attract citations, especially outside its own cluster. Additionally, we find that some clusters, such as Health, Electricity, Control theory and Governance, have received no significant influence from Blockchain in all these years, even if, Governance, for instance, has been second only to Blockchain in terms of number of papers.

Secondly, in Fig. 4(b) we investigate the impact of the different clusters on Blockchain's literature. Cybersecurity, which has been one of the clusters that grew the most among all STEM clusters from the 1980s to the 2010s, has been stably the most influential cluster on Blockchain. The other top positions have instead changed from the first period considered, with Smart grids, which did not even have any influence on Blockchain at first, and Social network analysis becoming the most important clusters after Cybersecurity. Notice also how Robot swarms and Investments have experienced an increase in knowledge flow towards Blockchain, while the opposite has happened for Telecommunication, Optimization, Governance and Health.

Comparing the two plots in Fig. 4, we find examples of only unidirectional influences between Blockchain and the other clusters. The cluster of Social network analysis, third in position since 2015 to influence Blockchain, has not been influenced by it during the same period, which is also the case of Robot swarms and Governance. A similar situation is found for Wireless technologies, that has been strongly influenced by Blockchain over time, but only in recent years it has had a small impact on it.

Finally, we have conducted a similar analysis on the Governance cluster in Fig. S10 in the SM. In this case, we consider three different periods of times: 1950–1980, that is the early stage when it was the most important cluster overall; 1981-1990, when the amount of knowledge flow from Governance stopped to increase, still remaining among the top in terms of ranking; and 1991–2000, in which its role diminished and got surpassed by almost all other clusters by the end of the period. We do not find many noticeable differences between the first two periods. Most clusters have no significant knowledge flow from and to Governance, showing how *(de)centralization* developed independently in this cluster at first. Differently from Blockchain, the top clusters to have interactions with governance are Environment, Social network analysis and Investments. Wireless technologies, Blockchain and Robot swarms have also been influenced by Governance, but not vice-versa, apart from the sporadic case of Wireless technologies in the middle period. We can also see that the influence from Governance has increased over time on clusters like Blockchain, Optimization and Robot swarms, showing how the last years of the last century have been important milestones for the future of these clusters.

4 Discussion

In this paper we have developed and presented a framework that allowed us to quantitatively investigate how different topics have risen in the *(de)centralization* literature and have influenced it. By exploiting the S2AG corpus, we have shown that the literature on *(de)centralization* has exponentially increased in the past 70 years, with an author in 154 contributing to articles on the topic in 2021. We have observed a diversification of research fields engaging in (de)centralization studies, starting from Governance as the most prolific field and gradually expanding to include various other disciplines. Furthermore, analyzing the evolution of knowledge flows between clusters, we have revealed a gradual increase of influence between different fields. Initially, the various fields operated in isolation, with minimal cross-disciplinary interaction. However, as the literature developed, we observed a growing interconnection among these fields, with high knowledge flows especially within STEM subjects. Finally, we have shown how Governance has lost its leading role in the *(de)centralization* literature in favour of Blockchain. In fact, if Governance has remained mostly independent after the 1990s, without influencing or being influenced by other clusters, Blockchain has been the most influential cluster for the last ten years, and has recently become the most productive one.

A significant aspect of the framework we have developed is its versatility. The methodology can be used to examine the evolution of any scholarly term or concept within academic literature. For example, future work could use it to investigate the unfolding of such important topics as "gender inequality" and "artificial intelligence". Additionally, it can be utilized to explore the interplay between collective and independent innovation in the field of science. Our pipeline relies on two key methods, the multilayer hierarchical stochastic block model [21] and knowledge flows [23]. On the one hand, we employ the first one to cluster documents and words in the dataset to identify different themes and topics, using information of both citations between papers and of the words used in each document. On the other hand, knowledge flows allow us to identify significant influences between clusters over time. With the present paper, we publicly release the pipeline code to allow other researchers to perform similar analyses on other concepts.

Our study presents some limitations which also represent directions for future work. First, we only consider academic papers that directly mention the word *(de)centralization* or one of its variants (e.g. "centralised", "centralizing", etc.). A broadened analysis could also include all articles cited by these papers, in order to further understand the roots of this topic in the different fields. Moreover, we have limited the semantic information to the words of document titles. Future studies could build on state of the art large scale language models and Natural Language Processing techniques to extract more information from the articles text (i.e., abstract and/or full text) and offer more detailed insights of their content. For example, a set of keywords could be used instead of the plain text of the title [33], so as to better characterize papers and disambiguate terms that could have different meanings in different contexts. In fact, the stochastic block model assigns one block to each node of the network [21, 34], thus indirectly assigning the main meaning of a word to a certain topic, disregarding other nuances of the word. A possible way to take this into account is to consider mixed-membership stochastic block models [35, 36], where each node is assigned to a distribution or mixture of categories.

Finally, our methodology is able to identify direct flows of knowledge between two fields but misses less straightforward chains of interaction. For example, a field a could indi-

rectly influence field *b* if there is a direct knowledge flow between field *a* and another field *c*, which in turn influences field *b*. The inclusion of temporally and causally compatible higher order interactions (i.e., more than pairwise) is therefore an obvious route to improve on the current work. Moreover, it would be interesting to investigate the presence of citation biases or other spurious effects that have affected the evolution of science. For instance, the rise of interdisciplinary knowledge flows could have been favoured by the presence of online easily-accessible papers. Furthermore, even though the combination of clustering and knowledge flows are a powerful tool to detect statistically significant influence across fields and time, it is worth noticing that citations can be noisy, and are not always a definitive indicator of cross-pollination. We hence emphasize the need for complementary methods and approaches, such as qualitative historical analyses and expert interviews, to provide a more comprehensive understanding of the development of the concept of "decentralization" in different disciplines. We leave such research questions to future work.

Overall, our analysis provides new insights in the origin and evolution of the ubiquitous concept of *(de)centralization*, shedding light on the academic roots and influence of the blockchain technology. Additionally, our pipeline can be used to analyse quantitatively any other concept in the academic literature, and can be easily combined with other text and network analysis tools. We therefore anticipate that our results will be of interest to researchers working in a vast array of disciplines.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00418-1.

Additional file 1. The interested reader can find more analyses in the Supplementary Material about the dataset (Section S1) and the hierarchical clustering method (Section S2), as well as other results about the importance of Governance and Blockchain in the history of (de)centralization (Section S3) (PDF 2.7 MB)

Acknowledgements
The authors wish to thank the anonymous referees for their valuable input on an earlier version of the manuscript.

Abbreviations
COVID-19, COronaVIrus Disease of 2019; NFT, Non-Fungible Token; STEM, Science, Technology, Engineering, Mathematics; S2AG, Semantic Scholar Academic Graph; MAG, Microsoft Academic Graph; hSBM, hierarchical Stochastic Block Model; SM, Supplementary Material.

Availability of data and materials
All the code used for the pipeline presented in this paper can be freely accessed and used through the Github repository available at https://github.com/alberto-bracci/decentralization. The data used in this work can be obtained applying the pipeline to the open-access S2AG corpus, available at https://www.semanticscholar.org/. All computational work has been conducted on high memory nodes of the HPC cluster of Queen Mary University of London [37].

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
GDB, ABr, NP, VL and ABa designed the study. GDB and ABr carried out the data collection. GDB and ABr performed the measurements. GDB, ABr, NP, VL and ABa analysed the data, discussed the results, and contributed to the final manuscript. All authors read and approved the final manuscript.

Author details
[1]School of Mathematical Sciences, Queen Mary University of London, Mile End Road, E1 4NS London, United Kingdom. [2]CNRS, GEMASS, 59 rue Pouchet, F-75017 Paris, France. [3]Sony Computer Science Laboratories Rome, Joint Initiative

CREF-Sony, Centro Ricerche Enrico Fermi, Via Panisperna 89/A, I-00184 Rome, Italy. [4]Department of Mathematics, City, University of London, Northampton Square, EC1V 0HB London, United Kingdom. [5]Complexity Science Hub Vienna, Josefstädter Str. 39, A-1080 Vienna, Austria. [6]Dipartimento di Fisica ed Astronomia, Università di Catania and INFN, Via S. Sofia, 64, I-95123 Catania, Italy. [7]The Alan Turing Institute, British Library, 96 Euston Road, NW1 2DB London, United Kingdom. [8]UCL Centre for Blockchain Technologies, University College London, Malet Place, WC1E 6BT London, United Kingdom.

Received: 10 March 2023 Accepted: 18 September 2023 Published online: 03 October 2023

References

1. London H (1975) The meaning of decentralization. Soc Stud 66(2):55–59. https://doi.org/10.1080/00220973.1943.11019391
2. Schneider N (2019) Decentralization: an incomplete ambition. J Cult Econ 12(4):265–285. https://doi.org/10.1080/17530350.2019.1589553
3. Vaudenay S (2020) Centralized or decentralized? The contact tracing dilemma. Cryptology eprint archive, paper 2020/531. https://eprint.iacr.org/2020/531
4. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. Decentralized Business Review
5. Nadini M, Alessandretti L, Di Giacinto F, Martino M, Aiello LM, Baronchelli A (2021) Mapping the nft revolution: market trends, trade networks, and visual features. Sci Rep 11(1):1–11. https://doi.org/10.1038/s41598-021-00053-8
6. ElBahrawy A, Alessandretti L, Kandler A, Pastor-Satorras R, Baronchelli A (2017) Evolutionary dynamics of the cryptocurrency market. R Soc Open Sci 4(11):170623. https://doi.org/10.1098/rsos.170623
7. Mekacher A, Bracci A, Nadini M, Martino M, Alessandretti L, Aiello LM, Baronchelli A (2022) How rarity shapes the nft market. arXiv preprint. arXiv:2204.10243
8. Treisman D (2007) The architecture of government: rethinking political decentralization. Cambridge University Press, New York
9. Bray M (1999) Control of education: issues and tensions in centralization and decentralization. In: Comparative education: the dialectic of the global and the local, pp 207–232
10. Ebel RD, Yilmaz S (2002) On the measurement and impact of fiscal decentralization, vol 2809. World Bank Publications, Washington
11. Saltman R, Busse R, Figueras J (2006) Decentralization in health care: strategies and outcomes. McGraw-Hill, New York
12. Galloway AR (2004) Protocol: how control exists after decentralization. MIT Press, Cambridge
13. Tanner HG, Kumar A (2005) Towards decentralization of multi-robot navigation functions. In: Proceedings of the 2005 IEEE international conference on robotics and automation. IEEE, pp 4132–4137
14. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239. https://doi.org/10.1016/0378-8733(78)90021-7
15. Ammar W, Groeneveld D, Bhagavatula C, Beltagy I, Crawford M, Downey D, Dunkelberger J, Elgohary A, Feldman S, Ha V, Kinney R, Kohlmeier S, Lo K, Murray T, Ooi H-H, Peters M, Power J, Skjonsberg S, Wang LL, Wilhelm C, Yuan Z, van Zuylen M, Etzioni O (2018) Construction of the literature graph in semantic scholar. In: Proceedings of the 2018 annual conference of the North American chapter of the association for computational linguistics
16. Governance. Accessed June 6, 2022. https://dictionary.cambridge.org/dictionary/english/governance
17. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu B-J, Wang K (2015) An overview of Microsoft academic service (mas) and applications. In: Proceedings of the 24th international conference on world wide web, pp 243–246
18. Next steps for Microsoft academic – expanding into new horizons. Accessed June 6, 2022. https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/
19. Semantic scholar's paper field of study classifier. https://github.com/allenai/s2_fos. Accessed June 8, 2022
20. Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. Sci Adv 4(7):1360. https://doi.org/10.1126/sciadv.aaq1360
21. Hyland CC, Tao Y, Azizi L, Gerlach M, Peixoto TP, Altmann EG (2021) Multilayer networks for text analysis with multiple data types. EPJ Data Sci 10(1):33. https://doi.org/10.1140/epjds/s13688-021-00288-5
22. Milojević S (2015) Quantifying the cognitive extent of science. J Informetr 9(4):962–973. https://doi.org/10.1016/j.joi.2015.10.005
23. Sun Y, Latora V (2020) The evolution of knowledge within and across fields in modern physics. Sci Rep 10(1):12097. https://doi.org/10.1038/s41598-020-68774-w
24. Bornmann L, Haunschild R, Mutz R (2021) Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun 8(1):224. https://doi.org/10.1057/s41599-021-00903-w
25. Smith JT (1851) Local self-government and centralization: the characteristics of each; and its practical tendencies as affecting social, moral, and political welfare and progress, including comprehensive outlines of the English constitution. In: Moral, and political welfare and progress (London, 1851), pp 27–51
26. Buchanan JM (1954) Individual choice in voting and the market. J Polit Econ 62(4):334–343
27. Kaufman HF (1959) Toward an interactional conception of community. Soc Forces 38:8. https://doi.org/10.2307/2574010
28. Bachrach P, Baratz MS (1962) Two faces of power. Am Polit Sci Rev 56(4):947–952. https://doi.org/10.2307/1952796
29. Šiljak DD (1978) On decentralized control of large scale systems. In: Proceedings of the 7th triennial world congress of the IFAC on a link between science and applications of automatic control, Helsinki, Finland, 12-16 June, pp 1849–1856. https://doi.org/10.1016/S1474-6670(17)66158-5
30. Columbus L Gartner hype cycle for emerging technologies, 2016 adds blockchain &. machine learning for first time. Accessed June 9, 2022. https://www.forbes.com/sites/louiscolumbus/2016/08/21/gartner-hype-cycle-for-emerging-technologies-2016-adds-blockchain-machine-learning-for-first-time/?sh=3d9bc1973f82
31. McLean S, Deane-Johns S (2016) Demystifying blockchain and distributed ledger technology–hype or hero? Comput Law Rev Int 17(4):97–102. https://doi.org/10.9785/cri-2016-0402

32. Ametrano FM (2016) Bitcoin, blockchain, and distributed ledgers: between hype and reality. https://doi.org/10.2139/ssrn.2832249. Available at SSRN 2832249
33. Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A (2020) Yake! Keyword extraction from single documents using multiple local features. Inf Sci 509:257–289. https://doi.org/10.1016/j.ins.2019.09.013
34. Bouveyron C, Latouche P, Zreik R (2018) The stochastic topic block model for the clustering of vertices in networks with textual edges. Stat Comput 28:11–31. https://doi.org/10.1007/s11222-016-9713-7
35. Airoldi EM, Blei D, Fienberg S, Xing E (2008) Mixed membership stochastic blockmodels. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. Curran Associates, Vancouver
36. Zhu Y, Yan X, Getoor L, Moore C (2013) Scalable text and link analysis with mixed-topic link models. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 473–481. https://doi.org/10.1145/2487575.2487693
37. King T, Butcher S, Zalewski L (2017) Apocrita - high performance computing cluster for Queen Mary University of London. https://doi.org/10.5281/zenodo.438045

Publisher's Note